# Statistical Regression Analysis

Larry Winner
University of Florida
Department of Statistics

August 19, 2021

# Contents

# Chapter 1

# Probability Distributions, Estimation, and Testing

## 1.1 Introduction

Here we introduce probability distributions, and basic estimation/testing methods. **Random variables** are outcomes of an experiment or data-generating process, where the outcome is not known in advance, although the set of possible outcomes is. Random variables can be **discrete** or **continuous**. Discrete random variables can take on only a finite or countably infinite set of possible outcomes. Continuous random variables can take on values along a continuum. In many cases, variables of one type may be treated as or reported as the other type. In the introduction, we will use upper-case letters (such as $Y$) to represent random variables, and lower-case letters (such as $y$) to represent specific outcomes. Not all (particularly applied statistics) books follow this convention.

### 1.1.1 Discrete Random Variables/Probability Distributions

In many applications, the result of the data-generating process is the count of a number of events of some sort. In some cases, a certain number of trials are conducted, and the outcome of each trial is observed as a "Success" or "Failure" (binary outcomes). In these cases, the number of trials ending in Success is observed. Alternatively, a series of trials may be conducted until a pre-selected number of Successes are observed. In other settings, the number of events of interest may be counted in a fixed amount of time or space, without actually breaking the domain into a set of distinct trials.

For discrete random variables, we will use $p(y)$ to represent the probability that the random variable $Y$ takes on the value $y$. We require that all such probabilities be bounded between 0 and 1 (inclusive), and that they sum to 1:

$$P\{Y = y\} = p(y) \qquad 0 \leq p(y) \leq 1 \qquad \sum_y p(y) = 1$$

The **cumulative distribution function** is the probability that a random variable takes on a value less than or equal to a specific value $y^*$. It is an increasing function that begins at 0 and increases to 1, and we will denote it as $F(y^*)$. For discrete random variables it is a step function, taking a step at each point where $p(y) > 0$:

$$F(y^*) = P(Y \leq y^*) = \sum_{y \leq y^*} p(y)$$

The **mean** or **Expected Value** ($\mu$) of a random variable is it's long-run average if the experiment was conducted repeatedly ad infinitum. The **variance** $\left(\sigma^2\right)$ is the average squared difference between the random variable and its mean, measuring the dispersion within the distribution. The **standard deviation** ($\sigma$) is the positive square root of the variance, and is in the same units as the data.

$$\mu_Y = E\{Y\} = \sum_y yp(y) \qquad \sigma_Y^2 = V\{Y\} = E\left\{(Y-\mu_Y)^2\right\} = \sum_y (y-\mu_Y)^2 p(y) \qquad \sigma_Y = +\sqrt{\sigma_Y^2}$$

Note that for any function of $Y$, the expected value and variance of the function is computed as follows:

$$E\{g(Y)\} = \sum_y g(y)p(y) = \mu_{g(Y)} \qquad V\{g(Y)\} = E\left\{\left(g(Y)-\mu_{g(Y)}\right)^2\right\} = \sum_y \left(g(y)-\mu_{g(Y)}\right)^2 p(y)$$

For any constants $a$ and $b$, we have the mean and variance of the linear function $a + bY$:

$$E\{a+bY\} = \sum_y ap(y) + \sum_y byp(y) = a\sum_y p(y) + b\sum_y yp(y) = a(1) + bE\{Y\} = a + b\mu_Y$$

$$V\{a+bY\} = \sum_y ((a+by) - (a+b\mu_Y))^2 p(y) = b^2 \sum_y (y-\mu_Y)^2 p(y) = b^2\sigma_Y^2$$

A very useful result in mathematical statistics is the following:

$$\sigma_Y^2 = V\{Y\} = E\left\{(Y-\mu_Y)^2\right\} = E\left\{Y^2 - 2\mu_Y Y + \mu_2\right\} = E\left\{Y^2\right\} - 2\mu_Y E\{Y\} + \mu_Y^2 = E\left\{Y^2\right\} - \mu_Y^2$$

Thus, $E\left\{Y^2\right\} = \sigma_Y^2 + \mu_Y^2$. Also, from this result we obtain: $E\{Y(Y-1)\} = \sigma_Y^2 + \mu_Y^2 - \mu_Y$. From this, we can obtain $\sigma_Y^2 = E\{Y(Y-1)\} - \mu_Y^2 + \mu_Y$, which is useful for some discrete probability distributions.

Next, we consider several families of discrete probability distributions: the Binomial, Poisson, and Negative Binomial families.

### Binomial Distribution

When an experiment consists of $n$ independent trials, each of which can end in one of two outcomes: Success or Failure with constant probability of success, we refer to this as a **binomial experiment**. The random variable $Y$ is the number of Successes in the $n$ trials, and can take on the values $y = 0, 1, \ldots, n$. Note that in some settings, the "Success" can be a negative attribute. We denote the probability of success as $\pi$, which lies between 0 and 1. We use the notation: $Y \sim B(n, \pi)$. The probability distribution, mean and variance of $Y$ depend on the sample size $n$ and probability of success $\pi$.

$$p(y) = \binom{n}{y}\pi^y(1-\pi)^{n-y} \qquad E\{Y\} = \mu_Y = n\pi \qquad V\{Y\} = \sigma_Y^2 = n\pi(1-\pi)$$

where $\binom{n}{y} = \frac{n!}{y!(n-y)!}$. In practice, $\pi$ will be unknown, and estimated from sample data. Note that to obtain the mean and variance, we have:

$$E\{Y\} = \mu_Y = \sum_{y=0}^{n} yp(y) = \sum_{y=0}^{n} y\frac{n!}{y!(n-y)!}\pi^y(1-\pi)^{n-y} = \sum_{y=1}^{n} y\frac{n!}{y!(n-y)!}\pi^y(1-\pi)^{n-y} =$$

$$= n\pi\sum_{y=1}^{n}\frac{(n-1)!}{(y-1)!(n-y)!}\pi^{y-1}(1-\pi)^{n-y} = n\pi\sum_{y^*=0}^{n-1}\binom{n-1}{y^*}\pi^{y^*}(1-\pi)^{n-1-y^*} = n\pi\sum_{y^*} p(y^*) = n\pi \qquad y^* = y-1$$

To obtain the variance, we use the result from above, $\sigma_Y^2 = E\{Y(Y-1)\} - \mu_Y^2 + \mu_Y$:

$$E\{Y(Y-1)\} = \sum_{y=0}^{n} y(y-1)p(y) = \sum_{y=0}^{n} y(y-1)\frac{n!}{y!(n-y)!}\pi^y(1-\pi)^{n-y} = \sum_{y=2}^{n} y(y-1)\frac{n!}{y!(n-y)!}\pi^y(1-\pi)^{n-y} =$$

$$= n(n-1)\pi^2\sum_{y=2}^{n}\frac{(n-2)!}{(y-2)!(n-y)!}\pi^{y-2}(1-\pi)^{n-y} = n(n-1)\pi^2\sum_{y^{**}=0}^{n-2}\binom{n-2}{y^{**}}\pi^{y^{**}}(1-\pi)^{n-2-y^{**}}$$

$$n(n-1)\pi^2\sum_{y^{**}} p(y^{**}) = n(n-1)\pi^2 \qquad y^{**} = y-2$$

$$\Rightarrow \quad \sigma_Y^2 = n(n-1)\pi^2 - n^2\pi^2 + n\pi = n\pi - n\pi^2 = n\pi(1-\pi)$$

**Poisson Distribution**

In many applications, researchers observe the counts of a random process in some fixed amount of time or space. The random variable $Y$ is a count that can take on any non-negative integer. One important aspect of the Poisson family is that the mean and variance are the same. This is one aspect that does not work for all applications. We use the notation: $Y \sim \text{Poi}(\lambda)$. The probability distribution, mean and variance of $Y$ are:

$$p(y) = \frac{e^{-\lambda}\lambda^y}{y!} \qquad E\{Y\} = \mu_Y = \lambda \qquad V\{Y\} = \sigma_Y^2 = \lambda$$

Note that $\lambda > 0$. The Poisson arises by dividing the time/space into $n$ infinitely small areas, each having either 0 or 1 Success, with Success probability $\pi = \lambda/n$. Then $Y$ is the number of areas having a success.

$$p(y) = \frac{n!}{y!(n-y)!}\left(\frac{\lambda}{n}\right)^y\left(1-\frac{\lambda}{n}\right)^{n-y} = \frac{n(n-1)\cdots(n-y+1)}{y!}\left(\frac{\lambda}{n}\right)^y\left(1-\frac{\lambda}{n}\right)^{n-y} =$$

$$= \frac{1}{y!}\left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\cdots\left(\frac{n-y+1}{n}\right)\lambda^y\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-y}$$

The limit as $n$ goes to $\infty$ is:

$$\lim_{n\to\infty} p(y) = \frac{1}{y!}(1)(1)\cdots(1)\lambda^y e^{-\lambda}(1) = p(y) = \frac{e^{-\lambda}\lambda^y}{y!}$$

To obtain the mean of $Y$ for the Poisson distribution, we have:

$$E\{Y\} = \mu_Y = \sum_{y=0}^{\infty} yp(y) = \sum_{y=0}^{\infty} y\frac{e^{-\lambda}\lambda^y}{y!} = \sum_{y=1}^{\infty} y\frac{e^{-\lambda}\lambda^y}{y!} = \sum_{y=1}^{\infty}\frac{e^{-\lambda}\lambda^y}{(y-1)!} =$$

$$\lambda\sum_{y=1}^{\infty}\frac{e^{-\lambda}\lambda^{(y-1)}}{(y-1)!}=\lambda\sum_{y^*=0}^{\infty}\frac{e^{-\lambda}\lambda^{y^*}}{y^*!}=\lambda\sum_{y^*}p\left(y^*\right)=\lambda$$

We use the same result as that for the binomial to obtain the variance for the Poisson distribution:

$$E\left\{Y(Y-1)\right\}=\sum_{y=0}^{\infty}y(y-1)p(y)=\sum_{y=0}^{\infty}y(y-1)\frac{e^{-\lambda}\lambda^{y}}{y!}=\sum_{y=2}^{\infty}y(y-1)\frac{e^{-\lambda}\lambda^{y}}{y!}=\lambda^2\sum_{y=2}^{\infty}\frac{e^{-\lambda}\lambda^{(y-2)}}{(y-2)!}=\lambda^2\sum_{y^{**}=0}^{\infty}\frac{e^{-\lambda}\lambda^{y^{**}}}{y^{**}!}=\lambda^2$$

$$\Rightarrow\quad\sigma_Y^2=\lambda^2-\lambda^2+\lambda=\lambda$$

**Negative Binomial Distribution**

The negative binomial distribution is used in two quite different contexts. The first is where a binomial type experiment is being conducted, except instead of having a fixed number of trials, the experiment is completed when the $r^{th}$ success occurs. The random variable $Y$ is the number of trials needed until the $r^{th}$ success, and can take on any integer value greater than or equal to $r$. The probability distribution, its mean and variance are:

$$p(y)=\binom{y-1}{r-1}\pi^r\left(1-\pi\right)^y-r\qquad E\left\{Y\right\}=\mu_Y=\frac{r}{\pi}\qquad V\left\{Y\right\}=\sigma_Y^2=\frac{r\left(1-\pi\right)}{\pi^2}.$$

A second use of the negative binomial distribution is as a model for count data. It arises from a mixture of Poisson models. In this setting it has 2 parameters and is more flexible than the Poisson (which has the variance equal to the mean), and can take on any non-negative integer value. In this form, the negative binomial distribution and its mean and variance can be written as (see e.g. Agresti (2002) and Cameron and Trivedi (2005)).

$$f\left(y;\mu,\alpha\right)=\frac{\Gamma\left(\alpha^{-1}+y\right)}{\Gamma\left(\alpha^{-1}\right)\Gamma\left(y+1\right)}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{y}\qquad\Gamma(w)=\int_0^{\infty}x^{w-1}e^{-x}dx=(w-1)\,\Gamma\left(w-1\right).$$

The mean and variance of this form of the Negative Binomial distribution are

$$E\left\{Y\right\}=\mu\qquad V\left\{Y\right\}=\mu\left(1+\alpha\mu\right).$$

### 1.1.2 Continuous Random Variables/Probability Distributions

Random variables that can take on any value along a continuum are continuous. Here, we consider the normal, gamma, $t$, and $F$ families. Special cases of the gamma family include the exponential and chi-squared distributions. Continuous distributions are density functions, as opposed to probability mass functions. Their density is always non-negative, and integrates to 1. We will use the notation $f(y)$ for density functions. The mean and variance for continuous distributions are obtained in a similar manner as discrete distributions, with integration replacing summation.

$$E\left\{Y\right\}=\mu_Y=\int_{-\infty}^{\infty}yf(y)dy\qquad\qquad V\left\{Y\right\}=\sigma_Y^2=\int_{-\infty}^{\infty}\left(y-\mu_Y\right)^2f(y)dy$$

In general, for any function $g(Y)$, we have:

$$E\left\{g(Y)\right\}=\int_{-\infty}^{\infty}g(y)f(y)dy=\mu_{g(Y)}\qquad V\left\{g(Y)\right\}=E\left\{\left(g(Y)-\mu_{g(Y)}\right)^2\right\}=\int_{-\infty}^{\infty}\left(g(y)-\mu_{g(Y)}\right)^2f(y)dy$$

### Normal Distribution

The normal distributions, also known as the Gaussian distributions, are a family of symmetric mound-shaped distributions. The distribution has 2 parameters: the mean $\mu$ and the variance $\sigma^2$, although often it is indexed by its standard deviation $\sigma$. We use the notation $Y \sim N\left(\mu, \sigma^2\right)$. The probability density function, the mean and variance are:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \qquad E\{Y\} = \mu_Y = \mu \qquad V\{Y\} = \sigma_Y^2 = \sigma^2$$

The mean $\mu$ defines the center (median and mode) of the distribution, and the standard deviation $\sigma$ is a measure of the spread ($\mu - \sigma$ and $\mu + \sigma$ are the inflection points). Despite the differences in location and spread of the different distributions in the normal family, probabilities with respect to standard deviations from the mean are the same for all normal distributions. For $-\infty < z_1 < z_2 < \infty$, we have:

$$P\left(\mu + z_1\sigma \leq Y \leq \mu + z_2\sigma\right) = \int_{\mu+z_1\sigma}^{\mu+z_2\sigma} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) dy = \int_{z1}^{z_2} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \Phi(z_2) - \Phi(z_1)$$

Where $Z$ is **standard normal**, a normal distribution with mean 0, and variance (standard deviation) 1. Here $\Phi(z^*)$ is the cumulative distribution function of the standard normal distribution, up to the point $z^*$:

$$\Phi(z^*) = \int_{-\infty}^{z^*} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

These probabilities and critical values can be obtained directly or indirectly from standard tables, statistical software, or spreadsheets. Note that:

$$Y \sim N\left(\mu, \sigma^2\right) \qquad \Rightarrow \qquad Z = \frac{Y-\mu}{\sigma} \sim N(0,1)$$

This makes it possible to use the standard normal table for any normal distribution. Plots of three normal distributions are given in Figure 1.1.

### Gamma Distribution

The gamma family of distributions are used to model non-negative random variables that are often right-skewed. There are two widely used parameterizations. The first given here is in terms of *shape* and *scale* parameters:

$$f(y) = \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-y/\beta} \quad y \geq 0, \alpha > 0, \beta > 0 \qquad E\{Y\} = \mu_Y = \alpha\beta \qquad V\{Y\} = \sigma_Y^2 = \alpha\beta^2$$

Here, $\Gamma(\alpha)$ is the gamma function $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy$ and is built-in to virtually all statistical packages and spreadsheets. It also has two simple properties:

$$\alpha > 1: \quad \Gamma(\alpha) = (\alpha-1)\,\Gamma(\alpha-1) \qquad\qquad \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}.$$

Thus, if $\alpha$ is an integer, $\Gamma(\alpha) = (\alpha-1)!$. The second version given here is in terms of *shape* and *rate* parameters.

$$f(y) = \frac{\theta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-y\theta} \quad y \geq 0, \alpha > 0, \theta > 0 \qquad E\{Y\} = \mu_Y = \frac{\alpha}{\theta} \qquad V\{Y\} = \sigma_Y^2 = \frac{\alpha}{\theta^2}$$

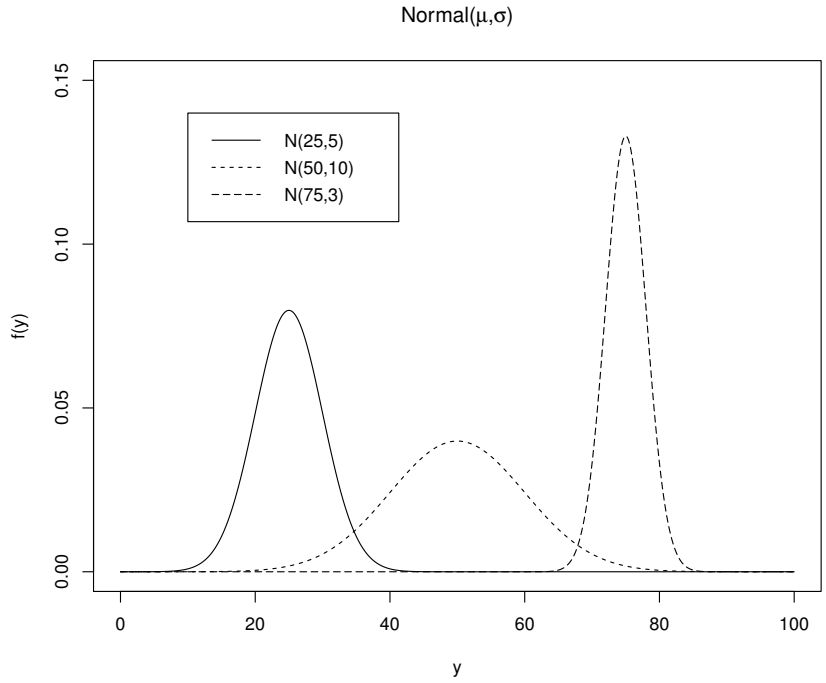


Figure 1.1: Three Normal Densities

Note that different software packages use different parameterizations in generating samples and giving tail-areas and critical values. For instance, EXCEL uses the first parameterization and $R$ uses the second. Figure 1.2 displays three gamma densities of various shapes.

Two special cases are the exponential family, where $\alpha = 1$ and the chi-squared family, with $\alpha = \nu/2$ and $\beta = 2$ for integer valued $\nu$. For the exponential family, based on the second parameterization:

$$f(y) = \theta e^{-y\theta} \qquad E\{Y\} = \mu_Y = \frac{1}{\theta} \qquad V\{Y\} = \sigma_Y^2 = \frac{1}{\theta^2}$$

Probabilities for the exponential distribution are trivial to obtain as $F(y^*) = 1 - e^{-y^*\theta}$. Figure 1.3 gives three exponential distributions.

For the chi-square family, based on the first parameterization:

$$f(y) = \frac{1}{\Gamma\left(\frac{\nu}{2}\right) 2^{\nu/2}} y^{\frac{\nu}{2}-1} e^{-y/2} \qquad E\{Y\} = \mu_Y = \nu \qquad V\{Y\} = \sigma_Y^2 = 2\nu$$

Here, $\nu$ is the **degrees of freedom** and we denote the distribution as: $Y \sim \chi^2_\nu$. Upper and lower critical values of the chi-square distribution are available in tabular form, and in statistical packages and spreadsheets. Probabilities can be obtained with statistical packages and spreadsheets. Figure 1.4 gives three Chi-square distributions.

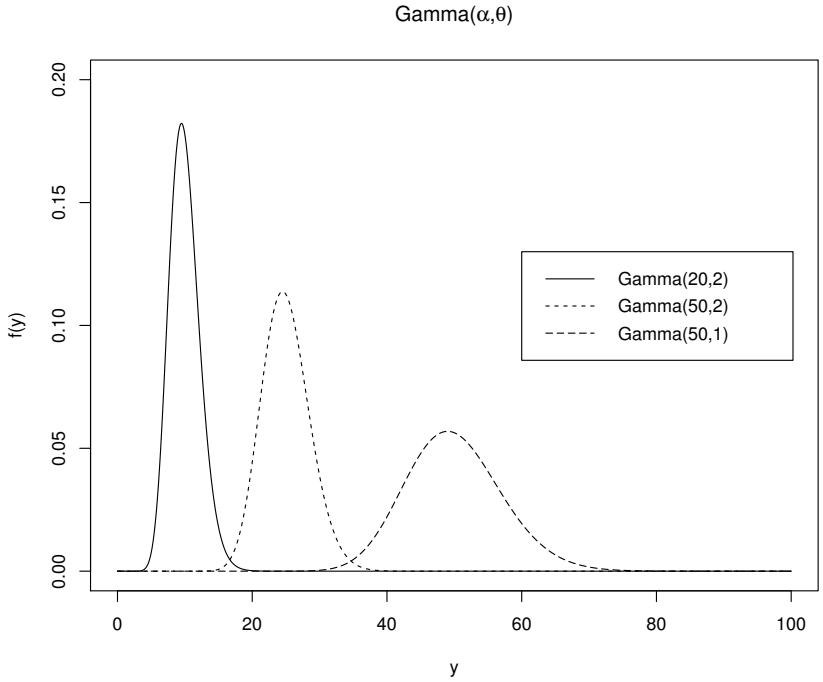


Figure 1.2: Three Gamma Densities

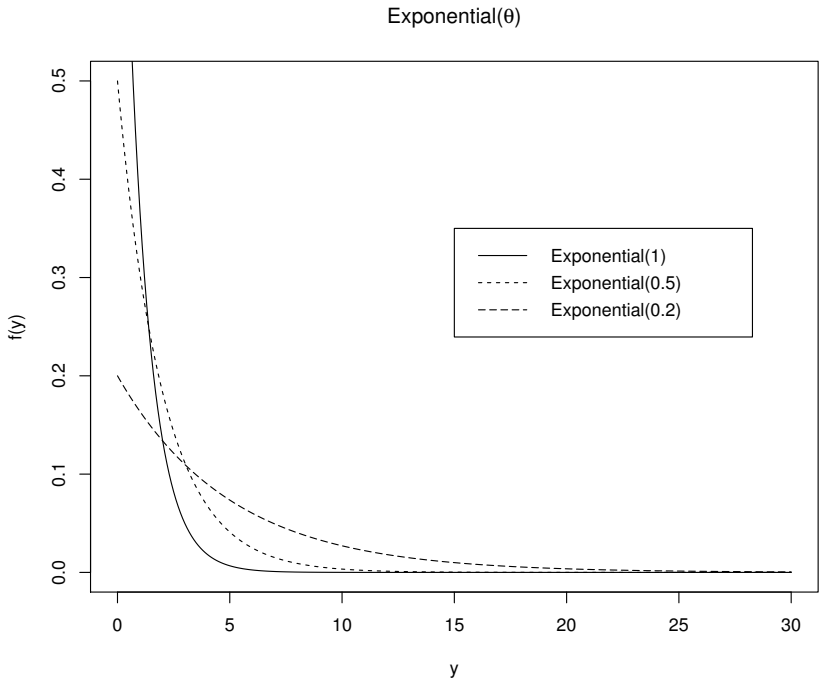


Figure 1.3: Three Exponential Densities

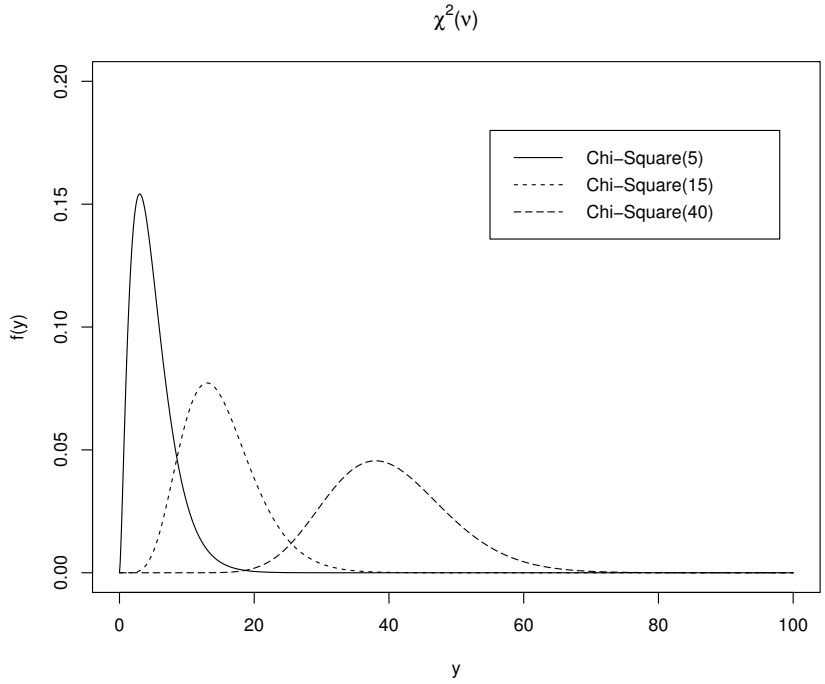


Figure 1.4: Three Chi-Square Densities

## Beta Distribution

The Beta distribution can be used to model data that are proportions (or percentages divided by 100). The traditional model for the Beta distribution is given below.

$$f(y;\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}y^{\alpha-1}(1-y)^{\beta-1}\qquad 0<y<1;\quad \alpha>0,\beta>0\qquad \int_0^1 w^a(1-w)^b\,dw=\frac{\Gamma(a+1)\Gamma(b+1)}{\Gamma(a+b+2)}$$

Note that the Uniform distribution is a special case, with $\alpha=\beta=1$. The mean and variance of the Beta distribution are obtained as follows:

$$E\{Y\}=\int_0^1 y\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}y^{\alpha-1}(1-y)^{\beta-1}\,dy=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 y^{\alpha}(1-y)^{\beta-1}\,dy=$$

$$=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)}=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\alpha\Gamma(\alpha)\Gamma(\beta)}{(\alpha+\beta)\Gamma(\alpha+\beta)}=\frac{\alpha}{\alpha+\beta}$$

By extending this logic, we can obtain:

$$E\left\{Y^2\right\}=\int_0^1 y^2\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}y^{\alpha-1}(1-y)^{\beta-1}\,dy=\frac{\alpha(\alpha+1)}{(\alpha+\beta+1)(\alpha+\beta)}\quad\Rightarrow\quad V\{Y\}=\frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$$

An alternative formulation of the distribution involves setting re-parameterizing as follows:

$$\mu=\frac{\alpha}{\alpha+\beta}\qquad\phi=\alpha+\beta\quad\Rightarrow\quad\alpha=\mu\phi\qquad\beta=(1-\mu)\phi.$$

Figure 1.5 gives three Beta distributions.

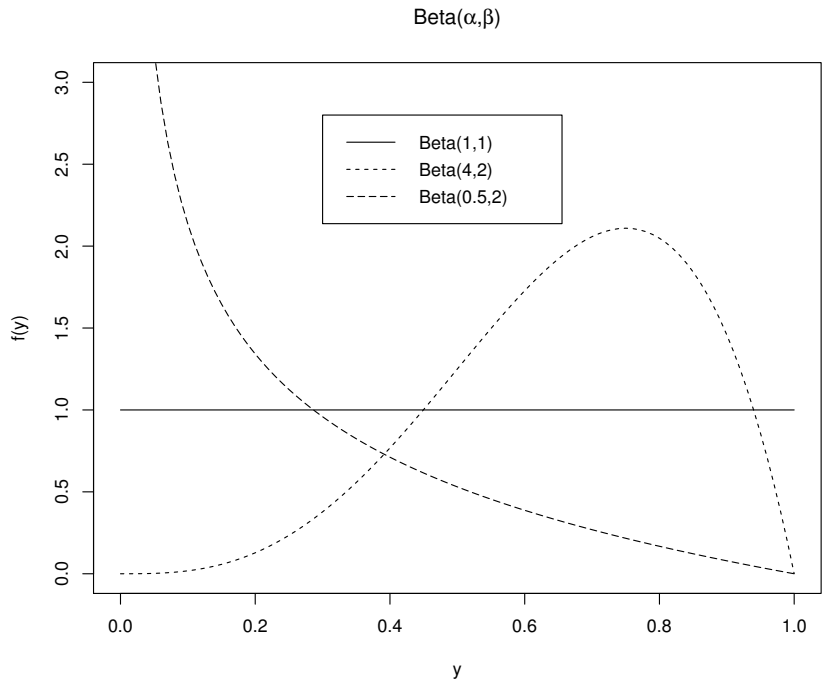


Figure 1.5: Three Beta Densities

## 1.2 Linear Functions of Multiple Random Variables

Suppose we simultaneously observe two random variables: $X$ and $Y$. Their joint probability distribution can be discrete, continuous, or mixed (one discrete, the other continuous). We consider the **joint distribution** and the **marginal distributions** for the discrete case below.

$$p(x,y) = P\{X = x, Y = y\} \qquad p_X(x) = P\{X = x\} = \sum_y p(x,y) \qquad p_Y(y) = P\{Y = y\} = \sum_x p(x,y)$$

For the continuous case, we have the following joint and marginal densities and cumulative distribution function.

$$\text{Joint Density when } X = x, Y = y : f(x,y) \qquad f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy \qquad f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx$$

$$F(a,b) = P\{X \leq a, Y \leq b\} = \int_{-\infty}^{b}\int_{-\infty}^{a} f(x,y)dxdy$$

Note the following.

$$\text{Discrete: } \sum_x \sum_y p(x,y) = \sum_x p_x(x) = \sum_y p_Y(y) = 1$$

$$\text{Continuous: } \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = \int_{-\infty}^{\infty} f_X(x)dx = \int_{-\infty}^{\infty} f_Y(y)dy = 1$$

The **conditional probability** that $X = x$, given $Y = y$ (or $Y = y$ given $X = x$)is denoted as follows.

$$p(x|y) = P\{X = x|Y = y\} = \frac{p(x,y)}{p_Y(y)} \qquad p(y|x) = P(Y = y|X = x) = \frac{p(x,y)}{p_X(x)}$$

assuming $p_Y(y) > 0$ and $p_X(x) > 0$. This simply implies the probability that both occur divided by the probability that $Y = y$ or that $X = x$. $X$ and $Y$ are said to be independent if $p(x|y) = p(x)$ for all $y$, and that $p(y|x) = p(y)$ for all $x$. The conditional densities for continuous random variables are similarly defined based on the joint and marginal densities.

$$f(x|y) = \frac{f(x,y)}{f_Y(y)} \qquad f_Y(y) > 0 \qquad f(y|x) = \frac{f(x,y)}{f_X(x)} \qquad f_X(x) > 0$$

The **conditional mean** and **variance** are the mean and variance of the conditional distribution (density), and are often functions of the conditioning variable.

$$\text{Discrete: } E\{Y|X = x\} = \mu_{Y|x} = \sum_y yp(y|x) \qquad \text{Continuous: } E\{Y|X = x\} = \mu_{Y|x} = \int_{-\infty}^{\infty} yf(y|x)dy$$

$$\text{Discrete: } V\{Y|X = x\} = \sigma^2_{Y|x} = \sum_y \left(y - \mu_{Y|x}\right)^2 p(y|x)$$

$$\text{Continuous: } V\{Y|X = x\} = \sigma^2_{Y|x} = \int_{-\infty}^{\infty} \left(y - \mu_{Y|x}\right)^2 f(y|x)dy$$

Next we consider the **variance of the conditional mean** and the **mean of the conditional variance** for the continuous case (with integration being replaced by summation for the discrete case).

$$V_X\{E\{Y|x\}\} = \int_{-\infty}^{\infty} \left(\mu_{Y|x} - \mu_Y\right)^2 f_X(x)dx$$

$$E_X\{V\{Y|x\}\} = \int_{-\infty}^{\infty} \sigma^2_{Y|x} f_X(x)dx$$

Note that we can partition the variance of $Y$ into the sum of the variance of the conditional mean and mean of the conditional variance:

$$V\{Y\} = V_X\{E\{Y|x\}\} + E_X\{V\{Y|x\}\}$$

The **covariance** $\sigma_{XY}$ between $X$ and $Y$ is the average product of deviations from the mean for $X$ and $Y$. For the discrete case, we have the following.

$$\sigma_{XY} = E\{(X - \mu_X)(Y - \mu_Y)\} = \sum_x \sum_y (x - \mu_X)(y - \mu_Y)p(x,y) = \sum_x \sum_y (xy - x\mu_Y - \mu_X y + \mu_X\mu_Y)p(x,y) =$$

$$= \sum_x \sum_y xyp(x,y) - \mu_Y \sum_x \sum_y xp(x,y) - \mu_X \sum_x \sum_y yp(x,y) + \mu_X\mu_Y = E\{XY\} - \mu_X\mu_Y$$

For the continuous case, replace summation with integration. If $X$ and $Y$ are independent, $\sigma_{XY} = 0$, but the converse is not typically the case. Covariances can be either positive or negative, depending on the association (if any) between $X$ and $Y$. The covariance is unbounded, and depends on the scales of measurement for $X$ and $Y$. The **correlation** $\rho_{XY}$ is a measure that is unit-less, is not affected by linear transformations of $X$ and/or $Y$, and is bounded between -1 and 1.

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$$

where $\sigma_X$ and $\sigma_Y$ are the standard deviations of the marginal distributions of $X$ and $Y$, respectively.

The mean and variance of any **linear function** of $X$ and $Y$: $W = aX + bY$ for fixed constants $a$ and $b$ are for the discrete case:

$$E\{W\} = E\{aX + bY\} = \sum_x \sum_y (ax + by)p(x, y) = a\sum_x xp_X(x) + b\sum_y yp_Y(y) = a\mu_X + b\mu_Y$$

$$V\{W\} = V\{aX + bY\} = \sum_x \sum_y \left[(ax + by) - (a\mu_x + b\mu_y)\right]^2 p(x, y) =$$

$$\sum_x \sum_y \left[a^2 (x - \mu_X)^2 + b^2 (y - \mu_Y)^2 + 2ab (x - \mu_X)(y - \mu_Y)\right] p(x, y) = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY}.$$

For the continuous case, replace summation with integration.

In general, if $Y_1, \ldots, Y_n$ are a sequence of random variables, and $a_1, \ldots, a_n$ are a sequence of constants, we have the following results.

$$E\left\{\sum_{i=1}^n a_i Y_i\right\} = \sum_{i=1}^n a_i E\{Y_i\} = \sum_{i=1}^n a_i\mu_i$$

$$V\left\{\sum_{i=1}^n a_i Y_i\right\} = \sum_{i=1}^n a_i^2\sigma_i^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^n a_i a_j \sigma_{ij}$$

Here $\mu_i$ is the mean of $Y_i$, $\sigma_i^2$ is the variance of $Y_i$, and $\sigma_{ij}$ is the covariance of $Y_i$ and $Y_j$.

If we have two linear functions of the same sequence of random variables, say $W_1 = \sum_{i=1}^n a_i Y_i$ and $W_2 = \sum_{i=1}^n b_i Y_i$, we can obtain their covariance as follows.

$$\text{COV}\{W_1, W_2\} = \sum_{i=1}^n \sum_{j=1}^n a_i b_j \sigma_{ij} = \sum_{i=1}^n a_i b_i \sigma_i^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^n a_i b_j \sigma_{ij}$$

The last term drops out when $Y_1, \ldots, Y_n$ are independent.

## 1.3 Functions of Normal Random Variables

First, note that if $Z \sim N(0, 1)$, then $Z^2 \sim \chi_1^2$. Many software packages present $Z$-tests as (Wald) $\chi^2$-tests. See the section on testing below.

Suppose $Y_1, ..., Y_n$ are independent with $Y_i \sim N\left(\mu, \sigma^2\right)$ for $i = 1, \ldots, n$. Then the sample mean and sample variance are computed as follow.

$$\overline{Y} = \frac{\sum_{i=1}^n Y_i}{n} \qquad S^2 = \frac{\sum_{i=1}^n (Y_i - \overline{Y})^2}{n - 1}$$

In this case, we obtain the following sampling distributions for the mean and a function of the variance.

$$\overline{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad \frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{\sigma^2} \sim \chi^2_{n-1} \qquad \overline{Y}, \quad \frac{(n-1)S^2}{\sigma^2} \text{ are independent.}$$

Note that in general, if $Y_1, ..., Y_n$ are normally distributed (and not necessarily with the same mean and/or variance), any linear function of them will be normally distributed, with mean and variance given in the previous section.

Two distributions associated with the normal and chi-squared distributions are **Student's** $t$ and $F$. Student's $t$-distribution is similar to the standard normal ($N(0,1)$), except that is indexed by its degrees of freedom and that it has heavier tails than the standard normal. As its degrees of freedom approach infinity, its distribution converges to the standard normal. Let $Z \sim N(0,1)$ and $W \sim \chi^2_\nu$, where $Z$ and $W$ are independent. Then, we get:

$$Y \sim N\left(\mu, \sigma^2\right) \quad \Rightarrow \quad Z = \frac{Y-\mu}{\sigma} \sim N(0,1) \qquad T = \frac{Z}{\sqrt{W/\nu}} \sim t_\nu$$

where the probability density, mean, and variance for Student's $t$-distribution are:

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \qquad E\{T\} = \mu_T = 0 \qquad V\{T\} = \frac{\nu}{\nu-2} \quad \nu > 2$$

and we use the notation $T \sim t_\nu$. Three $t$-distributions, along with the standard normal ($z$) distribution are shown in Figure 1.6.

Now consider the sample mean and variance, and the fact they are independent.

$$\overline{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \Rightarrow \quad Z = \frac{\overline{Y}-\mu}{\sqrt{\frac{\sigma^2}{n}}} = \sqrt{n}\frac{\overline{Y}-\mu}{\sigma} \sim N(0,1)$$

$$W = \frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{\sigma^2} \sim \chi^2_{n-1} \quad \Rightarrow \quad \sqrt{\frac{W}{\nu}} = \sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}} = \frac{S}{\sigma}$$

$$\Rightarrow \quad T = \frac{Z}{\sqrt{W/\nu}} = \frac{\sqrt{n}\frac{\overline{Y}-\mu}{\sigma}}{\frac{S}{\sigma}} = \sqrt{n}\frac{\overline{Y}-\mu}{S} \sim t_\nu$$

The $F$-distribution arises often in Regression and Analysis of Variance applications. If $W_1 \sim \chi^2_{\nu_1}$, $W_2 \sim \chi^2_{\nu_2}$, and $W_1$ and $W_2$ are independent, then:

$$F = \frac{\left[\frac{W_1}{\nu_1}\right]}{\left[\frac{W_2}{\nu_2}\right]} \sim F_{\nu_1,\nu_2}.$$

where the probability density, mean, and variance for the $F$-distribution are given below as a function of the specific point $F = f$.

$$f(f) = \left[\frac{\Gamma\left(\left(\frac{\nu_1+\nu_2}{2}\right)\right)\nu_1^{\nu_1/2}\nu_2^{\nu_2/2}}{\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)}\right]\left[\frac{f^{\nu_1/2-1}}{(\nu_1 f + \nu_2)^{(\nu_1+\nu_2)/2}}\right]$$

$$E\{F\} = \mu_F = \frac{\nu_2}{\nu_2-2} \quad \nu_2 > 2 \qquad V\{F\} = \frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)} \quad \nu_2 > 4$$

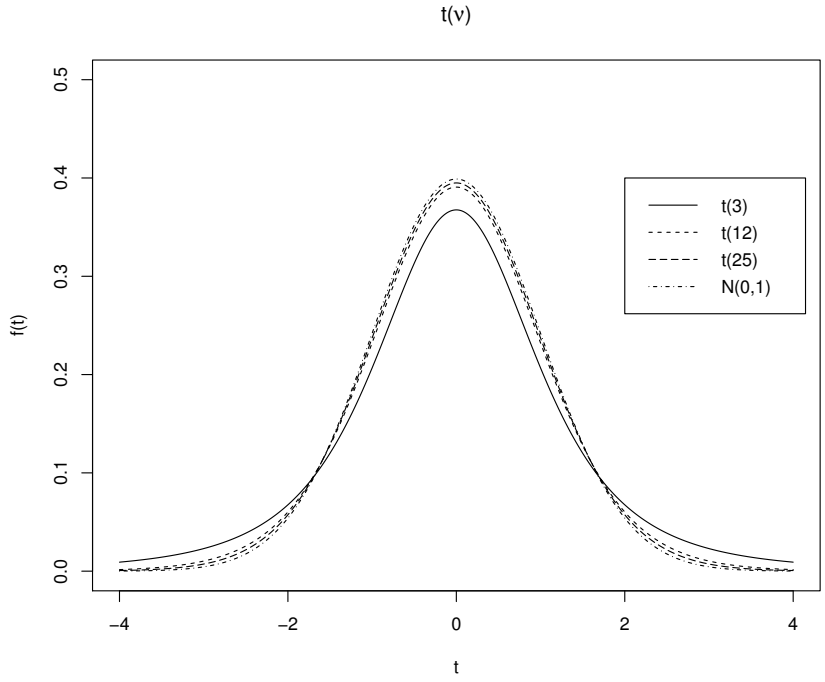


Figure 1.6: Three t-densities and z

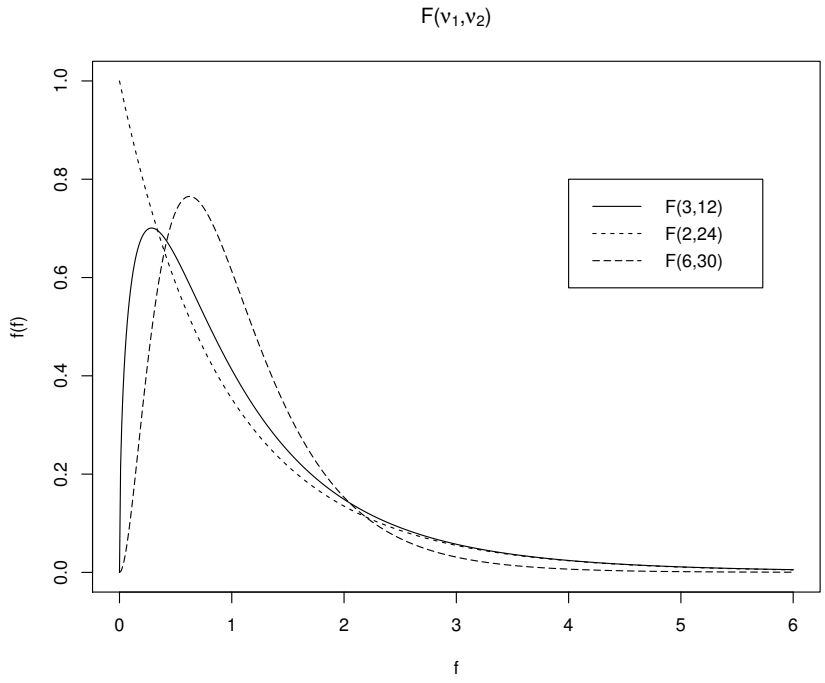


Figure 1.7: Three F-densities

Three $F$-distributions are given in Figure 1.7.

Critical values for the $t$ and $F$-distributions are given in statistical textbooks. Probabilities can be obtained from many statistical packages and spreadsheets. Technically, the $t$ and $F$ distributions described here are **central** $t$ and **central** $F$ distributions.

### Inferences Regarding $\mu$ and $\sigma^2$

We can test hypotheses concerning $\mu$ and obtain confidence intervals based on the sample mean and standard deviation when the data are independent $N\left(\mu, \sigma^2\right)$. Let $t\left(\alpha/2, \nu\right)$ be the value such that if:

$$T \sim t_\nu \quad \Rightarrow \quad P\left(T \geq t_{\alpha/2,\nu}\right) = \alpha/2.$$

then we get the following probability statement that leads to a $(1-\alpha)100\%$ confidence interval for $\mu$.

$$1-\alpha = P\left(-t_{\alpha/2,n-1} \leq \sqrt{n}\frac{\overline{Y}-\mu}{S} \leq t_{\alpha/2,n-1}\right) = P\left(-t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} \leq \overline{Y}-\mu \leq t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right) =$$

$$= P\left(\overline{Y} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{Y} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right)$$

Once a sample is observed with $Y_1 = y_1, \ldots, Y_n = y_n$, and the sample mean, $\overline{y}$ and sample standard deviation, $s$ are obtained, the Confidence Interval for $\mu$ is formed as follows.

$$\overline{y} \pm t_{\alpha/2,n-1}s_{\overline{Y}} \qquad \equiv \qquad \overline{y} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

A 2-sided test of whether $\mu = \mu_0$ is set up as follows, where $TS$ is the test statistic, and $RR$ is the rejection region.

$$H_0 : \mu = \mu_0 \quad H_A : \mu \neq \mu_0 \qquad TS : t_{obs} = \sqrt{n}\frac{\overline{y}-\mu_0}{s} \qquad RR : |t_{obs}| \geq t_{\alpha/2,n-1}$$

with $P$-value $= 2P\left(t_{n-1} \geq |t_{obs}|\right)$.

To make inferences regarding $\sigma^2$, we will make use of the following notational convention.

$$W \sim \chi^2_\nu \quad \Rightarrow \quad P\left(W \geq \chi^2_{\alpha/2,\nu}\right) = \alpha/2$$

Since the $\chi^2$ distribution is not symmetric around 0, as Student's $t$ is, we will have to also obtain $\chi^2_{1-\alpha/2,\nu}$, representing the lower tail of the distribution having area=$\alpha/2$. Then, we can obtain a $(1-\alpha)100\%$ Confidence interval for $\sigma^2$, based on the following probability statements.

$$1-\alpha = P\left(\chi^2_{1-\alpha/2,\nu} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{\alpha/2,\nu}\right) = P\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2,\nu}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,\nu}}\right)$$

Based on the observed sample variance $s^2$, the Confidence Intervals for $\sigma^2$ and $\sigma$ are formed as follow.

$$\sigma^2 : \left(\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}\right) \qquad \sigma : \left(\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}}, \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}}\right)$$

To obtain a $(1-\alpha)100\%$ Confidence interval for $\sigma$, take the positive square roots of the end points. To test $H_0 : \sigma^2 = \sigma^2_0$ versus $H_A : \sigma^2 \neq \sigma^2_0$, simply check whether $\sigma^2_0$ lies in the confidence interval for $\sigma^2$.

## 1.4 Likelihood Functions and Maximum Likelihood Estimation

Suppose we take a random sample of $n$ items from a probability mass (discrete) or probability density (continuous) function. We can write the marginal probability density (mass) for the each observation (say $y_i$) as a function of one or more parameters ($\theta$):

$$\text{Discrete: } p\left(y_i|\theta\right) \qquad \text{Continuous: } f\left(y_i|\theta\right).$$

If the data are independent, then we get the joint density (mass) functions as the product of the individual (marginal) functions:

$$\text{Discrete: } p\left(y_1,\ldots,y_n|\theta\right)=\prod_{i=1}^{n}p\left(y_i|\theta\right) \qquad \text{Continuous: } f\left(y_1,\ldots,y_n|\theta\right)=\prod_{i=1}^{n}f\left(y_i|\theta\right).$$

Consider the following cases: Binomial, Poisson, Negative Binomial, Normal, Gamma, and Beta Distributions. For the binomial case, suppose we consider $n$ individual trials, where each trial can end in Success (with probability $\pi$) or Failure (with probability $1-\pi$). Note that each $y_i$ will equal 1 (S) or 0 (F). This is referred to as a **Bernoulli distribution** when each trial is considered individually.

$$p\left(y_i|\pi\right)=\pi^{y_i}\left(1-\pi\right)^{1-y_i} \quad \Rightarrow \quad p\left(y_1,\ldots,y_n|\pi\right)=\prod_{i=1}^{n}p\left(y_i|\pi\right)=\pi^{\sum y_i}\left(1-\pi\right)^{n-\sum y_i}$$

For the Poisson model, we have:

$$p\left(y_i|\lambda\right)=\frac{e^{-\lambda}\lambda^{y_i}}{y_i!} \quad \Rightarrow \quad p\left(y_1,\ldots,y_n|\lambda\right)=\prod_{i=1}^{n}p\left(y_i|\lambda\right)=\frac{e^{-n\lambda}\lambda^{\sum y_i}}{\prod y_i!}.$$

For the Negative Binomial distribution, we have:

$$f\left(y_i|\mu,\alpha\right)=\frac{\Gamma\left(\alpha^{-1}+y_i\right)}{\Gamma\left(\alpha^{-1}\right)\Gamma\left(y_i+1\right)}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{y_i}.$$

For the Normal distribution, we obtain:

$$f\left(y_i|\mu,\sigma^2\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\left(y_i-\mu\right)^2}{2\sigma^2}\right) \quad \Rightarrow$$

$$f\left(y_1,\ldots,y_n|\mu,\sigma^2\right)=\prod_{i=1}^{n}f\left(y_i|\mu,\sigma^2\right)=\left(2\pi\sigma^2\right)^{-n/2}\exp\left[-\frac{\sum\left(y_i-\mu\right)^2}{2\sigma^2}\right].$$

For the Gamma model, we have:

$$f\left(y_i|\alpha,\theta\right)=\frac{\theta^{\alpha}}{\Gamma(\alpha)}y_i^{\alpha-1}e^{-y\theta}.$$

For the Beta distribution, we have:

$$f\left(y_i|\alpha,\beta\right)=\frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}y_i^{\alpha-1}\left(1-y_i\right)^{\beta-1}$$

An alternative formulation of the distribution involves setting re-parameterizing as follows:

$$\mu=\frac{\alpha}{\alpha+\beta} \qquad \phi=\alpha+\beta \quad \Rightarrow \quad \alpha=\mu\phi \qquad \beta=\left(1-\mu\right)\phi$$

The re-parameterized model, and mean and variance are:

$$f(y_i;\mu,\phi)=\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}y_i^{\mu\phi-1}(1-y_i)^{(1-\mu)\phi-1}\qquad 0<\mu<1\quad \phi>0$$

$$E\{Y\}=\mu\qquad V\{Y\}=\frac{\mu(1-\mu)}{\phi+1}$$

Note that in each of these cases (and for other distributions as well), once we have collected the data, the joint distribution can be thought of as a function of unknown parameter(s). This is referred to as the **likelihood function**. Our goal is to choose parameter value(s) that maximize the likelihood function. These are referred to as **maximum likelihood estimators (MLEs)**. For most distributions, it is easier to maximize the log of the likelihood function, which is a monotonic function of the likelihood.

$$\text{Likelihood: } L(\theta|y_1,\ldots,y_n)=f(y_1,\ldots,y_n|\theta)\qquad \text{Log-Likelihood: } l=\ln(L)$$

To obtain the MLE(s), we take the derivative of the log-likelihood with respect to the parameter(s) $\theta$, set to zero, and solve for $\hat{\theta}$. Under commonly met regularity conditions, the maximum likelihood estimator $\hat{\theta}_{ML}$ is asymptotically normal, with mean equal to the true parameter(s) $\theta$, and variance (or variance-covariance matrix when the number of parameters, $p>1$) equal to:

$$V\left\{\hat{\theta}_{ML}\right\}=-\left(E\left\{\frac{\partial^2 l}{\partial\theta\partial\theta'}\right\}\right)^{-1}=I^{-1}(\theta)$$

where:

$$\theta=\begin{bmatrix}\theta_1\\ \vdots\\ \theta_p\end{bmatrix}\qquad \frac{\partial^2 l}{\partial\theta\partial\theta'}=\begin{bmatrix}\frac{\partial^2 l}{\partial\theta_1^2} & \cdots & \frac{\partial^2 l}{\partial\theta_1\partial\theta_p}\\ \vdots & \ddots & \vdots\\ \frac{\partial^2 l}{\partial\theta_p\partial\theta_1} & \cdots & \frac{\partial^2 l}{\partial\theta_p^2}\end{bmatrix}.$$

The estimated variance (or variance-covariance matrix) replaces the unknown parameter values $\theta$ with their ML estimates $\hat{\theta}_{ML}$. The **standard error** is the standard deviation of its sampling distribution, the square root of the variance.

We can construct approximate large-sample Confidence Intervals for the parameter(s) $\theta$, based on the asymptotic normality of the MLEs:

$$\hat{\theta}_{ML}\pm z_{\alpha/2}\sqrt{\hat{V}\left\{\hat{\theta}_{ML}\right\}}\qquad\qquad P\left\{Z\geq z_{\alpha/2}\right\}=\frac{\alpha}{2}.$$

Now, we consider the 6 distributions described above.

**Bernoulli/Binomial Distribution**

$$L(\pi|y_1,\ldots,y_n)=\pi^{\sum y_i}(1-\pi)^{n-\sum y_i}\quad\Rightarrow\quad l=\ln(L)=\sum y_i\ln(\pi)+\left(n-\sum y_i\right)\ln(1-\pi)$$

Taking the derivative of $l$ with respect to $\pi$, setting to 0, and solving for $\hat{\pi}$, we get:

$$\frac{\partial l}{\partial\pi}=\frac{\sum y_i}{\pi}-\frac{n-\sum y_i}{1-\pi}\overset{\text{set}}{=}0\quad\Rightarrow\quad\hat{\pi}=\frac{\sum y_i}{n}.$$

$$\Rightarrow \quad \frac{\partial^2 l}{\partial \pi^2} = -\frac{\sum y_i}{\pi^2} - \frac{n - \sum y_i}{(1-\pi)^2} \qquad \Rightarrow \qquad E\left\{\frac{\partial^2 l}{\partial \pi^2}\right\} = -\frac{n\pi}{\pi^2} - \frac{n(1-\pi)}{(1-\pi)^2} = -n\left(\frac{1}{\pi} + \frac{1}{1-\pi}\right) = -\frac{n}{\pi(1-\pi)}$$

$$\Rightarrow \quad V\{\hat{\pi}_{ML}\} = -\left(-\frac{n}{\pi(1-\pi)}\right)^{-1} = \frac{\pi(1-\pi)}{n} \quad \Rightarrow \quad \hat{V}\{\hat{\pi}_{ML}\} = \frac{\hat{\pi}(1-\hat{\pi})}{n}$$

**Example: WNBA Free Throw Shooting - Maya Moore**

We would like to estimate WNBA star Maya Moore's true probability of making a free throw, treating her 2014 season attempts as a random sample from her underlying population of all possible (in game) free throw attempts. We treat her individual attempts as independent Bernoulli trials with probability of success $\pi$. Further, we would like to test whether her underlying proportion is $\pi = \pi_0 = 0.80$ (80%). Over the course of the season, she attempted 181 free throws, and made 160 of them.

$$\hat{\pi} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{160}{181} = 0.884 \qquad \hat{V}(\hat{\pi}) = \frac{\hat{\pi}(1-\hat{\pi})}{n} = \frac{0.884(0.116)}{181} = .0005665$$

A 95% Confidence Interval for $\pi$ is:

$$\hat{\pi} \pm z_{.025}\sqrt{\hat{V}(\hat{\pi})} \quad \equiv \quad 0.884 \pm 1.96\sqrt{0.0005665} \quad \equiv \quad 0.884 \pm 0.047 \quad \equiv \quad (0.837, 0.931)$$

$\nabla$

**Poisson Distribution**

$$L(\lambda|y_1, \ldots, y_n) = \frac{e^{-n\lambda}\lambda^{\sum y_i}}{\prod y_i!} \quad \Rightarrow \quad l = \ln(L) = -n\lambda + \sum y_i \ln(\lambda) - \sum \ln(y_i!)$$

$$\frac{\partial l}{\partial \lambda} = -n + \frac{\sum y_i}{\lambda} \stackrel{\text{set}}{=} 0 \quad \Rightarrow \quad \hat{\lambda} = \frac{\sum y_i}{n}$$

$$\Rightarrow \quad \frac{\partial^2 l}{\partial \lambda^2} = -\frac{\sum y_i}{\lambda^2} \quad \Rightarrow \quad V\left\{\hat{\lambda}_{ML}\right\} = -\left(-\frac{n\lambda}{\lambda^2}\right)^{-1} = \frac{\lambda}{n} \qquad \Rightarrow \qquad \hat{V}\left\{\hat{\lambda}_{ML}\right\} = \frac{\hat{\lambda}}{n}$$

**Example: English Premier League Football Total Goals per Game - 2012/13 Season**

We are interested in estimating the population mean combined goals per game among the 2012/13 English Premier League (EPL) teams, based on the sample of games played in the season (380 total games). There are 20 teams, and each team plays each other team twice, one at Home, one Away. Assuming a Poisson model (which may not be reasonable, as different teams play in different games), we will estimate the underlying population mean $\lambda$. There were 380 games, with a total of 1063 goals, and sample mean and variance of 2.797 and 3.144, respectively. The number of goals and frequencies are given in Table 1.1.

| Goals | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Games | 35 | 61 | 72 | 91 | 64 | 32 | 13 | 9 | 1 | 0 | 2 |

Table 1.1: Frequency Tabulation for EPL 2012/2013 Total Goals per Game

| Goals | Observed | Expected(Poisson) | Expected(Neg Bin) |
|---|---|---|---|
| 0 | 35 | 23.41 | 27.29 |
| 1 | 61 | 65.24 | 67.74 |
| 2 | 72 | 90.92 | 87.90 |
| 3 | 91 | 84.46 | 79.34 |
| 4 | 64 | 58.85 | 55.94 |
| 5 | 32 | 32.80 | 32.81 |
| 6 | 13 | 15.24 | 16.65 |
| $\geq 7$ | 12 | 9.08 | 12.33 |

Table 1.2: Frequency Tabulation and Expected Counts for EPL 2012/2013 Total Goals per Game

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{1063}{380} = 2.797 \qquad \hat{V}\left(\hat{\lambda}\right) = \frac{\hat{\lambda}}{n} = \frac{2.797}{380} = 0.007361$$

A 95% Confidence Interval for $\lambda$ is:

$$\hat{\lambda} \pm z_{.025}\sqrt{\hat{V}\left(\hat{\lambda}\right)} \quad \equiv \quad 2.797 \pm 1.96\sqrt{0.007361} \quad \equiv \quad 2.797 \pm 0.168 \quad \equiv \quad (2.629, 2.965)$$

Table 1.2 gives the categories (goals), observed and expected counts, for the Poisson and Negative Binomial (next subsection) and the Chi-Square Goodness-of-fit tests for the two distributions. The goodness-of-fit test statistics are computed as follows, where $O_i$ is the **Observed** count for the $i^{th}$ category, $E_i$ is the **Expected** count for the $i^{th}$ category, and $N$ is the total number of observations.

$$\hat{\pi}_i = P\left(Y = i\right) = \frac{e^{\hat{\lambda}}\hat{\lambda}^i}{i!} \quad i = 0\ldots, 6 \qquad E_i = N \cdot \hat{\pi}_i \quad i = 0\ldots, 6 \qquad O_7 = N - \sum_{i=0}^{6} O_i \quad E_7 = N - \sum_{i=0}^{6} E_i$$

$$X^2_{\text{GOF}} = \sum_{i=0}^{7} \frac{(O_i - E_i)^2}{E_i}$$

The degrees of freedom for the Chi-Square Goodness-of-Fit test is the number of categories minus the number of estimated parameters. In the case of the EPL Total goals per game with a Poisson distribution, we have 8 categories $(0, 1, \ldots, 7^+)$ and one estimated parameter $(\lambda)$, for 8-1=7 degrees of freedom.

$$X^2_{\text{GOF-Poi}} = \frac{(35 - 23.41)^2}{23.41} + \cdots + \frac{(12 - 9.08)^2}{9.08} = 12.197 \qquad \chi^2\left(0.05, 7\right) = 14.067 \qquad P\left(\chi^2_7 \geq 12.197\right) = .0943$$

We fail to reject the hypothesis that total goals per game follows a Poisson distribution. We compare the Poisson model to the Negative Binomial model below.

$\nabla$

**Negative Binomial Distribution**

The probability distribution function for the Negative Binomial distribution can be written as (where $Y$ is the number of occurrences of an event):

$$f\left(y_i|\mu,\alpha\right)=\frac{\Gamma\left(\alpha^{-1}+y\right)}{\Gamma\left(\alpha^{-1}\right)\Gamma\left(y+1\right)}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{y}.$$

The mean and variance of this form of the Negative Binomial distribution are:

$$E\left\{Y\right\}=\mu \qquad V\left\{Y\right\}=\mu\left(1+\alpha\mu\right).$$

Note that $\mu$ and $\alpha^{-1}$ must be strictly positive, we re-parameterize so that $\alpha^{*}=\ln\alpha^{-1}$, $\mu^{*}=\ln\mu$. This way $\alpha^{*}$ and $\mu^{*}$ do not need to be positive in the estimation algorithm. The likelihood function for the $i^{th}$ observation is given below, which will be used to obtain estimators of parameters and their estimated variances.

$$L_i\left(\mu,\alpha\right)=f\left(y_i;\mu,\alpha\right)=\frac{\Gamma\left(\alpha^{-1}+y_i\right)}{\Gamma\left(\alpha^{-1}\right)\Gamma\left(y_i+1\right)}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{y_i}.$$

Note that, due to the recursive pattern of the $\Gamma\left(\cdot\right)$ function, we have:

$$\frac{\Gamma\left(\alpha^{-1}+y_i\right)}{\Gamma\left(\alpha^{-1}\right)\Gamma\left(y_i+1\right)}=\frac{\left(\alpha^{-1}+y_i-1\right)\left(\alpha^{-1}+y_i-2\right)\ldots\left(\alpha^{-1}\right)\Gamma\left(\alpha^{-1}\right)}{\Gamma\left(\alpha^{-1}\right)\Gamma\left(y_i+1\right)}=\frac{\left(\alpha^{-1}+y_i-1\right)\left(\alpha^{-1}+y_i-2\right)\ldots\left(\alpha^{-1}\right)}{y_i!}.$$

Taking the logarithm of the likelihood for the $i^{th}$ observation, and replacing $\alpha^{-1}$ with $e^{\alpha^{*}}$ and $\mu$ with $e^{\mu^{*}}$, we get:

$$l_i=\ln L_i\left(\mu,\alpha\right)=\sum_{j=0}^{y_i-1}\ln(\alpha^{-1}+j)-\ln\left(y_i!\right)+\alpha^{-1}\ln\alpha^{-1}+y_i\ln\mu-\left(\alpha^{-1}+y_i\right)\ln\left(\alpha^{-1}+\mu\right)=$$

$$\sum_{j=0}^{y_i-1}\ln\left(e^{\alpha^{*}}+j\right)-\ln\left(y_i!\right)+e^{\alpha^{*}}\ln e^{\alpha^{*}}+y_i\ln e^{\mu^{*}}-\left(e^{\alpha^{*}}+y_i\right)\ln\left(e^{\alpha^{*}}+e^{\mu^{*}}\right).$$

Taking the derivative of the log likelihood with respect to $\alpha^{*}=\ln\alpha^{-1}$, we obtain the following result.

$$\frac{\partial l_i}{\partial\alpha^{*}}=e^{\alpha^{*}}\left[\sum_{j=0}^{y_i-1}\frac{1}{e^{\alpha^{*}}+j}+\ln e^{\alpha^{*}}+1-\ln\left(e^{\alpha^{*}}+e^{\mu^{*}}\right)-\frac{e^{\alpha^{*}}+y_i}{e^{\alpha^{*}}+e^{\mu^{*}}}\right]$$

The derivative of the log likelihood with respect to $\mu^{*}=\ln\mu$ is

$$\frac{\partial l_i}{\partial\mu^{*}}=y_i-\frac{e^{\mu^{*}}\left(e^{\alpha^{*}}+y_i\right)}{e^{\alpha^{*}}+e^{\mu^{*}}}.$$

The second derivatives, used to obtain estimated variances are

$$\frac{\partial^2 l_i}{\partial\left(\alpha^{*}\right)^2}=e^{\alpha^{*}}\left[\sum_{j=0}^{y_i-1}\frac{1}{e^{\alpha^{*}}+j}+\ln e^{\alpha^{*}}+1-\ln\left(e^{\alpha^{*}}+e^{\mu^{*}}\right)-\frac{e^{\alpha^{*}}+y_i}{e^{\alpha^{*}}+e^{\mu^{*}}}-\sum_{j=0}^{y_i-1}\frac{e^{\alpha^{*}}}{\left(e^{\alpha^{*}}+j\right)^2}+1-\frac{e^{\alpha^{*}}}{e^{\alpha^{*}}+e^{\mu^{*}}}+\frac{e^{\alpha^{*}}\left(y_i-\mu\right)}{\left(e^{\alpha^{*}}+\mu\right)^2}\right]$$

$$\frac{\partial^2 l_i}{\partial\left(\mu^{*}\right)^2}=-\frac{e^{\alpha^{*}}e^{\mu^{*}}\left(e^{\alpha^{*}}+y_i\right)}{\left(e^{\alpha^{*}}+e^{\mu^{*}}\right)^2} \qquad \frac{\partial^2 l_i}{\partial\alpha^{*}\partial\mu^{*}}=\frac{e^{\alpha^{*}}e^{\mu^{*}}\left(y_i-e^{\mu^{*}}\right)}{\left(e^{\alpha^{*}}+e^{\mu^{*}}\right)^2}.$$

Once these have been computed (and note that they are functions of the unknown parameters), compute the following quantities.

$$g_{\alpha^*} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \alpha^*} \qquad g_{\mu^*} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \mu^*}$$

$$G_{\alpha^*} = \sum_{i=1}^{n} \frac{\partial^2 l_i}{\partial (\alpha^*)^2} \qquad G_{\mu^*} = \sum_{i=1}^{n} \frac{\partial^2 l_i}{\partial (\mu^*)^2}$$

We define $\theta$ as the vector containing elements $\mu^*$ and $\alpha^*$ and create a vector of first partial derivatives $g_\theta$, a matrix of second partial derivatives $G_\theta$.

$$\theta = \begin{bmatrix} \mu^* \\ \alpha^* \end{bmatrix} \qquad g_\theta = \begin{bmatrix} g_{\mu *} \\ g_{\alpha^*} \end{bmatrix} \qquad G_\theta = \begin{bmatrix} G_{\mu^*} & \sum_{i=1}^{n} \frac{\partial^2 l_i}{\partial \alpha^* \partial \mu^*} \\ \sum_{i=1}^{n} \frac{\partial^2 l_i}{\partial \alpha^* \partial \mu^*} & G_{\alpha^*} \end{bmatrix}$$

The Newton-Raphson algorithm is used to obtain (iterated) estimates of $\mu^*$ and $\alpha^*$. Then $\alpha^*$ is back-transformed to $\alpha^{-1}$ and $\mu^*$ to $\mu$.

In the first step, set $\alpha^* = 0$, which corresponds to $\alpha^{-1} = 1$, and obtain an iterated estimate of $\mu^*$. A reasonable starting value for $\mu^*$ is $\ln\left(\overline{Y}\right)$.

$$\tilde{\mu^*}^{(k)} = \tilde{\mu^*}^{(k-1)} - \frac{g_{\tilde{\mu^*}^{(k-1)}}}{G_{\tilde{\mu^*}^{(k-1)}}}$$

Iterate to convergence. In the second step, use your estimate of $\mu^*$, and obtain an iterated estimate of $\alpha^*$. A reasonable starting value is $\ln\left[\exp\left(\hat{\mu}^*\right)^2 / \left(\exp\left(\hat{\mu}^*\right) + s_y^2\right)\right]$, where $s_y^2$ is the sample variance of $y_1, \ldots, y_n$.

$$\tilde{\alpha^*}^{(k)} = \tilde{\alpha^*}^{(k-1)} - \frac{g_{\tilde{\alpha^*}^{(k-1)}}}{G_{\tilde{\alpha^*}^{(k-1)}}}$$

In the final step, we use the estimates from steps 1 and 2 as starting values to get a combined estimate and estimated variance for $\theta$:

$$\tilde{\theta}^{(k)} = \tilde{\theta}^{(k-1)} - \left[G_{\tilde{\theta}^{(k-1)}}\right]^{-1} g_{\tilde{\theta}^{(k-1)}}$$

After iterating to convergence, we get:

$$\hat{V}\left\{\hat{\theta}\right\} = -\left[E\left\{G_{\hat{\theta}}\right\}\right]^{-1} \qquad \hat{\alpha^{-1}} = e^{\hat{\alpha^*}} \qquad \hat{\mu} = e^{\hat{\mu}^*}$$

$$\hat{V}\left\{\hat{\alpha^{-1}}\right\} = \hat{V}\left\{\hat{\alpha^*}\right\}\left(\frac{\partial \hat{\alpha^{-1}}}{\partial \hat{\alpha^*}}\right)^2 \qquad \hat{V}\left\{\hat{\mu}\right\} = \hat{V}\left\{\hat{\mu}^*\right\}\left(\frac{\partial \hat{\mu}}{\partial \hat{\mu}^*}\right)^2$$

**Example: English Premier League Football Total Goals per Game - 2012/13 Season**

For this data, we obtain estimates of $\mu$ and $\alpha^{-1}$ as follows:

$$\hat{\theta} = \begin{bmatrix} \hat{\mu}^* \\ \hat{\alpha}^* \end{bmatrix} = \begin{bmatrix} 1.02868 \\ 3.09385 \end{bmatrix} \qquad \hat{V}\left\{\hat{\theta}\right\} = -E\left\{\left[G_{\hat{\theta}}\right]^{-1}\right\} = \begin{bmatrix} 0.001060 & 0.000000 \\ 0.000000 & 0.453907 \end{bmatrix}$$

$$\hat{\mu} = e^{\hat{\mu}^*} = e^{1.02868} = 2.79737 \qquad \hat{V}\left\{\hat{\mu}\right\} = \hat{V}\left\{\hat{\mu}^*\right\}\left(e^{\hat{\mu}^*}\right)^2 = 0.001060\,(2.79737)^2 = 0.00829 \quad \hat{\sigma}_{\hat{\mu}} = \sqrt{0.00829} = 0.0911$$

$$\hat{\alpha^{-1}} = e^{\hat{\alpha^*}} = e^{3.0939} = 22.0619 \qquad \hat{V}\left\{\hat{\alpha}^{-1}\right\} = \hat{V}\left\{\hat{\alpha}^*\right\}\left(e^{\hat{\alpha}^*}\right)^2 = 0.453907\,(22.0619)^2 = 220.928$$

The goodness-of-fit test statistics are computed as follows, where $O_i$ is the **Observed** count for the $i^{th}$ category, $E_i$ is the **Expected** count for the $i^{th}$ category, and $N$ is the total number of observations:

$$\hat{\pi}_i = P\left(Y=i\right) = \frac{\Gamma\left(\hat{\alpha}^{-1}+i\right)}{\Gamma\left(\hat{\alpha}^{-1}\right)\Gamma\left(i+1\right)}\left(\frac{\hat{\alpha}^{-1}}{\hat{\mu}+\hat{\alpha}^{-1}}\right)^{\hat{\alpha}^{-1}}\left(\frac{\hat{\mu}}{\hat{\mu}+\hat{\alpha}^{-1}}\right)^{i} \quad i=0\ldots,6$$

$$E_i = N\cdot\hat{\pi}_i \quad i=0\ldots,6 \qquad O_7 = N-\sum_{i=0}^{6}O_i \quad E_7 = N-\sum_{i=0}^{6}E_i$$

The degrees of freedom for the Chi-Square Goodness-of-Fit test is the number of categories minus the number of estimated parameters. In the case of the EPL Total goals per game with a Negative Binomial distribution, we have 8 categories $(0,1,\ldots,7^+)$ and two estimated parameters $\left(\alpha^{-1},\mu\right)$, for 8-2=6 degrees of freedom. The expected counts are given in Table 1.2.

$$X^2_{\text{GOF-NB}} = \frac{(35-27.29)^2}{27.29}+\cdots+\frac{(12-12.33)^2}{12.33} = 9.435 \qquad \chi^2\left(0.05,6\right) = 12.592 \qquad P\left(\chi^2_6 \geq 12.197\right) = .1506$$

$\nabla$

**Normal Distribution**

$$L\left(\theta|y_1,\ldots,y_n\right) = \left(2\pi\sigma^2\right)^{-n/2}\exp\left[-\frac{\sum\left(y_i-\mu\right)^2}{2\sigma^2}\right] \quad \Rightarrow \quad l = \ln(L) = -\frac{n}{2}\left[\ln(2\pi)+\ln\left(\sigma^2\right)\right]-\frac{\sum\left(y_i-\mu\right)^2}{2\sigma^2}$$

$$\frac{\partial l}{\partial\mu} = \frac{\sum\left(y_i-\mu\right)}{\sigma^2} \overset{\text{set}}{=} 0 \quad \Rightarrow \quad \hat{\mu} = \frac{\sum y_i}{n}$$

$$\frac{\partial l}{\partial\sigma^2} = -\frac{n}{2\sigma^2}+\frac{\sum\left(y_i-\mu\right)^2}{2\sigma^4} \overset{\text{set}}{=} 0 \quad \Rightarrow \quad \hat{\sigma}^2 = \frac{\sum\left(y_i-\hat{\mu}\right)^2}{n}.$$

For the normal model, we have, where $E\left\{Y_i\right\} = \mu$ and $E\left\{\left(Y_i-\mu\right)^2\right\} = \sigma^2$:

$$l = -\frac{n}{2}\left[\ln(2\pi)+\ln\left(\sigma^2\right)\right]-\frac{\sum\left(y_i-\mu\right)^2}{2\sigma^2}$$

$$\frac{\partial l}{\partial\mu} = \frac{\sum\left(y_i-\mu\right)}{\sigma^2}$$

$$\frac{\partial l}{\partial\sigma^2} = -\frac{n}{2\sigma^2}+\frac{\sum\left(y_i-\mu\right)^2}{2\sigma^4}$$

$$\frac{\partial^2 l}{\partial \mu^2} = -\frac{n}{\sigma^2} \qquad \Rightarrow \qquad E\left\{\frac{\partial^2 l}{\partial \mu^2}\right\} = -\frac{n}{\sigma^2}$$

$$\frac{\partial^2 l}{\partial \left(\sigma^2\right)^2} = \frac{n}{2\sigma^4} - \frac{2\sum\left(y_i-\mu\right)^2}{2\sigma^6} \quad \Rightarrow \quad E\left\{\frac{\partial^2 l}{\partial \left(\sigma^2\right)^2}\right\} = \frac{n}{2\sigma^4} - \frac{2n\sigma^2}{2\sigma^6} = -\frac{n}{2\sigma^4}$$

$$\frac{\partial^2 l}{\partial \mu \partial \sigma^2} = -\frac{\sum\left(y_i-\mu\right)}{\sigma^4} \qquad \Rightarrow \qquad E\left\{\frac{\partial^2 l}{\partial \mu \partial \sigma^2}\right\} = 0$$

$$\Rightarrow \quad V\left\{\begin{bmatrix} \hat{\mu}_{ML} \\ \hat{\sigma}^2_{ML} \end{bmatrix}\right\} = -\begin{bmatrix} -\frac{n}{\sigma^2} & 0 \\ 0 & -\frac{n}{2\sigma^4} \end{bmatrix}^{-1} = \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$

$$\Rightarrow \quad \hat{V}\left\{\begin{bmatrix} \hat{\mu}_{ML} \\ \hat{\sigma}^2_{ML} \end{bmatrix}\right\} = \begin{bmatrix} \frac{\hat{\sigma}^2}{n} & 0 \\ 0 & \frac{2\hat{\sigma}^4}{n} \end{bmatrix}$$

$$\Rightarrow \quad V\left\{\hat{\mu}_{ML}\right\} = \frac{\sigma^2}{n} \qquad \Rightarrow \qquad \hat{V}\left\{\hat{\mu}_{ML}\right\} = \frac{\hat{\sigma}^2}{n}$$

$$\Rightarrow \quad V\left\{\hat{\sigma}^2_{ML}\right\} = \frac{2\sigma^4}{n} \qquad \Rightarrow \qquad \hat{V}\left\{\hat{\sigma}^2_{ML}\right\} = \frac{2\hat{\sigma}^4}{n}.$$

Also note that the covariance of $\hat{\mu}$ and $\hat{\sigma}^2$ is zero.

**Gamma Distribution**

The Gamma distribution can be used to model continuous random variables that take on positive values and have long right tails (skewed). There are several parameterizations of the model, this version is based on $\alpha$ being a shape parameter and $\beta$ being a rate parameter.

$$f\left(y;\alpha,\beta\right) = \frac{\beta^\alpha}{\Gamma\left(\alpha\right)} y^{\alpha-1} e^{-\beta y} \qquad y>0; \quad \alpha,\beta>0 \qquad \int_0^\infty w^a e^{-bw} dw = \frac{\Gamma\left(a+1\right)}{b^{a+1}}$$

The mean and variance of $Y$ can be obtained as follows:

$$E\left\{Y\right\} = \frac{\beta^\alpha}{\Gamma\left(\alpha\right)} \int_0^\infty y^\alpha e^{-\beta y} dy = \frac{\beta^\alpha}{\Gamma\left(\alpha\right)} \frac{\Gamma\left(\alpha+1\right)}{\beta^{\alpha+1}} = \frac{\alpha}{\beta}$$

$$E\left\{Y^2\right\} = \frac{\beta^\alpha}{\Gamma\left(\alpha\right)} \int_0^\infty y^{\alpha+1} e^{-\beta y} dy = \frac{\beta^\alpha}{\Gamma\left(\alpha\right)} \frac{\Gamma\left(\alpha+2\right)}{\beta^{\alpha+2}} = \frac{\alpha\left(\alpha+1\right)}{\beta^2} \quad \Rightarrow \quad V\left\{Y\right\} = \frac{\alpha\left(\alpha+1\right)}{\beta^2} - \left(\frac{\alpha}{\beta}\right)^2 = \frac{\alpha}{\beta^2}$$

An alternative parameterization for the Gamma distribution is:

$$L_i\left(\mu,\phi;y_i\right) = f\left(y_i;\mu,\phi\right) = \frac{1}{y_i \Gamma\left(\frac{1}{\phi}\right)} \left(\frac{y_i}{\mu\phi}\right)^{1/\phi} e^{-\frac{y_i}{\mu\phi}} \qquad \alpha = \frac{1}{\phi} \qquad \beta = \frac{1}{\mu\phi}$$

$$\Rightarrow \quad E\left\{Y\right\} = \frac{\alpha}{\beta} = \mu \qquad V\left\{Y\right\} = \frac{\alpha}{\beta^2} = \mu^2\phi$$

Since $\mu$ and $\phi$, must be positive, we transform to $\mu^*$ and $\phi^*$:

$$\mu^* = -\ln\left(\mu\right) \qquad \phi^* = -\ln\left(\phi\right) \quad \Rightarrow \quad \mu = e^{-\mu^*} \qquad \phi = e^{-\phi^*}$$

$$L_i\left(\mu^*,\phi^*;y_i\right)=f\left(y_i;\mu^*,\phi^*\right)=\frac{1}{y_i\Gamma\left(e^{\phi^*}\right)}\left(y_ie^{\mu^*}e^{\phi^*}\right)^{e^{\phi^*}}\ e^{-y_ie^{\mu^*}e^{\phi^*}}$$

The log Likelihood for the $i^{th}$ observation is:

$$l_i=-\ln\left(y_i\right)-\ln\Gamma\left(e^{\phi^*}\right)+e^{\phi^*}\left[\ln\left(y_i\right)+\mu^*+\phi^*\right]-y_ie^{\mu^*}e^{\phi^*}$$

The relevant derivatives to obtain maximum likelihood estimates and their estimated variance-covariance matrix are given below, with the following definition:

$$\psi\left(w\right)=\frac{d\ln\Gamma\left(w\right)}{dw}$$

$$\frac{\partial l_i}{\partial\mu^*}=e^{\phi^*}-y_ie^{\mu^*}e^{\phi^*}$$

$$\frac{\partial l_i}{\partial\phi^*}=-\psi\left(e^{\phi^*}\right)e^{\phi^*}+e^{\phi^*}\left[\ln\left(y_i\right)+\mu^*+\phi^*\right]+e^{\phi^*}-y_ie^{\mu^*}e^{\phi^*}$$

Setting each derivative equal to 0 and summing over all observations gives:

$$\sum_{i=1}^{n}y_i=ne^{-\mu^*}\qquad n\psi\left(e^{\phi^*}\right)=n+\sum_{i=1}^{n}\ln\left(y_i\right)+n\mu^*+n\phi^*-e^{\mu^*}\sum_{i=1}^{n}y_i$$

The second derivatives are:

$$\frac{\partial^2 l_i}{\partial\mu^*\partial\phi^*}=e^{\phi^*}-y_ie^{\mu^*}e^{\phi^*}\quad\Rightarrow\quad E\left\{\sum_{i=1}^{n}\frac{\partial^2 l_i}{\partial\mu^*\partial\phi^*}\right\}=ne^{\phi^*}-ne^{-\mu^*}e^{\mu^*}e^{\phi^*}=0$$

$$\frac{\partial^2 l_i}{\partial\left(\mu^*\right)^2}=-y_ie^{\mu^*}e^{\phi^*}\quad\Rightarrow\quad E\left\{\sum_{i=1}^{n}\frac{\partial^2 l_i}{\partial\left(\mu^*\right)^2}\right\}=-ne^{-\mu^*}e^{\mu^*}e^{\phi^*}=-ne^{\phi^*}$$

$$\frac{\partial^2 l_i}{\partial\left(\phi^*\right)^2}=-\psi'\left(e^{\phi^*}\right)\left(e^{\phi^*}\right)^2-\psi\left(e^{\phi^*}\right)e^{\phi^*}+e^{\phi^*}\left[\ln\left(y_i\right)+\mu^*+\phi^*\right]+2e^{\phi^*}-y_ie^{\mu^*}e^{\phi^*}$$

$$\Rightarrow\quad E\left\{\sum_{i=1}^{n}\frac{\partial^2 l_i}{\partial\left(\phi^*\right)^2}\right\}=-n\psi'\left(e^{\phi^*}\right)\left(e^{\phi^*}\right)^2+ne^{\phi^*}=ne^{\phi^*}\left[1-\psi'\left(e^{\phi^*}\right)e^{\phi^*}\right]$$

The estimation process begins with choosing starting values for $\mu^*$ and $\phi^*$.

$$E\left\{Y\right\}=\mu=e^{-\mu^*}\quad\Rightarrow\quad\tilde{\mu}^{*(0)}=-\ln\left(\overline{Y}\right)$$

$$V\left\{Y\right\}=\sigma_Y^2=\mu^2\phi=\left(e^{-\mu^*}\right)^2e^{-\phi^*}\quad\Rightarrow\quad e^{-\phi^*}=\frac{\sigma^2}{\mu^2}\quad\Rightarrow\quad\tilde{\phi}^{*(0)}=2\ln\overline{Y}-\ln s_Y^2$$

The Newton-Raphson algorithm is obtained as follows:

$$\theta^*=\begin{bmatrix}\mu^*\\ \phi^*\end{bmatrix}\qquad g_{\theta^*}=\begin{bmatrix}\sum_{i=1}^{n}\frac{\partial l_i}{\partial\mu^*}\\ \sum_{i=1}^{n}\frac{\partial l_i}{\partial\phi^*}\end{bmatrix}$$

$$G_{\theta^*} = \begin{bmatrix} \sum_{i=1}^n \frac{\partial^2 l_i}{\partial(\mu^*)^2} & \sum_{i=1}^n \frac{\partial^2 l_i}{\partial\mu^*\partial\phi^*} \\ \sum_{i=1}^n \frac{\partial^2 l_i}{\partial\mu^*\partial\phi^*} & \sum_{i=1}^n \frac{\partial^2 l_i}{\partial(\phi^*)^2} \end{bmatrix}$$

We then iterate to convergence:

$$\tilde{\theta}^{*(k)} = \tilde{\theta}^{*(k-1)} - \left[G_{\tilde{\theta}^{*(k-1)}}\right]^{-1} g_{\tilde{\theta}^{*(k-1)}}$$

After iterating to convergence, we obtain the following variances for the ML estimates, and their back-transformed values.

$$\hat{V}\left\{\hat{\theta}^*\right\} = -\left[E\left\{G_{\hat{\theta}^*}\right\}\right]^{-1} = -\begin{bmatrix} -ne^{\hat{\phi}^*} & 0 \\ 0 & ne^{\hat{\phi}^*}\left(1-\psi'\left(e^{\hat{\phi}^*}\right)e^{\hat{\phi}^*}\right) \end{bmatrix}^{-1} = \begin{bmatrix} ne^{\hat{\phi}^*}\left(\psi'\left(e^{\hat{\phi}^*}\right)e^{\hat{\phi}^*}-1\right) & 0 \\ 0 & ne^{\hat{\phi}^*} \end{bmatrix}$$

$$\hat{\mu} = e^{-\hat{\mu}^*} \qquad \hat{\phi} = e^{-\hat{\phi}^*} \qquad \hat{V}\{\hat{\mu}\} = \hat{V}\{\hat{\mu}^*\}\left(\frac{\partial\hat{\mu}}{\partial\hat{\mu}^*}\right)^2 \qquad \hat{V}\left\{\hat{\phi}\right\} = \hat{V}\left\{\hat{\phi}^*\right\}\left(\frac{\partial\hat{\phi}}{\partial\hat{\phi}^*}\right)^2$$

Note that there is an alternative means of estimating $\phi$, based on the **Method of Moments**, that some software packages, such as R use. The method makes use of the following results. In this setting, all observations have the same mean $\mu$, in regression models described later, they will have individual means, based on one or more covariates.

$$E\{Y_i\} = \mu \quad \Rightarrow \quad E\left\{\frac{Y_i}{\mu}\right\} = 1 \qquad V\{Y_i\} = \phi\mu^2 \quad \Rightarrow \quad V\left\{\frac{Y_i}{\mu}\right\} = \phi$$

$$E\left\{\frac{Y_i}{\mu}-1\right\} = E\left\{\frac{Y_i-\mu}{\mu}\right\} = 0 \qquad V\left\{\frac{Y_i}{\mu}-1\right\} = V\left\{\frac{Y_i-\mu}{\mu}\right\} = V\left\{\frac{Y_i}{\mu}\right\} = \phi$$

$$\Rightarrow \quad E\left\{\left(\frac{Y_i}{\mu}-1\right)^2\right\} = E\left\{\left(\frac{Y_i-\mu}{\mu}\right)^2\right\} = \phi$$

$$\tilde{\phi} = \frac{1}{n-1}\sum_{i=1}^n \left(\frac{Y_i-\hat{\mu}}{\hat{\mu}}\right)^2$$

The $n-1$ represents that 1 mean parameter has been estimated ($\mu$).

**Example: Running Speeds Among Females at a Marathon**

The running speeds (miles per hour) among $n = 1045$ females who completed the Rock and Roll Marathon in Washington are all positive and are seen to be skewed right. The histogram and corresponding gamma distribution are shown in Figure 1.8. Here we are treating these times as a random sample of times from a larger conceptual population.

The ML estimates and estimated variance-covariance matrix are given below. The mean and variance of the speeds are 5.839839 and 0.6906284, respectively. This leads to starting values of $-\ln(5.839839) = 1.7647$ for $\mu^*$ and $2\ln(5.839839) - \ln(0.6906284) = 3.8996$ for $\phi^*$.

$$\hat{\theta}^* = \begin{bmatrix} \hat{\mu}^* \\ \hat{\phi}^* \end{bmatrix} = \begin{bmatrix} -1.764703 \\ 3.936980 \end{bmatrix} \qquad \hat{\theta} = \begin{bmatrix} e^{-\hat{\mu}^*} \\ e^{-\hat{\phi}^*} \end{bmatrix} = \begin{bmatrix} 5.839838 \\ 0.019507 \end{bmatrix}$$

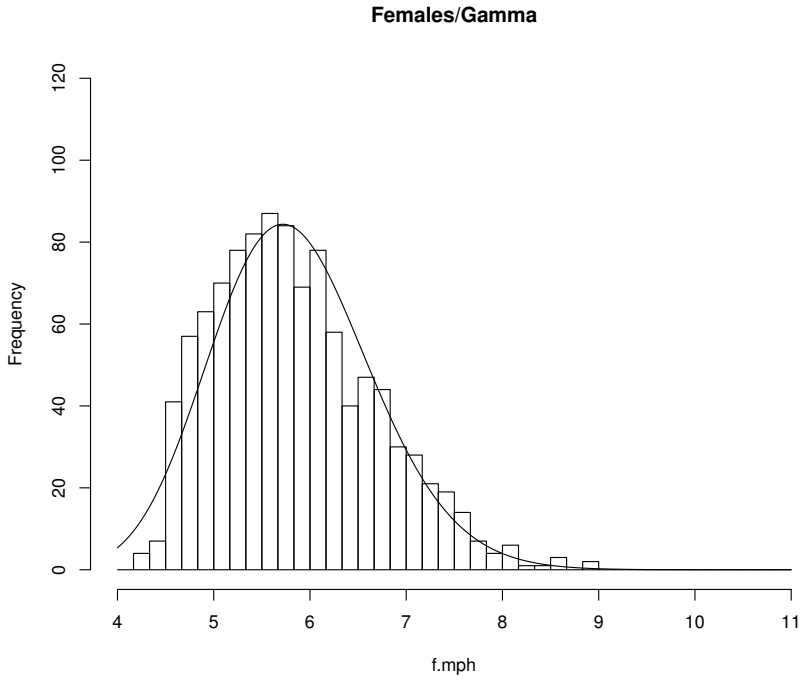


Figure 1.8: Histogram of female speeds and Gamma density - 2015 Rock and Roll marathon

$$\hat{V}\left\{\hat{\theta}^*\right\} = \begin{bmatrix} 0.00001866702 & 0 \\ 0 & 0.001901512 \end{bmatrix}$$

$$\hat{V}\left\{\hat{\mu}\right\} = 0.00001866702\left(-e^{-\hat{\mu}^*}\right)^2 = 0.00001866702\left(-5.839838\right)^2 = 0.000636615$$

An approximate 95% Confidence Interval for $\mu$ is

$$5.839838 \pm 1.96\sqrt{0.000636615} \equiv 5.839838 \pm 1.96\left(0.025231\right) \equiv 5.839838 \pm 0.049453 \equiv \left(5.790385, 5.889291\right).$$

The method of moments estimator for $\phi$ is computed below.

$$\tilde{\phi} = \frac{1}{1045-1}\sum_{i=1}^{1045}\left(\frac{y_i - 5.839838}{5.839838}\right)^2 = 0.02025082$$

Returning to the "original" form of the Gamma distribution with $E\left\{Y\right\} = \mu = \frac{\alpha}{\beta}$ and $V\left\{Y\right\} = \mu^2\phi = \frac{\alpha}{\beta^2}$, we obtain (based on the ML estimator for $\mu$ and the moments estimator of $\phi$)

$$\hat{\alpha} = \frac{1}{\tilde{\phi}} = \frac{1}{0.02025082} = 49.38 \qquad \hat{\beta} = \frac{1}{\hat{\mu}\tilde{\phi}} = \frac{1}{5.839838(0.02025082)} = 8.46$$

$\nabla$

**Beta Distribution**

The Beta distribution can be used to model data that are proportions (or percentages divided by 100). The traditional model for the Beta distribution is:

$$f(y;\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}y^{\alpha-1}(1-y)^{\beta-1}\qquad 0<y<1;\quad \alpha>0,\beta>0$$

$$E\{Y\}=\frac{\alpha}{\alpha+\beta}\qquad V\{Y\}=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

The alternative formulation of the distribution involves setting re-parameterizing as follows:

$$\mu=\frac{\alpha}{\alpha+\beta}\qquad \phi=\alpha+\beta\quad\Rightarrow\quad \alpha=\mu\phi\qquad \beta=(1-\mu)\phi.$$

The re-parameterized model, and mean and variance are:

$$L_i(\mu,\phi;y_i)=f(y_i|\mu,\phi)=\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}y_i^{\mu\phi-1}(1-y_i)^{(1-\mu)\phi-1}\qquad 0<\mu<1\quad\phi>0$$

$$E\{Y\}=\mu\qquad V\{Y\}=\frac{\mu(1-\mu)}{\phi+1}$$

To estimate the parameters by maximum likelihood, we make the following transformations, so that the estimated parameters have no restrictions on being positive or bounded between 0 and 1.

$$\mu=\frac{e^\gamma}{1+e^\gamma}\quad\Rightarrow\quad\gamma=\ln\left(\frac{\mu}{1-\mu}\right)\qquad\qquad \phi=e^{\phi^*}\quad\Rightarrow\quad\phi^*=\ln(\phi)$$

This leads to the log-likelihood function:

$$l_i=\ln L_i=\ln\Gamma(\phi)-\ln\Gamma(\mu\phi)-\ln\Gamma((1-\mu)\phi)+(\mu\phi-1)\ln y_i+((1-\mu)\phi-1)\ln(1-y_i)=$$

$$=\ln\Gamma\left(e^{\phi^*}\right)-\ln\Gamma\left(\frac{e^\gamma}{1+e^\gamma}e^{\phi^*}\right)-\ln\Gamma\left(\frac{1}{1+e^\gamma}e^{\phi^*}\right)+\left(\frac{e^\gamma}{1+e^\gamma}e^{\phi^*}-1\right)\ln y_i+\left(\frac{1}{1+e^\gamma}e^{\phi^*}-1\right)\ln(1-y_i).$$

The relevant derivatives to obtain maximum likelihood estimates and their estimated variance-covariance matrix are given below, with the following definitions.

$$\psi(w)=\frac{d\ln\Gamma(w)}{dw}\quad y_i^*=\ln y_i-\ln(1-y_i)=\ln\left(\frac{y_i}{1-y_i}\right)\quad \mu^*=\psi\left(\frac{e^\gamma}{1+e^\gamma}e^{\phi^*}\right)-\psi\left(\frac{1}{1+e^\gamma}e^{\phi^*}\right)$$

$$\frac{\partial l_i}{\partial\phi^*}=\psi\left(e^{\phi^*}\right)e^{\phi^*}-\psi\left(\frac{e^\gamma}{1+e^\gamma}e^{\phi^*}\right)\frac{e^\gamma}{1+e^\gamma}e^{\phi^*}-\psi\left(\frac{1}{1+e^\gamma}e^{\phi^*}\right)\frac{1}{1+e^\gamma}e^{\phi^*}+\frac{e^\gamma}{1+e^\gamma}e^{\phi^*}\ln y_i+\frac{1}{1+e^\gamma}e^{\phi^*}\ln(1-y_i)$$

Setting the derivative equal to 0 and summing over all observations gives:

$$\left(\frac{e^\gamma}{1+e^\gamma}e^{\phi^*}\right)\sum_{i=1}^n\ln y_i+\left(\frac{1}{1+e^\gamma}e^{\phi^*}\right)\sum_{i=1}^n\ln(1-y_i)=$$

$$= n\psi\left(\frac{e^{\gamma}}{1+e^{\gamma}}e^{\phi^*}\right)\frac{e^{\gamma}}{1+e^{\gamma}}e^{\phi^*} + n\psi\left(\frac{1}{1+e^{\gamma}}e^{\phi^*}\right)\frac{1}{1+e^{\gamma}}e^{\phi^*} - n\psi\left(e^{\phi^*}\right)e^{\phi^*}.$$

$$\frac{\partial l_i}{\partial \gamma} = \frac{e^{\gamma}}{\left(1+e^{\gamma}\right)^2}e^{\phi^*}\left(y_i^* - \mu^*\right)$$

Setting the derivative equal to 0 and summing over all observations gives:

$$\sum_{i=1}^{n} y_i^* = n\mu^*.$$

$$\frac{\partial^2 l_i}{\partial\gamma\partial\phi^*} = \frac{e^{\gamma}}{\left(1+e^{\gamma}\right)^2}e^{\phi^*}\left(y_i^* - \mu^*\right) + \frac{e^{\gamma}}{\left(1+e^{\gamma}\right)^3}\left(e^{\phi^*}\right)^2\psi'\left(\frac{1}{1+e^{\gamma}}e^{\phi^*}\right) - \frac{\left(e^{\gamma}\right)^2}{\left(1+e^{\gamma}\right)^3}\left(e^{\phi^*}\right)^2\psi'\left(\frac{e^{\gamma}}{1+e^{\gamma}}e^{\phi^*}\right)$$

$$\frac{\partial^2 l}{\partial\gamma\partial\phi^*} = n\left[\frac{e^{\gamma}}{\left(1+e^{\gamma}\right)^3}\left(e^{\phi^*}\right)^2\psi'\left(\frac{1}{1+e^{\gamma}}e^{\phi^*}\right) - \frac{\left(e^{\gamma}\right)^2}{\left(1+e^{\gamma}\right)^3}\left(e^{\phi^*}\right)^2\psi'\left(\frac{e^{\gamma}}{1+e^{\gamma}}e^{\phi^*}\right)\right]$$

$$\frac{\partial^2 l_i}{\partial\left(\phi^*\right)^2} = \psi\left(e^{\phi^*}\right)e^{\phi^*} + \psi'\left(e^{\phi^*}\right)\left(e^{\phi^*}\right)^2 - \psi\left(\frac{e^{\gamma}}{1+e^{\gamma}}e^{\phi^*}\right)\frac{e^{\gamma}}{1+e^{\gamma}}e^{\phi^*} - \psi'\left(\frac{e^{\gamma}}{1+e^{\gamma}}e^{\phi^*}\right)\left(\frac{e^{\gamma}}{1+e^{\gamma}}e^{\phi^*}\right)^2$$
$$-\psi\left(\frac{1}{1+e^{\gamma}}e^{\phi^*}\right)\frac{1}{1+e^{\gamma}}e^{\phi^*} - \psi'\left(\frac{1}{1+e^{\gamma}}e^{\phi^*}\right)\left(\frac{1}{1+e^{\gamma}}e^{\phi^*}\right)^2 + \frac{e^{\gamma}}{1+e^{\gamma}}e^{\phi^*}\ln y_i + \frac{1}{1+e^{\gamma}}e^{\phi^*}\ln\left(1-y_i\right)$$

$$\frac{\partial^2 l}{\partial\left(\phi^*\right)^2} = n\left[\psi'\left(e^{\phi^*}\right)\left(e^{\phi^*}\right)^2 - \psi'\left(\frac{e^{\gamma}}{1+e^{\gamma}}e^{\phi^*}\right)\left(\frac{e^{\gamma}}{1+e^{\gamma}}e^{\phi^*}\right)^2 - \psi'\left(\frac{1}{1+e^{\gamma}}e^{\phi^*}\right)\left(\frac{1}{1+e^{\gamma}}e^{\phi^*}\right)^2\right]$$

$$\frac{\partial^2 l_i}{\partial\gamma^2} = \frac{e^{\gamma}e^{\phi^*}\left(1-e^{\gamma}\right)}{\left(1+e^{\gamma}\right)^3}\left(y_i^* - \mu^*\right) - \left[\psi'\left(\frac{e^{\gamma}}{1+e^{\gamma}}e^{\phi^*}\right) + \psi'\left(\frac{1}{1+e^{\gamma}}e^{\phi^*}\right)\right]\left(\frac{e^{\gamma}}{1+e^{\gamma}}\frac{1}{1+e^{\gamma}}e^{\phi^*}\right)^2$$

$$\frac{\partial^2 l}{\partial\gamma^2} = -n\left[\psi'\left(\frac{e^{\gamma}}{1+e^{\gamma}}e^{\phi^*}\right) + \psi'\left(\frac{1}{1+e^{\gamma}}e^{\phi^*}\right)\right]\left(\frac{e^{\gamma}}{1+e^{\gamma}}\frac{1}{1+e^{\gamma}}e^{\phi^*}\right)^2$$

The estimation process begins with choosing starting values for $\phi^*$ and $\gamma$.

$$E\{Y\} = \mu = \frac{e^{\gamma}}{1+e^{\gamma}} \quad \Rightarrow \quad \gamma = \ln\left(\frac{\mu}{1-\mu}\right) \quad \Rightarrow \quad \tilde{\gamma}^{(0)} = \ln\left(\frac{\overline{Y}}{1-\overline{Y}}\right)$$

$$V\{Y\} = \sigma_Y^2 = \frac{\mu(1-\mu)}{\phi+1} = \frac{\mu(1-\mu)}{e^{\phi^*}+1} \quad \Rightarrow \quad e^{\phi^*} = \frac{\mu(1-\mu)}{\sigma^2} - 1 \quad \Rightarrow \quad \tilde{\phi}^{*(0)} = \ln\left(\frac{\overline{Y}(1-\overline{Y})}{s_Y^2} - 1\right)$$

The Newton-Raphson algorithm is obtained as follows:

$$\theta = \begin{bmatrix} \gamma \\ \phi^* \end{bmatrix} \qquad g_\theta = \begin{bmatrix} \sum_{i=1}^n \frac{\partial l_i}{\partial \gamma} \\ \sum_{i=1}^n \frac{\partial l_i}{\partial \phi^*} \end{bmatrix}$$

$$G_\theta = \begin{bmatrix} \sum_{i=1}^n \frac{\partial^2 l_i}{\partial \gamma^2} & \sum_{i=1}^n \frac{\partial^2 l_i}{\partial \gamma \partial \phi^*} \\ \sum_{i=1}^n \frac{\partial^2 l_i}{\partial \gamma \partial \phi^*} & \sum_{i=1}^n \frac{\partial^2 l_i}{\partial (\phi^*)^2} \end{bmatrix}$$

We then iterate to convergence:

$$\tilde{\theta}^{(k)} = \tilde{\theta}^{(k-1)} - \left[G_{\tilde{\theta}^{(k-1)}}\right]^{-1} g_{\tilde{\theta}^{(k-1)}}$$

After iterating to convergence, we get:

$$\hat{V}\left\{\hat{\theta}\right\} = -\left[E\left\{G_{\hat{\theta}}\right\}\right]^{-1} \qquad \hat{\mu} = \frac{e^{\hat{\gamma}}}{1+e^{\hat{\gamma}}} \qquad \hat{\phi} = e^{\hat{\phi}^*}$$

$$\hat{V}\{\hat{\mu}\} = \hat{V}\{\hat{\gamma}\}\left(\frac{\partial \hat{\mu}}{\partial \hat{\gamma}}\right)^2 \qquad \hat{V}\left\{\hat{\phi}\right\} = \hat{V}\left\{\hat{\phi}^*\right\}\left(\frac{\partial \hat{\phi}}{\partial \hat{\phi}^*}\right)^2$$

**Example: NASCAR Proportion of Prize Money Won by Fords - 1995 Season Races**

During the 1995 NASCAR Winston Cup season, there were $n$=31 races. The average proportion of prize money won by Fords was $\overline{Y} = 0.5060$ with a standard deviation of $s_Y = 0.0475$. We treat this as a sample from a conceptual population of races that could have been held that year. For starting values, we use:

$$\tilde{\gamma} = \ln\left(\frac{\overline{Y}}{1-\overline{Y}}\right) = \ln\left(\frac{0.5060}{1-0.5060}\right) = 0.0240$$

$$\tilde{\phi}^* = \ln\left(\frac{\overline{Y}(1-\overline{Y})}{s_Y^2} - 1\right) = \ln\left(\frac{0.5060(1-0.5060)}{0.0475^2} - 1\right) = \ln(109.7874) = 4.6985$$

After 109 iterations, we obtain:

$$\hat{\theta} = \begin{bmatrix} \hat{\gamma} \\ \hat{\phi}^* \end{bmatrix} = \begin{bmatrix} 0.02402206 \\ 4.72667962 \end{bmatrix}$$

$$\hat{V}\left\{\hat{\theta}\right\} = -\left[E\left\{G_{\hat{\theta}}\right\}\right]^{-1} = \begin{bmatrix} 0.001132761 & 0.00001364478 \\ 0.00001364478 & 0.06394990 \end{bmatrix}$$

Transforming back to obtain estimates for $\mu$ and $\phi$ gives:

$$\mu = \frac{e^{\gamma}}{1+e^{\gamma}} \quad \Rightarrow \quad \hat{\mu} = \frac{e^{0.02402206}}{1+e^{0.02402206}} = 0.506005226 \qquad \frac{\partial \mu}{\partial \gamma} = \frac{e^{\hat{\gamma}}}{(1+e^{\hat{\gamma}})^2}$$

$$\hat{V}\{\hat{\mu}\} = \hat{V}\{\hat{\gamma}\}\left[\frac{\partial \mu}{\partial \gamma}\right]^2 |_{\gamma=\hat{\gamma}} = 0.001132761\left[\frac{e^{0.02402206}}{(1+e^{0.02402206})^2}\right]^2 = 0.00007077717$$

An approximate 95% Confidence Interval for $\mu$ is:

$$\hat{\mu} \pm 1.96\sqrt{\hat{V}\{\hat{\mu}\}} \quad \equiv \quad 0.506005226 \pm 1.96\sqrt{0.00007077717} \quad \equiv \quad 0.5060 \pm 0.0165 \quad \equiv \quad (0.4895, 0.5225)$$

$$\hat{\phi} = e^{\hat{\phi}^*} = e^{4.72667962} = 112.92 \qquad \hat{V}\left\{\hat{\phi}\right\} = \hat{V}\left\{\hat{\phi}^*\right\}\left[e^{\phi^*}\right]^2 |_{\hat{\phi}^*} = 0.06394990(112.92)^2 = 815.4205$$

Thus, the estimated standard error of $\hat{\phi}$ is $\sqrt{815.4205} = 28.5556$.

$\nabla$

## 1.5 Likelihood Ratio, Wald, and Score (Lagrange Multiplier) Tests

When we wish to test hypotheses regarding value(s) of parameter(s) $\theta$, there are 3 general classes of tests that make use of the likelihood function and MLEs. These are referred to as **Likelihood Ratio**, **Wald**, and **Score (Lagrange Multiplier)** tests. Asymptotically, they are equivalent. In small-samples, their properties can differ. We consider first the case of a single parameter, then the case of multiple parameters.

The likelihood ratio test is based on the difference in the log-likelihood function $l(\theta) = \ln L(\theta|y_1, \ldots, y_n)$ at its maximum, evaluated at $\theta = \hat{\theta}$ and when it is evaluated at the null value $\theta = \theta_0$.

The Wald test is based on the difference between the maximized value $\hat{\theta}$ and the null value $\theta_0$ in terms of the estimated standard error (square root of the variance) of $\hat{\theta}$.

The score (Lagrange Multiplier) test is based on a function of the derivative (slope) of the likelihood function evaluated at the null value $\theta_0$. It does not depend on the MLE $\hat{\theta}$, so is often used in complex estimation problems.

### 1.5.1 Single Parameter Models

For one parameter families (such as the Binomial (Bernoulli), Poisson, and Exponential), the procedures are conducted as follows. Note that a Normal with known variance is also a case, but rare in actual practice.

We wish to test a point null hypothesis $H_0 : \theta = \theta_0$ versus an alternative $H_A : \theta \neq \theta_0$. Note that if $\theta_0$ is at the edge of the parameter space, critical values will need to be adjusted.

The **Likelihood Ratio Test** is conducted as follows:

1. Identify the parameter space $\Omega$, such as $\Omega \equiv \{\theta : 0 < \theta < 1\}$ for Binomial or $\Omega \equiv \{\theta : \theta > 0\}$ for the Poisson.
2. Identify the parameter space under $H_0 : \Omega_0 \equiv \{\theta : \theta = \theta_0\}$
3. Evaluate the maximum log-likelihood (terms not involving $\theta$ can be ignored)
4. Evaluate the log-likelihood under $H_0$ (terms not involving $\theta$ can be ignored)
5. Compute $X^2_{\text{LR}} = -2\left[l\left(\theta_0\right) - l\left(\hat{\theta}\right)\right]$
6. Under the null hypothesis, $X^2_{\text{LR}}$ is asymptotically distributed as $\chi^2(1)$, where the 1 degree of freedom refers to the number of restrictions under $H_0$
7. Reject $H_0$ for large values of $X^2_{\text{LR}}$ $\qquad \left(X^2_{\text{LR}} \geq \chi^2_{\alpha,1}\right)$.

The **Wald Test** makes use of the ML estimate, and its standard error, and asymptotic normality to conduct the test. First, consider the variance of the ML estimator described above (using slightly different notation):

$$V\left\{\hat{\theta}\right\} = I^{-1}(\theta) \qquad\qquad I(\theta) = -E\left\{\frac{\partial^2 l(\theta)}{\partial \theta^2}\right\}$$

where $E\left\{\frac{\partial^2 l(\theta)}{\partial \theta^2}\right\}$ is called the **Fisher Information**. Then we obtain the Wald statistic, which is the square of a large-sample $Z$-statistic (note the use of the estimated variance):

$$X^2_{\text{W}} = \frac{\left(\hat{\theta} - \theta_0\right)^2}{\hat{V}\left\{\hat{\theta}\right\}} = I\left(\hat{\theta}\right)\left(\hat{\theta} - \theta_0\right)^2$$

As with the Likelihood Ratio Test, under the null hypothesis, $X^2_{\text{W}}$ is asymptotically $\chi^2_1$ and we use the same rejection region: $\left(X^2_{\text{W}} \geq \chi^2_{\alpha,1}\right)$

The **Score (Lagrange Multiplier) Test** is based on the derivative of the log-likelihood, and actually does not make use of the ML estimate $\hat{\theta}$, which can be an advantage in complex estimation problems.

First, compute the first derivative of the log-likelihood, evaluated at the null value $\theta_0$. Note that this will only equal 0 if $\theta_0 = \hat{\theta}$ (the maximum likelihood estimate). This value is called the **score**:

$$s\left(\theta, y\right) = \frac{\partial l(\theta)}{\partial \theta} \qquad s\left(\theta_0, y\right) = \left.\frac{\partial l(\theta)}{\partial \theta}\right|_{\theta=\theta_0}$$

Next, multiply the score squared by the variance of the ML estimate, evaluated at the null value $\theta_0$, to obtain the score statistic:

$$V\left\{\hat{\theta}\right\}\Big|_{\theta=\theta_0} = \frac{1}{I\left(\theta_0\right)} \qquad \Rightarrow \qquad X^2_{\text{LM}} = \frac{s\left(\theta_0, y\right)^2}{I\left(\theta_0\right)}$$

As with the Likelihood Ratio and Wald statistics, we reject the null if $X^2_{LM} \geq \chi^2_{\alpha,1}$.

In the case of the Exponential distribution, where $\hat{\theta} = \frac{n}{\sum y_i} = \frac{1}{\bar{Y}}$, and $\mu_Y = \frac{1}{\theta}$:

$$L\left(\theta | y_1, \ldots, y_n\right) = \theta^n e^{-\theta \sum y_i} \qquad \Rightarrow \qquad l(\theta) = n\ln(\theta) - \theta \sum y_i$$

$$\frac{\partial l(\theta)}{\partial \theta} = \frac{n}{\theta} - \sum y_i \qquad \frac{\partial^2 l(\theta)}{\partial \theta^2} = -\frac{n}{\theta^2} \qquad I(\theta) = -E\left\{-\frac{n}{\theta^2}\right\} = \frac{n}{\theta^2}.$$

For the Likelihood Ratio Test, we obtain:

$$l\left(\hat{\theta}\right) = n\ln\left(\hat{\theta}\right) - \hat{\theta}\sum y_i = n\ln\left(\hat{\theta}\right) - \frac{n}{\sum y_i}\sum y_i = n\ln\left(\hat{\theta}\right) - n$$

$$l\left(\theta_0\right) = n\ln\left(\theta_0\right) - \theta_0 \sum y_i = n\ln\left(\theta_0\right) - \theta_0\left(\frac{n}{\hat{\theta}}\right)$$

So that the Likelihood Ratio statistic is:

$$X^2_{\text{LR}} = -2\left[l\left(\theta_0\right) - l\left(\hat{\theta}\right)\right] = -2\left[\left(n\ln\left(\theta_0\right) - \theta_0\left(\frac{n}{\hat{\theta}}\right)\right) - n\ln\left(\left(\hat{\theta}\right) - n\right)\right] = -2n\left[\ln\left(\frac{\theta_0}{\hat{\theta}}\right) - \left(\frac{\theta_0}{\hat{\theta}} - 1\right)\right]$$

For the Wald Test, we get the statistic:

$$X^2_{\text{W}} = \frac{\left(\hat{\theta} - \theta_0\right)^2}{\hat{V}\left\{\hat{\theta}\right\}} = I\left(\hat{\theta}\right)\left(\hat{\theta} - \theta_0\right)^2 = \frac{\left(\hat{\theta} - \theta_0\right)^2}{\hat{\theta}^2}$$

For the Score (Lagrange Multiplier) Test, we obtain the statistic:

$$s(\theta_0, y) = \frac{n}{\theta_0} - \sum y_i = \frac{n - \theta_0 \sum y_i}{\theta_0} \qquad I(\theta_0) = \frac{n}{\theta_0^2}$$

$$\Rightarrow \qquad X^2_{\text{LM}} = \frac{s(\theta_0, y)^2}{I(\theta_0)} = \frac{\left(\frac{n - \theta_0 \sum y_i}{\theta_0}\right)^2}{\left(\frac{n}{\theta_0^2}\right)} = \frac{\theta_0^2}{n}\left(\frac{n - \theta_0 n\overline{Y}}{\theta_0}\right)^2 = \frac{\theta_0^2 n^2}{n\theta_0^2}\left(1 - \theta_0\overline{Y}\right)^2 = n\left(1 - \theta_0\overline{Y}\right)^2$$

**Example: WNBA Free Throw Shooting - Maya Moore**

Consider again WNBA star Maya Moore's true probability of making a free throw, treating her 2014 season attempts as a random sample from her underlying population of all possible (in game) free throw attempts. We treat her individual attempts as independent Bernoulli trials with probability of success $\pi$. We would like to test whether her underlying proportion is $\pi = \pi_0 = 0.80$ (80%). Over the course of the season, she attempted 181 free throws, and made 160 of them.

$$\hat{\pi} = \frac{\sum_{i=1}^n y_i}{n} = \frac{160}{181} = 0.884 \qquad \hat{V}(\hat{\pi}) = \frac{\hat{\pi}(1 - \hat{\pi})}{n} = \frac{0.884(0.116)}{181} = .0005665$$

$$l(\pi) = \ln L(\pi) = \left(\sum_{i=1}^n y_i\right)\ln\pi + \left(n - \sum_{i=1}^n y_i\right)\ln(1 - \pi)$$

$$\Rightarrow \quad l(\hat{\pi}) = \ln L(\hat{\pi}) = 160\,(\ln(0.884)) + (181 - 160)\,(\ln(1 - 0.884)) = -64.965$$

$$\Rightarrow \quad l(\pi_0) = \ln L(\pi_0) = 160\,(\ln(0.80)) + (181 - 160)\,(\ln(1 - 0.80)) = -69.501$$

$$\frac{\partial l}{\partial \pi} = \frac{\sum_{i=1}^n y_i}{\pi} - \frac{n - \sum_{i=1}^n y_i}{1 - \pi}$$

$$\Rightarrow \quad s(\pi_0, y) = \frac{\partial l}{\partial \pi_0} = \frac{160}{0.80} - \frac{181 - 160}{1 - 0.80} = 95$$

$$\frac{\partial^2 l}{\partial \pi^2} = -\frac{\sum_{i=1}^n y_i}{\pi^2} - \frac{n - \sum_{i=1}^n y_i}{(1 - \pi)^2}$$

$$\Rightarrow \quad E\left\{\frac{\partial^2 l}{\partial \pi^2}\right\} - \frac{n\pi}{\pi^2} - \frac{n - n\pi}{(1 - \pi)^2} = -\frac{n}{\pi(1 - \pi)} \quad \Rightarrow \quad I(\pi_0) = \frac{n}{\pi_0(1 - \pi_0)} = \frac{181}{0.80\,(1 - 0.80)} = 1131.25$$

We now compute the Likelihood Ratio, Wald and Score (Lagrange Multiplier) Test Statistics for testing $H_0 : \pi = 0.80$. The critical value for $\alpha = 0.05$ level tests is $\chi^2(0.05, 1) = 3.841$.

$$X^2_{\text{LR}} = -2\left[\ln L(\pi_0) - \ln L(\hat{\pi})\right] = -2\left[(-69.501) - (-64.965)\right] = 9.072$$

$$X^2_{\text{W}} = \frac{(\hat{\pi} - \pi_0)^2}{\hat{V}(\hat{\pi})} = \frac{(0.884 - 0.800)^2}{0.0005665} = 12.455$$

$$X^2_{\text{LM}} = \frac{(s(\pi_0, y))^2}{I(\pi_0)} = \frac{95^2}{1131.25} = 7.978$$

$\nabla$

### 1.5.2 Multiple Parameter Models

For models with multiple parameters, all three tests can be extended to make tests among the parameters (not necessarily all of them). For instance, in a Normal model, we may wish to test $H_0 : \mu = 100, \sigma^2 = 400$ against the alternative that either $\mu \neq 100$ and/or $\sigma^2 \neq 400$. Another possibility is that we may be simultaneously modeling a Poisson model among 3 populations and wish to test $H_0 : \lambda_1 = \lambda_2 = \lambda_3$ versus the alternative that the Poisson parameters are not the same among the populations.

Suppose we have $p$ parameters to be estimated. We have $g \leq p$ **linearly independent** linear hypotheses among the parameters. For instance, we cannot test $H_0 : \mu = 100, \mu = 120$. Note, for an introduction to matrix algebra, see the Regression notes. We can write the null hypothesis as follows:

$$\text{Parameter Vector: } \theta = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \end{bmatrix} \qquad H_0 : R\theta = r \qquad R = \begin{bmatrix} R_{11} & \cdots & R_{1p} \\ \vdots & \ddots & \vdots \\ R_{g1} & \cdots & R_{gp} \end{bmatrix} \qquad r = \begin{bmatrix} r_1 \\ \vdots \\ r_g \end{bmatrix}$$

where $R$ and $r$ are a matrix and vector of constants that define the restrictions from the null hypothesis.

For the Normal model example, we have (with $g = 2$ restrictions):

$$\theta = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} \qquad R = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad r = \begin{bmatrix} 100 \\ 400 \end{bmatrix}$$

For the Poisson example, there are various ways we could test $H_0 : \lambda_1 = \lambda_2 = \lambda_3$, but keep in mind there are only 2 linearly independent restrictions (we are not testing what value they are, just that they are equal). One possibility is:

$$H_{01} : \lambda_1 = \lambda_2, \lambda_1 = \lambda_3 \qquad \Rightarrow \qquad \lambda_1 - \lambda_2 = 0 \quad \lambda_1 - \lambda_3 = 0$$

Note that with these two statements, we imply that $\lambda_2 = \lambda_3$, and including that would cause a redundancy. A second possibility is:

$$H_{02} : \lambda_1 = \lambda_2, \lambda_2 = \lambda_3 \qquad \Rightarrow \qquad \lambda_1 - \lambda_2 = 0 \quad \lambda_2 - \lambda_3 = 0$$

Again, this implies that $\lambda_1 = \lambda_3$.

For these hypotheses, we have:

$$\theta = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} \qquad R_1 = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \qquad R_2 = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \qquad r_1 = r_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Defining $l(\theta_1, \ldots, \theta_p | y) = l(\theta, y)$ as the log-likelihood, and $n_\bullet$ as the overall sample size (summed across group sizes if comparing several populations), we obtain the following quantities for the three tests:

$$\hat{\theta} \equiv \text{ MLE over entire parameter space} \qquad \tilde{\theta} \equiv \text{ MLE over constraint } H_0$$

$$s_i(\theta, y) = \frac{\partial l(\theta)}{\partial \theta_i} \qquad I_{ij}(\theta) = -E\left\{\frac{\partial^2 l(\theta)}{\partial \theta_i \partial \theta_j}\right\}$$

with:

$$s(\theta, y) = \begin{bmatrix} s_1(\theta, y) \\ \vdots \\ s_p(\theta, y) \end{bmatrix} \qquad I(\theta) = \begin{bmatrix} I_{11}(\theta) & \cdots & I_{1p}(\theta) \\ \vdots & \ddots & \vdots \\ I_{p1}(\theta) & \cdots & I_{pp}(\theta) \end{bmatrix}.$$

Each of the chi-squared statistics will be asymptotically $\chi^2_g$, under the null hypothesis, where $g$ is the number of restrictions (rows of $R$ and $r$). The statistics are obtained as follow.

$$\text{Likelihood Ratio:} \qquad X^2_{\text{LR}} = -2\left[l\left(\tilde{\theta}, y\right) - l\left(\hat{\theta}, y\right)\right]$$

$$\text{Wald:} \qquad X^2_{\text{W}} = \left(R\hat{\theta} - r\right)'\left(R\left(I\left(\hat{\theta}\right)\right)^{-1}R'\right)^{-1}\left(R\hat{\theta} - r\right)$$

$$\text{Score (LM):} \qquad X^2_{\text{LM}} = s\left(\tilde{\theta}, y\right)'\left(I\left(\tilde{\theta}\right)\right)^{-1}s\left(\tilde{\theta}, y\right)$$

**Example: Goals per Game for 5 European Premier Football Leagues**

Consider goals per game for the following 5 European Premier Football Leagues for the 2004 seasons. In each league, all pairs of teams play each opponent twice, once Home and Away. Thus, if there are $t_i$ teams in the in the $i^{th}$ league, there will be $n_i = t_i(t_i - 1)$ total games. The 5 leagues are: England ($t_1 = 20, n_1 = 380$), France ($t_2 = 20, n_2 = 380$), Germany ($t_3 = 18, n_3 = 306$), Italy ($t_4 = 20, n_4 = 380$), and Spain ($t_5 = 20, n_5 = 380$). All games are 90 minutes, with no overtime. We treat the observed games as a sample from a conceptual of all games that could be played among the teams within the leagues. We model the total goals per game as a Poisson random variable, with possibly different means among the leagues. Further, we assume independence among all games within each league. We define the model of scores below. Our goal is to test $H_0 : \theta_1 = \cdots = \theta_5 = \theta$.

$$Y_{ij} \sim \text{Poi}(\theta_i) \qquad i = 1, \ldots, 5; \quad j = 1, \ldots, n_i \qquad p(y_{ij}|\theta_i) = \frac{e^{-\theta_i}\theta_i^{y_{ij}}}{y_{ij}!}$$

The ML estimates of the parameters are obtained below under the null ($H_0$) and alternative ($H_A$) hypotheses, along with other quantities needed for the tests, where $y_{i\bullet}$ and $y_{\bullet\bullet}$ represent sums over one or both subscripts.

$$L_A = \prod_{i=1}^{5}\prod_{j=1}^{n_i}\frac{e^{-\theta_i}\theta_i^{y_{ij}}}{y_{ij}!} = \frac{e^{-\sum_{i=1}^{5} n_i\theta_i}\prod_{i=1}^{5}\theta_i^{y_{i\bullet}}}{\prod_{i=1}^{5}\prod_{j=1}^{y_{ij}}} \qquad L_0 = \prod_{i=1}^{5}\prod_{j=1}^{n_i}\frac{e^{-\theta}\theta^{y_{ij}}}{y_{ij}!} = \frac{e^{-n_\bullet\theta}\theta^{y_{\bullet\bullet}}}{\prod_{i=1}^{5}\prod_{j=1}^{n_i} y_{ij}!}$$

$$l_A = \ln L_A = -\sum_{i=1}^{5} n_i\theta_i + \sum_{i=1}^{5} y_{i\bullet}\ln\theta_i - \ln\left(\prod_{i=1}^{5}\prod_{j=1}^{n_i} y_{ij}!\right) \qquad l_0 = \ln L_0 = -n_\bullet\theta + y_{\bullet\bullet}\ln\theta - \ln\left(\prod_{i=1}^{5}\prod_{j=1}^{n_i} y_{ij}!\right)$$

$$\frac{\partial l_A}{\partial\theta_i} = -n_i + \frac{y_{i\bullet}}{\theta_i} \quad \Rightarrow \quad \hat{\theta}_i = \frac{y_{i\bullet}}{n_i} \qquad \frac{\partial l_0}{\partial\theta} = -n_\bullet + \frac{y_{\bullet\bullet}}{\theta} \quad \Rightarrow \quad \tilde{\theta} = \frac{y_{\bullet\bullet}}{n_\bullet}$$

$$\frac{\partial^2 l_A}{\partial\theta_i^2} = -\frac{y_{i\bullet}}{\theta_i^2} \qquad \frac{\partial^2 l_A}{\partial\theta_i\theta_{i'}} = 0 \quad i \neq i' \qquad \frac{\partial^2 l_0}{\partial\theta^2} = -\frac{y_{\bullet\bullet}}{\theta^2}$$

$$E\left\{\frac{\partial^2 l_A}{\partial\theta_i^2}\right\} = -\frac{n_i\theta_i}{\theta_i^2} = -\frac{n_i}{\theta_i} \qquad E\left\{\frac{\partial^2 l_0}{\partial\theta^2}\right\} = -\frac{n_\bullet\theta}{\theta^2} = -\frac{n_\bullet}{\theta}$$

| League (i) | $n_i$ | $y_{i\bullet}$ | $\hat{\theta}_i$ | $-n_i\hat{\theta}_i + y_{i\bullet}\ln\hat{\theta}_i$ |
|---|---|---|---|---|
| England (1) | 380 | 975 | 2.5658 | -56.2904 |
| France (2) | 380 | 826 | 2.1737 | -184.6742 |
| Germany (3) | 306 | 890 | 2.9805 | 60.1964 |
| Italy (4) | 380 | 960 | 2.5263 | -70.3085 |
| Spain (5) | 380 | 980 | 2.5789 | -51.5663 |
| Overall | 1826 | 4631 | $\tilde{\theta} = 2.5361$ | $n_\bullet\tilde{\theta} + y_{\bullet\bullet}\ln\tilde{\theta} = -321.1818$ |

Table 1.3: Frequency Tabulation and Expected Counts for EPL 2012/2013 Total Goals per Game

The data and estimates are given in Table 1.3, followed with the test for equal means.

$$R = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 \\ 1 & 0 & 0 & 0 & -1 \end{bmatrix} \qquad r = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad \hat{\theta} = \begin{bmatrix} 2.5658 \\ 2.1737 \\ 2.9805 \\ 2.5263 \\ 2.5789 \end{bmatrix}$$

$$I\left(\hat{\theta}\right) = \begin{bmatrix} \frac{380}{2.5658} & 0 & 0 & 0 & 0 \\ 0 & \frac{380}{2.1737} & 0 & 0 & 0 \\ 0 & 0 & \frac{306}{2.9805} & 0 & 0 \\ 0 & 0 & 0 & \frac{380}{2.5263} & 0 \\ 0 & 0 & 0 & 0 & \frac{380}{2.5789} \end{bmatrix} = \begin{bmatrix} 148.1026 & 0 & 0 & 0 & 0 \\ 0 & 174.8184 & 0 & 0 & 0 \\ 0 & 0 & 105.2090 & 0 & 0 \\ 0 & 0 & 0 & 150.4167 & 0 \\ 0 & 0 & 0 & 0 & 147.3469 \end{bmatrix}$$

$$I\left(\tilde{\theta}\right) = \begin{bmatrix} \frac{380}{2.5361} & 0 & 0 & 0 & 0 \\ 0 & \frac{380}{2.5361} & 0 & 0 & 0 \\ 0 & 0 & \frac{306}{2.5361} & 0 & 0 \\ 0 & 0 & 0 & \frac{380}{2.5361} & 0 \\ 0 & 0 & 0 & 0 & \frac{380}{2.5361} \end{bmatrix} = \begin{bmatrix} 149.8337 & 0 & 0 & 0 & 0 \\ 0 & 149.8337 & 0 & 0 & 0 \\ 0 & 0 & 120.6556 & 0 & 0 \\ 0 & 0 & 0 & 149.8337 & 0 \\ 0 & 0 & 0 & 0 & 149.8337 \end{bmatrix}$$

$$R\hat{\theta} - r = R\hat{\theta} = \begin{bmatrix} 0.3921 \\ -0.3427 \\ 0.0395 \\ -0.0131 \end{bmatrix} \qquad RI^{-1}\left(\hat{\theta}\right)R' = \begin{bmatrix} 132.7165 & -22.3377 & -36.2251 & -38.4859 \\ -22.3377 & 89.9603 & -21.8010 & -21.3561 \\ -36.2251 & -21.8010 & 119.2479 & -30.5326 \\ -35.4859 & -21.3561 & -30.5326 & 117.4374 \end{bmatrix}$$

$$s\left(\tilde{\theta}, y\right) = \begin{bmatrix} -380 + \frac{975}{2.5361} \\ -380 + \frac{826}{2.5361} \\ -306 + \frac{890}{2.5361} \\ -380 + \frac{960}{2.5361} \\ -380 + \frac{980}{2.5361} \end{bmatrix} = \begin{bmatrix} 4.4418 \\ -54.3088 \\ 44.9264 \\ -1.4727 \\ 6.4133 \end{bmatrix}$$

We now compute the 3 test statistics, keeping in mind there are $g = 4$ parameter restictions under the null hypothesis, and that $\chi^2(4, 0.05) = 9.488$.

$$\text{Likelihood Ratio:} \qquad X^2_{\text{LR}} = -2\left[l\left(\tilde{\theta}, y\right) - l\left(\hat{\theta}, y\right)\right] =$$

$$-2\left[-321.1818 - (-56.2904 - 184.6742 + 60.1964 - 70.3085 - 51.5663)\right] = -2\left[-321.1818 - (-302.6430)\right] = 37.078$$

$$\text{Wald:} \qquad X^2_{\text{W}} = \left(R\hat{\theta} - r\right)'\left(R\left(I\left(\hat{\theta}\right)\right)^{-1}R'\right)^{-1}\left(R\hat{\theta} - r\right) = 37.660$$

$$\text{Score (LM):} \qquad X^2_{\text{LM}} = s\left(\tilde{\theta}, y\right)'\left(I\left(\tilde{\theta}\right)\right)^{-1}s\left(\tilde{\theta}, y\right) = 36.834$$

$\nabla$

## 1.6 Sampling Distributions and an Introduction to the Bootstrap

Previously we described the sampling distributions of various estimators derived from independent and normally distributed random variables. Also, we considered the large-sample properties of maximum likelihood estimators, that inherently meant we know the underlying distribution of the data.

One useful tool for obtaining the exact distribution of linear functions of random variables (when it even exists) is the **moment-generating function** or mgf. This function serves 2 primary purposes. First, it can be used to obtain the non-central moments of a distribution: $E\{Y\}, E\{Y^2\}, E\{Y^3\}, \ldots$. The moment-generating function (if it exists) for a distribution can be obtained as follows:

$$\text{Discrete Distribution: } M_Y(t) = E\{e^{tY}\} = \sum_y e^{ty}p(y)$$

$$\text{Continuous Distribution: } M_Y(t) = E\{e^{tY}\} = \int_{-\infty}^{\infty} e^{ty}f(y)dy$$

Without going through the derivations, we obtain (most involve rules of sums or competing the square or change of variables in integration):

$$\text{Binomial: } M_Y(t) = \sum_{y=0}^{n} e^{ty}\frac{n!}{y!(n-y)!}\pi^y(1-\pi)^{n-y} = \sum_{y=0}^{n}\frac{n!}{y!(n-y)!}\left(\pi e^t\right)^y(1-\pi)^{n-y} = \left(\pi e^t + (1-\pi)\right)^n$$

$$\text{Poisson: } M_Y(t) = \sum_{y=0}^{\infty} e^{ty}\frac{e^{-\lambda}\lambda^y}{y!} = \sum_{y=0}^{\infty}\frac{e^{-\lambda}\left(\lambda e^t\right)^y}{y!} = e^{\lambda\left(e^t-1\right)}$$

$$\text{Normal: } M_Y(t) = \int_{-\infty}^{\infty} e^{ty}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]dy = \exp\left[\mu t + \frac{t^2\sigma^2}{2}\right]$$

$$\text{Gamma: } M_Y(t) = \int_{-\infty}^{\infty} e^{ty}\frac{1}{\Gamma(\alpha)\beta^\alpha}y^{\alpha-1}e^{-y/\beta}dy = (1-\beta t)^{-\alpha}$$

Note that for the other formulation of the Gamma, we would have $M_Y(t) = \left(1-\frac{t}{\theta}\right)^{-\alpha}$. Further, for the Exponential, we would have $M_Y(t) = \left(1-\frac{t}{\theta}\right)^{-1}$ and for the chi-square, we would have $(1-2t)^{-\nu/2}$.

The mgf can be used to obtain the non-central moments as follows, based on a series expansion $e^{tY} = \sum_{i=0}^{\infty}\frac{(tY)^i}{i!}$

$$\left.\frac{dM(t)}{dt}\right|_{t=0} = M'(0) = E\{Y\} \qquad \left.\frac{d^2M(t)}{dt^2}\right|_{t=0} = M''(0) = E\{Y^2\}$$

and so on such that $M^{(k)}(0) = E\{Y^k\}$.

More importantly here, if we obtain a sum or linear function of **independent** random variables, we can use the uniqueness of the mgf to obtain the distribution of the sum or linear function. Consider $W =$

$Y_1 + \cdots + Y_n$, a sum of independent random variables:

$$M_W(t) = E\left\{e^{tW}\right\} = E\left\{e^{t(Y_1+\cdots+Y_n)}\right\} = \prod_{i=1}^{n} E\left\{e^{tY_i}\right\} = \prod_{i=1}^{n} M_{Y_i}(t).$$

The independence of $Y_1, \ldots, Y_n$ is why we can use this result.

Consider $m$ Binomial random variables, each with success probability $\pi$, but with possibly varying sample sizes $n_i$:

$$Y_i \sim \text{Bin}\left(n_i, \pi\right) \quad i = 1, \ldots, m \qquad M_{Y_i}(t) = \left(\pi e^t + (1-\pi)\right)^{n_i}$$

Thus if we let $W = Y_1 + \cdots + Y_m$, we have:

$$M_W(t) = \prod_{i=1}^{m} M_{Y_i}(t) = \prod_{i=1}^{m} \left(\pi e^t + (1-\pi)\right)^{n_i} = \left(\pi e^t + (1-\pi)\right)^{\sum n_i} \qquad \Rightarrow \qquad W \sim \text{Binomial}\left(\sum n_i, \pi\right)$$

Thus, the sum of independent Binomial random variables with common success probability is Binomial with the same success probability, and a sample size equal to the sum of the individual sample sizes.

Similar results hold for independent Poisson random variables, where $Y_i \sim \text{Poisson}(\lambda_i)$. Let $W = Y_1 + \cdots + Y_n$:

$$M_W(t) = \prod_{i=1}^{n} M_{Y_i}(t) = \prod_{i=1}^{n} e^{\lambda_i\left(e^t-1\right)} = \exp\left[\left(\sum \lambda_i\right)\left(e^t - 1\right)\right] \qquad \Rightarrow \qquad W \sim \text{Poisson}\left(\sum \lambda_i\right)$$

For a sum of independent Gammas, with common $\beta$ or $\theta$, that is $Y_i \sim \text{Gamma}\left(\alpha_i, \beta\right)$. Let $W = Y_1 + \cdots + Y_n$:

$$M_W(t) = \prod_{i=1}^{n} M_{Y_i}(t) = \prod_{i=1}^{n} (1-\beta t)^{-\alpha_i} = (1-\beta t)^{-\sum \alpha_i} \qquad \Rightarrow \qquad W \sim \text{Gamma}\left(\sum \alpha_i, \beta\right)$$

Now consider any linear function $U = a_1 Y_1 + \cdots + a_n Y_n$, for constants $a_1, \ldots, a_n$. This will not work for many distributions in general.

$$M_U(t) = E\left\{e^{tU}\right\} = E\left\{e^{t(a_1Y_1+\cdots+a_nY_n)}\right\} = \prod_{i=1}^{n} E\left\{e^{ta_iY_i}\right\} = \prod_{i=1}^{n} M_{Y_i}(a_i t)$$

Now consider independent Normals, with $Y_i \sim N\left(\mu_i, \sigma_i^2\right)$. Let $U = a_1 Y_1 + \cdots + a_n Y_n$:

$$M_U(t) = \prod_{i=1}^{n} M_{Y_i}(a_i t) = \prod_{i=1}^{n} \exp\left[\mu_i a_i t + \frac{a_i^2 t^2 \sigma_i^2}{2}\right] = \exp\left[t\left(\sum a_i \mu_i\right) + \frac{t^2\left(\sum a_i^2 \sigma_i^2\right)}{2}\right]$$

$$\Rightarrow \quad U \sim N\left(a_i\mu_i, \sum a_i^2 \sigma_i^2\right)$$

So, in special circumstances, when we know the exact distribution of data, we can obtain the exact distribution of some specific estimators. Due to the **Central Limit Theorem**, we can also state that

sample means of independent observations have sampling distributions that asymptotically converge to the Normal (assuming finite variance). Thus, in many cases:

$$\sqrt{n}\frac{\overline{Y}-\mu}{S} \overset{\text{approx}}{\sim} N(0,1) \quad \Rightarrow \quad \overline{Y} \overset{\text{approx}}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

However, in many settings, estimators either are very complex and no sampling distribution can be derived, or samples are not large enough to rely on large-sample asymptotics. In these settings, the **bootstrap method** is applied to a statistic to obtain a Confidence Interval for the underlying parameter. The method assumes the sample is representative of the underlying population (e.g. no inherent bias). Also, if the sampling plan has any specific patterns to it, such as clusters, the bootstrap should reflect that.

The algorithm works as follows.

1. Obtain a sample of size $N$ from the population of interest.
2. Generate a method (function) to compute the statistic of interest.
3. Generate a random sample with replacement from the original sample, apply the function, and save the result.
4. Repeat this process over many samples.
5. Obtain a $(1-\alpha)100\%$ Confidence Interval for the parameter of interest.

The last step can be conducted various ways, the most common way is to select the cut-off values of the middle $(1-\alpha)100\%$ bootstrap sample results. Other ways, particularly bias-corrected methods are implemented in standard statistical software packages, and make use of the mean and standard deviation (standard error) of the bootstrap estimates. This version is referred to as the **non-parametric bootstrap**, which makes no assumptions on the underlying distribution of the data.

**Example: Modeling of Sheer Strength of Reinforced Concrete Beams**

Colotti (2016) describes a model for predicting sheer strength of reinforced concrete beams. Predictions are made of the breaking strength of $n = 200$ beams that have been measured in the academic literature, and compared with the actual (experimental) breaking strength. One reported measure is the **Coefficent of Variation (CV)** of the ratios of the experimental to model predicted breaking strength. The CV is computed as $100\left(\frac{S_Y}{\overline{Y}}\right)$, and measures the percentage of standard deviation relative to the mean. Note that this is a complicated function to obtain the sampling distribution of, particularly if the data are not normally distributed. The mean and standard deviation of the ratios are $\overline{Y} = 1.1011$ and $S_Y = 0.2237$, yielding $CV_Y = 100(0.2237/1.1011) = 20.3161$. That is, the standard deviation is about 20% as large as the mean. Note that in many measurement reliability studies, researchers need very small CVs. A histogram of the ratios is given in Figure 1.9, demonstrating a symmetric distribution, possibly flatter than a Normal, centered around 1.1.

To understand the uncertainty associated with the point estimate of $CV_Y$, we apply the nonparametric bootstrap. We obtain 100,000 samples of size 200 (with replacement) from the original sample, computing

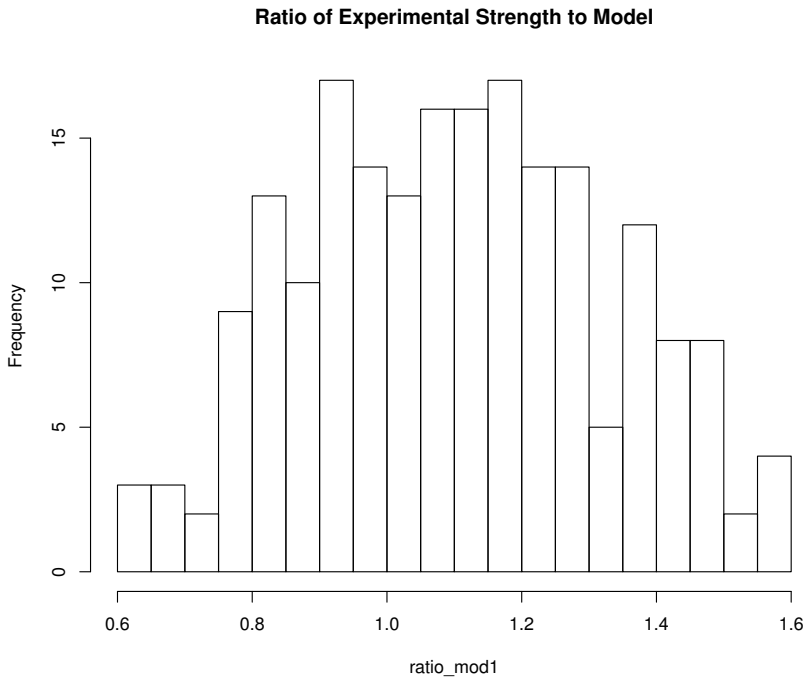


Figure 1.9: Ratio of Experiment/Model Prediction of Reinforced Concrete Beam Sheer Strength

| Mean | SD | Min | 2.5%-ile | 25%-ile | Median | 75%-ile | 97.5%-ile | Max |
|---|---|---|---|---|---|---|---|---|
| 20.25 | 0.85 | 16.31 | 18.59 | 19.69 | 20.26 | 20.83 | 21.91 | 24.00 |

Table 1.4: Summary Statistics of $CV_Y$ for 100000 Bootstrap Samples

$CV_Y$ for each sample. Some of the summary statistics of the samples are given in Table 1.4. A histogram is given in Figure 1.10.

$\nabla$

Another possibility is when you are confident about the underlying distribution, but unsure of parameter values. Then, the parameters can be estimated (based on methods such as ML in previous sections), and then many samples can be generated using random number generators from the corresponding distribution. The Confidence Interval and mean and standard error of the estimator can be obtained as well. This is referred to as the **parametric bootstrap**.

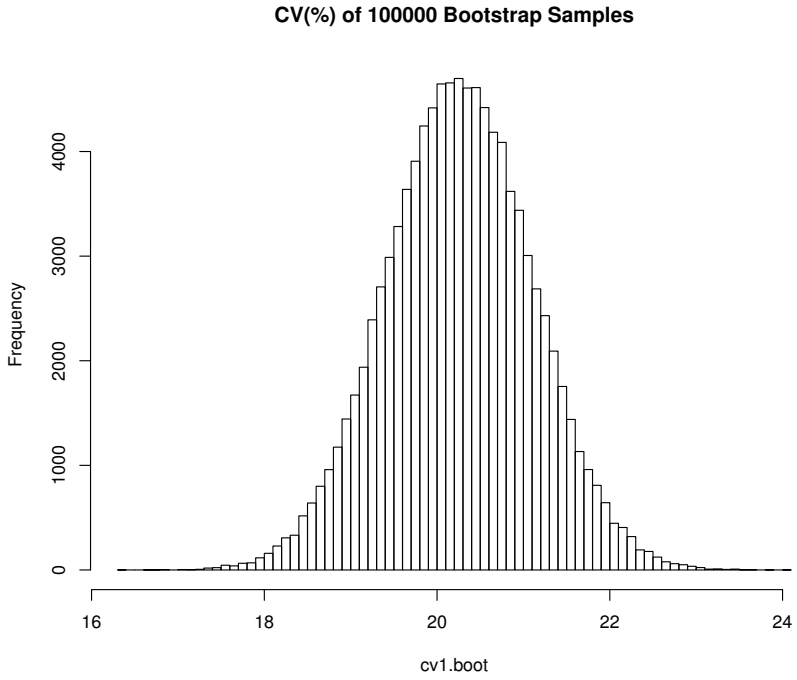


Figure 1.10: $CV_Y$ for 100,000 Bootstrap Samples

# Chapter 2

# Simple Linear Regression

## 2.1 Introduction

Linear regression is used when we have a numeric response variable and numeric (and possibly categorical) predictor (explanatory) variable(s). The mean of the response variable is to be related to the predictor(s) with random error terms assumed to be independent and normally distributed with constant variance. The fitting of linear regression models is very flexible, allowing for fitting curvature and interactions between factors.

When there is a single numeric predictor, we refer to the model as **Simple Regression**. The response variable is denoted as $Y$ and the predictor variable is denoted as $X$ which is assumed to be a fixed constant. The assumed model is given below.

$$Y = \beta_0 + \beta_1 X + \epsilon \qquad \epsilon \sim N(0, \sigma^2) \text{ independent} \qquad \Rightarrow \qquad Y \sim N(\beta_0 + \beta_1 X, \sigma^2) \text{ independent}$$

Here $\beta_0$ is the intercept (mean of $Y$ when $X$=0) and $\beta_1$ is the slope (the change in the mean of $Y$ when $X$ increases by 1 unit). Of primary concern is whether $\beta_1 = 0$, which implies the mean of $Y$ is constant ($\beta_0$), and thus $Y$ and $X$ are not associated.

Note that this model assumes:

$$E\{\epsilon\} = 0 \qquad V\{\epsilon\} = E\{\epsilon^2\} = \sigma^2 \qquad \text{COV}\{\epsilon_i, \epsilon_j\} = E\{\epsilon_i \epsilon_j\} = 0 \quad i \neq j$$

In practice the variance may not be constant, and the errors may not be independent. These assumptions will be checked after fitting the regression model.

## Estimation of Model Parameters

We obtain a sample of pairs $(X_i, Y_i) \quad i = 1, \ldots, n$. Our goal is to choose estimators of $\beta_0$ and $\beta_1$ that minimize the error sum of squares: $Q = \sum_{i=1}^n \epsilon_i^2$. The resulting estimators are derived below by taking derivatives of $Q$ with respect to $\beta_0$ and $\beta_1$, setting each to zero, and solving for the **Ordinary Least Squares (OLS) Estimators**, $\hat{\beta}_0$ and $\hat{\beta}_1$.

$$Q = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n [Y_i - (\beta_0 + \beta_1 X_i)]^2 \quad \Rightarrow$$

$$\frac{\partial Q}{\partial \beta_0} = 2\sum_{i=1}^n [Y_i - (\beta_0 + \beta_1 X_i)]^{2-1}(-1) = -2\left[\sum_{i=1}^n Y_i - n\beta_0 - \beta_1 \sum_{i=1}^n X_i\right]$$

$$\frac{\partial Q}{\partial \beta_1} = 2\sum_{i=1}^n [Y_i - (\beta_0 + \beta_1 X_i)]^{2-1}(-X_i) = -2\left[\sum_{i=1}^n X_i Y_i - \beta_0 \sum_{i=1}^n X_i - \beta_1 \sum_{i=1}^n X_i^2\right]$$

Simultaneously setting both of these derivatives to zero, we solve for $\hat{\beta}_1$ and $\hat{\beta}_0$, making use of the so called **normal equations**, which ironically have nothing to do with the normal distribution.

$$\frac{\partial Q}{\partial \beta_0} = 0 \quad \Rightarrow \quad \sum_{i=1}^n Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n X_i$$

$$\frac{\partial Q}{\partial \beta_1} = 0 \quad \Rightarrow \quad \sum_{i=1}^n X_i Y_i = \hat{\beta}_0 \sum_{i=1}^n X_i + \hat{\beta}_1 \sum_{i=1}^n X_i^2$$

Next, multiply the first equation by $\sum_{i=1}^n X_i$, the second by $n$, and subtract the first from the second to obtain the estimator $\hat{\beta}_1$ for $\beta_1$.

$$\sum_{i=1}^n Y_i \sum_{i=1}^n X_i = n\hat{\beta}_0 \sum_{i=1}^n X_i + \hat{\beta}_1 \left(\sum_{i=1}^n X_i\right)^2 \qquad n\sum_{i=1}^n X_i Y_i = n\hat{\beta}_0 \sum_{i=1}^n X_i + n\hat{\beta}_1 \sum_{i=1}^n X_i^2 \quad \Rightarrow$$

$$n\sum_{i=1}^n X_i Y_i - \sum_{i=1}^n X_i \sum_{i=1}^n Y_i = n\hat{\beta}_1 \sum_{i=1}^n X_i^2 - \hat{\beta}_1 \left(\sum_{i=1}^n X_i\right)^2 \quad \Rightarrow$$

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^n X_i Y_i - \sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n\sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i\right)^2} = \frac{\sum_{i=1}^n X_i Y_i - \frac{\sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n}}{\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}} = \frac{\sum_{i=1}^n X_i Y_i - n\overline{XY}}{\sum_{i=1}^n X_i^2 - n\overline{X}^2}$$

The last form can be simplified as follows.

$$\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^n X_i Y_i - \overline{Y}\sum_{i=1}^n X_i - \overline{X}\sum_{i=1}^n Y_i + n\overline{XY} = \sum_{i=1}^n X_i Y_i - n\overline{XY} - n\overline{XY} + n\overline{XY} = \sum_{i=1}^n X_i Y_i - n\overline{XY}$$

$$\sum_{i=1}^n (X_i - \overline{X})^2 = \sum_{i=1}^n X_i^2 - \overline{X}\sum_{i=1}^n X_i - \overline{X}\sum_{i=1}^n X_i + n\overline{X}^2 = \sum_{i=1}^n X_i^2 - n\overline{X}^2 - n\overline{X}^2 + n\overline{X}^2 = \sum_{i=1}^n X_i^2 - n\overline{X}^2$$

From the first of the normal equations, we easily obtain the OLS estimator $\hat{\beta}_0$ for $\beta_0$ as a function of $\hat{\beta}_1$, $\overline{Y}$, and $\overline{X}$.

$$\sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i \quad \Rightarrow \quad \hat{\beta}_0 = \frac{\sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i}{n} \quad = \quad \overline{Y} - \hat{\beta}_1 \overline{X}$$

Here we summarize the OLS estimators for the simple linear regression model.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \qquad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

Two useful results that create ways of writing $\hat{\beta}_1$ and $\hat{\beta}_0$ as linear functions of $Y_1$,...,$Y_n$ are as follow.

$$\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^{n}\left(X_i - \overline{X}\right) Y_i - \sum_{i=1}^{n}\left(X_i - \overline{X}\right)\overline{Y} = \sum_{i=1}^{n}\left(X_i - \overline{X}\right) Y_i - 0 = \sum_{i=1}^{n}\left(X_i - \overline{X}\right) Y_i$$

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}\left(X_i - \overline{X}\right) X_i - \sum_{i=1}^{n}\left(X_i - \overline{X}\right)\overline{X} = \sum_{i=1}^{n}\left(X_i - \overline{X}\right) X_i - 0 = \sum_{i=1}^{n}\left(X_i - \overline{X}\right) X_i$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \sum_{i=1}^{n} \frac{\left(X_i - \overline{X}\right)}{\sum_{i=1}^{n}(X_i - \overline{X})^2} Y_i$$

Once we have the least squares estimates, we obtain **fitted (predicted) values** and **residuals** for each observation. The **error sum of squares (SSE)** is obtained as the sum of the squared residuals.

Fitted Values: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = \overline{Y} + \hat{\beta}_1\left(X_i - \overline{X}\right)$ Residuals: $e_i = Y_i - \hat{Y}_i$ $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

$$SSE = \sum_{i=1}^{n}\left[Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)\right]^2 = \sum_{i=1}^{n}\left[Y_i^2 + \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)^2 - 2Y_i\left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)\right]$$

$$= \sum_{i=1}^{n} Y_i^2 + \left(\overline{Y} + \hat{\beta}_1\left(X_i - \overline{X}\right)\right)^2 - 2Y_i\left(\overline{Y} + \hat{\beta}_1\left(X_i - \overline{X}\right)\right)$$

$$= \sum_{i=1}^{n} Y_i^2 + \sum_{i=1}^{n}\left(\overline{Y}^2 + \hat{\beta}_1^2\left(X_i - \overline{X}\right)^2 + 2\overline{Y}\hat{\beta}_1\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\right) - 2\sum_{i=1}^{n} Y_i\left(\overline{Y} + \hat{\beta}_1\left(X_i - \overline{X}\right)\right)$$

$$= \sum_{i=1}^{n} Y_i^2 + n\overline{Y}^2 + \hat{\beta}_1^2\sum_{i=1}^{n}(X_i - \overline{X})^2 + 2\overline{Y}\hat{\beta}_1\sum_{i=1}^{n}\left(X_i - \overline{X}\right) - 2n\overline{Y}^2 - 2\hat{\beta}_1\frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$= \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2 - \hat{\beta}_1^2 \sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 - \hat{\beta}_1^2 \sum_{i=1}^{n}(X_i - \overline{X})^2$$

The (unbiased) estimate of the error variance $\sigma^2$ is $s^2 = MSE = \frac{SSE}{n-2}$, where $MSE$ is the **Mean Square Error**. The subtraction of 2 can be thought of as the fact that we have estimated two parameters: $\beta_0$ and $\beta_1$. The derivation of $E\{MSE\}$ is given in the Analysis of Variance section.

**Example: Cysteic Acid Content and Age of Carpets**

Csapo, et al (1995) report results of a study relating age of carpet specimens ($Y$, in years) to cysteic acid levels ($X$). There were $n = 23$ specimens used to fit the regression model, and 2 specimens with only cysteic acid levels, but no age given. We will use the regression model to predict age of specimens 24 and 25. The data, and calculations for the least squares estimates are given in Table 2.1. A plot of the data and estimated regression equation are given in Figure 2.1.

$$\hat{\beta}_1 = \frac{18550.7913}{39.6970} = 467.3096531 \qquad \hat{\beta}_0 = 1017.73913 - 467.3096531(2.89522) = 1017.73913 - 1352.96425 = -335.22512$$

$$\hat{Y}_i = -335.22512 + 467.30965X_i = 1017.73913 + 467.30965\,(X_i - 2.89522)$$

$$SSE = 8733546.43 - (467.30965)^2(39.6970) = 8733546.43 - 8668963.73 = 64582.90 \qquad MSE = \frac{64582.90}{23-2} = 3075.38$$

A set of commands and output for the "Brute-Force" computations in R are given below. The program, based on the **lm** function for the full analysis is given at the end of the chapter.

```
### Commands
carpet <- read.csv("http://www.stat.ufl.edu/~winner/data/carpet_age.csv",
      header=T)
attach(carpet); names(carpet)

### f ==> full data  m==> missing age
age_f <- age[1:23]; age_m <- age[24:25]
cys_acid_f <- cys_acid[1:23]; cys_acid_m <- cys_acid[24:25]

(n <- length(age_f))
(ybar <- mean(age_f))
(xbar <- mean(cys_acid_f))
(SS_XX <- sum((cys_acid_f - xbar)^2))
(SS_XY <- sum((cys_acid_f - xbar) * (age_f - ybar)))
(SS_YY <- sum((age_f - ybar)^2))

(beta1_hat <- SS_XY / SS_XX)
(beta0_hat <- ybar - beta1_hat * xbar)
Y_hat <- beta0_hat + beta1_hat * cys_acid_f

(SS_ERR <- sum((age_f - Y_hat)^2));  (df_ERR <- n-2); (MS_ERR <- SS_ERR/df_ERR)

### Output
> (n <- length(age_f))
[1] 23
> (ybar <- mean(age_f))
```

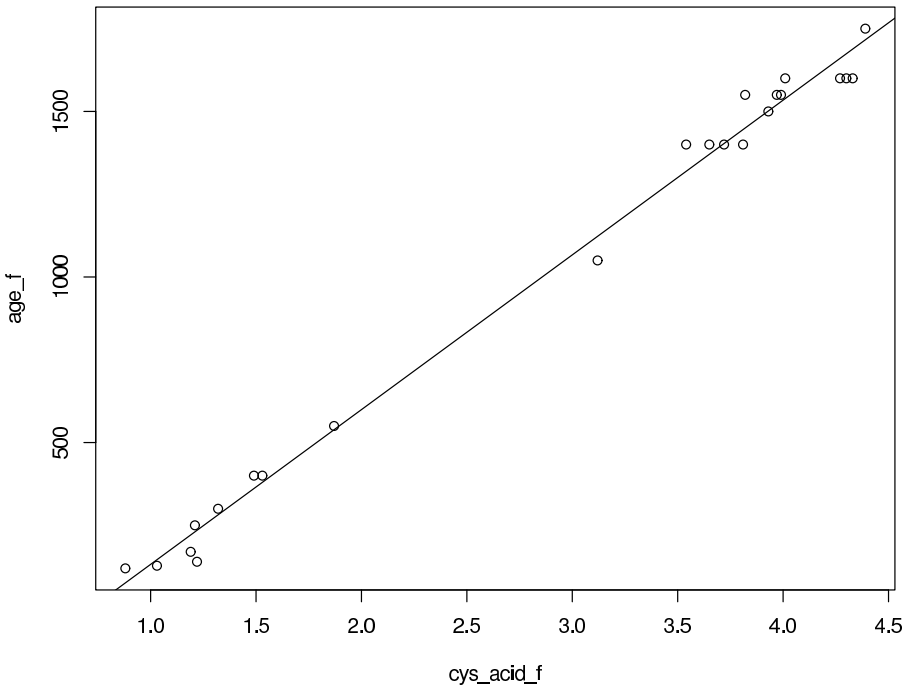


Figure 2.1: Plot of Age ($Y$) versus Cysteic Acid ($X$)

```
[1] 1017.739
> (xbar <- mean(cys_acid_f))
[1] 2.895217
> (SS_XX <- sum((cys_acid_f - xbar)^2))
[1] 39.69697
> (SS_XY <- sum((cys_acid_f - xbar) * (age_f - ybar)))
[1] 18550.79
> (SS_YY <- sum((age_f - ybar)^2))
[1] 8733546
>
> (beta1_hat <- SS_XY / SS_XX)
[1] 467.31
> (beta0_hat <- ybar - beta1_hat * xbar)
[1] -335.2248
> Y_hat <- beta0_hat + beta1_hat * cys_acid_f
>
> (SS_ERR <- sum((age_f - Y_hat)^2));  (df_ERR <- n-2); (MS_ERR <- SS_ERR/df_ERR)
[1] 64576.89
[1] 21
[1] 3075.09
```

$\nabla$

Making use of the normal equations, we obtain the following useful results regarding the residuals.

$$\sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i \quad \Rightarrow \quad \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right) = \sum_{i=1}^{n} \left[Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)\right] = 0$$

| specimen id | age ($Y$) | cysteic acid ($X$) | $Y-\overline{Y}$ | $X-\overline{X}$ | $\left(Y-\overline{Y}\right)^2$ | $\left(X-\overline{X}\right)^2$ | $\left(X-\overline{X}\right)\left(Y-\overline{Y}\right)$ |
|---|---|---|---|---|---|---|---|
| 1 | 1750 | 4.39 | 732.26 | 1.4948 | 536205.98 | 2.2344 | 1094.5708 |
| 2 | 1600 | 4.30 | 582.26 | 1.4048 | 339027.72 | 1.9734 | 817.9499 |
| 3 | 1600 | 4.27 | 582.26 | 1.3748 | 339027.72 | 1.8900 | 800.4821 |
| 4 | 1600 | 4.33 | 582.26 | 1.4348 | 339027.72 | 2.0586 | 835.4178 |
| 5 | 1600 | 4.01 | 582.26 | 1.1148 | 339027.72 | 1.2427 | 649.0943 |
| 6 | 1550 | 3.99 | 532.26 | 1.0948 | 283301.63 | 1.1985 | 582.7099 |
| 7 | 1550 | 3.97 | 532.26 | 1.0748 | 283301.63 | 1.1552 | 572.0647 |
| 8 | 1550 | 3.82 | 532.26 | 0.9248 | 283301.63 | 0.8552 | 492.2256 |
| 9 | 1500 | 3.93 | 482.26 | 1.0348 | 232575.55 | 1.0708 | 499.0352 |
| 10 | 1400 | 3.81 | 382.26 | 0.9148 | 146123.37 | 0.8368 | 349.6856 |
| 11 | 1400 | 3.54 | 382.26 | 0.6448 | 146123.37 | 0.4157 | 246.4752 |
| 12 | 1400 | 3.72 | 382.26 | 0.8248 | 146123.37 | 0.6803 | 315.2821 |
| 13 | 1400 | 3.65 | 382.26 | 0.7548 | 146123.37 | 0.5697 | 288.5239 |
| 14 | 1050 | 3.12 | 32.26 | 0.2248 | 1040.76 | 0.0505 | 7.2517 |
| 15 | 550 | 1.87 | -467.74 | -1.0252 | 218779.89 | 1.0511 | 479.5343 |
| 16 | 400 | 1.49 | -617.74 | -1.4052 | 381601.63 | 1.9746 | 868.0578 |
| 17 | 400 | 1.53 | -617.74 | -1.3652 | 381601.63 | 1.8638 | 843.3482 |
| 18 | 300 | 1.32 | -717.74 | -1.5752 | 515149.46 | 2.4813 | 1130.5952 |
| 19 | 250 | 1.21 | -767.74 | -1.6852 | 589423.37 | 2.8400 | 1293.8073 |
| 20 | 170 | 1.19 | -847.74 | -1.7052 | 718661.63 | 2.9078 | 1445.5795 |
| 21 | 140 | 1.22 | -877.74 | -1.6752 | 770425.98 | 2.8064 | 1470.4039 |
| 22 | 128 | 1.03 | -889.74 | -1.8652 | 791635.72 | 3.4790 | 1659.5569 |
| 23 | 120 | 0.88 | -897.74 | -2.0152 | 805935.55 | 4.0611 | 1809.1395 |
| 24 | N/A | 1.92 | N/A | N/A | N/A | N/A | N/A |
| 25 | N/A | 1.44 | N/A | N/A | N/A | N/A | N/A |
| Mean(1-23) | 1017.73913 | 2.89522 | 0.00 | 0.0000 | 379719.41 | 1.7260 | 806.5561 |
| Sum(1-23) | 23408 | 66.59 | 0.00 | 0.0000 | 8733546.43 | 39.6970 | 18550.7913 |

Table 2.1: Data and sums of squares and cross-products for Carpet Age / Cysteic Acid Study

$$\sum_{i=1}^{n} X_i Y_i = \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 \quad \Rightarrow \quad \sum_{i=1}^{n} X_i e_i = \sum_{i=1}^{n} \left( X_i Y_i - X_i \hat{Y}_i \right) = \sum_{i=1}^{n} \left[ X_i Y_i - \left( \hat{\beta}_0 X_i + \hat{\beta}_1 X_i^2 \right) \right] = 0$$

$$\Rightarrow \quad \sum_{i=1}^{n} \hat{Y}_i e_i = \hat{\beta}_0 \sum_{i=1}^{n} e_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i e_i = 0 + 0 = 0$$

The estimators $\hat{\beta}_1$ and $\hat{\beta}_0$ can be written as linear functions of $Y_1, \ldots, Y_n$:

$$\hat{\beta}_1 = \sum_{i=1}^{n} a_i Y_i \qquad \hat{\beta}_0 = \sum_{i=1}^{n} b_i Y_i \quad \text{where} \quad a_i = \frac{X_i - \overline{X}}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \qquad b_i = \frac{1}{n} - \frac{\overline{X} \left( X_i - \overline{X} \right)}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

and thus using the following basic rules of mathematical statistics given in Chapter 1, we have the following results.

$$E \left\{ \sum_{i=1}^{n} a_i Y_i \right\} = \sum_{i=1}^{n} a_i E \left\{ Y_i \right\} \qquad V \left\{ \sum_{i=1}^{n} a_i Y_i \right\} = \sum_{i=1}^{n} a_i^2 V\{Y_i\} + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_i a_j \text{COV}\{Y_i, Y_j\}$$

The last term of the variance drops out when the data are independent.

$$E \left\{ \hat{\beta}_1 \right\} = \sum_{i=1}^{n} \frac{X_i - \overline{X}}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \left( \beta_0 + \beta_1 X_i \right) = \frac{\beta_0}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \sum_{i=1}^{n} \left( X_i - \overline{X} \right) + \frac{\beta_1}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \sum_{i=1}^{n} \left( X_i - \overline{X} \right) X_i$$

$$= 0 + \frac{\beta_1}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \sum_{i=1}^{n} \left( X_i - \overline{X} \right)^2 = \beta_1$$

$$V \left\{ \hat{\beta}_1 \right\} = \sum_{i=1}^{n} \left[ \frac{X_i - \overline{X}}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \right]^2 \sigma^2 = \frac{\sigma^2}{\left( \sum_{i=1}^{n} (X_i - \overline{X})^2 \right)^2} \sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{\sigma^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

$$E \left\{ \hat{\beta}_0 \right\} = \sum_{i=1}^{n} \left[ \frac{1}{n} - \frac{\overline{X} \left( X_i - \overline{X} \right)}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \right] \left( \beta_0 + \beta_1 X_i \right)$$

$$= n \frac{1}{n} \beta_0 + \frac{1}{n} \beta_1 \sum_{i=1}^{n} X_i - \frac{\overline{X} \beta_0}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \sum_{i=1}^{n} \left( X_i - \overline{X} \right) - \frac{\beta_1 \overline{X}}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \sum_{i=1}^{n} \left( X_i - \overline{X} \right) X_i$$

$$= \beta_0 + \beta_1 \overline{X} - 0 - \beta_1 \overline{X} = \beta_0$$

$$V \left\{ \hat{\beta}_0 \right\} = \sum_{i=1}^{n} \left[ \frac{1}{n} - \frac{\overline{X} \left( X_i - \overline{X} \right)}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \right]^2 \sigma^2$$

$$= \sigma^2 \left[ n \left( \frac{1}{n} \right)^2 + \left( \frac{\overline{X}}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \right)^2 \sum_{i=1}^{n} (X_i - \overline{X})^2 - 2 \frac{\overline{X}}{n \sum_{i=1}^{n} (X_i - \overline{X})^2} \sum_{i=1}^{n} \left( X_i - \overline{X} \right) \right]$$

$$= \sigma^2 \left[ \frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2} \right]$$

$$\text{COV}\left\{\hat{\beta}_0, \hat{\beta}_1\right\} = \sum_{i=1}^n \left( \frac{1}{n} - \frac{\overline{X}\left(X_i - \overline{X}\right)}{\sum_{i=1}^n (X_i - \overline{X})^2} \right) \left( \frac{X_i - \overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2} \right) \sigma^2$$

$$= \sigma^2 \sum_{i=1}^n \left[ \frac{1}{n} \frac{X_i - \overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2} - \frac{\overline{X}\left(X_i - \overline{X}\right)}{\sum_{i=1}^n (X_i - \overline{X})^2} \frac{X_i - \overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2} \right] = -\frac{\overline{X}\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2}$$

Thus, the sampling distributions of $\hat{\beta}_1$ and $\hat{\beta}_0$, assuming independent, normal errors with constant variance are:

$$\hat{\beta}_1 \sim N\left( \beta_1, \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2} \right) \qquad \hat{\beta}_0 \sim N\left( \beta_0, \sigma^2 \left[ \frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2} \right] \right)$$

The standard error is the square root of the variance, and the estimated standard error is the standard error with the unknown $\sigma^2$ replaced by $MSE$.

$$\hat{SE}\{\hat{\beta}_1\} = \sqrt{\frac{MSE}{\sum_{i=1}^n (X_i - \overline{X})^2}} \qquad \hat{SE}\{\hat{\beta}_0\} = \sqrt{MSE \left[ \frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2} \right]}$$

Making use of these derivations, we obtain the following useful results, which are simple to show in matrix form.

$$\text{COV}\left\{\overline{Y}, \hat{\beta}_1\right\} = \sigma^2 \sum_{i=1}^n \left( \frac{1}{n} \right) \left( \frac{X_i - \overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2} \right) = 0$$

$$\text{COV}\left\{Y_i, \hat{Y}_i\right\} = \text{COV}\left\{Y_i, \overline{Y} + \hat{\beta}_1\left(X_i - \overline{X}\right)\right\} = \sigma^2 \left[ \frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{\sum_{i=1}^n (X_i - \overline{X})^2} \right]$$

$$\text{COV}\left\{Y_i, e_i\right\} = \text{COV}\left\{Y_i, Y_i - \hat{Y}_i\right\} = \sigma^2 \left[ 1 - \left( \frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{\sum_{i=1}^n (X_i - \overline{X})^2} \right) \right]$$

$$\text{COV}\left\{\hat{Y}_i, e_i\right\} = \text{COV}\left\{\hat{Y}_i, Y_i - \hat{Y}_i\right\} = \sigma^2 \left[ \frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{\sum_{i=1}^n (X_i - \overline{X})^2} \right] - \sigma^2 \left[ \frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{\sum_{i=1}^n (X_i - \overline{X})^2} \right] = 0$$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = \overline{Y} + \hat{\beta}_1\left(X_i - \overline{X}\right) \quad \Rightarrow \quad E\left\{\hat{Y}_i\right\} = \beta_0 + \beta_1 X_i \qquad V\left\{\hat{Y}_i\right\} = \sigma^2 \left[ \frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{\sum_{i=1}^n (X_i - \overline{X})^2} \right]$$

$$e_i = Y_i - \hat{Y}_i \quad \Rightarrow \quad E\{e_i\} = 0$$

$$V\{e_i\} = \sigma^2 \left[ 1 + \left( \frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{\sum_{i=1}^n (X_i - \overline{X})^2} \right) - 2 \left( \frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{\sum_{i=1}^n (X_i - \overline{X})^2} \right) \right] = \sigma^2 \left[ 1 - \left( \frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{\sum_{i=1}^n (X_i - \overline{X})^2} \right) \right]$$

## 2.2 Inference Regarding $\beta_1$

Primarily of interest is inferences regarding $\beta_1$. Note that if $\beta_1 = 0$, $Y$ and $X$ are not associated. We can test hypotheses and construct confidence intervals based on the estimate $\beta_1$ and its estimated standard error. The $t$-test is conducted as follows. Note that the null value $\beta_{10}$ is almost always 0, and that software packages that report these tests always are treating $\beta_{10}$ as 0. Here, and in all other tests, $TS$ represents Test Statistic, and $RR$ represents Rejection Region.

$$H_0 : \beta_1 = \beta_{10} \qquad H_A : \beta_1 \neq \beta_{10} \quad TS : t_{obs} = \frac{\hat{\beta}_1 - \beta_{10}}{\hat{SE}\{\hat{\beta}_1\}} \qquad RR : |t_{obs}| \geq t_{\alpha/2,n-2} \qquad P\text{-value} : 2P(t_{n-2} \geq |t_{obs}|)$$

One-sided tests use the same test statistic, but the Rejection Region and $P$-value are changed to reflect the alternative hypothesis.

$$H_A^+ : \beta_1 > \beta_{10} \qquad RR : t_{obs} \geq t_{\alpha,n-2} \qquad P\text{-value} : P(t_{n-2} \geq t_{obs})$$

$$H_A^- : \beta_1 < \beta_{10} \qquad RR : t_{obs} \leq -t_{\alpha,n-2} \qquad P\text{-value} : P(t_{n-2} \leq t_{obs})$$

A $(1-\alpha)100\%$ confidence interval for $\beta_1$ is obtained as:

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2}\hat{SE}\{\hat{\beta}_1\}$$

Note that the confidence interval represents the values of $\beta_{10}$ that the two-sided test:

$$H_0 : \beta_1 = \beta_{10} \qquad H_A : \beta_1 \neq \beta_{10}$$

fails to reject the null hypothesis.

Inferences regarding $\beta_0$ are rarely of interest, but can be conducted in analogous manner, using the estimate $\hat{\beta}_0$ and its estimated standard error $\hat{SE}\{\hat{\beta}_0\}$.

**Example: Cysteic Acid Content and Age of Carpets**

$$\hat{\beta}_1 = 467.310 \quad \hat{\beta}_0 = -335.225 \quad n = 23 \quad MSE = 3075.38 \quad \overline{X} = 2.8952 \quad \sum_{i=1}^{n}(X_i - \overline{X})^2 = 39.6970$$

$$\Rightarrow \qquad \hat{SE}\{\hat{\beta}_1\} = \sqrt{\frac{3075.38}{39.6970}} = 8.802 \qquad \hat{SE}\{\hat{\beta}_0\} = \sqrt{3075.38\left[\frac{1}{23} + \frac{(2.8952)^2}{39.6970}\right]} = 27.984$$

$$H_0 : \beta_1 = 0 \qquad H_A : \beta_1 \neq 0 \qquad TS : t_{obs} = \frac{467.310}{8.802} = 53.091 \qquad RR : |t_{obs}| \geq t_{.025,23-1} = 2.080$$

$$P\text{-value} : 2P(t_{23-2} \geq |53.091|) \approx 0 \qquad 95\%\text{CI for } \beta_1 : \quad 467.310 \pm 2.080(8.802) \quad \equiv \quad (449.002, 485.618)$$

$$H_0 : \beta_0 = 0 \qquad H_A : \beta_0 \neq 0 \qquad TS : t_{obs} = \frac{-335.225}{27.984} = -11.980 \qquad RR : |t_{obs}| \geq t_{.025,23-1} = 2.080$$

$$P\text{-value} : 2P(t_{23-2} \geq |-11.980|) < .0001 \qquad 95\%\text{CI for } \beta_0 : \quad -335.225 \pm 2.080(27.984) \quad \equiv \quad (-393.432, -277.018)$$

The R commands and output are given below.

```
### Commands

(beta1_hat <- SS_XY / SS_XX)
(beta0_hat <- ybar - beta1_hat * xbar)
(SS_ERR <- sum((age_f - Y_hat)^2));  (df_ERR <- n-2); (MS_ERR <- SS_ERR/df_ERR)
(SE_beta1_hat <- sqrt(MS_ERR / SS_XX))
(SE_beta0_hat <- sqrt(MS_ERR * ((1/n) + xbar^2/SS_XX)))
(t_beta1 <- beta1_hat / SE_beta1_hat)
(t_beta0 <- beta0_hat / SE_beta0_hat)
(t_crit <- qt(.975,n-2))
(P_beta1 <- 2*(1-pt(abs(t_beta1),n-2)))
(P_beta0 <- 2*(1-pt(abs(t_beta0),n-2)))
(CI95_beta1 <- beta1_hat + qt(c(.025,.975),n-2) * SE_beta1_hat)
(CI95_beta0 <- beta0_hat + qt(c(.025,.975),n-2) * SE_beta0_hat)

### Output

> (beta1_hat <- SS_XY / SS_XX)
[1] 467.31
> (beta0_hat <- ybar - beta1_hat * xbar)
[1] -335.2248
> (SE_beta1_hat <- sqrt(MS_ERR / SS_XX))
[1] 8.801368
> (SE_beta0_hat <- sqrt(MS_ERR * ((1/n) + xbar^2/SS_XX)))
[1] 27.98259
> (t_beta1 <- beta1_hat / SE_beta1_hat)
[1] 53.09515
> (t_beta0 <- beta0_hat / SE_beta0_hat)
[1] -11.97976
> (t_crit <- qt(.975,n-2))
[1] 2.079614
> (P_beta1 <- 2*(1-pt(abs(t_beta1),n-2)))
[1] 0
> (P_beta0 <- 2*(1-pt(abs(t_beta0),n-2)))
[1] 7.511813e-11
> (CI95_beta1 <- beta1_hat + qt(c(.025,.975),n-2) * SE_beta1_hat)
[1] 449.0065 485.6134
> (CI95_beta0 <- beta0_hat + qt(c(.025,.975),n-2) * SE_beta0_hat)
[1] -393.4178 -277.0318
```

$\nabla$

## 2.3 Estimating a Mean and Predicting a New Observation @ $X = X^*$

We may want to estimate the mean response at a specific level $X^*$. The parameter of interest is $\mu^* = \beta_0 + \beta_1 X^*$. The point estimate for the conditional mean of $Y$, given $X = X^*$, its mean and variance are as follow.

$$\hat{Y}^* = \hat{\beta}_0 + \hat{\beta}_1 X^* \qquad E\left\{\hat{Y}^*\right\} = E\left\{\hat{\beta}_0 + \hat{\beta}_1 X^*\right\} = \beta_0 + \beta_1 X^*$$

$$V\left\{\hat{Y}^*\right\} = V\left\{\hat{\beta}_0 + \hat{\beta}_1 X^*\right\} = \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2} + (X^*)^2 \frac{1}{\sum_{i=1}^n (X_i - \overline{X})^2} + 2X^*\left(-\frac{\overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)\right]$$

$$= \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2} + \frac{(X^*)^2}{\sum_{i=1}^n (X_i - \overline{X})^2} - \frac{2X^*\overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2}\right] = \sigma^2\left[\frac{1}{n} + \frac{\left(X^* - \overline{X}\right)^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right]$$

The point estimate, standard error, and $(1-\alpha)100\%$ Confidence Interval are given below:

$$\hat{Y}^* = \hat{\beta}_0 + \hat{\beta}_1 X^* \qquad \hat{SE}\left\{\hat{Y}^*\right\} = \sqrt{MSE\left[\frac{1}{n} + \frac{\left(X^* - \overline{X}\right)^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right]} \qquad (1-\alpha)100\%\text{ CI } : \hat{Y}^* \pm t_{\alpha/2,n-2}\hat{SE}\left\{\hat{Y}^*\right\}$$

To obtain a $(1-\alpha)100\%$ Confidence Interval for the entire regression line (not just a single point), we can use the Working-Hotelling method.

$$\hat{Y}^* \pm \sqrt{2F_{\alpha,2,n-2}}\hat{SE}\left\{\hat{Y}^*\right\}$$

If we are interested in predicting a new observation when $X = X^*$, we have uncertainty with respect to estimating the mean (as seen by the Confidence Interval above), and the random error for the new case (with standard deviation $\sigma$). The point prediction is the same as for the mean, $\hat{Y}^*$. The prediction error, the difference between the actual value $Y_{\text{New}}$ and its prediction $\hat{Y}^*$, and its variance are as follow. Note that the new observation will be independent of its prediction, as it was not used in the calibration of the regression model.

$$\text{Prediction Error:} \quad Y_{\text{New}} - \hat{Y}^*$$

$$V\left\{Y_{\text{New}} - \hat{Y}^*\right\} = V\left\{Y_{\text{New}}\right\} + V\left\{\hat{Y}^*\right\} = \sigma^2\left[1 + \frac{1}{n} + \frac{\left(X^* - \overline{X}\right)^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right]$$

The prediction, estimated standard error of prediction, and $(1-\alpha)100\%$ Prediction Interval are given below.

$$\hat{Y}^* = \hat{\beta}_0 + \hat{\beta}_1 X^* \qquad \hat{SE}\left\{\hat{Y}^*_{\text{New}}\right\} = \sqrt{MSE\left[1 + \frac{1}{n} + \frac{\left(X^* - \overline{X}\right)^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right]}$$

$$(1-\alpha)100\%\text{ PI } : \hat{Y}^* \pm t_{\alpha/2,n-2}\hat{SE}\left\{\hat{Y}^*_{\text{New}}\right\}$$

Note that the Prediction Interval will tend to be much wider than the Confidence Interval for the mean when $MSE$ is not very small.

**Example: Cysteic Acid Content and Age of Carpets**

The Confidence Interval for the Mean and Prediction Interval for a single observation are based on the following quantities as a function of $X^*$.

$$95\%\text{ CI for Mean: } (-335.225 + 467.310X^*) \pm 2.080\sqrt{3075.38\left[\frac{1}{23} + \frac{(X^* - 2.8952)^2}{39.6970}\right]}$$

$$95\%\text{ PI for Individual Specimen: } (-335.225 + 467.310X^*) \pm 2.080\sqrt{3075.38\left[1 + \frac{1}{23} + \frac{(X^* - 2.8952)^2}{39.6970}\right]}$$

In particular, specimens 24 and 25 had $X_{24} = 1.92$ and $X_{25} = 1.44$, respectively, with no age reported. We construct 95% Prediction Intervals for their ages.

$$\hat{Y}_{24} = -335.225 + 467.310(1.92) = 562.01 \qquad 562.01 \pm 2.080\sqrt{3075.38\left[1 + \frac{1}{23} + \frac{(1.92 - 2.8952)^2}{39.6970}\right]}$$

$$\equiv \quad 562.02 \pm 119.17 \quad \equiv \quad (442.85, 681.19)$$

$$\hat{Y}_{24} = -335.225 + 467.310(1.44) = 337.70 \qquad 337.70 \pm 2.080\sqrt{3075.38\left[1 + \frac{1}{23} + \frac{(1.44 - 2.8952)^2}{39.6970}\right]}$$

$$\equiv \quad 337.70 \pm 120.80 \quad \equiv \quad (216.90, 458.50)$$

The data, fitted equation, and pointwise (not simultaneous) Confidence and Prediction Intervals are given in Figure 2.2. Note that for the Working-Hotelling simultaneous confidence intervals, we would replace $t_{.025,21} = 2.080$ with $\sqrt{2F_{.05,2,21}} = \sqrt{2(3.467)} = 2.633$. The R commands and output are given below.

```
### Commands
X_s <- seq(0,5,0.01)
yhat_h <- beta0_hat + beta1_hat * X_s
```

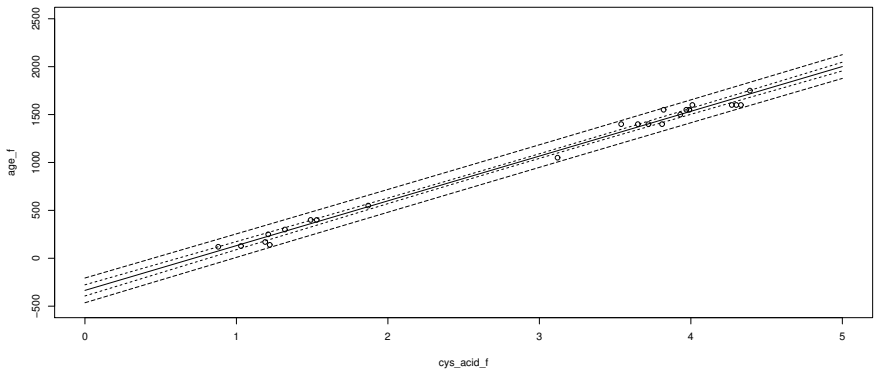


Figure 2.2: Plot of Age ($Y$) versus Cysteic Acid ($X$), Fitted Equation, Confidence Interval for Mean, and Prediction Interval for Individual

```
CI_LO <- yhat_h + qt(.025,n-2) * sqrt(MS_ERR*((1/n)+(X_s-xbar)^2/SS_XX))
CI_HI <- yhat_h + qt(.975,n-2) * sqrt(MS_ERR*((1/n)+(X_s-xbar)^2/SS_XX))
PI_LO <- yhat_h + qt(.025,n-2) * sqrt(MS_ERR*(1 + (1/n)+(X_s-xbar)^2/SS_XX))
PI_HI <- yhat_h + qt(.975,n-2) * sqrt(MS_ERR*(1 + (1/n)+(X_s-xbar)^2/SS_XX))

plot(cys_acid_f,age_f,xlim=c(0,5),ylim=c(-500,2500))
lines(X_s,yhat_h,lty=1)
lines(X_h,CI_LO,lty=2)
lines(X_h,CI_HI,lty=2)
lines(X_h,PI_LO,lty=5)
lines(X_h,PI_HI,lty=5)

(yhat_miss <- beta0_hat + beta1_hat * cys_acid_m)
(PE_miss <- sqrt(MS_ERR * (1 + (1/n) + (cys_acid_m - xbar)^2/SS_XX)))
(PI_age_24 <- yhat_miss[1] + qt(c(.025,.975),n-2) * PE_miss[1])
(PI_age_25 <- yhat_miss[2] + qt(c(.025,.975),n-2) * PE_miss[2])

### Output
> (yhat_miss <- beta0_hat + beta1_hat * cys_acid_m)
[1] 562.0103 337.7015
> (PE_miss <- sqrt(MS_ERR * (1 + (1/n) + (cys_acid_m - xbar)^2/SS_XX)))
[1] 57.29277 58.07609
> (PI_age_24 <- yhat_miss[1] + qt(c(.025,.975),n-2) * PE_miss[1])
[1] 442.8635 681.1572
> (PI_age_25 <- yhat_miss[2] + qt(c(.025,.975),n-2) * PE_miss[2])
[1] 216.9257 458.4774
```

$\nabla$

## 2.4 Analysis of Variance

When there is no association between $Y$ and $X$ ($\beta_1 = 0$), the best predictor of each observation is $\overline{Y} = \hat{\beta}_0$ (in terms of minimizing sum of squares of prediction errors). In this case, the total variation can be denoted as

$TSS = \sum_{i=1}^n (Y_i - \overline{Y})^2$, the **Total (Corrected) Sum of Squares**. Note that some software packages also print the **Total (Uncorrected) Sum of Squares**, $USS = \sum_{i=1}^n Y_i^2$. Unless specifically stated, in these notes, whenever referring to the Total Sum of Squares, we mean the corrected version.

When there is an association between $Y$ and $X$ ($\beta_1 \neq 0$), the best predictor of each observation is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = \overline{Y} + \hat{\beta}_1 \left(X_i - \overline{X}\right)$ (in terms of minimizing sum of squares of prediction errors). In this case, the error variation can be denoted as $SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$, the **Error (Residual) Sum of Squares**.

The difference between $TSS$ and $SSE$ is the variation "explained" by the regression of $Y$ on $X$ (as opposed to having ignored $X$). It represents the difference between the fitted values and the mean: $SSR = \sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2$, the **Regression Sum of Squares**.

$$Y_i - \overline{Y} = \left(Y_i - \hat{Y}_i\right) + \left(\hat{Y}_i - \overline{Y}\right) \quad \Rightarrow \quad \left(Y_i - \overline{Y}\right)^2 = \left(Y_i - \hat{Y}_i\right)^2 + \left(\hat{Y}_i - \overline{Y}\right)^2 + 2\left(Y_i - \hat{Y}_i\right)\left(\hat{Y}_i - \overline{Y}\right)$$

$$\sum_{i=1}^n \left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^n \left(Y_i - \hat{Y}_i\right)^2 + \sum_{i=1}^n \left(\hat{Y}_i - \overline{Y}\right)^2 + 2\sum_{i=1}^n e_i \left(\overline{Y} + \hat{\beta}_1 \left(X_i - \overline{X}\right) - \overline{Y}\right) =$$

$$\sum_{i=1}^n \left(Y_i - \hat{Y}_i\right)^2 + \sum_{i=1}^n \left(\hat{Y}_i - \overline{Y}\right)^2 + 0$$

The last portion is 0 as $\sum_{i=1}^n e_i = \sum_{i=1}^n X_i e_i = 0$.

$$\Rightarrow \quad TSS = SSE + SSR \qquad \sum_{i=1}^n (Y_i - \overline{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2$$

These sums of squares can be expanded as follow.

$$\sum_{i=1}^n (Y_i - \overline{Y})^2 = \sum_{i=1}^n \left[Y_i^2 + \overline{Y}^2 - 2Y_i\overline{Y}\right] = \sum_{i=1}^n Y_i^2 + n\overline{Y}^2 - 2n\overline{Y}^2 = \sum_{i=1}^n Y_i^2 - n\overline{Y}^2$$

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n \left[Y_i^2 + \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)^2 - 2Y_i\left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)\right]$$

$$= \sum_{i=1}^n \left[Y_i^2 + \left(\overline{Y} + \hat{\beta}_1 \left(X_i - \overline{X}\right)\right)^2 - 2Y_i\left(\overline{Y} + \hat{\beta}_1 \left(X_i - \overline{X}\right)\right)\right]$$

$$= \sum_{i=1}^n Y_i^2 + n\overline{Y}^2 + \hat{\beta}_1^2 \sum_{i=1}^n (X_i - \overline{X})^2 + 2\hat{\beta}_1\overline{Y}\sum_{i=1}^n \left(X_i - \overline{X}\right) - 2n\overline{Y}^2 - 2\hat{\beta}_1 \sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})$$

$$= \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2 + \left( \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \right)^2 \sum_{i=1}^{n}(X_i - \overline{X})^2 - 2\frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

$$= \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2 - \frac{\left(\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})\right)^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = TSS - \hat{\beta}_1^2 \sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = \sum_{i=1}^{n}\left[\overline{Y} + \hat{\beta}_1\left(X_i - \overline{X}\right) - \overline{Y}\right]^2 = \hat{\beta}_1^2 \sum_{i=1}^{n}(X_i - \overline{X})^2$$

From these expansions we can derive the expected sums of squares for error and regression.

$$E\left\{Y_i^2\right\} = \sigma^2 + (\beta_0 + \beta_1 X_i)^2 \qquad E\left\{\sum_{i=1}^{n} Y_i^2\right\} = n\sigma^2 + n\beta_0^2 + \beta_1^2 \sum_{i=1}^{n} X_i^2 + 2\beta_0\beta_1 \sum_{i=1}^{n} X_i$$

$$E\left\{\overline{Y}^2\right\} = \frac{\sigma^2}{n} + \left(\beta_0 + \beta_1\overline{X}\right)^2 \qquad E\left\{n\overline{Y}^2\right\} = \sigma^2 + n\beta_0^2 + n\beta_1^2\overline{X}^2 + 2n\beta_0\beta_1\overline{X}$$

$$E\left\{\hat{\beta}_1^2\right\} = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2} + \beta_1^2 \qquad E\left\{\hat{\beta}_1^2 \sum_{i=1}^{n}(X_i - \overline{X})^2\right\} = \sigma^2 + \beta_1^2 \sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$E\{TSS\} = \left(n\sigma^2 + n\beta_0^2 + \beta_1^2 \sum_{i=1}^{n} X_i^2 + 2\beta_0\beta_1 \sum_{i=1}^{n} X_i\right) - \left(\sigma^2 + n\beta_0^2 + n\beta_1^2\overline{X}^2 + 2n\beta_0\beta_1\overline{X}\right)$$

$$= (n-1)\sigma^2 + \beta_1^2 \sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$E\{SSE\} = E\{SSE\} = \left((n-1)\sigma^2 + \beta_1^2 \sum_{i=1}^{n}(X_i - \overline{X})^2\right) - \left(\sigma^2 + \beta_1^2 \sum_{i=1}^{n}(X_i - \overline{X})^2\right) = (n-2)\sigma^2$$

$$E\{SSR\} = \sigma^2 + \beta_1^2 \sum_{i=1}^{n}(X_i - \overline{X})^2$$

Each sum of squares has a **degrees of freedom** associated with it. The **Total Degrees of Freedom** is $df_{\text{Total}} = n - 1$. The **Error Degrees of Freedom** are $df_{\text{Error}} = n - 2$ (for simple regression). The **Regression Degrees of Freedom** are $df_{\text{Regression}} = 1$ (for simple regression).

| Source | $df$ | $SS$ | $MS$ | $F_{obs}$ | $P$-value |
|---|---|---|---|---|---|
| Regression (Model) | 1 | $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$ | $MSR = \frac{SSR}{1}$ | $F_{obs} = \frac{MSR}{MSE}$ | $P(F_{1,n-2} \geq F_{obs})$ |
| Error (Residual) | $n-2$ | $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ | $MSE = \frac{SSE}{n-2}$ | | |
| Total (Corrected) | $n-1$ | $TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$ | | | |

Table 2.2: Analysis of Variance Table for Simple Linear Regression

$$df_{\text{Total}} = df_{\text{Error}} + df_{\text{Regression}} \qquad n-1 = n-2+1$$

The Error and Regression sums of squares each have a **Mean Square**, which is the sum of squares divided by its corresponding degrees of freedom: $MSE = SSE/(n-2)$ and $MSR = SSR/1$. These mean squares have the following **Expected Values**, average values in repeated sampling at the same observed $X$ levels.

$$E\{MSE\} = \sigma^2 \qquad E\{MSR\} = \sigma^2 + \beta_1^2 \sum_{i=1}^{n}(X_i - \overline{X})^2$$

Note that when $\beta_1 = 0$, then $E\{MSR\} = E\{MSE\}$, otherwise $E\{MSR\} > E\{MSE\}$. A second way of testing whether $\beta_1 = 0$ is by the $F$-test:

$$H_0 : \beta_1 = 0 \qquad H_A : \beta_1 \neq 0 \quad TS : F_{obs} = \frac{MSR}{MSE} \qquad RR : F_{obs} \geq F_{\alpha,1,n-2} \qquad P\text{-value} : P(F_{1,n-2} \geq F_{obs})$$

The Analysis of Variance is typically set up in a table as in Table 2.2.

Another sum of squares that is often computed by software packages is the sum of squares for the intercept, $SS\mu$, which is computed as follows, with the following relationships with the 2 versions of the total sum of squares. The degrees of freedom for the intercept is 1.

$$SS\mu = n\overline{Y}^2 \qquad TSS = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2 = USS - SS\mu$$

A measure often reported from a regression analysis is the **Coefficient of Determination** or $r^2$. This represents the variation in $Y$ "explained" by $X$, divided by the total variation in $Y$.

$$r^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = \frac{SSR}{TSS} = 1 - \frac{SSE}{TSS} \qquad 0 \leq r^2 \leq 1$$

The interpretation of $r^2$ is the proportion of variation in $Y$ that is"explained" by $X$, and is often reported as a percentage ($100r^2$).

**Example: Cysteic Acid Content and Age of Carpets**

Table 2.3 gives the observed data, fitted values, residuals, and the computations to obtain the sums of squares and the Analysis of Variance. The ANOVA table is in Table 2.4.

$$TSS = 8733546.4 \qquad SSE = 64576.9 \qquad SSR = 8668969.7 \qquad r^2 = \frac{8668969.7}{8733546.4} = .9926$$

The R commands and output are given below.

```
### Commands
(SS_YY <- sum((age_f - ybar)^2))
(SS_ERR <- sum((age_f - Y_hat)^2));  (df_ERR <- n-2); (MS_ERR <- SS_ERR/df_ERR)
(SS_REG <- sum((Y_hat - ybar)^2)); (df_REG <- 1); (MS_REG <- SS_REG/df_REG)
(F_obs <- MS_REG / MS_ERR)
(F_crit <- qf(.95,1,n-2))
(P_F <- 1 - pf(F_obs,1,n-2))
(r_square <- SS_REG / SS_YY)

### Output
> (SS_YY <- sum((age_f - ybar)^2))
[1] 8733546
> (SS_ERR <- sum((age_f - Y_hat)^2));  (df_ERR <- n-2); (MS_ERR <- SS_ERR/df_ERR)
[1] 64576.89
[1] 21
[1] 3075.09
> (SS_REG <- sum((Y_hat - ybar)^2)); (df_REG <- 1); (MS_REG <- SS_REG/df_REG)
[1] 8668970
[1] 1
[1] 8668970
> (F_obs <- MS_REG / MS_ERR)
[1] 2819.095
> (F_crit <- qf(.95,1,n-2))
[1] 4.324794
> (P_F <- 1 - pf(F_obs,1,n-2))
[1] 0
> (r_square <- SS_REG / SS_YY)
[1] 0.9926059
```

$\nabla$

## 2.5 Correlation

The regression coefficient $\beta_1$ depends on the units of $Y$ and $X$. It also depends on which variable is the dependent variable and which is the independent variable. A second widely reported measure is the **Pearson Product Moment Coefficient of Correlation**. It is invariant to linear transformations of $Y$ and/or $X$, and does not distinguish which is the dependent and which is the independent variable. This makes it a widely reported measure when researchers are interested in how 2 random variables vary together in a

| id | age ($Y$) | cys acid ($X$) | $\hat{Y}$ | $e = Y - \hat{Y}$ | $(Y - \overline{Y})^2$ | $e^2$ | $(\hat{Y} - \overline{Y})^2$ |
|---|---|---|---|---|---|---|---|
| 1 | 1750 | 4.39 | 1716.266 | 33.73407 | 536206 | 1137.987 | 487939.7 |
| 2 | 1600 | 4.3 | 1674.208 | -74.208 | 339027.7 | 5506.833 | 430951.4 |
| 3 | 1600 | 4.27 | 1660.189 | -60.1887 | 339027.7 | 3622.684 | 412741.5 |
| 4 | 1600 | 4.33 | 1688.227 | -88.2273 | 339027.7 | 7784.063 | 449554.4 |
| 5 | 1600 | 4.01 | 1538.688 | 61.31185 | 339027.7 | 3759.143 | 271387.9 |
| 6 | 1550 | 3.99 | 1529.342 | 20.65805 | 283301.6 | 426.7551 | 261737.4 |
| 7 | 1550 | 3.97 | 1519.996 | 30.00425 | 283301.6 | 900.2551 | 252261.7 |
| 8 | 1550 | 3.82 | 1449.899 | 100.1007 | 283301.6 | 10020.16 | 186762.4 |
| 9 | 1500 | 3.93 | 1501.303 | -1.30335 | 232575.5 | 1.698722 | 233834.4 |
| 10 | 1400 | 3.81 | 1445.226 | -45.2262 | 146123.4 | 2045.405 | 182745.2 |
| 11 | 1400 | 3.54 | 1319.052 | 80.94753 | 146123.4 | 6552.503 | 90789.73 |
| 12 | 1400 | 3.72 | 1403.168 | -3.16826 | 146123.4 | 10.03786 | 148555.6 |
| 13 | 1400 | 3.65 | 1370.457 | 29.54344 | 146123.4 | 872.8148 | 124409.6 |
| 14 | 1050 | 3.12 | 1122.782 | -72.7823 | 1040.764 | 5297.261 | 11034.06 |
| 15 | 550 | 1.87 | 538.6448 | 11.35517 | 218779.9 | 128.9398 | 229531.3 |
| 16 | 400 | 1.49 | 361.067 | 38.93295 | 381601.6 | 1515.775 | 431218.2 |
| 17 | 400 | 1.53 | 379.7594 | 20.24055 | 381601.6 | 409.68 | 407018.1 |
| 18 | 300 | 1.32 | 281.6244 | 18.37565 | 515149.5 | 337.6644 | 541865 |
| 19 | 250 | 1.21 | 230.2203 | 19.77974 | 589423.4 | 391.2382 | 620186 |
| 20 | 170 | 1.19 | 220.8741 | -50.8741 | 718661.6 | 2588.17 | 634993.9 |
| 21 | 140 | 1.22 | 234.8934 | -94.8934 | 770426 | 9004.749 | 612847.5 |
| 22 | 128 | 1.03 | 146.1045 | -18.1045 | 791635.7 | 327.7717 | 759747 |
| 23 | 120 | 0.88 | 76.00797 | 43.99203 | 805935.5 | 1935.299 | 886857.6 |
| 24 | N/A | 1.92 | N/A | NA | NA | NA | NA |
| 25 | N/A | 1.44 | N/A | NA | NA | NA | NA |
| Mean(1-23) | 1017.739 | 2.895217 | 1017.739 | .0000 | 379719.4 | 2807.691 | 376911.7 |
| Sum(1-23) | 23408 | 66.59 | 23408 | .0000 | 8733546.4 | 64576.9 | 8668969.7 |

Table 2.3: Data and ANOVA computations for Carpet Age / Cysteic Acid Study

| Source | *df* | *SS* | *MS* | $F_{obs}$ | $F_{.05}$ | $P$-value |
|---|---|---|---|---|---|---|
| Regression (Model) | 1 | 8668969.7 | $\frac{8668969.7}{1} = 8668969.7$ | $\frac{8668969.7}{3075.1} = 2819.1$ | 4.325 | .0000 |
| Error (Residual) | $23 - 2 = 21$ | 64576.9 | $\frac{64576.9}{21} = 3075.1$ | | | |
| Total (Corrected) | $23 - 1 = 22$ | 8733546.4 | | | | |

Table 2.4: Analysis of Variance Table for Carpet Aging Study

population. The population correlation coefficient is labeled $\rho$, and the sample correlation is labeled $r$, and is computed as:

$$r = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2 \sum_{i=1}^n (Y_i - \overline{Y})^2}} = \left(\frac{S_X}{S_Y}\right)\hat{\beta}_1.$$

where $S_X$ and $S_Y$ are the standard deviations of $X$ and $Y$, respectively. While $\hat{\beta}_1$ can take on any value, $r$ lies between $-1$ and $+1$, taking on the extreme values if all of the points fall on a straight line. The test of whether $\rho = 0$ is mathematically equivalent to the $t$-test for testing whether $\beta_1 = 0$. The 2-sided test is given below.

$$H_0 : \rho = 0 \qquad H_A : \rho \neq 0 \quad TS : t_{obs} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \qquad RR : |t_{obs}| \geq t_{\alpha/2,n-2} \qquad P - \text{value} : 2P(t_{n-2} \geq |t_{obs}|)$$

To construct a large-sample confidence interval, we use **Fisher's $z$ transform** to make $r$ approximately normal. We then construct a confidence interval on the transformed correlation, then "back transform" the end points.

$$z' = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \qquad (1-\alpha)100\% \text{ CI for } \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) : \qquad z' \pm z_{\alpha/2}\sqrt{\frac{1}{n-3}}$$

Labeling the endpoints of the Confidence Interval as $(a, b)$, we obtain:

$$(1-\alpha)100\% \text{ Confidence Interval for } \rho : \qquad \left(\frac{e^{2a}-1}{e^{2a}+1}, \frac{e^{2b}-1}{e^{2b}+1}\right)$$

**Example: Cysteic Acid Content and Age of Carpets**

For the Carpet Aging analysis, we obtain the following calculations.

$$r = \frac{18550.7913}{\sqrt{(39.6970)(8733546.4)}} = \frac{18550.7913}{18619.7635} = .9963$$

$$H_0 : \rho = 0 \qquad H_A : \rho \neq 0 \qquad TS : t_obs = \frac{.9963}{\sqrt{\frac{1-.9963^2}{23-2}}} = 53.10 \qquad RR : |t_{obs}| \geq t_{.025,23-2} = 2.080$$

Clearly, the P-value is 0. Next, we compute a 95% Confidence Interval for $\rho$ based on Fisher's $z$ transform and back-transforming.

$$z' = \frac{1}{2}\ln\left(\frac{1+.9963}{1-.9963}\right) = 3.1454 \qquad 95\% \text{ CI for } \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) : \qquad 3.1454 \pm 1.96\sqrt{\frac{1}{22-3}} \quad \equiv \quad (2.6957, 3.5951)$$

$$95\% \text{ Confidence Interval for } \rho: \quad \left(\frac{e^{2(2.6957)}-1}{e^{2(2.6957}+1}, \frac{e^{2(3.5951}-1}{e^{2(3.5951}+1}\right) \quad \equiv \quad (.9909, .9985)$$

The R commands and output are given below.

```
### Commands
### Correlation Test/CI
(r <- SS_XY / sqrt(SS_XX * SS_YY))
(t_r <- sqrt(n-2) * (r / sqrt(1 - r^2)))
(t_r_crit <- qt(.975,n-2))
(P_r <- 2*(1 - pt(abs(t_r),n-2)))

(z_r <- 0.5 * log((1+r) / (1-r)))
(CI_z_rho <- z_r + qnorm(c(.025,.975),0,1) * sqrt(1 / (n-3)))
(CI_rho <- (exp(2*CI_z_rho) - 1) / (exp(2*CI_z_rho) + 1))

### Output
> (r <- SS_XY / sqrt(SS_XX * SS_YY))
[1] 0.9962961
> (t_r <- sqrt(n-2) * (r / sqrt(1 - r^2)))
[1] 53.09515
> (t_r_crit <- qt(.975,n-2))
[1] 2.079614
> (P_r <- 2*(1 - pt(abs(t_r),n-2)))
[1] 0

> (z_r <- 0.5 * log((1+r) / (1-r)))
[1] 3.144829
> (CI_z_rho <- z_r + qnorm(c(.025,.975),0,1) * sqrt(1 / (n-3)))
[1] 2.706567 3.583090
> (CI_rho <- (exp(2*CI_z_rho) - 1) / (exp(2*CI_z_rho) + 1))
[1] 0.9911243 0.9984567
```

$\nabla$

## 2.6 Regression Through the Origin

In some applications, it is believed that $E\{Y\}$ is proportional to $X$. That is, $E\{Y|X=0\}=0$. Note that some people interpret this as $Y=0$ when $X=0$. However, that would imply that $V\{Y|X=0\}=0$, which is not consistent with the model. We consider regression through the origin here, as it is useful in specific applications and is also used in many diagnostic tests for model adequacy.

$$Y=\beta_1 X+\epsilon \qquad \epsilon \sim N(0,\sigma^2) \text{ independent} \quad \Rightarrow \quad Y \sim N(\beta_1 X, \sigma^2) \text{ independent}$$

The least squares estimator is derived below.

$$Q=\sum_{i=1}^{n}\epsilon_i^2=\sum_{i=1}^{n}[Y_i-\beta_1 X_i]^2 \qquad \Rightarrow$$

$$\frac{\partial Q}{\partial \beta_1} = 2\sum_{i=1}^{n}\left[Y_i - \beta_1 X_i\right]^{2-1}(-X_i) \quad = \quad -2\left[\sum_{i=1}^{n} X_i Y_i - \beta_1 \sum_{i=1}^{n} X_i^2\right]$$

$$\Rightarrow \quad \sum_{i=1}^{n} X_i Y_i = \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 \quad \Rightarrow \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$$

The mean and variance of $\hat{\beta}_1$ are obtained below, along with its sampling distribution.

$$E\left\{\hat{\beta}_1\right\} = \frac{1}{\sum_{i=1}^{n} X_i^2}\sum_{i=1}^{n} X_i \beta_1 X_i = \beta_1 \frac{\sum_{i=1}^{n} X_i^2}{\sum_{i=1}^{n} X_i^2} = \beta_1$$

$$V\left\{\hat{\beta}_1\right\} = \frac{1}{\left(\sum_{i=1}^{n} X_i^2\right)^2}\sum_{i=1}^{n} X_i^2 \sigma^2 = \frac{\sigma^2}{\sum_{i=1}^{n} X_i^2} \qquad \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n} X_i^2}\right)$$

A second, also unbiased, estimator of $\beta_1$ is given below.

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i} \quad \Rightarrow \quad E\left\{\tilde{\beta}_1\right\} = \frac{1}{\sum_{i=1}^{n} X_i}\sum_{i=1}^{n} \beta_1 X_i = \beta_1$$

$$V\left\{\tilde{\beta}_1\right\} = \frac{1}{\left(\sum_{i=1}^{n} X_i\right)^2}\sum_{i=1}^{n} \sigma^2 = \frac{n\sigma^2}{\left(\sum_{i=1}^{n} X_i\right)^2}$$

Note that the least squares estimator has a smaller variance than the second estimator. This can be seen as considering the difference between their reciprocals.

$$\frac{1}{V\left\{\hat{\beta}_1\right\}} - \frac{1}{V\left\{\tilde{\beta}_1\right\}} = \frac{1}{\sigma^2}\left[\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}\right] = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} > 0$$

$$\Rightarrow \quad \frac{1}{V\left\{\hat{\beta}_1\right\}} > \frac{1}{V\left\{\tilde{\beta}_1\right\}} \quad \Rightarrow \quad V\left\{\hat{\beta}_1\right\} < V\left\{\tilde{\beta}_1\right\}$$

For this model, the error sum of squares, with $n-1$ degrees of freedom, and its Expected Mean Square are given below.

$$SSE = \sum_{i=1}^{n}\left(Y_i - \hat{\beta}_1 X_i\right)^2 = \sum_{i=1}^{n} Y_i^2 + \hat{\beta}_1^2 \sum_{i=1}^{n} X_i^2 - 2\hat{\beta}_1 \sum_{i=1}^{n} X_i Y_i =$$

$$\sum_{i=1}^{n} Y_i^2 + \left(\frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}\right)^2 \sum_{i=1}^{n} X_i^2 - 2\frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} \sum_{i=1}^{n} X_i Y_i = \sum_{i=1}^{n} Y_i^2 - \hat{\beta}_1^2 \sum_{i=1}^{n} X_i^2$$

$$E\{SSE\} = n\sigma^2 + \beta_1^2 \sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} X_i^2 \left(\frac{\sigma^2}{\sum_{i=1}^{n} X_i^2} + \beta_1^2\right) = (n-1)\sigma^2 \quad \Rightarrow \quad E\{MSE\} = E\left\{\frac{SSE}{n-1}\right\} = \sigma^2$$

$$\Rightarrow \quad \hat{SE}\left\{\hat{\beta}_1\right\} = \sqrt{\frac{MSE}{\sum_{i=1}^{n} X_i^2}}$$

The t-test and Confidence Interval for $\beta_1$ are obtained as follow.

$$H_0 : \beta_1 = \beta_{10} \quad H_A : \beta_1 \neq \beta_{10} \quad TS : t_{obs} = \frac{\hat{\beta}_1 - \beta_{10}}{\hat{SE}\left\{\hat{\beta}_1\right\}} \quad RR : |t_{obs}| \geq t_{\alpha/2,n-1}$$

$$(1-\alpha)\,100\% \text{ Confidence Interval for } \beta_1 : \quad \hat{\beta}_1 \pm t_{\alpha/2,n-1} \hat{SE}\left\{\hat{\beta}_1\right\}$$

Be wary of the Analysis of Variance and coefficient of determination reported by statistical software packages. Since the regression line does not necessarily go through the point $(\overline{X}, \overline{Y})$, the Total Sum of Squares, Regression Sum of Squares, and coefficient of determination are computed as follow.

$$TSS = \sum_{i=1}^{n} Y_i^2 \qquad SSR = \sum_{i=1}^{n} \hat{Y}_i^2 \qquad r^2 = \frac{\sum_{i=1}^{n} \hat{Y}_i^2}{\sum_{i=1}^{n} Y_i^2} \quad \neq \quad \frac{\sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}$$

**Example: In-vivo and in-vitro Assessments of Sunscreen Product SPF**

Miksa, Lutz, and Guy (2013) reported a study of 3 in-vitro methods of assessing $n = 32$ suntan products' SPF. There were 3 in-vitro methods that were compared with in-vivo. The in-vitro methods were:

- MPP = Molded polymethyl methacylate (PMMA) plates
- MSSP = Molded skin-mimicking PMMA plates
- SPP = sand-blasted PMMA plates

The researchers fit regressions through the origin where $Y$ was the in-vitro method (1-at-a-time) and $X$ was the in-vivo measurement. The data (means for each method for each product) are given in Table 2.5,

and a plot of MPP in-vitro versus in-vivo assessments along with the fitted regression through the origin. Calculations for the estimated slope for in-vitro MPP ($Y1$) are shown here, based on results from the spreadsheet generated Table 2.5.

$$\sum_{i=1}^{n} X_i Y_i = 74775.34 \qquad \sum_{i=1}^{n} X_i^2 = 75513.74 \qquad \hat{\beta}_1 = \frac{74775.34}{75513.74} = 0.990222 \qquad \hat{Y}_i = 0.99022 X_i$$

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = 16167.06 \qquad s^2 = MSE = \frac{16167.06}{32-1} = 521.52 \qquad \hat{SE}\left\{\hat{\beta}_1\right\} = \sqrt{\frac{521.52}{75513.74}} = 0.0831$$

$$95\% \text{ CI for } \beta_1 : \qquad 0.9902 \pm 2.040(0.0831) \quad \equiv \quad 0.9902 \pm 0.1695 \quad \equiv \quad (0.8207, 1.1597)$$

$$SSR = 74044.16 \qquad TSS = 90211.22 \qquad r^2 = \frac{74044.16}{90211.22} = .8208$$

The Analysis of Variance is given in Table 2.6. Direct calculations for all three methods in R are obtained below.

```
### Commands
spf <- read.csv("http://www.stat.ufl.edu/~winner/data/spf_substrate.csv",
     header=T)
attach(spf); names(spf)

### Scalar form Analysis

n <- length(SPF1)
X <- SPF1; Y1 <- SPF2;  Y2 <- SPF3;  Y3 <- SPF4

(S_XX <- sum(X^2))
(S_XY1 <- sum(X*Y1))
(S_XY2 <- sum(X*Y2))
(S_XY3 <- sum(X*Y3))

(beta1_hat <- S_XY1 / S_XX)
(beta2_hat <- S_XY2 / S_XX)
(beta3_hat <- S_XY3 / S_XX)

Y1_hat <- X * beta1_hat
Y2_hat <- X * beta2_hat
Y3_hat <- X * beta3_hat

(SS_ERR1 <- sum((Y1 - Y1_hat)^2))
(SS_ERR2 <- sum((Y2 - Y2_hat)^2))
(SS_ERR3 <- sum((Y3 - Y3_hat)^2))

(s2_1 <- SS_ERR1 / (n-1))
(s2_2 <- SS_ERR2 / (n-1))
(s2_3 <- SS_ERR3 / (n-1))

(SE_beta1_hat <- sqrt(s2_1 / S_XX))
(SE_beta2_hat <- sqrt(s2_2 / S_XX))
(SE_beta3_hat <- sqrt(s2_3 / S_XX))

(CI_beta1 <- beta1_hat + qt(c(.025,.975),n-1) * SE_beta1_hat)
(CI_beta2 <- beta2_hat + qt(c(.025,.975),n-1) * SE_beta2_hat)
(CI_beta3 <- beta3_hat + qt(c(.025,.975),n-1) * SE_beta3_hat)
```

```
(SS_REG1 <- sum(Y1_hat^2))
(SS_REG2 <- sum(Y2_hat^2))
(SS_REG3 <- sum(Y3_hat^2))

(r_square1 <- SS_REG1 / sum(Y1^2))
(r_square2 <- SS_REG2 / sum(Y2^2))
(r_square3 <- SS_REG3 / sum(Y3^2))

### Output

> (S_XX <- sum(X^2))
[1] 75513.74
> (S_XY1 <- sum(X*Y1))
[1] 74775.34
> (S_XY2 <- sum(X*Y2))
[1] 116792.4
> (S_XY3 <- sum(X*Y3))
[1] 130454.6
>
> (beta1_hat <- S_XY1 / S_XX)
[1] 0.9902216
> (beta2_hat <- S_XY2 / S_XX)
[1] 1.546638
> (beta3_hat <- S_XY3 / S_XX)
[1] 1.727561
>
> Y1_hat <- X * beta1_hat
> Y2_hat <- X * beta2_hat
> Y3_hat <- X * beta3_hat
>
> (SS_ERR1 <- sum((Y1 - Y1_hat)^2))
[1] 16167.06
> (SS_ERR2 <- sum((Y2 - Y2_hat)^2))
[1] 86836.99
> (SS_ERR3 <- sum((Y3 - Y3_hat)^2))
[1] 35947.7
> (s2_1 <- SS_ERR1 / (n-1))
[1] 521.5181
> (s2_2 <- SS_ERR2 / (n-1))
[1] 2801.193
> (s2_3 <- SS_ERR3 / (n-1))
[1] 1159.603
> (SE_beta1_hat <- sqrt(s2_1 / S_XX))
[1] 0.08310395
> (SE_beta2_hat <- sqrt(s2_2 / S_XX))
[1] 0.192601
> (SE_beta3_hat <- sqrt(s2_3 / S_XX))
[1] 0.1239201
> (CI_beta1 <- beta1_hat + qt(c(.025,.975),n-1) * SE_beta1_hat)
[1] 0.820730 1.159713
> (CI_beta2 <- beta2_hat + qt(c(.025,.975),n-1) * SE_beta2_hat)
[1] 1.153826 1.939451
> (CI_beta3 <- beta3_hat + qt(c(.025,.975),n-1) * SE_beta3_hat)
[1] 1.474824 1.980297
> (SS_REG1 <- sum(Y1_hat^2))
[1] 74044.16
> (SS_REG2 <- sum(Y2_hat^2))
[1] 180635.7
> (SS_REG3 <- sum(Y3_hat^2))
[1] 225368.2
> (r_square1 <- SS_REG1 / sum(Y1^2))
[1] 0.8207866
> (r_square2 <- SS_REG2 / sum(Y2^2))
[1] 0.6753425
```

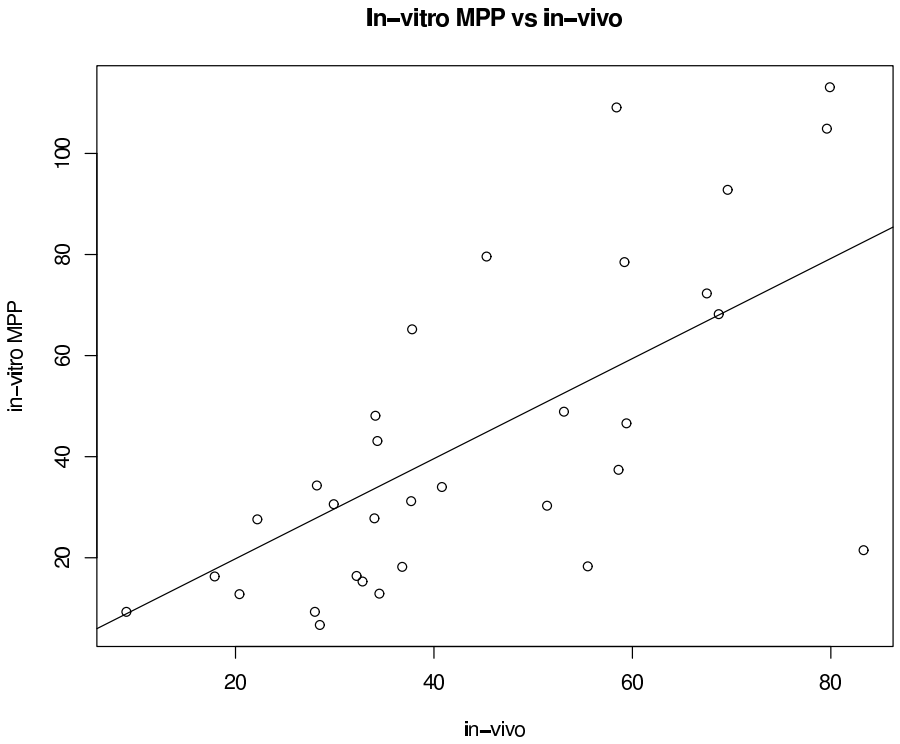


Figure 2.3: Plot of in-vitro MPP vs in-vivo with Regression through the Origin

```
> (r_square3 <- SS_REG3 / sum(Y3^2))
[1] 0.8624358
```

## 2.7 Case of Random Independent Variable

So far, $X$ has been considered to be a fixed constant, which typically occurs in controlled experiments. When $X$ is a random variable, but it is independent of the random error term $\epsilon$, everything derived previously holds, but the inference is conditional of the observed $X$ levels in the sample. In the first model considered, $Y$ and $X$ are jointly distributed as bivariate normal random variables. In the second model, $X$ is assumed to be random and $Y$ is related to $X$, with the error term being independent of $X$.

### 2.7.1 Bivariate Normal $Y$ and $X$

In the "classical simple regression model," $Y$ and $X$ are jointly distributed as bivariate normal as defined below.

$$E\{X\} = \mu_X \quad V\{X\} = \sigma_X^2 \quad E\{Y\} = \mu_Y \quad V\{Y\} = \sigma_Y^2 \quad \text{CORR}\{X, Y\} = \rho$$

| Product | in-vivo ($X$) | MPP($Y1$) | MSSP($Y2$) | SPP($Y3$) | $X^2$ | $XY1$ | $\hat{Y}1$ | $e1$ |
|---|---|---|---|---|---|---|---|---|
| P1 | 9 | 9.3 | 14.7 | 16.5 | 81 | 83.7 | 8.91 | 0.39 |
| P2 | 20.4 | 12.8 | 27.3 | 29.3 | 416.16 | 261.12 | 20.20 | -7.40 |
| P3 | 32.2 | 16.4 | 27.6 | 48.1 | 1036.84 | 528.08 | 31.89 | -15.49 |
| P4 | 40.8 | 34 | 58.3 | 64.6 | 1664.64 | 1387.2 | 40.40 | -6.40 |
| P5 | 32.8 | 15.3 | 23.1 | 44 | 1075.84 | 501.84 | 32.48 | -17.18 |
| P6 | 28.2 | 34.3 | 31.4 | 120 | 795.24 | 967.26 | 27.92 | 6.38 |
| P7 | 83.3 | 21.5 | 42.7 | 69.4 | 6938.89 | 1790.95 | 82.49 | -60.99 |
| P8 | 79.9 | 113.1 | 201.1 | 96.1 | 6384.01 | 9036.69 | 79.12 | 33.98 |
| P9 | 79.6 | 104.9 | 308.1 | 173.3 | 6336.16 | 8350.04 | 78.82 | 26.08 |
| P10 | 69.6 | 92.8 | 128.4 | 153.9 | 4844.16 | 6458.88 | 68.92 | 23.88 |
| P11 | 68.7 | 68.2 | 32.7 | 88.4 | 4719.69 | 4685.34 | 68.03 | 0.17 |
| P12 | 67.5 | 72.3 | 56.1 | 172.4 | 4556.25 | 4880.25 | 66.84 | 5.46 |
| P13 | 59.4 | 46.6 | 36.8 | 80.1 | 3528.36 | 2768.04 | 58.82 | -12.22 |
| P14 | 59.2 | 78.5 | 154.3 | 113.7 | 3504.64 | 4647.2 | 58.62 | 19.88 |
| P15 | 58.6 | 37.4 | 69.2 | 140.5 | 3433.96 | 2191.64 | 58.03 | -20.63 |
| P16 | 58.4 | 109.1 | 186.1 | 80.8 | 3410.56 | 6371.44 | 57.83 | 51.27 |
| P17 | 55.5 | 18.3 | 37.5 | 81.9 | 3080.25 | 1015.65 | 54.96 | -36.66 |
| P18 | 53.1 | 48.9 | 90.9 | 64.5 | 2819.61 | 2596.59 | 52.58 | -3.68 |
| P19 | 51.4 | 30.3 | 28 | 100.4 | 2641.96 | 1557.42 | 50.90 | -20.60 |
| P20 | 45.3 | 79.6 | 84 | 98.1 | 2052.09 | 3605.88 | 44.86 | 34.74 |
| P21 | 37.8 | 65.2 | 61.7 | 95.5 | 1428.84 | 2464.56 | 37.43 | 27.77 |
| P22 | 37.7 | 31.2 | 53.8 | 39.1 | 1421.29 | 1176.24 | 37.33 | -6.13 |
| P23 | 36.8 | 18.2 | 30.4 | 63.2 | 1354.24 | 669.76 | 36.44 | -18.24 |
| P24 | 34.5 | 12.9 | 9.6 | 46.2 | 1190.25 | 445.05 | 34.16 | -21.26 |
| P25 | 34.3 | 43.1 | 78 | 62.2 | 1176.49 | 1478.33 | 33.96 | 9.14 |
| P26 | 34.1 | 48.1 | 25 | 51.6 | 1162.81 | 1640.21 | 33.77 | 14.33 |
| P27 | 34 | 27.8 | 38.2 | 38.4 | 1156 | 945.2 | 33.67 | -5.87 |
| P28 | 29.9 | 30.6 | 26.9 | 151.5 | 894.01 | 914.94 | 29.61 | 0.99 |
| P29 | 28.5 | 6.7 | 4.5 | 34.3 | 812.25 | 190.95 | 28.22 | -21.52 |
| P30 | 28 | 9.3 | 17.8 | 53 | 784 | 260.4 | 27.73 | -18.43 |
| P31 | 22.2 | 27.6 | 60.7 | 56.2 | 492.84 | 612.72 | 21.98 | 5.62 |
| P32 | 17.9 | 16.3 | 14.6 | 31.5 | 320.41 | 291.77 | 17.72 | -1.42 |
| Sum | 1428.6 | 1380.6 | 2059.5 | 2558.7 | 75513.74 | 74775.34 | 1414.63 | -34.03 |
| SumSq | 75513.74 | 90211.22 | 267472.65 | 261315.85 | | | 74044.16 | 16167.06 |

Table 2.5: in-vivo and 3 in-vitro assessments of SPF for 32 suntan products

| Source | $df$ | $SS$ | $MS$ | $F_{obs}$ | $F_{.05}$ | $P$-value |
|---|---|---|---|---|---|---|
| Regression | 1 | 74044.16 | $\frac{74044.16}{1} = 74044.16$ | $\frac{74044.16}{521.52} = 141.98$ | 4.160 | .0000 |
| Error | $32 - 1 = 31$ | 16167.06 | $\frac{16167.06}{31} = 521.52$ | | | |
| Total (Uncorrected) | 32 | 90211.22 | | | | |

Table 2.6: Analysis of Variance Table for Suntan Product SPF Study

$$f(x,y)=(2\pi)^{-1}\left(\sigma_X^2\sigma_Y^2\left(1-\rho^2\right)\right)^{-1/2}\exp\left\{-\frac{1}{2\left(1-\rho^2\right)}\left[\frac{\left(x-\mu_X\right)^2}{\sigma_X^2}-\frac{2\rho\left(x-\mu_X\right)\left(y-\mu_Y\right)}{\sigma_X\sigma_Y}+\frac{\left(y-\mu_Y\right)^2}{\sigma_Y^2}\right]\right\}$$

$$-\infty<x,y<\infty,-\infty<\mu_X,\mu_Y<\infty,\sigma_X,\sigma_Y>0,-1\leq\rho\leq1$$

To obtain the **marginal distribution** of $X$, integrate $Y$ out of the joint density. Note that the final marginal distribution for $X$ is normal with mean $\mu_X$ and variance $\sigma_X^2$. This makes use of "completing the square" and forming a new normal density.

$$f_X(x)=\int_{-\infty}^{\infty}f(x,y)dy=(2\pi)^{-1}\left(\sigma_X^2\sigma_Y^2\left(1-\rho^2\right)\right)^{-1/2}\times$$

$$\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2\left(1-\rho^2\right)}\left[\frac{\left(x-\mu_X\right)^2\sigma_Y^2-2\rho\left(x-\mu_X\right)\left(y-\mu_Y\right)\sigma_X\sigma_Y+\left(y-\mu_Y\right)^2\sigma_X^2}{\sigma_X^2\sigma_Y^2}\right]\right\}dy$$

$$=(2\pi)^{-1}\left(\sigma_X^2\sigma_Y^2\left(1-\rho^2\right)\right)^{-1/2}\times$$

$$\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2\left(1-\rho^2\right)\sigma_X^2\sigma_Y^2}\left[\left(x-\mu_X\right)^2\sigma_Y^2-2\rho\left(x-\mu_X\right)\left(y-\mu_Y\right)\sigma_X\sigma_Y+\left(y-\mu_Y\right)^2\sigma_X^2+\left(x-\mu_X\right)^2\sigma_Y^2\rho^2-\left(x-\mu_X\right)^2\sigma_Y^2\rho^2\right]\right\}dy$$

$$=(2\pi)^{-1}\left(\sigma_X^2\sigma_Y^2\left(1-\rho^2\right)\right)^{-1/2}\times$$

$$\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2\left(1-\rho^2\right)\sigma_X^2\sigma_Y^2}\left[\left(x-\mu_X\right)^2\sigma_Y^2\rho^2-2\rho\left(x-\mu_X\right)\left(y-\mu_Y\right)\sigma_X\sigma_Y+\left(y-\mu_Y\right)^2\sigma_X^2+\left(x-\mu_X\right)^2\sigma_Y^2\left(1-\rho^2\right)\right]\right\}dy=$$

$$(2\pi)^{-1}\left(\sigma_X^2\sigma_Y^2\left(1-\rho^2\right)\right)^{-1/2}\times$$

$$\exp\left\{-\frac{\left(x-\mu_X\right)^2\sigma_Y^2\left(1-\rho^2\right)}{2\left(1-\rho^2\right)\sigma_X^2\sigma_Y^2}\right\}\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2\left(1-\rho^2\right)\sigma_X^2\sigma_Y^2}\left[\left(x-\mu_X\right)\sigma_Y\rho-\left(y-\mu_Y\right)\sigma_X\right]^2\right\}dy=$$

$$(2\pi)^{-1}\left(\sigma_X^2\sigma_Y^2\left(1-\rho^2\right)\right)^{-1/2}\times$$

$$\exp\left\{-\frac{\left(x-\mu_X\right)^2\sigma_Y^2\left(1-\rho^2\right)}{2\left(1-\rho^2\right)\sigma_X^2\sigma_Y^2}\right\}\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2\left(1-\rho^2\right)\sigma_Y^2}\left[\left(y-\mu_Y\right)-\rho\left(x-\mu_X\right)\frac{\sigma_Y}{\sigma_X}\right]^2\right\}dy=$$

$$(2\pi)^{-1}\left(\sigma_X^2\sigma_Y^2\left(1-\rho^2\right)\right)^{-1/2}\times$$

$$\exp\left\{-\frac{\left(x-\mu_X\right)^2\sigma_Y^2\left(1-\rho^2\right)}{2\left(1-\rho^2\right)\sigma_X^2\sigma_Y^2}\right\}\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2\left(1-\rho^2\right)\sigma_Y^2}\left[y-\left[\mu_Y+\rho\left(x-\mu_X\right)\frac{\sigma_Y}{\sigma_X}\right]\right]^2\right\}dy$$

In the integral, $Y$ is normal with mean $\mu_Y+\rho\left(x-\mu_X\right)\frac{\sigma_Y}{\sigma_X}$ and variance $\left(1-\rho^2\right)\sigma_Y^2$, so that the integral is equal to the reciprocal of the usual normalizing constant: $\sqrt{2\pi\left(1-\rho^2\right)\sigma_Y^2}$. This implies that $f_X(x)$ can be written as follows.

$$f_X(x)=(2\pi)^{-1}\left(\sigma_X^2\sigma_Y^2\left(1-\rho^2\right)\right)^{-1/2}\times$$

$$\exp\left\{-\frac{\left(x-\mu_X\right)^2\sigma_Y^2\left(1-\rho^2\right)}{2\left(1-\rho^2\right)\sigma_X^2\sigma_Y^2}\right\}\sqrt{2\pi\left(1-\rho^2\right)\sigma_Y^2}=$$

$$\frac{1}{\sqrt{2\pi\sigma_X^2}}\exp\left\{-\frac{(x-\mu_X)^2}{2\sigma_X^2}\right\} \quad -\infty < x < \infty$$

The **conditional distribution** of $Y$ given $X = x$, is the ratio of the joint density divided by the marginal density of $X$ (each evaluated at $X = x$).

$$f(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{\sqrt{2\pi\sigma_X^2}}{2\pi\sqrt{\sigma_X^2\sigma_Y^2(1-\rho^2)}}\times$$

$$\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right] + \frac{1}{2}\frac{(x-\mu_X)^2}{\sigma_X^2}\right\}$$

Multiplying the last term in the exponent by $\left(1-\rho^2\right)/\left(1-\rho^2\right)$ yields the following.

$$f(y|x) = \frac{1}{\sqrt{2\pi\sigma_Y^2(1-\rho^2)}}\times$$

$$\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2}\left(1-\left(1-\rho^2\right)\right) - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right\} =$$

$$\frac{1}{\sqrt{2\pi\sigma_Y^2(1-\rho^2)}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{y-\mu_Y}{\sigma_Y} - \frac{(x-\mu_X)\rho}{\sigma_X}\right]^2\right\} =$$

$$\frac{1}{\sqrt{2\pi\sigma_Y^2(1-\rho^2)}}\exp\left\{-\frac{1}{2\sigma_Y^2(1-\rho^2)}\left[(y-\mu_Y) - \frac{(x-\mu_X)\rho\sigma_Y}{\sigma_X}\right]^2\right\} =$$

$$\frac{1}{\sqrt{2\pi\sigma_Y^2(1-\rho^2)}}\exp\left\{-\frac{1}{2\sigma_Y^2(1-\rho^2)}\left[y - \left[\mu_Y + \frac{(x-\mu_X)\rho\sigma_Y}{\sigma_X}\right]\right]^2\right\} \quad -\infty < y < \infty$$

Given $X = x$, $Y$ is normally distributed with conditional mean $\mu_Y + (x-\mu_X)\rho\sigma_Y/\sigma_X$ and conditional variance $\sigma_Y^2\left(1-\rho^2\right)$. Notice that in terms of the linear regression model, the following results are obtained.

$$E\{Y|X=x\} = \left(\mu_Y - \mu_X\frac{\rho\sigma_Y}{\sigma_X}\right) + x\frac{\rho\sigma_Y}{\sigma_X} \qquad \rho = \frac{\text{COV}(X,Y)}{\sigma_X\sigma_Y} \quad \Rightarrow$$

$$\frac{\rho\sigma_Y}{\sigma_X} = \frac{\text{COV}(X,Y)}{\sigma_X^2} = \beta_1 \qquad \mu_Y - \mu_X\frac{\rho\sigma_Y}{\sigma_X} = \mu_Y - \beta_1\mu_X = \beta_0$$

The conditional variance of $Y$, given $X = x$ is the unconditional variance of $Y$ ($\sigma_Y^2$) times $1-\rho^2$, and does not depend on $X$.

## 2.8 R Programs and Output Based on lm Function

For both of the datasets used in this chapter, the R programs based on use of the built-in **lm** function are given below.

### 2.8.1 Carpet Aging Analysis

#### R Program

```
carpet <- read.csv("http://www.stat.ufl.edu/~winner/data/carpet_age.csv",
      header=T)
attach(carpet); names(carpet)

### f ==> full data  m==> missing age
age_f <- age[1:23]; age_m <- age[24:25]
cys_acid_f <- cys_acid[1:23]; cys_acid_m <- cys_acid[24:25]

########## lm function
carpet.mod1 <- lm(age_f ~ cys_acid_f)
summary(carpet.mod1)
anova(carpet.mod1)
confint(carpet.mod1)
predict(carpet.mod1,list(cys_acid_f=cys_acid_m),int="p")

plot(cys_acid_f, age_f)
abline(carpet.mod1)

cor.test(cys_acid_f, age_f)
```

#### R Output

```
> carpet.mod1 <- lm(age_f ~ cys_acid_f)
> summary(carpet.mod1)

Call:
lm(formula = age_f ~ cys_acid_f)

Residuals:
   Min     1Q Median     3Q    Max
-94.89 -48.05  18.38  31.87 100.10

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -335.225     27.983  -11.98 7.51e-11 ***
cys_acid_f   467.310      8.801   53.09  < 2e-16 ***

Residual standard error: 55.45 on 21 degrees of freedom
Multiple R-squared:  0.9926,    Adjusted R-squared:  0.9923
F-statistic:  2819 on 1 and 21 DF,  p-value: < 2.2e-16

> anova(carpet.mod1)
```

```
Analysis of Variance Table

Response: age_f
           Df  Sum Sq Mean Sq F value    Pr(>F)
cys_acid_f  1 8668970 8668970  2819.1 < 2.2e-16 ***
Residuals  21   64577    3075

> confint(carpet.mod1)
                2.5 %    97.5 %
(Intercept) -393.4178 -277.0318
cys_acid_f   449.0065  485.6134
> predict(carpet.mod1,list(cys_acid_f=cys_acid_m),int="p")
       fit      lwr      upr
1 562.0103 442.8635 681.1572
2 337.7015 216.9257 458.4774
>
>
> cor.test(cys_acid_f, age_f)

        Pearson's product-moment correlation

data:  cys_acid_f and age_f
t = 53.0951, df = 21, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9911243 0.9984567
sample estimates:
      cor
0.9962961
```

### 2.8.2 Suntan Product SPF Assessments

#### R Program

```
spf <- read.csv("http://www.stat.ufl.edu/~winner/data/spf_substrate.csv",
     header=T)
attach(spf); names(spf)

spf.mod2 <- lm(SPF2 ~ SPF1 - 1)
summary(spf.mod2)
anova(spf.mod2)
confint(spf.mod2)

plot(SPF1, SPF2,main="In-vitro MPP vs in-vivo",xlab="in-vivo",
   ylab="in-vitro MPP")
abline(spf.mod2)
```

#### R Output

```
> spf.mod2 <- lm(SPF2 ~ SPF1 - 1)
> summary(spf.mod2)

Call:
lm(formula = SPF2 ~ SPF1 - 1)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-60.985 -17.444  -2.553  10.435  51.271

Coefficients:
     Estimate Std. Error t value Pr(>|t|)
SPF1   0.9902     0.0831   11.91 4.17e-13 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 22.84 on 31 degrees of freedom
Multiple R-squared:  0.8208,    Adjusted R-squared:  0.815
F-statistic:   142 on 1 and 31 DF,  p-value: 4.169e-13

> anova(spf.mod2)
Analysis of Variance Table

Response: SPF2
          Df Sum Sq Mean Sq F value    Pr(>F)
SPF1       1  74044   74044  141.98 4.169e-13 ***
Residuals 31  16167     522
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
> confint(spf.mod2)
       2.5 %   97.5 %
SPF1 0.82073 1.159713
```

# Chapter 3

# Matrix Form of Simple Linear Regression

We can write out the regression model in a more concise form using the **matrix form**. This is particularly helpful when we have multiple predictors. We first "string out" the dependent variable ($Y$), and the predictor variable ($X$) into **arrays**. In fact, we augment the $X^s$ with a column of $1^s$ for the intercept:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

We can make use of some basic matrix rules to simplify the algebra of regression models. Note that matrices with one row or column are referred to as **vectors**. Matrices with the same number of rows and columns are referred to as **square matrices**. When referring to elements of matrices, the row represents the first subscript, and the column is the second subscript. Vector elements have one subscript.

The **transpose** of a matrix or vector, is the matrix or vector obtained by interchanging its rows and columns (turning it on its side, counterclockwise, and flipping it upside down). It is typically written with a "prime" or "T" as a superscript.

$$\mathbf{Y}' = \begin{bmatrix} Y_1 & Y_2 & \cdots & Y_n \end{bmatrix} \quad \mathbf{X}' = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{bmatrix} \quad \boldsymbol{\beta}' = \begin{bmatrix} \beta_0 & \beta_1 \end{bmatrix} \quad \boldsymbol{\varepsilon}' = \begin{bmatrix} \epsilon_1 & \epsilon_2 & \cdots & \epsilon_n \end{bmatrix}$$

**Matrix Addition/Subtraction:** If two matrices are of the same dimension (numbers of rows and columns), then the matrix formed by adding/subtracting each of the elements within the given rows and columns is the addition/subtraction of the two matrices.

$$\mathbf{A} = \begin{bmatrix} 4 & 8 \\ 2 & -4 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 3 & 1 \\ 8 & 6 \end{bmatrix} \quad \Rightarrow$$

$$\mathbf{A}+\mathbf{B} = \begin{bmatrix} 4+3 & 8+1 \\ 2+8 & -4+6 \end{bmatrix} = \begin{bmatrix} 7 & 9 \\ 10 & 2 \end{bmatrix} \qquad \mathbf{A}-\mathbf{B} = \begin{bmatrix} 4-3 & 8-1 \\ 2-8 & -4-6 \end{bmatrix} = \begin{bmatrix} 1 & 7 \\ -6 & -10 \end{bmatrix}$$

**Matrix Multiplication:** Unlike Addition/Subtraction, Multiplication takes sums of products of matrix elements. The number of **columns** of the **left-hand** matrix must be equal to the number of **rows** of the **right-hand** matrix. The resulting matrix has the same number of rows of the left-hand matrix and the number of columns as the right-hand matrix. Note that multiplication of square matrices of common dimensions will result in a square matrix of the same dimension. The elements of a matrix created by multiplication are the sums of products of elements in the rows of the left-hand matrix with the elements of the columns of the right-hand matrix.

$$\mathbf{AB} = \begin{bmatrix} 4 & 8 \\ 2 & -4 \end{bmatrix} \begin{bmatrix} 3 & 1 \\ 8 & 6 \end{bmatrix} = \begin{bmatrix} 76 & 52 \\ -26 & -22 \end{bmatrix}$$

Note the computation of elements of $\mathbf{AB}$.

$$\mathbf{AB}_{11} = 4(3)+8(8) = 12+64 = 76 \qquad \mathbf{AB}_{12} = 4(1)+8(6) = 4+48 = 52$$

$$\mathbf{AB}_{21} = 2(3)+(-4)(8) = 6-32 = -26 \qquad \mathbf{AB}_{22} = 2(1)+(-4)(6) = 2-24 = -22$$

$$\text{In General: } \mathbf{AB}_{ij} = \sum_{k=1}^{c_\mathbf{A}=r_\mathbf{B}} a_{ik}b_{kj} \qquad i = 1,\ldots,r_\mathbf{A}; \quad j = 1,\ldots,c_\mathbf{B}$$

Important matrix multiplications for the simple linear regression model are:

$$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 1(\beta_0)+X_1(\beta_1) \\ 1(\beta_0)+X_2(\beta_1) \\ \vdots \\ 1(\beta_0)+X_n(\beta_1) \end{bmatrix} = \begin{bmatrix} \beta_0+\beta_1X_1 \\ \beta_0+\beta_1X_2 \\ \vdots \\ \beta_0+\beta_1X_n \end{bmatrix}$$

The statistical model in matrix form (which easily generalizes to multiple predictor variables) is written as:

$$y_i = \mathbf{x_i}'\boldsymbol{\beta} = \begin{bmatrix} 1 & X_i \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \epsilon_i \quad \Rightarrow \quad \mathbf{Y} = \mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon} \quad \text{where } \mathbf{X} = \begin{bmatrix} \mathbf{x_1}' \\ \vdots \\ \mathbf{x_n}' \end{bmatrix}.$$

Other matrices used in model estimation are:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{bmatrix} \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} = \begin{bmatrix} 1(1)+\cdots+1(1) & 1(X_1)+\cdots+1(X_n) \\ X_1(1)+\cdots+X_n(1) & X_1^2+\cdots+X_n^2 \end{bmatrix}$$

$$\Rightarrow \quad \mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{bmatrix} \qquad \mathbf{X}'\mathbf{Y} = \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_iY_i \end{bmatrix} \qquad \mathbf{Y}'\mathbf{Y} = \left[\sum_{i=1}^n Y_i^2\right]$$

**Identity and Unit Matrices:** The **identity** (or **I**) matrix is a square matrix with $1^s$ on the main diagonal, and $0^s$ elsewhere. When the identity matrix is multiplied by any multiplication-compatible matrix, it reproduces the multiplied matrix. Thus, it acts like 1 in scalar arithmetic. The **unit** (or **J**) matrix is a matrix of $1^s$ in all cells. When the unit matrix is multiplied by a multiplication-compatible matrix, it sums the elements of each column (and reproduces the sums for each row).

$$\mathbf{IA} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 4 & 8 \\ 2 & -4 \end{bmatrix} = \begin{bmatrix} 1(4)+0(2) & 1(8)+0(-4) \\ 0(4)+1(2) & 0(8)+1(-4) \end{bmatrix} = \begin{bmatrix} 4 & 8 \\ 2 & -4 \end{bmatrix} = \mathbf{A}$$

$$\mathbf{JA} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 4 & 8 \\ 2 & -4 \end{bmatrix} = \begin{bmatrix} 1(4)+1(2) & 1(8)+1(-4) \\ 1(4)+1(2) & 1(8)+1(-4) \end{bmatrix} = \begin{bmatrix} 6 & 4 \\ 6 & 4 \end{bmatrix}$$

**Matrix Inversion:** If a matrix is square and of full rank (no linear functions of a set of columns/rows are equal to another column/row), then an inverse exists. Note that in simple regression, this simply means that the $X$ levels are not all the same among observations. When we have more than one predictor variable, it means that none of the predictors is a linear function of the other predictors. When a square, full rank matrix is multiplied by its inverse, we obtain the identity matrix. This is analogous to the scalar operation: $a(1/a) = 1$, assuming $a \neq 0$. For a $2 \times 2$ matrix, the inverse is simple to compute. For larger matrices, we will use computers to obtain them.

$$\mathbf{D} = \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix} \Rightarrow \mathbf{D}^{-1} = \frac{1}{D_{11}D_{22} - D_{12}D_{21}} \begin{bmatrix} D_{22} & -D_{12} \\ -D_{21} & D_{11} \end{bmatrix}$$

Note that if **D** is not full rank (its columns/rows) are multiples of each other, $D_{11}D_{22} - D_{12}D_{21} = 0$, and its inverse does not exist.

$$\mathbf{A} = \begin{bmatrix} 4 & 8 \\ 2 & -4 \end{bmatrix} \Rightarrow \mathbf{A^{-1}} = \frac{1}{4(-4) - 8(2)} \begin{bmatrix} -4 & -8 \\ -2 & 4 \end{bmatrix} = \frac{1}{-32} \begin{bmatrix} -4 & -8 \\ -2 & 4 \end{bmatrix} = \begin{bmatrix} \frac{1}{8} & \frac{1}{4} \\ \frac{1}{16} & -\frac{1}{8} \end{bmatrix}$$

Confirm that $\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. Serious rounding errors can occur when the division of the determinant $\frac{1}{D_{11}D_{22}-D_{12}D_{21}}$ is rounded down to too few decimal places.

Some useful results are as follow, assuming matrices are compatible for the operations, which always holds when each is square and of the same dimension. The second result also assumes that each is full rank, meaning their inverses exist.

$$(\mathbf{AB})' = \mathbf{B}'\mathbf{A}' \qquad (\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1} \qquad \text{trace}(\mathbf{AB}) = \text{trace}(\mathbf{BA})$$

The last result involves the **trace** of the matrix, which constitutes the sum of the diagonal elements of a square matrix. Each of the results extends to 3 or more matrices in a similar manner.

$$(\mathbf{ABC})' = \mathbf{C}'\mathbf{B}'\mathbf{A}' \qquad (\mathbf{ABC})^{-1} = \mathbf{C}^{-1}\mathbf{B}^{-1}\mathbf{A}^{-1} \qquad \text{trace}(\mathbf{ABC}) = \text{trace}(\mathbf{CBA})$$

An important application in regression is as follows. The **normal equations** that are obtained from ordinary least squares are: $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}$. This is a result from calculus as we try and minimize the error sum of squares.

$$Q = \sum_{i=1}^{n} \epsilon_i^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}\mathbf{X}'\mathbf{Y} + \boldsymbol{\beta}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

Making use of the following calculus results for column vectors $\mathbf{a}$ and $\mathbf{w}$, and symmetric matrix $\mathbf{A}$:

$$\frac{\partial \mathbf{a}'\mathbf{w}}{\partial \mathbf{w}} = \mathbf{a} \qquad \frac{\partial \mathbf{w}'\mathbf{A}'\mathbf{w}}{\partial \mathbf{w}} = 2\mathbf{A}\mathbf{w}$$

we obtain, the derivative for $Q$ with respect to $\boldsymbol{\beta}$, set it to 0, and solve for $\hat{\boldsymbol{\beta}}$. First, we demonstrate the results for a simple case where $\mathbf{w}$ is $2 \times 1$, $\mathbf{a}$ is $2 \times 1$, and $\mathbf{A}$ is $2 \times 2$.

$$\mathbf{a}'\mathbf{w} = \begin{bmatrix} a_1 & a_2 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = a_1 w_1 + a_2 w_2 \quad \Rightarrow \quad \frac{\partial \mathbf{a}'\mathbf{w}}{\partial \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \mathbf{a}$$

$$\mathbf{w}'\mathbf{A}\mathbf{w} = \begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{12} & A_{22} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = A_{11}w_1^2 + 2A_{12}w_1w_2 + A_{22}w_2^2$$

$$\Rightarrow \quad \frac{\partial \mathbf{w}'\mathbf{A}\mathbf{w}}{\partial \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}} = \begin{bmatrix} 2A_{11}w_1 + 2A_{12}w_2 \\ 2A_{12}w_1 + 2A_{22}w_2 \end{bmatrix} = \mathbf{2A}\mathbf{w}$$

$$\frac{\partial Q}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \stackrel{\text{set}}{=} 0 \quad \Rightarrow \quad \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y} \quad \Rightarrow \quad \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

From this result, we can obtain the vectors of fitted values and residuals, and the sums of squares for the ANOVA from the data matrices and vectors:

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 + \hat{\beta}_1 X_1 \\ \hat{\beta}_0 + \hat{\beta}_1 X_2 \\ \vdots \\ \hat{\beta}_0 + \hat{\beta}_1 X_n \end{bmatrix} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

$$\boldsymbol{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} Y_1 - \hat{Y}_1 \\ Y_2 - \hat{Y}_2 \\ \vdots \\ Y_n - \hat{Y}_n \end{bmatrix} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y}$$

$$\overline{\mathbf{Y}} = \begin{bmatrix} \overline{Y} \\ \overline{Y} \\ \vdots \\ \overline{Y} \end{bmatrix} = \frac{1}{n} \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n Y_i \\ \vdots \\ \sum_{i=1}^n Y_i \end{bmatrix} = \frac{1}{n}\mathbf{J}\mathbf{Y}$$

The matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is very important in regression analysis. It is **symmetric** and **idempotent**. That is, it is equal to its transpose, and when you multiply it by itself (square it), you obtain it again. It is called the **hat** or **projection** matrix, and is often denotes as $\mathbf{H}$ or $\mathbf{P}$. Here, we will use $\mathbf{P}$ to denote the projection matrix.

The sums of squares can be written in terms of **quadratic forms** of the data vector $\mathbf{Y}$. First however note the following results involving matrices used in their construction:

$$\mathbf{PX} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{XI} = \mathbf{X} \qquad \mathbf{PP} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{XI}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{P}$$

Note that if the model has an intercept ($\beta_0$), then the first column of $\mathbf{X}$, is a column of $1^s$. Then, since $\mathbf{PX} = \mathbf{X}$, that implies $\mathbf{PJ} = \mathbf{J}$, since $\mathbf{J}$ is a $n \times n$ matrix of $1^s$.

$$\mathbf{PJ} = \mathbf{J} \quad \Rightarrow \quad \mathbf{P}\frac{1}{n}\mathbf{J} = \frac{1}{n}\mathbf{J} \qquad \mathbf{JJ} = n\mathbf{J} \quad \Rightarrow \quad \frac{1}{n}\mathbf{J}\frac{1}{n}\mathbf{J} = \frac{1}{n}\mathbf{J}$$

Now, we re-introduce the sums of squares, and write them in matrix form. The Total (corrected) sum of squares is:

$$TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \left(\mathbf{Y} - \overline{\mathbf{Y}}\right)'\left(\mathbf{Y} - \overline{\mathbf{Y}}\right) \quad = \quad \mathbf{Y}'(\mathbf{I} - \frac{1}{n}\mathbf{J})'(\mathbf{I} - \frac{1}{n}\mathbf{J})\mathbf{Y} \quad = \quad \mathbf{Y}'(\mathbf{I} - \frac{1}{n}\mathbf{J})\mathbf{Y}$$

This is partitioned into the error ($SSE$) and regression ($SSR$) sums of squares:

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \left(\mathbf{Y} - \hat{\mathbf{Y}}\right)'\left(\mathbf{Y} - \hat{\mathbf{Y}}\right) \quad = \quad \mathbf{Y}'(\mathbf{I} - \mathbf{P})'(\mathbf{I} - \mathbf{P})\mathbf{Y} \quad = \quad \mathbf{Y}'(\mathbf{I} - \mathbf{P})\mathbf{Y}$$

$$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = \left(\hat{\mathbf{Y}} - \overline{\mathbf{Y}}\right)'\left(\hat{\mathbf{Y}} - \overline{\mathbf{Y}}\right) \quad = \quad \mathbf{Y}'(\mathbf{P} - \frac{1}{n}\mathbf{J})'(\mathbf{P} - \frac{1}{n}\mathbf{J})\mathbf{Y} \quad = \quad \mathbf{Y}'(\mathbf{P} - \frac{1}{n}\mathbf{J})\mathbf{Y}$$

Some useful matrix simplifications are helpful in making computations from spreadsheet calculations that avoid working directly with $n \times n$ matrices "inside" the sums of squares.

$$\mathbf{Y}'\mathbf{IY} \quad = \quad \sum_{i=1}^{n} Y_i^2 \qquad\qquad \mathbf{Y}'\frac{1}{n}\mathbf{JY} \quad = \quad \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$$

$$\mathbf{Y}'\mathbf{PY} \quad = \quad \mathbf{Y}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} \quad = \quad \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}$$

**Example: Cysteic Acid Content and Age of Carpets**

For the Carpet Aging analysis, Table 3.1 contains the computations used to obtain the elements of the matrices **X'X**, **X'Y**, and **Y'Y**. We use these to obtain the matrix form of the least squares estimate (based on specimens 1-23), and the sums of squares for the Analysis of Variance.

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 23 & 66.59 \\ 66.59 & 232.4895 \end{bmatrix} \qquad \mathbf{X}'\mathbf{Y} = \begin{bmatrix} 23408 \\ 86322.04 \end{bmatrix} \qquad \mathbf{Y}'\mathbf{Y} = [32556784]$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{23(232.4895) - (66.59)^2}\begin{bmatrix} 232.4895 & -66.59 \\ -66.59 & 23 \end{bmatrix} = \frac{1}{913.0304}\begin{bmatrix} 232.4895 & -66.59 \\ -66.59 & 23 \end{bmatrix}$$

$$\Rightarrow \qquad \hat{\boldsymbol{\beta}} = \frac{1}{913.0304}\begin{bmatrix} 232.4895 & -66.59 \\ -66.59 & 23 \end{bmatrix}\begin{bmatrix} 23408 \\ 86322.04 \end{bmatrix}$$

$$= \frac{1}{913.0304}\begin{bmatrix} 232.4895(23408) + (-66.59)(86322.04) \\ (-66.59)(23408) + 23(86322.04) \end{bmatrix} = \frac{1}{913.0304}\begin{bmatrix} -306070.43 \\ 426668.2 \end{bmatrix}$$

$$= \begin{bmatrix} -335.225 \\ 467.310 \end{bmatrix}$$

$$\mathbf{Y}'\mathbf{I}\mathbf{Y} = 32556784 \qquad \mathbf{Y}'\frac{1}{n}\mathbf{J}\mathbf{Y} = \frac{(23408)^2}{23} = 23823237.57$$

$$\mathbf{Y}'\mathbf{P}\mathbf{Y} = -335.225(23408) + 467.310(86322.04) = 32492205.71$$

$$\Rightarrow \quad TSS = 32556784 - 23823237.57 = 8733546.43$$

$$\Rightarrow \quad SSE = 32556784 - 32492205.71 = 64578.29$$

$$\Rightarrow \quad SSR = 32492205.71 - 23823237.57 = 8668968.14$$

**R Program for Matrix Computations**

```
carpet <- read.csv("http://www.stat.ufl.edu/~winner/data/carpet_age.csv",
      header=T)
attach(carpet); names(carpet)

### f ==> full data  m==> missing age
age_f <- age[1:23]; age_m <- age[24:25]
cys_acid_f <- cys_acid[1:23]; cys_acid_m <- cys_acid[24:25]

################ Matrix form  (.m represents matrix form)
(n <- length(age_f))
```

| specimen ($i$) | age($Y$) | cys acid($X$) | $Y^2$ | $X^2$ | $XY$ |
|---|---|---|---|---|---|
| 1 | 1750 | 4.39 | 3062500 | 19.2721 | 7682.5 |
| 2 | 1600 | 4.3 | 2560000 | 18.49 | 6880 |
| 3 | 1600 | 4.27 | 2560000 | 18.2329 | 6832 |
| 4 | 1600 | 4.33 | 2560000 | 18.7489 | 6928 |
| 5 | 1600 | 4.01 | 2560000 | 16.0801 | 6416 |
| 6 | 1550 | 3.99 | 2402500 | 15.9201 | 6184.5 |
| 7 | 1550 | 3.97 | 2402500 | 15.7609 | 6153.5 |
| 8 | 1550 | 3.82 | 2402500 | 14.5924 | 5921 |
| 9 | 1500 | 3.93 | 2250000 | 15.4449 | 5895 |
| 10 | 1400 | 3.81 | 1960000 | 14.5161 | 5334 |
| 11 | 1400 | 3.54 | 1960000 | 12.5316 | 4956 |
| 12 | 1400 | 3.72 | 1960000 | 13.8384 | 5208 |
| 13 | 1400 | 3.65 | 1960000 | 13.3225 | 5110 |
| 14 | 1050 | 3.12 | 1102500 | 9.7344 | 3276 |
| 15 | 550 | 1.87 | 302500 | 3.4969 | 1028.5 |
| 16 | 400 | 1.49 | 160000 | 2.2201 | 596 |
| 17 | 400 | 1.53 | 160000 | 2.3409 | 612 |
| 18 | 300 | 1.32 | 90000 | 1.7424 | 396 |
| 19 | 250 | 1.21 | 62500 | 1.4641 | 302.5 |
| 20 | 170 | 1.19 | 28900 | 1.4161 | 202.3 |
| 21 | 140 | 1.22 | 19600 | 1.4884 | 170.8 |
| 22 | 128 | 1.03 | 16384 | 1.0609 | 131.84 |
| 23 | 120 | 0.88 | 14400 | 0.7744 | 105.6 |
| 24 | N/A | 1.92 | N/A | N/A | N/A |
| 25 | N/A | 1.44 | N/A | N/A | N/A |
| Sum(1-23) | 23408 | 66.59 | 32556784 | 232.4895 | 86322.04 |

Table 3.1: Data and computations for matrix form of Carpet Age / Cysteic Acid Study

```
Y <- age_f
X <- cbind(rep(1,n), cys_acid_f)
(XPXI <- solve(t(X) %*% X))
ybar <- mean(Y)
ybar.m <- rep(ybar,n)

(beta_hat.m <- XPXI %*% t(X) %*% Y)
Y_hat.m <- X %*% beta_hat.m
e.m <- Y - Y_hat.m

(SS_ERR.m <- t(Y - Y_hat.m) %*% (Y - Y_hat.m))
(df_ERR <- n - ncol(X))
(MS_ERR.m <- SS_ERR.m / df_ERR)
(SS_REG.m <- t(Y_hat.m - ybar.m) %*% (Y_hat.m - ybar.m))
(df_REG <- ncol(X) - 1)
(MS_REG.m <- SS_REG.m / df_REG)
```

### R Output

```
> (n <- length(age_f))
[1] 23
> Y <- age_f
> X <- cbind(rep(1,n), cys_acid_f)
> (XPXI <- solve(t(X) %*% X))
                      cys_acid_f
            0.25463500 -0.07293295
cys_acid_f -0.07293295  0.02519084
> ybar <- mean(Y)
> ybar.m <- rep(ybar,n)
> (beta_hat.m <- XPXI %*% t(X) %*% Y)
           -335.2248
cys_acid_f  467.3100
> (SS_ERR.m <- t(Y - Y_hat.m) %*% (Y - Y_hat.m))
[1,] 64576.89
> (df_ERR <- n - ncol(X))
[1] 21
> (MS_ERR.m <- SS_ERR.m / df_ERR)
[1,] 3075.09
> (SS_REG.m <- t(Y_hat.m - ybar.m) %*% (Y_hat.m - ybar.m))
[1,] 8668970
> (df_REG <- ncol(X) - 1)
[1] 1
> (MS_REG.m <- SS_REG.m / df_REG)
[1,] 8668970
```

# Chapter 4

# Distributional Results

The model for the observed data (data generating process) can be thought of as $Y_i$ being a random variable with a mean (systematic component) of $\beta_0 + \beta_1 X_i$ and a random error term of $\epsilon_i$ that reflects all possible sources of variation beyond the predictor $X$. We assume that the error terms have mean 0, and variance $\sigma_i^2$. In general, the error terms may or may not be independent (uncorrelated). The expectation and variance-covariance matrix of the vector of error terms $\boldsymbol{\varepsilon}$ are:

$$E\{\boldsymbol{\varepsilon}\} = \begin{bmatrix} E\{\epsilon_1\} \\ E\{\epsilon_2\} \\ \vdots \\ E\{\epsilon_n\} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{0}$$

$$V\{\boldsymbol{\varepsilon}\} = E\left\{(\boldsymbol{\varepsilon} - E\{\boldsymbol{\varepsilon}\})(\boldsymbol{\varepsilon} - E\{\boldsymbol{\varepsilon}\})'\right\} = E\left\{(\boldsymbol{\varepsilon} - \mathbf{0})(\boldsymbol{\varepsilon} - \mathbf{0})'\right\} = E\{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\} =$$

$$\begin{bmatrix} E\{\epsilon_1^2\} & E\{\epsilon_1\epsilon_2\} & \cdots & E\{\epsilon_1\epsilon_n\} \\ E\{\epsilon_1\epsilon_2\} & E\{\epsilon_2^2\} & \cdots & E\{\epsilon_2\epsilon_n\} \\ \vdots & \vdots & \ddots & \vdots \\ E\{\epsilon_1\epsilon_n\} & E\{\epsilon_2\epsilon_n\} & \cdots & E\{\epsilon_n^2\} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1n} & \sigma_{2n} & \cdots & \sigma_n^2 \end{bmatrix}$$

The expectation and variance-covariance matrix of the data vector $\mathbf{Y}$ are:

$$E\{\mathbf{Y}\} = \begin{bmatrix} E\{Y_1\} \\ E\{Y_2\} \\ \vdots \\ E\{Y_n\} \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ \vdots \\ \beta_0 + \beta_1 X_n \end{bmatrix} = \mathbf{X}\boldsymbol{\beta}$$

$$\boldsymbol{\Sigma}_{\mathbf{Y}} = V\{\mathbf{Y}\} = E\left\{(\mathbf{Y} - E\{\mathbf{Y}\})(\mathbf{Y} - E\{\mathbf{Y}\})'\right\} = E\left\{(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\right\} =$$

$$\begin{bmatrix} E\left\{(Y_1 - E\{Y_1\})^2\right\} & E\{(Y_1 - E\{Y_1\})(Y_2 - E\{Y_2\})\} & \cdots & E\{(Y_1 - E\{Y_1\})(Y_n - E\{Y_n\})\} \\ E\{(Y_1 - E\{Y_1\})(Y_2 - E\{Y_2\})\} & E\left\{(Y_2 - E\{Y_2\})^2\right\} & \cdots & E\{(Y_2 - E\{Y_2\})(Y_n - E\{Y_n\})\} \\ \vdots & \vdots & \ddots & \vdots \\ E\{(Y_1 - E\{Y_1\})(Y_n - E\{Y_n\})\} & E\{(Y_2 - E\{Y_2\})(Y_n - E\{Y_n\})\} & \cdots & E\left\{(Y_n - E\{Y_n\})^2\right\} \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1n} & \sigma_{2n} & \cdots & \sigma_n^2 \end{bmatrix}$$

where $\sigma_{ij}$ is the **covariance** between the $i^{th}$ and $j^{th}$ measurements. When the data are independent, but not necessarily of equal variance (heteroskedastic), we have:

$$V\{\mathbf{Y}\} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

When the data are independent with constant variance (homoskedastic), we have $V\{\mathbf{Y}\} = \sigma^2\mathbf{I}$. This is the common assumption underlying the model, which needs to be checked in practice.

For a random matrix $\mathbf{W}$, and a matrix of fixed constants $\mathbf{A}$ of compatible dimensions for multiplication, where the number of columns of $\mathbf{A}$ is equal to the number of rows of $\mathbf{W}$, we write the following.

$$\mathbf{A} = \begin{bmatrix} \mathbf{a_1'} \\ \mathbf{a_2'} \\ \vdots \\ \mathbf{a_{r_A}'} \end{bmatrix} \quad \mathbf{W} = \begin{bmatrix} \mathbf{w_1} & \mathbf{w_2} & \cdots & \mathbf{w_{c_W}} \end{bmatrix} \quad E\{\mathbf{W}\} = \begin{bmatrix} \mu_{\mathbf{w1}} & \mu_{\mathbf{w2}} & \cdots & \mu_{\mathbf{wc_W}} \end{bmatrix}$$

$$\mathbf{AW} = \begin{bmatrix} \mathbf{a_1'w_1} & \mathbf{a_1'w_2} & \cdots & \mathbf{a_1'w_{c_W}} \\ \mathbf{a_2'w_1} & \mathbf{a_2'w_2} & \cdots & \mathbf{a_2'w_{c_W}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{a_{r_A}'w_1} & \mathbf{a_{r_A}'w_2} & \cdots & \mathbf{a_{r_A}'w_{c_W}} \end{bmatrix}$$

$$E\{\mathbf{AW}\} = \begin{bmatrix} \mathbf{a_1'}\mu_{\mathbf{w1}} & \mathbf{a_1'}\mu_{\mathbf{w2}} & \cdots & \mathbf{a_1'}\mu_{\mathbf{wc_W}} \\ \mathbf{a_2'}\mu_{\mathbf{w1}} & \mathbf{a_2'}\mu_{\mathbf{w2}} & \cdots & \mathbf{a_2'}\mu_{\mathbf{wc_W}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{a_{r_A}'}\mu_{\mathbf{w1}} & \mathbf{a_{r_A}'}\mu_{\mathbf{w2}} & \cdots & \mathbf{a_{r_A}'}\mu_{\mathbf{wc_W}} \end{bmatrix} = \mathbf{A}E\{\mathbf{W}\}$$

$$V\{\mathbf{AW}\} = E\left\{(\mathbf{AW} - E\{\mathbf{AW}\})(\mathbf{AW} - E\{\mathbf{AW}\})'\right\} = E\left\{\mathbf{A}(\mathbf{W} - E\{\mathbf{W}\})(\mathbf{W} - E\{\mathbf{W}\})'\mathbf{A}'\right\} =$$

$$\mathbf{A}E\left\{(\mathbf{W} - E\{\mathbf{W}\})(\mathbf{W} - E\{\mathbf{W}\})'\right\}\mathbf{A}' = \mathbf{A}V\{\mathbf{W}\}\mathbf{A}'$$

When applied to the least squares estimate $\hat{\boldsymbol{\beta}}$, we obtain:

$$E\left\{\hat{\boldsymbol{\beta}}\right\} = E\left\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\right\} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E\{\mathbf{Y}\} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

$$V\left\{\hat{\boldsymbol{\beta}}\right\} = V\left\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\right\} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V\{\mathbf{Y}\}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{\mathbf{Y}}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

When the data are independent, with constant variance, $\boldsymbol{\Sigma}_{\mathbf{Y}} = \sigma^2\mathbf{I}$ then the variance of $\hat{\boldsymbol{\beta}}$ simplifies to:

$$V\left\{\hat{\boldsymbol{\beta}}\right\} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \quad \text{with estimated variance} \quad \hat{V}\left\{\hat{\boldsymbol{\beta}}\right\} = s^2(\mathbf{X}'\mathbf{X})^{-1} \quad s^2 = MSE = \frac{SSE}{n-2}$$

Further, if $\mathbf{Y}$ is (multivariate) normal, then so is $\hat{\boldsymbol{\beta}}$, and when based on large samples, $\hat{\boldsymbol{\beta}}$ is approximately normal, even when $\mathbf{Y}$ is not, based on Central Limit Theorem arguments. We also obtain the following results regarding the the vectors of fitted values $\left(\hat{\mathbf{Y}}\right)$ and residuals $(\boldsymbol{e})$, and the estimated mean when $X = X^*$ and the predicted value of $Y$ when $X = X^*$.

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{P}\mathbf{Y} \quad \Rightarrow \quad E\left\{\hat{\mathbf{Y}}\right\} = \mathbf{X}\boldsymbol{\beta} \qquad V\left\{\hat{\mathbf{Y}}\right\} = \mathbf{P}\sigma^2\mathbf{I}\mathbf{P}' = \sigma^2\mathbf{P}$$

$$\boldsymbol{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{P})\,\mathbf{Y} \quad \Rightarrow \quad E\{\boldsymbol{e}\} = (\mathbf{I} - \mathbf{P})\,\mathbf{X}\boldsymbol{\beta} = 0 \qquad V\{\boldsymbol{e}\} = (\mathbf{I} - \mathbf{P})\,\sigma^2\mathbf{I}\,(\mathbf{I} - \mathbf{P})' = \sigma^2\,(\mathbf{I} - \mathbf{P})$$

$$\hat{Y}^* = \mathbf{X}^{*'}\hat{\boldsymbol{\beta}} \quad \Rightarrow \quad E\left\{\hat{Y}^*\right\} = \mathbf{X}^{*'}\hat{\boldsymbol{\beta}} \qquad V\left\{\hat{Y}^*\right\} = \sigma^2\mathbf{X}^{*'}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}^* \qquad \mathbf{X}^* = \begin{bmatrix} 1 \\ X^* \end{bmatrix}$$

$$Y_{\text{New}} = \mathbf{X}^{*'}\hat{\boldsymbol{\beta}} \qquad V\left\{Y_{\text{New}} - \hat{Y}^*\right\} = V\left\{Y_{\text{New}}\right\} + V\left\{\hat{Y}^*\right\} = \sigma^2\left[1 + \mathbf{X}^{*'}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}^*\right]$$

**Example: Cysteic Acid Content and Age of Carpets**

Here we present R code and output for the matrix form of the Carpet Aging data. We first compute the estimated variance-covariance matrix of $\hat{\boldsymbol{\beta}}$, and test $H_0 : \beta_1 = 0$, then obtain Prediction Intervals for the ages of specimens 24 and 25. For specimens 24 and 25, the cysteic acid contents were $X^*_{24} = 1.92$ and $X^*_{25} = 1.44$, respectively.

```
################ Matrix form  (.m represents matrix form)
### Commands

carpet <- read.csv("http://www.stat.ufl.edu/~winner/data/carpet_age.csv",
      header=T)
attach(carpet); names(carpet)

### f ==> full data  m==> missing age
age_f <- age[1:23]; age_m <- age[24:25]
cys_acid_f <- cys_acid[1:23]; cys_acid_m <- cys_acid[24:25]

(n <- length(age_f))
Y <- age_f
X <- cbind(rep(1,n), cys_acid_f)
(XPXI <- solve(t(X) %*% X))
ybar <- mean(Y)
ybar.m <- rep(ybar,n)

(beta_hat.m <- XPXI %*% t(X) %*% Y)
Y_hat.m <- X %*% beta_hat.m
e.m <- Y - Y_hat.m

(SS_ERR.m <- t(Y - Y_hat.m) %*% (Y - Y_hat.m))
```

```
(df_ERR <- n - ncol(X))
(MS_ERR.m <- SS_ERR.m / df_ERR)
(SS_REG.m <- t(Y_hat.m - ybar.m) %*% (Y_hat.m - ybar.m))
(df_REG <- ncol(X) - 1)
(MS_REG.m <- SS_REG.m / df_REG)
(V_beta_hat.m <- MS_ERR.m[1,1] * XPXI)
(t_beta.m <- beta_hat.m / sqrt(diag(V_beta_hat.m)))
(P_t_beta.m <- 2*(1 - pt(abs(t_beta.m),n-2)))
X_miss24 <- matrix(c(1, cys_acid_m[1]),ncol=1)
X_miss25 <- matrix(c(1, cys_acid_m[2]),ncol=1)
(yhat24 <- t(X_miss24) %*% beta_hat.m)
(yhat25 <- t(X_miss25) %*% beta_hat.m)
(PE_yhat24 <- sqrt(MS_ERR.m * (1 + t(X_miss24) %*% XPXI %*% X_miss24)))
(PE_yhat25 <- sqrt(MS_ERR.m * (1 + t(X_miss25) %*% XPXI %*% X_miss25)))
(PI_miss24 <- yhat24 + qt(c(.025,.975),n-2) * PE_yhat24)
(PI_miss25 <- yhat25 + qt(c(.025,.975),n-2) * PE_yhat25)

### Output
> (V_beta_hat.m <- MS_ERR.m[1,1] * XPXI)
            783.0255 -224.27537
cys_acid_f -224.2754   77.46409
>
> (t_beta.m <- beta_hat.m / sqrt(diag(V_beta_hat.m)))
           -11.97976
cys_acid_f  53.09515
> (P_t_beta.m <- 2*(1 - pt(abs(t_beta.m),n-2)))
           7.511813e-11
cys_acid_f 0.000000e+00
> X_miss24 <- matrix(c(1, cys_acid_m[1]),ncol=1)
> X_miss25 <- matrix(c(1, cys_acid_m[2]),ncol=1)
>
> (yhat24 <- t(X_miss24) %*% beta_hat.m)
[1,] 562.0103
> (yhat25 <- t(X_miss25) %*% beta_hat.m)
[1,] 337.7015
>
> (PE_yhat24 <- sqrt(MS_ERR.m * (1 + t(X_miss24) %*% XPXI %*% X_miss24)))
[1,] 57.29277
> (PE_yhat25 <- sqrt(MS_ERR.m * (1 + t(X_miss25) %*% XPXI %*% X_miss25)))
[1,] 58.07609
> (PI_miss24 <- yhat24 + qt(c(.025,.975),n-2) * PE_yhat24)
[1] 442.8635 681.1572
> (PI_miss25 <- yhat25 + qt(c(.025,.975),n-2) * PE_yhat25)
[1] 216.9257 458.4774
```

For **Quadratic forms**, where we have a random column vector, $\mathbf{w}$, and a symmetric matrix of constants $\mathbf{A}$ we have the random scalar $\mathbf{w}'\mathbf{A}\mathbf{w}$. If $\mathbf{w}$ has mean $\mu_{\mathbf{W}}$ and variance-covariance matrix $\mathbf{\Sigma}_{\mathbf{W}}$, then:

$$E\left\{\mathbf{w}'\mathbf{A}\mathbf{w}\right\} = \text{trace}(\mathbf{A}\mathbf{\Sigma}_{\mathbf{W}}) + \mu'_{\mathbf{W}}\mathbf{A}\mu_{\mathbf{W}}.$$

Reconsider the simple case, with $\mathbf{w}$ being $2 \times 1$ and $\mathbf{A}$ being $2 \times 2$ considered previously.

$$\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \qquad \begin{bmatrix} A_{11} & A_{12} \\ A_{12} & A_{22} \end{bmatrix} \qquad \mu_{\mathbf{W}} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \qquad \mathbf{\Sigma}_{\mathbf{W}} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

$$\mathbf{w}'\mathbf{A}\mathbf{w} = \begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{12} & A_{22} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = A_{11}w_1^2 + 2A_{12}w_1w_2 + A_{22}w_2^2$$

$$E\left\{\mathbf{w}'\mathbf{A}\mathbf{w}\right\} = A_{11}E\{w_1^2\} + 2A_{12}E\{w_1w_2\} + A_{22}E\{w_2^2\} = A_{11}\left(\sigma_1^2+\mu_1^2\right) + 2A_{12}\left(\sigma_{12}+\mu_1\mu_2\right) + A_{22}\left(\sigma_2^2+\mu_2^2\right)$$

Now, evaluating the right-hand side of the equation.

$$(\mathbf{A}\mathbf{\Sigma_W}) = \begin{bmatrix} A_{11}\sigma_1^2 + A_{12}\sigma_{12} & A_{11}\sigma_{12} + A_{12}\sigma_2^2 \\ A_{12}\sigma_1^2 + A_{22}\sigma_{12} & A_{12}\sigma_{12} + A_{22}\sigma_2^2 \end{bmatrix} \quad \Rightarrow \quad \text{trace}(\mathbf{A}\mathbf{\Sigma_W}) = A_{11}\sigma_1^2 + 2A_{12}\sigma_{12} + A_{22}\sigma_2^2$$

$$\mu_{\mathbf{W}}'\mathbf{A}\mu_{\mathbf{W}} = \begin{bmatrix} \mu_1 & \mu_2 \end{bmatrix}\begin{bmatrix} A_{11} & A_{12} \\ A_{12} & A_{22} \end{bmatrix}\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \mu_1^2A_{11} + 2\mu_1\mu_2A_{12} + \mu_2^2A_{22}$$

This clearly generalizes to $\mathbf{w}$ of any $n \times 1$ and symmetric $\mathbf{A}$ of $n \times n$.

Consider the three primary quantities that make up the Analysis of Variance: $\mathbf{Y}'\mathbf{I}\mathbf{Y} = \mathbf{Y}'\mathbf{Y}$, $\mathbf{Y}'\mathbf{P}\mathbf{Y}$, $\mathbf{Y}'\left(\frac{1}{n}\right)\mathbf{J}\mathbf{Y}$. Here, we have $\mu_{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta}$ and $\mathbf{\Sigma_Y}$. Here, we consider the basic case (independent and constant variance, with $\mathbf{\Sigma_Y} = \sigma^2\mathbf{I}$). Further, recall that (when $\mathbf{AB}$ and $\mathbf{BA}$ are square) $\text{trace}(\mathbf{AB}) = \text{trace}(\mathbf{BA})$.

$$E\left\{\mathbf{Y}'\mathbf{I}\mathbf{Y}\right\} = \text{trace}(\mathbf{I}\mathbf{\Sigma_Y}) + \mu_{\mathbf{Y}}'\mathbf{I}\mu_{\mathbf{Y}} = \sigma^2\text{trace}(\mathbf{I}\mathbf{I}) + \mu_{\mathbf{Y}}'\mathbf{I}\mu_{\mathbf{Y}} = n\sigma^2 + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

Recalling that $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'$, we can obtain the trace of $\mathbf{P}$ as trace of $\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}} = \mathbf{I_2}$, which is 2. Further, recall that $\mathbf{P}\mathbf{X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{X} = \mathbf{X}$, so that $\boldsymbol{\beta}'\mathbf{X}'\mathbf{P}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$.

$$E\left\{\mathbf{Y}'\mathbf{P}\mathbf{Y}\right\} = \text{trace}(\mathbf{P}\mathbf{\Sigma_Y}) + \mu_{\mathbf{Y}}'\mathbf{P}\mu_{\mathbf{Y}} = \sigma^2\text{trace}(\mathbf{P}\mathbf{I}) + \mu_{\mathbf{Y}}'\mathbf{P}\mu_{\mathbf{Y}} = 2\sigma^2 + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

When we multiply $\mathbf{X}'\mathbf{J}\mathbf{X}$, we get:

$$\mathbf{X}'\mathbf{J} = \begin{bmatrix} n & n & \cdots & n \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i & \cdots & \sum_{i=1}^n X_i \end{bmatrix} \quad \Rightarrow \quad \mathbf{X}'\mathbf{J}\mathbf{X} = \begin{bmatrix} n^2 & n\sum_{i=1}^n X_i \\ n\sum_{i=1}^n X_i & \left(\sum_{i=1}^n X_i\right)^2 \end{bmatrix}.$$

$$E\left\{\mathbf{Y}'\left(\frac{1}{n}\right)\mathbf{J}\mathbf{Y}\right\} = \text{trace}\left(\frac{1}{n}\mathbf{J}\mathbf{\Sigma_Y}\right) + \mu_{\mathbf{Y}}'\left(\frac{1}{n}\right)\mathbf{J}\mu_{\mathbf{Y}} = \sigma^2\text{trace}\left(\frac{1}{n}\mathbf{J}\right) + \mu_{\mathbf{Y}}'\left(\frac{1}{n}\right)\mathbf{J}\mu_{\mathbf{Y}} = \sigma^2 + \boldsymbol{\beta}'\mathbf{X}'\frac{1}{n}\mathbf{J}\mathbf{X}\boldsymbol{\beta}$$

Now, we write the total, error and regression sums of squares.

$$TSS = \mathbf{Y}'\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y} \qquad SSE = \mathbf{Y}'\left(\mathbf{I} - \mathbf{P}\right)\mathbf{Y} \qquad SSR = \mathbf{Y}'\left(\mathbf{P} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y}$$

Consider $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\left(\frac{1}{n}\right)\mathbf{J}\mathbf{X}$:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{bmatrix} \quad \mathbf{X}'\left(\frac{1}{n}\right)\mathbf{J}\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \frac{\left(\sum_{i=1}^n X_i\right)^2}{n} \end{bmatrix}$$

$$\Rightarrow \quad \mathbf{X}'\mathbf{X} - \mathbf{X}'\left(\frac{1}{n}\right)\mathbf{J}\mathbf{X} = \begin{bmatrix} 0 & 0 \\ 0 & \sum_{i=1}^{n}(X_i - \overline{X})^2 \end{bmatrix}$$

This leads to the following Expected Sums of Squares.

$$E\{TSS\} = \left[n\sigma^2 + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}\right] - \left[\sigma^2 + \boldsymbol{\beta}'\mathbf{X}'\frac{1}{n}\mathbf{J}\mathbf{X}\boldsymbol{\beta}\right] = (n-1)\sigma^2 + \beta_1^2\sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$E\{SSE\} = \left[n\sigma^2 + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}\right] - \left[2\sigma^2 + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}\right] = (n-2)\sigma^2$$

$$E\{SSR\} = \left[2\sigma^2 + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}\right] - \left[\sigma^2 + \boldsymbol{\beta}'\mathbf{X}'\frac{1}{n}\mathbf{J}\mathbf{X}\boldsymbol{\beta}\right] = \sigma^2 + \beta_1^2\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Further, if $\mathbf{w}$ is normally distributed, and if $\mathbf{A}\boldsymbol{\Sigma}_\mathbf{W}\mathbf{A}\boldsymbol{\Sigma}_\mathbf{W} = \mathbf{A}\boldsymbol{\Sigma}_\mathbf{W}$ then we have the following results that are related to **Cochran's Theorem**, (see e.g. Result 5.15, p. 112, Monahan (2008)).

$$\mathbf{w}'\mathbf{A}\mathbf{w} \sim \chi^2\left(df_A, \Omega_A\right) \qquad df_A = \text{ rank}(A) \qquad \Omega_A = \frac{\mu_\mathbf{W}'\mathbf{A}\mu_\mathbf{W}}{2}$$

where $df_A$ and $\Omega_A$ are the **degrees of freedom** and **non-centrality parameter**, respectively. If $\Omega_A = 0$, then it is the standard (central) chi-square distribution. Two other important results are as follow.

$\mathbf{w}'\mathbf{A}\mathbf{w}$ and $\mathbf{w}'\mathbf{B}\mathbf{w}$ are independent if $\mathbf{A}\boldsymbol{\Sigma}_\mathbf{W}\mathbf{B} = \mathbf{0}$

$\mathbf{w}'\mathbf{A}\mathbf{w}$ and $\mathbf{B}\mathbf{w}$ are independent if $\mathbf{B}\boldsymbol{\Sigma}_\mathbf{W}\mathbf{A} = \mathbf{0}$

Note that with respect to the model with normal, independent errors of constant variance, we have:

$$\boldsymbol{\Sigma}_\mathbf{Y} = \sigma^2\mathbf{I} \qquad \mathbf{A}_E = \frac{1}{\sigma^2}(\mathbf{I} - \mathbf{P}) \qquad \mathbf{A}_R = \frac{1}{\sigma^2}\left(\mathbf{P} - \frac{1}{n}\mathbf{J}\right)$$

$$\mathbf{A}_E\boldsymbol{\Sigma}_\mathbf{Y}\mathbf{A}_E\boldsymbol{\Sigma}_\mathbf{Y} = \frac{1}{\sigma^2}(\mathbf{I} - \mathbf{P})\sigma^2\mathbf{I}\frac{1}{\sigma^2}(\mathbf{I} - \mathbf{P})\sigma^2\mathbf{I} = (\mathbf{I} - \mathbf{P})\mathbf{I}(\mathbf{I} - \mathbf{P})\mathbf{I} = (\mathbf{I} - \mathbf{P})\mathbf{I} = \mathbf{A}_E\boldsymbol{\Sigma}_\mathbf{Y}$$

$$\mathbf{A}_R\boldsymbol{\Sigma}_\mathbf{Y}\mathbf{A}_R\boldsymbol{\Sigma}_\mathbf{Y} = \frac{1}{\sigma^2}\left(\mathbf{P} - \frac{1}{n}\mathbf{J}\right)\sigma^2\mathbf{I}\frac{1}{\sigma^2}\left(\mathbf{P} - \frac{1}{n}\mathbf{J}\right)\sigma^2\mathbf{I} = \left(\mathbf{P} - \frac{1}{n}\mathbf{J}\right)\mathbf{I} = \mathbf{A}_R\boldsymbol{\Sigma}_\mathbf{Y}$$

$$\mathbf{A}_R\boldsymbol{\Sigma}_\mathbf{Y}\mathbf{A}_E = \frac{1}{\sigma^2}\left(\mathbf{P} - \frac{1}{n}\mathbf{J}\right)\sigma^2\mathbf{I}\frac{1}{\sigma^2}(\mathbf{I} - \mathbf{P}) = \mathbf{0}$$

This leads to the following important results:

$$\frac{SSE}{\sigma^2} \sim \chi^2(n-2, 0) \qquad \frac{SSR}{\sigma^2} \sim \chi^2\left(1, \frac{\beta_1^2\sum_{i=1}^{n}(X_i - \overline{X})^2}{2\sigma^2}\right).$$

Further, $SSE$ and $SSR$ are independent. Also $\hat{\boldsymbol{\beta}}$ and $SSE$ are independent, since:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \qquad SSE = \mathbf{Y}'(\mathbf{I} - \mathbf{P})\mathbf{Y}$$

$$\Rightarrow \quad (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_\mathbf{Y}\frac{\mathbf{1}}{\boldsymbol{\sigma^2}}(\mathbf{I} - \mathbf{P}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{0} \quad \text{ since } \mathbf{X}'\mathbf{P} = \mathbf{X}'.$$

The ratio of two independent chi-square random variables, each divided by its degrees of freedom, follows the $F$-distribution. If the numerator chi-square is non-central, and the denominator is a standard (central)

chi-square, it follows a non-central distribution, with the non-centrality parameter of the numerator chi-square. If both are central, the ratio follows a standard $F$-distribution. Thus, since $SSE$ and $SSR$ are independent:

$$\frac{SSR}{\sigma^2} \sim \chi^2\left(1, \frac{\beta_1^2 \sum_{i=1}^n (X_i - \overline{X})^2}{2\sigma^2}\right) \qquad \frac{SSE}{\sigma^2} \sim \chi^2(n-2, 0)$$

$$\Rightarrow \quad F = \frac{\frac{SSR}{\sigma^2}/1}{\frac{SSE}{\sigma^2}/(n-2)} = \frac{MSR}{MSE} \sim F_{1,n-2,\Omega} \qquad \Omega = \frac{\beta_1^2 \sum_{i=1}^n (X_i - \overline{X})^2}{2\sigma^2}.$$

When the null hypothesis $H_0 : \beta_1 = 0$ is true, the $F$-statistic follows the standard (central) $F$-distribution.

**The Power of the $F$-test**

We typically wish to test $H_0 : \beta_1 = 0$, based on the ANOVA $F$-test. When the null hypothesis is true, the $F$-statistic $F_{obs} = \frac{MSR}{MSE}$ is distributed as $F_{1,n-2}$, and we reject $H_0$ if $F_{obs} \geq F_{\alpha,1,n-2}$. If the null hypothesis is false ($\beta_1 \neq 0$), the $F$-statistic follows the non-central $F$-distribution. The power of the test can be obtained as the probability the $F$-statistic falls above the the critical $F$-value. That is, the power can be computed as a function of the "true" value of $\beta_1$.

$$\text{Power } = P(F^* \geq F_{\alpha,1,n-2} | F^* \sim F_{1,n-2,\Omega}) \qquad \Omega = \frac{\beta_1^2 \sum_{i=1}^n (X_i - \overline{X})^2}{2\sigma^2}$$

Note that SAS and R both have functions that compute densities, quantiles, and probabilities for central and non-central $F$-distributions. They parameterize the non-centrality parameter as $2\Omega$.

# Chapter 5

# Model Diagnostics and Influence Measures

The inferences regarding the simple linear regression model (tests and confidence intervals) are based on the following assumptions:

- Relation between $Y$ and $X$ is linear.
- Errors are normally distributed.
- Errors have constant variance.
- Errors are independent.

These assumptions can be checked graphically, as well as by statistical tests.

## 5.1 Checking Linearity

A plot of the residuals versus $X$ should be a random cloud of points centered at 0 (they sum to 0). A "U-shaped" or "inverted U-shaped" pattern is inconsistent with linearity.

A test for linearity can be conducted when there are repeat observations at certain $X$-levels (methods have also been developed to "group" $X$ values). Suppose we have $c$ distinct $X$-levels, with $n_j$ observations at the $j^{th}$ level. The data can be re-labeled as $Y_{ij}$ where $j$ represents the $X$ group, and $i$ represents the individual case within the group ($i = 1, \ldots, n_j$). We compute the following quantities.

$$\overline{Y}_j = \frac{\sum_{i=1}^{n_j} Y_{ij}}{n_j} \qquad \hat{Y}_j = \hat{\beta}_0 + \hat{\beta}_1 X_j$$

We then decompose the Error Sum of Squares into **Pure Error** and **Lack of Fit**.

$$Y_{ij} - \hat{Y}_j = \left(Y_{ij} - \overline{Y}_j\right) + \left(\overline{Y}_j - \hat{Y}_j\right) \qquad \Rightarrow$$

$$\left(Y_{ij} - \hat{Y}_j\right)^2 = \left(Y_{ij} - \overline{Y}_j\right)^2 + \left(\overline{Y}_j - \hat{Y}_j\right)^2 + 2\left(Y_{ij} - \overline{Y}_j\right)\left(\overline{Y}_j - \hat{Y}_j\right) \qquad \Rightarrow$$

$$\sum_{j=1}^{c}\sum_{i=1}^{n_j}\left(Y_{ij} - \hat{Y}_j\right)^2 = \sum_{j=1}^{c}\sum_{i=1}^{n_j}\left(Y_{ij} - \overline{Y}_j\right)^2 + \sum_{j=1}^{c}\sum_{i=1}^{n_j}\left(\overline{Y}_j - \hat{Y}_j\right)^2 + 2\sum_{j=1}^{c}\left(\overline{Y}_j - \hat{Y}_j\right)\sum_{i=1}^{n_j}\left(Y_{ij} - \overline{Y}_j\right) =$$

$$\sum_{j=1}^{c}\sum_{i=1}^{n_j}\left(Y_{ij} - \overline{Y}_j\right)^2 + \sum_{j=1}^{c} n_j\left(\overline{Y}_j - \hat{Y}_j\right)^2 + 0$$

This decomposition can be written as follows, where Pure Error represents variation in responses within the same X-levels, and Lack of Fit represents the difference between the fitted values for the linear regression model, and the cell means model.

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{j=1}^{c}\sum_{i=1}^{n_j}\left(Y_{ij} - \overline{Y}_j\right)^2 + \sum_{j=1}^{c} n_j\left(\overline{Y}_j - \hat{Y}_j\right)^2 \qquad SSE = SSPE + SSLF$$

We then partition the error degrees of freedom $(n-2)$ into Pure Error $(n-c)$ and Lack of Fit $(c-2)$. This leads to an $F$-test for testing $H_0$: Relation is Linear versus $H_A$: Relation is not Linear.

$$H_0 : E\{Y_{ij}\} = \beta_0 + \beta_1 X_j \qquad H_A : E\{Y_{ij}\} = \mu_j \neq \beta_0 + \beta_1 X_j$$

$$TS : F_{obs} = \frac{[SSLF/(c-2)]}{[SSPE/(n-c)]} = \frac{MSLF}{MSPE} \qquad RR : F_{obs} \geq F_{\alpha,c-2,n-c} \qquad P\text{-Value} : P\left(F_{c-2,n-c} \geq F_{obs}\right)$$

If the relationship is not linear, we can add polynomial terms to allow for "bends" in the relationship between $Y$ and $X$ using multiple regression, or fit a nonlinear regression model with a particular functional form.

The matrix form is helpful in understanding the distributional aspects of the test. Assume that the data are ordered by their specific $X$-levels, with $n_j$ observations at $X_j$.

$$H_0 : E\{\mathbf{Y}\} = \mathbf{X}\boldsymbol{\beta} \qquad H_A : E\{\mathbf{Y}\} = \mu \neq \mathbf{X}\boldsymbol{\beta}$$

$$\mathbf{X} = \begin{bmatrix} \mathbf{1}_{n_1} & X_1\mathbf{1}_{n_1} \\ \mathbf{1}_{n_2} & X_2\mathbf{1}_{n_2} \\ \vdots & \vdots \\ \mathbf{1}_{n_c} & X_c\mathbf{1}_{n_c} \end{bmatrix} \qquad \mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_c \end{bmatrix} \qquad \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \qquad \mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

$$\mathbf{P} = \begin{bmatrix} \mathbf{1}_{n_1} & X_1\mathbf{1}_{n_1} \\ \mathbf{1}_{n_2} & X_2\mathbf{1}_{n_2} \\ \vdots & \vdots \\ \mathbf{1}_{n_c} & X_c\mathbf{1}_{n_c} \end{bmatrix} (\mathbf{X}'\mathbf{X})^{-1} \begin{bmatrix} \mathbf{1}'_{n_1} & \mathbf{1}'_{n_2} & \cdots & \mathbf{1}'_{n_c} \\ X_1\mathbf{1}'_{n_1} & X_2\mathbf{1}'_{n_2} & \cdots & X_c\mathbf{1}'_{n_c} \end{bmatrix} =$$

$$\begin{bmatrix} P_{11}\mathbf{J}_{n_1\times n_1} & P_{12}\mathbf{J}_{n_1\times n_2} & \cdots & P_{1c}\mathbf{J}_{n_1\times n_c} \\ P_{21}\mathbf{J}_{n_2\times n_1} & P_{22}\mathbf{J}_{n_2\times n_2} & \cdots & P_{2c}\mathbf{J}_{n_2\times n_c} \\ \vdots & \vdots & \ddots & \vdots \\ P_{c1}\mathbf{J}_{n_c\times n_1} & P_{c2}\mathbf{J}_{n_c\times n_2} & \cdots & P_{cc}\mathbf{J}_{n_c\times n_c} \end{bmatrix} \qquad P_{ij} = \begin{bmatrix} 1 & X_i \end{bmatrix} (\mathbf{X}'\mathbf{X})^{-1} \begin{bmatrix} 1 \\ X_j \end{bmatrix}$$

$$\overline{\mathbf{Y}}_g = \begin{bmatrix} \overline{Y}_1\mathbf{1}_{n_1} \\ \overline{Y}_2\mathbf{1}_{n_2} \\ \vdots \\ \overline{Y}_c\mathbf{1}_{n_c} \end{bmatrix} = \frac{1}{n_g}\mathbf{J_g}\mathbf{Y} = \begin{bmatrix} n_1^{-1}\mathbf{J}_{n_1\times n_1} & \mathbf{0}_{n_1\times n_2} & \cdots & \mathbf{0}_{n_1\times n_c} \\ \mathbf{0}_{n_2\times n_1} & n_2^{-1}\mathbf{J}_{n_2\times n_2} & \vdots & \mathbf{0}_{n_2\times n_c} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{n_c\times n_1} & \mathbf{0}_{n_c\times n_2} & \cdots & n_c^{-1}\mathbf{J}_{n_c\times n_c} \end{bmatrix} \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_c \end{bmatrix}$$

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \mathbf{Y}'(\mathbf{I} - \mathbf{P})\mathbf{Y}$$

$$SSPE = \sum_{i=1}^{n_j}\sum_{j=1}^{c}\left(Y_{ij} - \overline{Y}_j\right)^2 = \mathbf{Y}'\left(\mathbf{I} - \frac{1}{n_g}\mathbf{J_g}\right)\mathbf{Y}$$

$$SSLF = \sum_{j=1}^{c}\sum_{i=1}^{n_j}\left(\overline{Y}_j - \hat{Y}_j\right)^2 = \mathbf{Y}'\left(\frac{1}{n_g}\mathbf{J_g} - \mathbf{P}\right)\mathbf{Y}$$

Due to the structures of $\mathbf{P}$ and $\frac{1}{n_g}\mathbf{J_g}$, we can easily see that $\mathbf{P}\frac{1}{n_g}\mathbf{J_g} = \mathbf{P}$ due to the $0^s$ in $\frac{1}{n_g}\mathbf{J_g}$. This is different from the linear regression model where $\mathbf{P}\frac{1}{n}\mathbf{J} = \mathbf{J}$. We still have that $\mathbf{PP} = \mathbf{P}$. The main results are given below.

$$\mathbf{P}\frac{1}{n_g}\mathbf{J_g} = \mathbf{P} \qquad \mathbf{PP} = \mathbf{P} \qquad \frac{1}{n_g}\mathbf{J_g}\frac{1}{n_g}\mathbf{J_g} = \frac{1}{n_g}\mathbf{J_g}$$

$$\Rightarrow \quad \left(\mathbf{I} - \frac{1}{n_g}\mathbf{J_g}\right)\left(\mathbf{I} - \frac{1}{n_g}\mathbf{J_g}\right) = \left(\mathbf{I} - \frac{1}{n_g}\mathbf{J_g}\right) \qquad \left(\frac{1}{n_g}\mathbf{J_g} - \mathbf{P}\right)\left(\frac{1}{n_g}\mathbf{J_g} - \mathbf{P}\right) = \left(\frac{1}{n_g}\mathbf{J_g} - \mathbf{P}\right)$$

This leads to the following distributional results for $SSPE$ and $SSLF$ for the standard regression model (normal, independent, and homoskedastic errors).

$$SSPE = \mathbf{Y}'\left(\mathbf{I} - \frac{1}{n_g}\mathbf{J_g}\right)\mathbf{Y} \sim \chi^2_{df_{PE},\Omega_{PE}} \qquad SSLF = \mathbf{Y}'\left(\frac{1}{n_g}\mathbf{J_g} - \mathbf{P}\right)\mathbf{Y} \sim \chi^2_{df_{LF},\Omega_{LF}}$$

The degrees of freedom and non-centrality parameters, as well their independence are given below.

$$df_{PE} = \text{trace}\left(\mathbf{I}-\frac{1}{n_g}\mathbf{J_g}\right) = n - \sum_{j=1}^{c}\frac{1}{n_j}n_j = n-c \qquad df_{LF} = \text{trace}\left(\frac{1}{n_g}\mathbf{J_g}-\mathbf{P}\right) = c-2$$

$$\Omega_{PE} = \frac{1}{2\sigma^2}\mu'\left(\mathbf{I}-\frac{1}{n_g}\mathbf{J_g}\right)\mu = \frac{1}{2\sigma^2}\left[\begin{array}{cccc}\mu_1\mathbf{1}'_{n_1} & \mu_2\mathbf{1}'_{n_2} & \cdots & \mu_c\mathbf{1}'_{n_c}\end{array}\right]\left(\mathbf{I}-\frac{1}{n_g}\mathbf{J_g}\right)\left[\begin{array}{c}\mu_1\mathbf{1}_{n_1}\\ \mu_2\mathbf{1}_{n_2}\\ \cdots\\ \mu_c\mathbf{1}_{n_c}\end{array}\right] = \frac{1}{2\sigma^2}\left(\mu'\mu-\mu'\mu\right) = 0$$

$$\Omega_{LF} = \frac{1}{2\sigma^2}\mu'\left(\frac{1}{n_g}\mathbf{J_g}-\mathbf{P}\right)\mu \geq 0 \Omega_{LF} = 0 \quad \Longleftrightarrow \quad \mu = \mathbf{X}\boldsymbol{\beta} \text{ or } \mu = \mathbf{0}$$

$$\left(\mathbf{I}-\frac{1}{n_g}\mathbf{J_g}\right)\left(\frac{1}{n_g}\mathbf{J_g}-\mathbf{P}\right) = \frac{1}{n_g}\mathbf{J_g}-\mathbf{P}-\frac{1}{n_g}\mathbf{J_g}+\mathbf{P} = 0 \quad \Rightarrow \quad SSPE \perp SSLF$$

$$\Rightarrow \quad F_{LF} = \frac{\left[\frac{SSLF}{c-2}\right]}{\left[\frac{SSPE}{n-c}\right]} = \frac{MSLF}{MSPE} \sim F_{c-2,n-c,\Omega_{LF}}$$

Under $H_0 : \mu = \mathbf{X}\boldsymbol{\beta}$, $F_{\text{LF}} \sim F_{c-2,n-c,\Omega_{LF}}$. Note that when there are $p$ predictor variables and the model has $p' = p+1$ parameters, we have the following degrees of freedom.

$$df_{\text{Err}} = n-p' \qquad df_{PE} = n-c \qquad df_{LF} = c-p'$$

**Example: Breaking Strength of Fibers**

A study was conducted, relating breaking strength of fibers (in the machine direction) to water pressure level (Ndaro, Jin, Chen, and Wu (2007)). There were $n = 30$ measurements, with $n_i = 5$ replicates at each of $c = 6$ water pressure levels ($X$=60, 80, 100, 120, 150, 200). The data are given in Table 5.1, along with computations needed to carry out the F-test for Lack of Fit. A plot of the data, fitted equation and grup means is given in Figure 5.1.

$$\hat{Y}_j = 229.0052+0.4640X_j \qquad SSPE = \sum_{j=1}^{c}\sum_{i=1}^{n_j}\left(Y_{ij}-\overline{Y}_j\right)^2 = \sum_{j=1}^{c}\sum_{i=1}^{n_j}\left(n_j-1\right)S_j^2 \qquad SSLF = \sum_{j=1}^{c}n_j\left(\overline{Y}_j-\hat{Y}_j\right)^2$$

$$SSPE = (5-1)\left[1586.495+312.086+1371.031+1058.175+1920.886+2075.293\right] = 4(8323.966) = 33295.864$$

$$SSLF = 5\left[(245.260 - 256.8449)^2 + \cdots + (321.962 - 321.8043)^2\right] = 5(345.8341) = 1729.1705$$

$$df_{PE} = n - c = 30 - 6 = 24 \qquad df_{LF} = c - 2 = 6 - 2 = 4$$

$$H_0 : \mu_j = \beta_0 + \beta_1 X_j \qquad H_A : \mu_j \neq \beta_0 + \beta_1 X_j \qquad TS : F_{LF} = \frac{1729.1705/4}{33295.864/24} = \frac{432.2926}{1387.3277} = 0.3116$$

$$RR : F_{LF} \geq F_{.05;4,24} = 2.776 \qquad P = P\left(F_{4,24} \geq 0.3116\right) = .8674$$

The R program and output are given below. Note that the **lm** function, used with pressure as a factor (nominal) variable fits the cell means model which produces the Pure Error Sum of Squares.

**R Program**

```
fiber1 <- read.table("http://www.stat.ufl.edu/~winner/data/fiber_strength.dat",
   header=F,col.names=c("pressure","strength"))
attach(fiber1)

pressure.f <- factor(pressure)

fiber.mod1 <- lm(strength ~ pressure)
summary(fiber.mod1)
anova(fiber.mod1)

fiber.mod2 <- lm(strength ~ pressure.f)
anova(fiber.mod2)

anova(fiber.mod1, fiber.mod2)

(y.mean <- aggregate(strength,list(pressure),mean))
colnames(y.mean) <- c("pressure","str.mean")

plot(pressure,strength, main="Strength vs Pressure for Fibre Experiment")
abline(fiber.mod1)
points(y.mean,pch=16)

### Matrix form

n <- length(strength)
Y <- strength
X0 <- rep(1,n)
X <- cbind(X0,pressure)

J1 <- matrix(rep(1,25),ncol=5);  J0 <- matrix(rep(0,25),ncol=5)
J.grp <- rbind(cbind(J1,J0,J0,J0,J0,J0),cbind(J0,J1,J0,J0,J0,J0),
   cbind(J0,J0,J1,J0,J0,J0),cbind(J0,J0,J0,J1,J0,J0),cbind(J0,J0,J0,J0,J1,J0),
   cbind(J0,J0,J0,J0,J0,J1))
J.grp <- (1/5) * J.grp

P.X <- X %*% solve(t(X) %*% X) %*% t(X)
I.n <- diag(n)

(SSE <- t(Y) %*% (I.n - P.X) %*% Y)
(dfE <- n-ncol(X))
(MSE <- SSE/dfE)

(SSPE <- t(Y) %*% (I.n - J.grp) %*% Y)
(dfPE <- n-length(unique(pressure)))
(MSPE <- SSPE/dfPE)

(SSLF <- t(Y) %*% (J.grp - P.X) %*% Y)
```

```
(dfLF <- length(unique(pressure))-ncol(X))
(MSLF <- SSLF/dfLF)

(F.LOF <- MSLF/MSPE)
(p.F.LOF <- 1 - pf(F.LOF,dfLF,dfPE))
```

### R Output

```
> summary(fiber.mod1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 229.0052    17.7091   12.93  2.5e-13 ***
pressure      0.4640     0.1394    3.33  0.00245 **

Residual standard error: 35.37 on 28 degrees of freedom
Multiple R-squared:  0.2836,    Adjusted R-squared:  0.2581
F-statistic: 11.09 on 1 and 28 DF,  p-value: 0.002447

> anova(fiber.mod1)
Analysis of Variance Table
Response: strength
          Df Sum Sq Mean Sq F value   Pr(>F)
pressure   1  13868 13868.4  11.087 0.002447 **
Residuals 28  35025  1250.9


> anova(fiber.mod2)
Analysis of Variance Table
Response: strength
           Df Sum Sq Mean Sq F value  Pr(>F)
pressure.f  5  15598  3119.5  2.2486 0.08215 .
Residuals  24  33296  1387.3

> anova(fiber.mod1, fiber.mod2)
Analysis of Variance Table
Model 1: strength ~ pressure
Model 2: strength ~ pressure.f
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     28 35025
2     24 33296  4    1729.2 0.3116 0.8674
>
### Matrix Form
> (SSE <- t(Y) %*% (I.n - P.X) %*% Y)
[1,] 35025.04
> (dfE <- n-ncol(X))
[1] 28
> (MSE <- SSE/dfE)
[1,] 1250.894
> (SSPE <- t(Y) %*% (I.n - J.grp) %*% Y)
[1,] 33295.87
> (dfPE <- n-length(unique(pressure)))
[1] 24
> (MSPE <- SSPE/dfPE)
[1,] 1387.328
> (SSLF <- t(Y) %*% (J.grp - P.X) %*% Y)
[1,] 1729.171
> (dfLF <- length(unique(pressure))-ncol(X))
[1] 4
> (MSLF <- SSLF/dfLF)
[1,] 432.2927
>
```

| $j$ | $X_j$ | $Y_{1j}$ | $Y_{2j}$ | $Y_{3j}$ | $Y_{4j}$ | $Y_{5j}$ | $\overline{Y}_j$ | $S_j^2$ | $\hat{Y}_j$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 60 | 225.60 | 189.25 | 245.86 | 284.25 | 281.34 | 245.260 | 1586.495 | 256.8449 |
| 2 | 80 | 294.22 | 250.71 | 272.36 | 287.13 | 262.89 | 273.462 | 312.086 | 266.1248 |
| 3 | 100 | 318.21 | 249.14 | 238.34 | 298.36 | 312.46 | 283.302 | 1371.031 | 275.4047 |
| 4 | 120 | 234.05 | 293.08 | 299.33 | 319.85 | 300.79 | 289.420 | 1058.175 | 284.6847 |
| 5 | 150 | 265.53 | 262.88 | 367.48 | 280.29 | 274.13 | 290.062 | 1920.886 | 298.6045 |
| 6 | 200 | 278.55 | 360.15 | 323.82 | 373.39 | 273.90 | 321.962 | 2075.293 | 321.8043 |

Table 5.1: Data and computations for the Lack of Fit test - Fiber strength data

```
> (F.LOF <- MSLF/MSPE)
[1,] 0.3116009
> (p.F.LOF <- 1 - pf(F.LOF,dfLF,dfPE))
[1,] 0.8673657
```

$\nabla$

## 5.2 Checking Normality

A normal probability plot of the ordered residuals versus their predicted values should fall approximately on a straight line. A histogram should be mound-shaped. Neither of these methods work well with small samples (even data generated from a normal distribution will not necessarily look like it is normal).

Various tests are computed directly by statistical computing packages. The Shapiro-Wilk and Kolmogorov-Smirnov tests are commonly reported, with $P$-values for testing $H_0$: Errors are normally distributed.

When data are not normally distributed, the **Box-Cox transformation** is often applied to the data. This involves fitting regression models for various power transformations of $Y$ on $X$, where:

$$Y_i^{(\lambda)} = \begin{cases} \frac{Y_i^{\lambda}-1}{\lambda(\dot{Y})^{(\lambda-1)}} & \lambda \neq 0 \\ \dot{Y}\ln(Y_i) & \lambda = 0 \end{cases}$$

Here $\dot{Y}$ is the geometric mean of $Y_1, \ldots, Y_n$, where all data are strictly positive (a constant can be added to all observations to assure this).

$$\dot{Y} = \left(\prod_{i=1}^{n} Y_i\right)^{1/n} = \exp\left\{\frac{\sum_{i=1}^{n}\ln(Y_i)}{n}\right\}$$

Values of $\lambda$ between -2 and 2 by 0.1 are typically run, and the value of $\lambda$ that has the smallest Error Sum of Squares (equivalently Maximum Likelihood) is identified. Standard statistical software packages will present an estimate and confidence interval for $\lambda$.

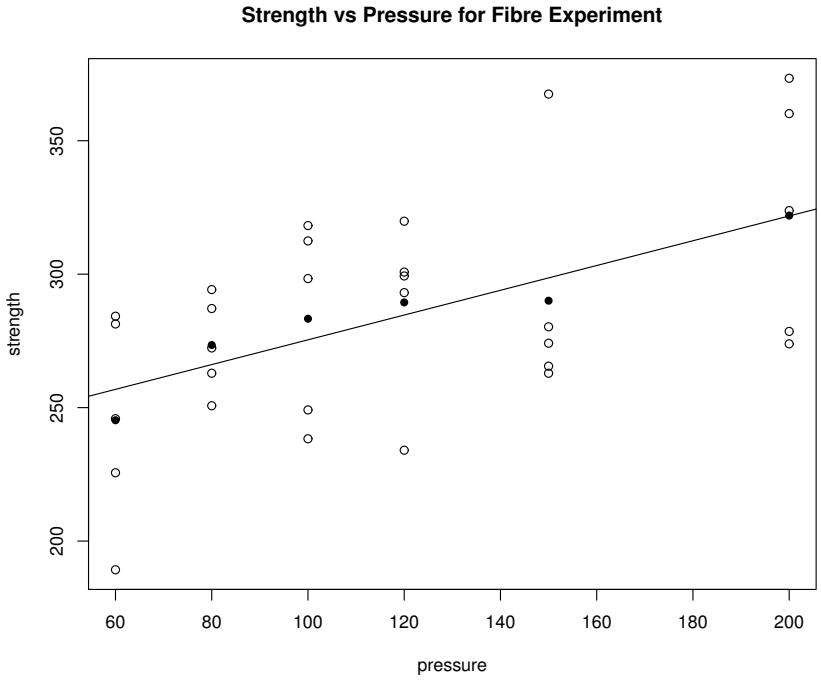


Figure 5.1: Plot of Fiber Strength versus Water Pressure, Fitted Equation, and Group Means (bold dots)

**Example: Bearing Capacity of Paddy Fields**

A study was conducted measuring the depth of soft layer ($X$) and bearing capacity at 10 cm depth measured by a penetrometer ($Y$) for $n = 51$ samples in paddy fields in Sputh China (Chanying and Junzheng (1998)). After fitting a simple linear regression, residuals are obtained and ranked. Then quantiles are assigned to the ranks, and $Z$ values corresponding to the quantiles are obtained based on the standard normal distribution, which represent the expected values of the residuals under normality when multiplied by the residual standard deviation. The correlation between the residuals and their expected values is .9953. The data, residuals, ranks, quantiles and expected values are given in Table 5.2. The fitted equation, $SSE$, $MSE$, and $s$ are given below. A plot of the data and regression line and residuals versus predicted values are given in Figure 5.2.

$$\hat{Y} = 290.2839 - 7.7359X \qquad SSE = \sum_{i=1}^{51} e_i^2 = 300506.8 \qquad s^2 = MSE = \frac{300506.8}{51-2} = 6132.79 \qquad s = \sqrt{6132.79} = 78.31$$

An R Program that fits the model, conducts the Shapiro-Wilk test for normality of errors, obtains a normal probability plot for the residuals, and obtains the Box-Cox transformation and its output are given below. Although there is absolutely no evidence of non-normality of errors, the Box-Cox regression suggests a square root transformation on $Y$, which is not pursued here. The normal probability plot is given in Figure 5.3 and the Box-Cox transformation is given in Figure 5.4.

**R Program**

```
paddy <- read.table("http://www.stat.ufl.edu/~winner/data/paddyfields_slr.dat",
     header=F,col.names=c("depth.pf", "bearcap.pf"))
attach(paddy)

paddy.mod1 <- lm(bearcap.pf ~ depth.pf)
summary(paddy.mod1)
paddy.e <- resid(paddy.mod1)
paddy.yhat <- predict(paddy.mod1)

par(mfrow=c(2,1))
plot(depth.pf,bearcap.pf)
abline(paddy.mod1)
plot(paddy.yhat, paddy.e)
abline(h=0)
par(mfrow=c(1,1))

shapiro.test(paddy.e)
qqnorm(paddy.e); qqline(paddy.e)

library(MASS)

bc.mod1 <- boxcox(paddy.mod1,plotit=T)
print(cbind(bc.mod1$x,bc.mod1$y))      # Print out results (lambda,log-like)
print(bc.mod1$x[which.max(bc.mod1$y)])     # Print out "best" lambda
ci.bc <- max(bc.mod1$y)-0.5*qchisq(0.95,1)   # Obtain cut-off for 95% CI (in log-like)
print(bc.mod1$x[bc.mod1$y>= ci.bc])    # Print Values of lambda in 95% CI
```

### R Output

```
> summary(paddy.mod1)

Call:
lm(formula = bearcap.pf ~ depth.pf)

Residuals:
     Min       1Q   Median       3Q      Max
-162.604  -54.972   -4.284   50.783  199.320

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  290.284     28.393  10.224 9.60e-14 ***
depth.pf      -7.736      1.715  -4.511 4.04e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 78.31 on 49 degrees of freedom
Multiple R-squared:  0.2935,    Adjusted R-squared:  0.279
F-statistic: 20.35 on 1 and 49 DF,  p-value: 4.039e-05

> shapiro.test(paddy.e)

        Shapiro-Wilk normality test

data:  paddy.e
W = 0.99008, p-value = 0.9446

> print(cbind(bc.mod1$x,bc.mod1$y))      # Print out results (lambda,log-like)
              [,1]       [,2]
  [1,] -2.00000000 -130.41873
  [2,] -1.95959596 -128.66654
  [3,] -1.91919192 -126.92862
```

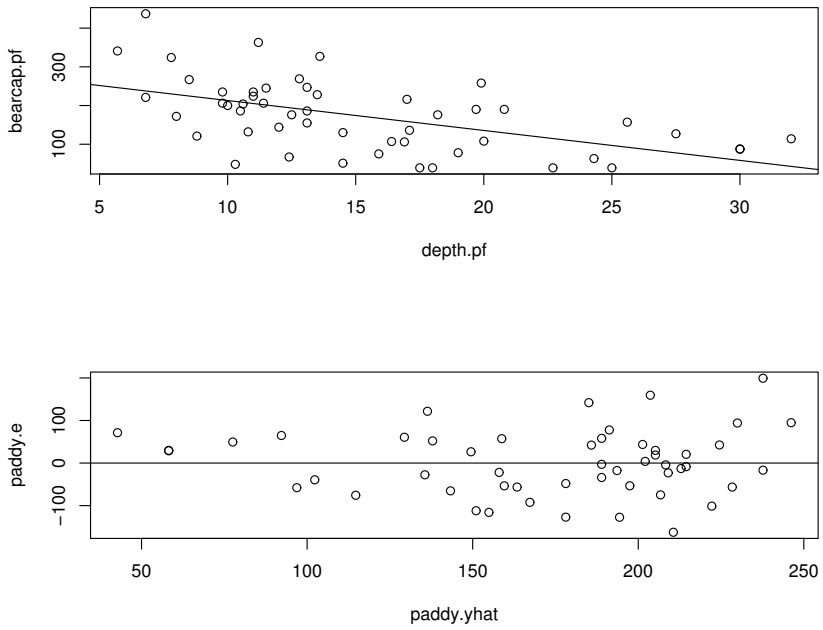


Figure 5.2: Plot of data and Regression Line (Top Panel) and Residuals versus Predicted Values (Lower Panel) - Paddy Field Bearing Capacity Analysis

```
     ...
 [98,]  1.91919192  -84.91787
 [99,]  1.95959596  -85.97515
[100,]  2.00000000  -87.05299
> print(bc.mod1$x[which.max(bc.mod1$y)])     # Print out "best" lambda
[1] 0.5454545
> ci.bc <- max(bc.mod1$y)-0.5*qchisq(0.95,1)   # Obtain cut-off for 95% CI (in log-like)
> print(bc.mod1$x[bc.mod1$y>= ci.bc])    # Print Values of lambda in 95% CI
 [1] 0.1818182 0.2222222 0.2626263 0.3030303 0.3434343 0.3838384 0.4242424
 [8] 0.4646465 0.5050505 0.5454545 0.5858586 0.6262626 0.6666667 0.7070707
[15] 0.7474747 0.7878788 0.8282828 0.8686869 0.9090909 0.9494949
```

$\nabla$

## 5.3 Checking Equal Variance

A plot of the residuals versus the fitted values should be a random cloud of points centered at 0. When the variances are unequal, the variance tends to increase with the mean, and we observe a funnel-type shape. Two tests for equal variance are the Brown-Forsyth test and the Breusch-Pagan (aka Cook-Weisberg) test.

| $i$ | Depth ($X$) | BearCap($Y$) | $\hat{Y}$ | $e$ | rank($e$) | quantile | $z$(qntl) | $E\{e\}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 10.6 | 204 | 208.28 | -4.28 | 26 | 0.5000 | 0.0000 | 0.0000 |
| 2 | 13.1 | 155 | 188.94 | -33.94 | 18 | 0.3439 | -0.4018 | -31.4686 |
| 3 | 11.4 | 206 | 202.09 | 3.91 | 28 | 0.5390 | 0.0980 | 7.6727 |
| 4 | 11.2 | 363 | 203.64 | 159.36 | 50 | 0.9683 | 1.8563 | 145.3687 |
| 5 | 9.8 | 235 | 214.47 | 20.53 | 30 | 0.5780 | 0.1969 | 15.4200 |
| 6 | 12.8 | 269 | 191.26 | 77.74 | 45 | 0.8707 | 1.1299 | 88.4815 |
| 7 | 12.4 | 67 | 194.36 | -127.36 | 2 | 0.0317 | -1.8563 | -145.368 |
| 8 | 12.0 | 144 | 197.45 | -53.45 | 15 | 0.2854 | -0.5670 | -44.4010 |
| 9 | 10.3 | 48 | 210.60 | -162.60 | 1 | 0.0122 | -2.2509 | -176.274 |
| 10 | 13.1 | 247 | 188.94 | 58.06 | 41 | 0.7927 | 0.8158 | 63.8844 |
| 11 | 20.8 | 190 | 129.38 | 60.62 | 42 | 0.8122 | 0.8860 | 69.3857 |
| 12 | 13.5 | 228 | 185.85 | 42.15 | 35 | 0.6756 | 0.4555 | 35.6678 |
| 13 | 24.3 | 63 | 102.30 | -39.30 | 17 | 0.3244 | -0.4555 | -35.6678 |
| 14 | 30.0 | 87 | 58.21 | 28.79 | 32 | 0.6171 | 0.2978 | 23.3216 |
| 15 | 32.0 | 114 | 42.74 | 71.26 | 44 | 0.8512 | 1.0417 | 81.5760 |
| 16 | 18.2 | 176 | 149.49 | 26.51 | 31 | 0.5976 | 0.2470 | 19.3462 |
| 17 | 16.4 | 107 | 163.42 | -56.42 | 12 | 0.2268 | -0.7493 | -58.6816 |
| 18 | 19.7 | 190 | 137.89 | 52.11 | 39 | 0.7537 | 0.6860 | 53.7259 |
| 19 | 17.1 | 136 | 158.00 | -22.00 | 21 | 0.4024 | -0.2470 | -19.3462 |
| 20 | 16.9 | 106 | 159.55 | -53.55 | 14 | 0.2659 | -0.6254 | -48.9766 |
| 21 | 19.9 | 258 | 136.34 | 121.66 | 48 | 0.9293 | 1.4704 | 115.1473 |
| 22 | 11.0 | 224 | 205.19 | 18.81 | 29 | 0.5585 | 0.1473 | 11.5322 |
| 23 | 7.8 | 324 | 229.94 | 94.06 | 46 | 0.8902 | 1.2278 | 96.1537 |
| 24 | 6.8 | 437 | 237.68 | 199.32 | 51 | 0.9878 | 2.2509 | 176.2748 |
| 25 | 6.8 | 221 | 237.68 | -16.68 | 23 | 0.4415 | -0.1473 | -11.5322 |
| 26 | 13.6 | 327 | 185.08 | 141.92 | 49 | 0.9488 | 1.6331 | 127.8949 |
| 27 | 11.0 | 235 | 205.19 | 29.81 | 34 | 0.6561 | 0.4018 | 31.4686 |
| 28 | 14.5 | 51 | 178.11 | -127.11 | 3 | 0.0512 | -1.6331 | -127.894 |
| 29 | 8.0 | 172 | 228.40 | -56.40 | 13 | 0.2463 | -0.6860 | -53.7259 |
| 30 | 10.8 | 132 | 206.74 | -74.74 | 9 | 0.1683 | -0.9609 | -75.2528 |
| 31 | 5.7 | 341 | 246.19 | 94.81 | 47 | 0.9098 | 1.3393 | 104.8799 |
| 32 | 9.8 | 206 | 214.47 | -8.47 | 25 | 0.4805 | -0.0489 | -3.8318 |
| 33 | 10.0 | 200 | 212.93 | -12.93 | 24 | 0.4610 | -0.0980 | -7.6727 |
| 34 | 14.5 | 130 | 178.11 | -48.11 | 16 | 0.3049 | -0.5104 | -39.9722 |
| 35 | 8.5 | 267 | 224.53 | 42.47 | 36 | 0.6951 | 0.5104 | 39.9722 |
| 36 | 8.8 | 121 | 222.21 | -101.21 | 6 | 0.1098 | -1.2278 | -96.1537 |
| 37 | 15.9 | 75 | 167.28 | -92.28 | 7 | 0.1293 | -1.1299 | -88.4815 |
| 38 | 12.5 | 176 | 193.59 | -17.59 | 22 | 0.4220 | -0.1969 | -15.4200 |
| 39 | 27.5 | 127 | 77.55 | 49.45 | 38 | 0.7341 | 0.6254 | 48.9766 |
| 40 | 30.0 | 88 | 58.21 | 29.79 | 33 | 0.6366 | 0.3493 | 27.3581 |
| 41 | 17.0 | 216 | 158.77 | 57.23 | 40 | 0.7732 | 0.7493 | 58.6816 |
| 42 | 13.1 | 186 | 188.94 | -2.94 | 27 | 0.5195 | 0.0489 | 3.8318 |
| 43 | 10.5 | 186 | 209.06 | -23.06 | 20 | 0.3829 | -0.2978 | -23.3216 |
| 44 | 25.6 | 157 | 92.25 | 64.75 | 43 | 0.8317 | 0.9609 | 75.2528 |
| 45 | 22.7 | 39 | 114.68 | -75.68 | 8 | 0.1488 | -1.0417 | -81.5760 |
| 46 | 18.0 | 39 | 151.04 | -112.04 | 5 | 0.0902 | -1.3393 | -104.879 |
| 47 | 25.0 | 39 | 96.89 | -57.89 | 11 | 0.2073 | -0.8158 | -63.8844 |
| 48 | 11.5 | 245 | 201.32 | 43.68 | 37 | 0.7146 | 0.5670 | 44.4010 |
| 49 | 17.5 | 39 | 154.91 | -115.91 | 4 | 0.0707 | -1.4704 | -115.147 |
| 50 | 19.0 | 78 | 143.30 | -65.30 | 10 | 0.1878 | -0.8860 | -69.3857 |
| 51 | 20.0 | 108 | 135.57 | -27.57 | 19 | 0.3634 | -0.3493 | -27.3581 |

Table 5.2: Data and computations for Residuals and their Expected Values - Paddy Field Bearing Capacity Analysis

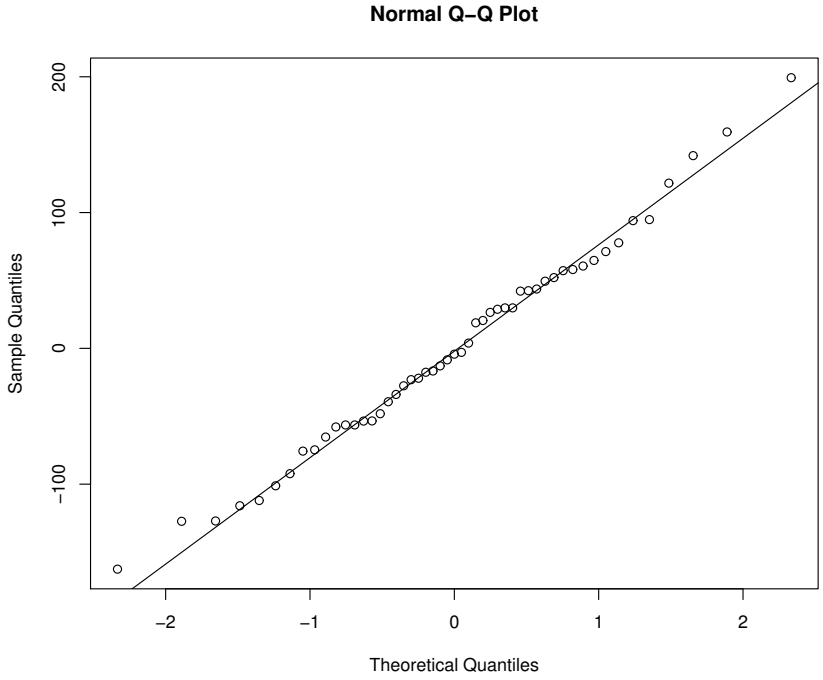


Figure 5.3: Normal Probability Plot of Residuals - Paddy Field Bearing Capacity Analysis

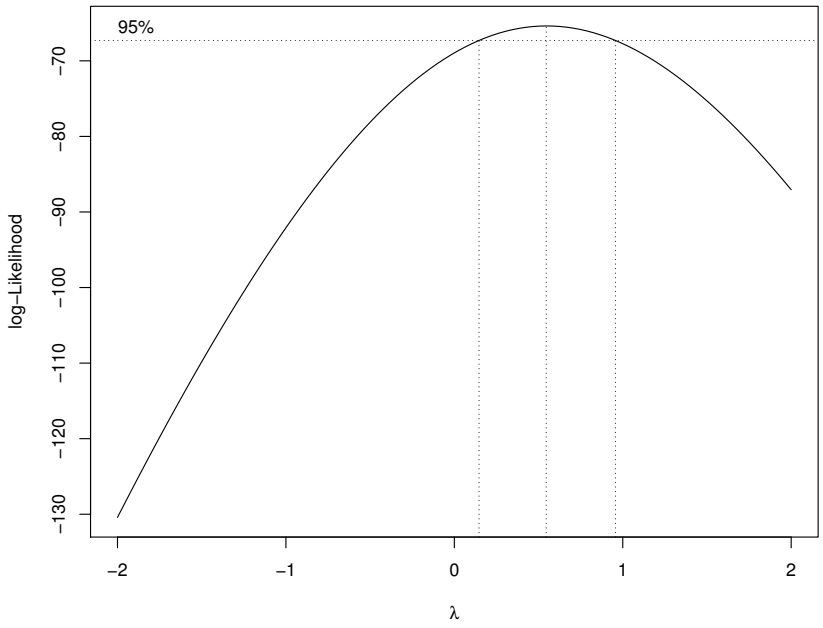


Figure 5.4: Box-Cox Transformation - Paddy Field Bearing Capacity Analysis

**Brown-Forsyth Test** - Splits data into two groups of approximately equal sample sizes based on their fitted values (any cases with the same fitted values should be in the same group). Then labeling the residuals $e_{11},\ldots,e_{1n_1}$ and $e_{21},\ldots,e_{2n_2}$, obtain the median residual for each group: $\tilde{e}_1$ and $\tilde{e}_2$, respectively. Then compute the following:

$$d_{ij}=|e_{ij}-\tilde{e}_i| \quad i=1,2;j=1,\ldots,n_i \qquad \overline{d}_i=\frac{\sum_{i=1}^{n_i}d_{ij}}{n_i} \qquad s_i^2=\frac{\sum_{i=1}^{n_i}\left(d_{ij}-\overline{d}_i\right)^2}{n_i-1} \qquad s_p^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$$

Then, a 2-sample $t$-test is conducted to test $H_0$: Equal Variances in the 2 groups:

$$TS: t_{obs}=\frac{\overline{d}_1-\overline{d}_2}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \qquad RR: |t_{obs}|\geq t_{\alpha/2,n-2} \qquad P\text{-value}=P\left(t_{n-2}\geq|t_{obs}|\right)$$

**Breusch-Pagan Test(aka Cook-Weisberg Test)** - Fits a regression of the squared residuals on $X$ and tests whether the variance is related to $X$ (this can be extended to multiple predictor variables).

$$H_0: V\left\{\epsilon_i\right\}=\sigma^2h\left(\gamma_0\right) \qquad H_A: V\left\{\epsilon_i\right\}=\sigma^2h\left(\gamma_0+\gamma_1X_i\right)$$

When the regression of the squared residuals is fit, we obtain $SSR_{e^2}$, the regression sum of squares. The test is conducted as follows, where $SSE$ is the Error Sum of Squares for the original regression of $Y$ on $X$.

$$TS: X_{obs}^2=\frac{(SSR_{e^2}/2)}{(SSE/n)^2} \qquad RR: X_{obs}^2\geq\chi_{\alpha,1}^2 \qquad P\text{-value: } P\left(\chi_1^2\geq X_{obs}^2\right)$$

When the variance is not constant, we can transform $Y$ (often the Box-Cox transformation will also to obtain approximately constant variance).

We can also use **Estimated Weighted Least Squares** by relating the standard deviation (or variance) of the errors to the mean. This is an iterative process, where the weights are re-weighted each iteration. The weights are the reciprocal of the estimated variance (as a function of the mean). Iteration continues until the regression coefficient estimates stabilize. This is described in detail in Chapter 6.

Another, simple to compute, method is to obtain robust standard errors of the OLS estimators based on the residuals from the linear regression (using the squared residuals as estimates of the variances for the individual cases). This method was originally proposed by White (1980). The estimated variance-covariance matrix (with resulting **robust to heteroskedasticity standard errors** for $\hat{\boldsymbol{\beta}}$ is):

$$\hat{V}\left\{\hat{\boldsymbol{\beta}}\right\}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{E}}_2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \qquad \hat{\mathbf{E}}_2=\begin{bmatrix} e_1^2 & 0 & \cdots & 0 \\ 0 & e_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e_n^2 \end{bmatrix}$$

When the variance is a power of the mean, Barlett's method can be used to obtain an approximate variance stabilizing transformation for $Y$. If the distribution of $Y$ is a from a known family (e.g. Binomial,

Poisson, Gamma), we can fit a **Generalized Linear Model**, or fit a "direct" regression model based on the specific distribution.

**Example: Bearing Capacity of Paddy Fields**

Applications of the Brown-Forsyth and Breusch-Pagan tests to the Paddy Field Bearing Capacity are given here. For the Brown-Forsyth test, the observations are broken into groups with $X \leq 13.1$ and $X \geq 13.5$, with $n_1 = 26$ and $n_2 = 25$. Note that these calculations were made using EXCEL, R code and output are given below.

$$\overline{d}_1 = 58.94 \qquad s_1 = 55.11 \qquad n_1 = 26 \qquad \overline{d}_2 = 64.70 \qquad s_2 = 39.09 \qquad n_2 = 25$$

$$s_p^2 = \frac{(26-1)(55.11)^2 + (25-1)(39.09)^2}{26+25-2} = \frac{112593.2}{49} = 2297.82$$

$$TS: t_{obs} = \frac{58.94 - 64.70}{\sqrt{2297.82\left(\frac{1}{26} + \frac{1}{25}\right)}} = \frac{-5.76}{13.43} = -0.429 \qquad RR: |t_{obs}| \geq t_{.025,49} = 2.010 \qquad 2P\left(t_{49} \geq |-0.429|\right) = .6698$$

For the Breusch-Pagan test, regress the squared residuals on $X$, and obtain the chi-square statistic as follows.

$$SSE = 300506.5 \qquad n = 51 \qquad \frac{300506.5}{51} = 5892.28 \qquad SSR_{e^2} = 132124650 \qquad \frac{132124650}{2} = 66062325$$

$$TS: X^2_{obs} = \frac{66062325}{(5892.28)^2} = 1.903 \qquad RR: X^2_{obs} \geq \chi^2_{.05,1} = 3.841 \qquad P\left(\chi^2_1 \geq 1.903\right) = .1677$$

The R program and output are given here.

**R Program**

```
paddy <- read.table("http://www.stat.ufl.edu/~winner/data/paddyfields_slr.dat",
     header=F,col.names=c("depth.pf", "bearcap.pf"))
attach(paddy)

paddy.mod1 <- lm(bearcap.pf ~ depth.pf)
summary(paddy.mod1)
paddy.e <- resid(paddy.mod1)

# Conduct Brown-Forsythe Test of Homogeneity of Variance
Y <- bearcap.pf
X <- depth.pf
Residuals <- paddy.e
group <- numeric(length(Y))

for (i in 1:length(Y)) {
if (X[i] <= 13.3) group[i]=1
else group[i]=2
}
```

```
d1 <- abs(Residuals[group==1] - median(Residuals[group==1]))
d2 <- abs(Residuals[group==2] - median(Residuals[group==2]))

n_d1 <- length(d1); n_d2 <- length(d2)
mean_d1 <- mean(d1);   mean_d2 <- mean(d2)
var_d1 <- var(d1); var_d2 <- var(d2)

s_BF <- sqrt(((n_d1-1)*var_d1 + (n_d2-1)*var_d2)/(n_d1+n_d2-2))

(t_BF <- (mean_d1 - mean_d2)/(s_BF*sqrt((1/n_d1)+(1/n_d2))))
(t_crit <- qt(.975,n_d1+n_d2-2))
(p_BF <- 2*(1-pt(abs(t_BF),n_d1+n_d2-2)))

# Conduct Breusch-Pagan Test of Homogeneity of Variance
# Brute Force Approach

E2 <- paddy.e^2                            # Compute e^2 for each observation
(SSE2E2 <- (length(E2)-1)*var(E2))           # Compute SSTO for e^2 values

paddy.mod2 <- lm(E2 ~ X)                    # Fit regression of e^2 on X
anova(paddy.mod2)
(SSE_E2 <- deviance(paddy.mod2))            # Obtain SSE from regression of e^2 on X
(SSR_E2 <- SSE2E2 - SSE_E2)                  # Compute SSR*

(X2_BP <- (SSR_E2/2)/(sum(E2)/length(E2))^2)   # Compute Breusch-Pagan test statistic
(X2_crit <- qchisq(.95,1))                     # Obtain critical value
(p_BP <- 1-pchisq(X2_BP,1))                    # Compute P-value

### Breusch-Pagan Test using lmtest package
install.packages("lmtest")
library(lmtest)
bptest(bearcap.pf ~ depth.pf,studentize=FALSE)
```

### R Output

```
> (t_BF <- (mean_d1 - mean_d2)/(s_BF*sqrt((1/n_d1)+(1/n_d2))))
[1] -0.4291169
> (t_crit <- qt(.975,n_d1+n_d2-2))
[1] 2.009575
> (p_BF <- 2*(1-pt(abs(t_BF),n_d1+n_d2-2)))
[1] 0.669719
>
> # Conduct Breusch-Pagan Test of Homogeneity of Variance
> # Brute Force Approach
>
> E2 <- paddy.e^2                            # Compute e^2 for each observation
> (SSE2E2 <- (length(E2)-1)*var(E2))           # Compute SSTO for e^2 values
[1] 3257558096
> anova(paddy.mod2)
Analysis of Variance Table
Response: E2
          Df     Sum Sq   Mean Sq F value Pr(>F)
X          1  132110903 132110903  2.0712 0.1565
Residuals 49 3125447193  63784637
> (SSE_E2 <- deviance(paddy.mod2))            # Obtain SSE from regression of e^2 on X
[1] 3125447193
> (SSR_E2 <- SSE2E2 - SSE_E2)                  # Compute SSR*
[1] 132110903
> (X2_BP <- (SSR_E2/2)/(sum(E2)/length(E2))^2)   # Compute Breusch-Pagan test statistic
[1] 1.902568
> (X2_crit <- qchisq(.95,1))                     # Obtain critical value
```

```
[1] 3.841459
> (p_BP <- 1-pchisq(X2_BP,1))                    # Compute P-value
[1] 0.1677911
> bptest(bearcap.pf ~ depth.pf,studentize=FALSE)
        Breusch-Pagan test
data:  bearcap.pf ~ depth.pf
BP = 1.9026, df = 1, p-value = 0.1678
```

## 5.4 Checking Independence

When the data are a time (or spatial) series, the errors can be correlated over time (or space), referred to as being **autocorrelated**. A plot of residuals versus time should be random, not displaying a trending pattern (linear or cyclical). If it does show these patterns, autocorrelation may be present.

The Durbin-Watson test is used to test for serial autocorrelation in the errors, where the null hypothesis is that the errors are uncorrelated. Unfortunately, the formal test can end in one of 3 possible outcomes: reject $H_0$, accept $H_0$, or inconclusive. Statistical software packages can report an approximate $P$-value, based on re-sampling of the regression residuals. The test is obtained as follows:

$$TS: DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

Decision Rule: $DW < d_L \quad \Rightarrow \quad$ Reject $H_0 \qquad DW > d_U \quad \Rightarrow \quad$ Accept $H_0 \qquad$ Otherwise Inconclusive

where tables of $d_L$ and $d_U$ are in standard regression texts and posted on the internet. These values are indexed by the number of predictor variables (1, in the case of simple regression) and the sample size ($n$).

Expanding the numerator helps understand the test. Recall that if the errors are independent (uncorrelated), then $E\{\epsilon_t \epsilon_{t-1}\} = 0$.

$$\sum (e_t - e_{t-1})^2 = \sum e_t^2 + \sum e_{t-1}^2 - 2\sum e_t e_{t-1} \quad \approx \quad 2\sum e_t^2 \qquad \text{if errors are uncorrelated}$$

Thus, we expect that the Durbin-Watson statistic should be around 2. If it is much smaller, that is evidence of positive autocorrelation; if it is much larger, evidence is for negative autocorrelation.

When errors are not independent (positively correlated), estimated standard errors of estimated regression coefficients tend to be too small, making the $t$-statistics artificially large and confidence intervals artificially narrow.

The **Cochrane-Orcutt** method transforms the $Y$ and $X$ variables, and fits the model based on the transformed responses. Another approach is to use **Estimated Generalized Least Squares (EGLS)**. This uses the estimated covariance structure of the observations to obtain estimates of the regression coefficients and their estimated standard errors. EGLS is described in detail in Chapter 6.

**Example: Silver from the New World to Spain 1720-1770**

A historical study reported the amount of silver situados minted in the New World ($X$, in millions) and payments made to Spain ($Y$ in millions) over the years 1720-1800 (Marichol and Mantecon (1994)). As

the American Revolution and its aftermath occurred during the final quarter of the century, we will only consider the years 1720-1770. The data are given in Table 5.3 and plots of the data and fitted and residuals versus predicted values are given in Figure 5.5. The fitted regression equation, error sum of squares, sum of squared differences among adjacent residuals and Durbin-Watson test are given below. There is no evidence of autocorrelation among the error terms for the model, as the Durnin-Watson statistic exceed $d_U$ and is close to 2. There does appear to be non-constant error variance, however.

$$\hat{Y}_t = -1.3839 + 0.2808X_t \qquad \sum_{t=1}^{51} e_t^2 = 36.0545 \qquad \sum_{t=2}^{51} (e_t - e_{t-1})^2 = 75.5092$$

$$TS: DW = \frac{75.5092}{36.0545} = 2.0943 \qquad d_L(p=1, n=51) \approx 1.50 \quad d_U(p=1, n=51) \approx 1.59$$

The R program and output are given below.

**R Program**

```
treas <- read.table("http://www.stat.ufl.edu/~winner/data/treas1700.dat",
   header=F,col.names=c("year","situados","minted","amerrev"))
attach(treas)

situados <- situados/1000000
minted <- minted/1000000
year.1770 <- subset(year, year <= 1770)
situados.1770 <- subset(situados, year <= 1770)
minted.1770 <- subset(minted, year <= 1770)

treas.mod1 <- lm(situados.1770 ~ minted.1770)
summary(treas.mod1)
anova(treas.mod1)
e.mod1 <- resid(treas.mod1)
yhat.mod1 <- predict(treas.mod1)
(SSE.mod1 <- sum(e.mod1^2))
treas.n <- length(situados.1770)
DW1.mod1 <- 0
for (t in 2:treas.n) {
DW1.mod1 <- DW1.mod1 + (e.mod1[t] - e.mod1[t-1])^2
}
DW1.mod1
(DW.mod1 <- DW1.mod1 / SSE.mod1)

#install.packages("car")
library(car)
durbinWatsonTest(treas.mod1)

plot(e.mod1, type="l")

par(mfrow=c(2,1))
plot(minted.1770, situados.1770)
abline(treas.mod1)
plot(yhat.mod1, e.mod1)
abline(h=0)
par(mfrow=c(1,1))
```

**R Output**

| year | paidSpn($Y$) | Minted ($X$) | year | paidSpn($Y$) | Minted ($X$) | year | paidSpn($Y$) | Minted ($X$) |
|---|---|---|---|---|---|---|---|---|
| 1720 | 1.226215 | 7.874315 | 1737 | 1.103957 | 8.122133 | 1754 | 1.775134 | 11.594 |
| 1721 | 1.04644 | 9.46073 | 1738 | 1.083357 | 9.49025 | 1755 | 2.06452 | 12.4865 |
| 1722 | 0.376311 | 8.823927 | 1739 | 0.524605 | 8.550686 | 1756 | 0.311502 | 12.2995 |
| 1723 | 0.921332 | 8.107343 | 1740 | 1.32067 | 9.55604 | 1757 | 3.687895 | 12.529 |
| 1724 | 0.928764 | 7.872819 | 1741 | 1.948375 | 8.644177 | 1758 | 2.313552 | 12.757591 |
| 1725 | 0.698335 | 7.369815 | 1742 | 0.672024 | 8.177 | 1759 | 1.910268 | 13.022 |
| 1726 | 0.69036 | 8.236645 | 1743 | 1.6223 | 8.619 | 1760 | 0.295113 | 11.9765 |
| 1727 | 0.702108 | 8.133081 | 1744 | 1.349189 | 10.285 | 1761 | 5.332595 | 11.781 |
| 1728 | 0.353345 | 9.228544 | 1745 | 1.493178 | 10.3275 | 1762 | 1.305525 | 10.11449 |
| 1729 | 0.640906 | 8.814968 | 1746 | 1.457684 | 11.509 | 1763 | 3.090352 | 11.775033 |
| 1730 | 0.714051 | 9.745871 | 1747 | 1.468279 | 12.002 | 1764 | 2.466581 | 9.792536 |
| 1731 | 0.961858 | 8.439871 | 1748 | 2.402106 | 11.628 | 1765 | 2.053284 | 11.604838 |
| 1732 | 1.063585 | 8.726466 | 1749 | 2.138665 | 11.8235 | 1766 | 2.620072 | 11.210047 |
| 1733 | 1.148936 | 10.009796 | 1750 | 1.61492 | 13.209 | 1767 | 2.340972 | 10.41511 |
| 1734 | 0.308749 | 8.506554 | 1751 | 0.848271 | 12.631 | 1768 | 2.573292 | 12.278956 |
| 1735 | 1.063897 | 7.922009 | 1752 | 0.91896 | 13.6255 | 1769 | 2.827777 | 11.938794 |
| 1736 | 0.880634 | 11.016 | 1753 | 2.227312 | 11.594 | 1770 | 3.222307 | 13.926324 |

Table 5.3: Silver Minted in the New World and Payments to Spain 1720-1770

```
> summary(treas.mod1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.38386    0.69661  -1.987   0.0526 .
minted.1770  0.28076    0.06608   4.249 9.58e-05 ***

Residual standard error: 0.8578 on 49 degrees of freedom
Multiple R-squared:  0.2692,    Adjusted R-squared:  0.2543
F-statistic: 18.05 on 1 and 49 DF,  p-value: 9.583e-05

> anova(treas.mod1)
Analysis of Variance Table
Response: situados.1770
            Df Sum Sq Mean Sq F value    Pr(>F)
minted.1770  1 13.283 13.2831  18.052 9.583e-05 ***
Residuals   49 36.055  0.7358

> (SSE.mod1 <- sum(e.mod1^2))
[1] 36.05453
> DW1.mod1
75.5092
> (DW.mod1 <- DW1.mod1 / SSE.mod1)
2.094306

> durbinWatsonTest(treas.mod1)
 lag Autocorrelation D-W Statistic p-value
   1     -0.05608488      2.094306   0.812
 Alternative hypothesis: rho != 0
```

$\nabla$

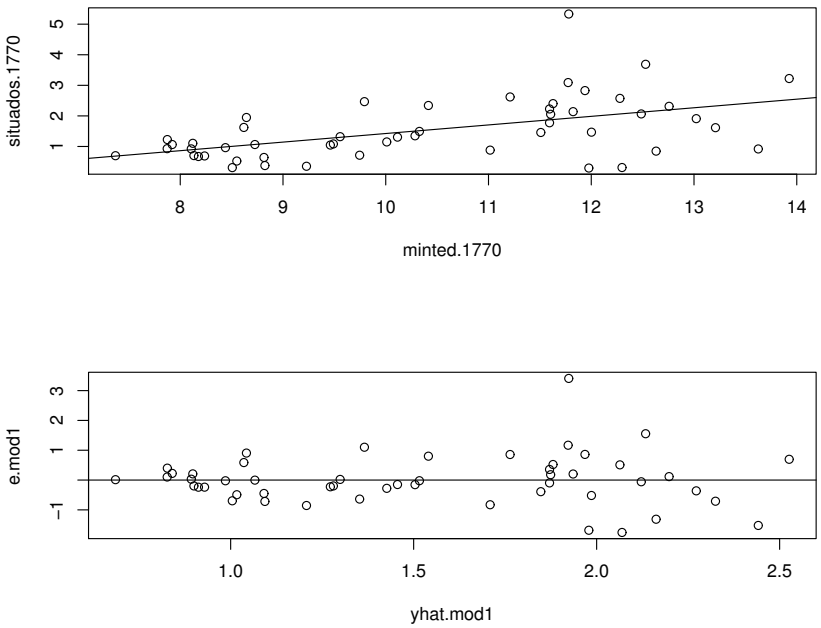


Figure 5.5: Data and Fitted Equation (Top Panel) and Residuals versus Fitted Values (Bottom Panel) - Spanish Treasure 1720-1770

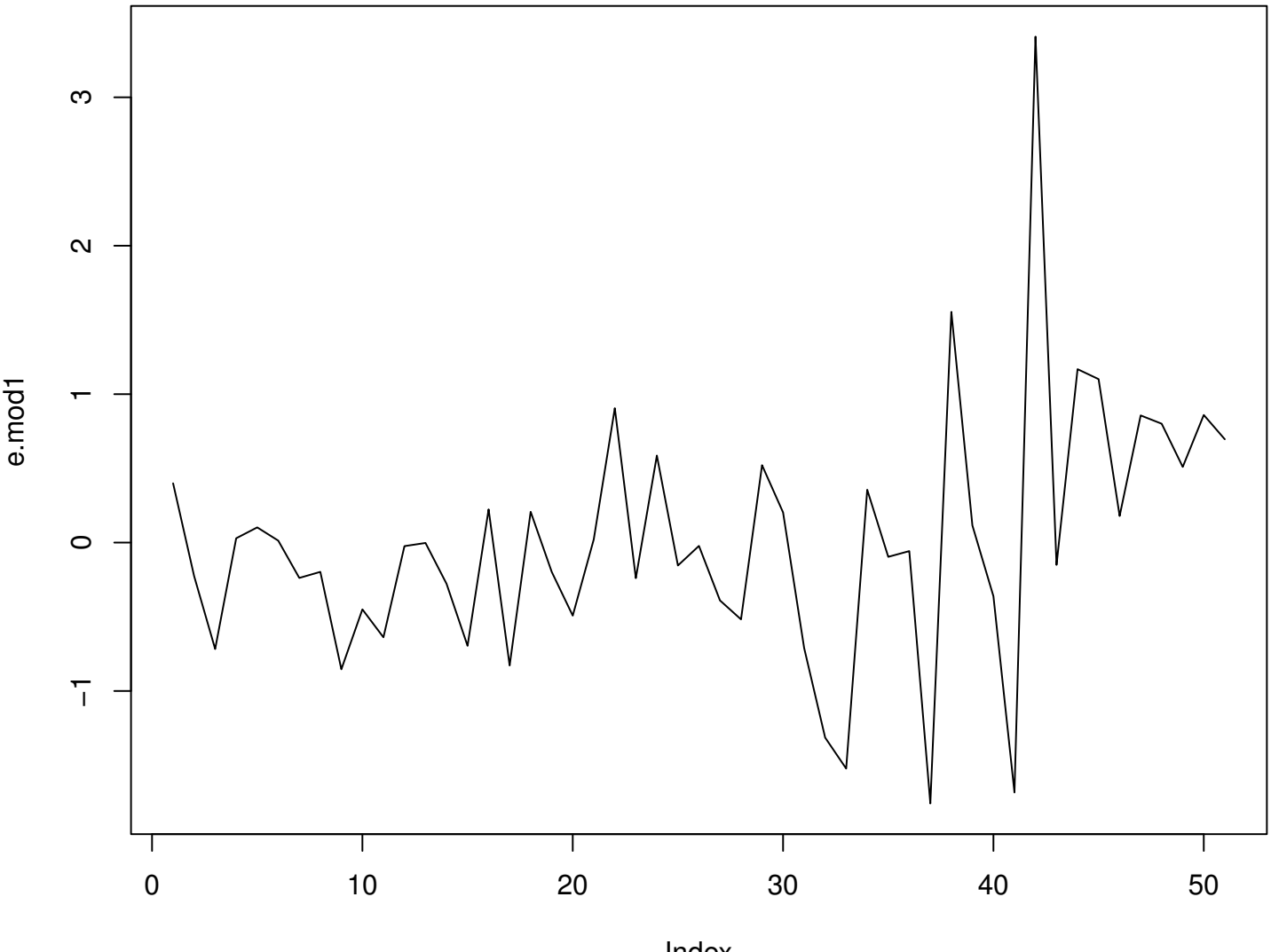


Figure 5.6: Residuals versus Time Order - Spanish Treasure 1720-1770

## 5.5 Detecting Outliers and Influential Observations

These measures are widely used in multiple regression, as well, when there are $p$ predictors, and $p' = p + 1$ parameters (including the intercept, $\beta_0$). Many of the "rules of thumb" are based on $p'$, which is 1+1=2 for simple regression. Most of these methods involve matrix algebra, but can be obtained from statistical software packages. Note that we will use $v_{ij}$ to denote the $i^{th}$ row $j^{th}$ column element of $\mathbf{P} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$. In particular, $v_{ii}$ is the $i^{th}$ diagonal element of $\mathbf{P}$.

Also, several of these methods make use of the estimated variance when the $i^{th}$ case was removed (to remove its effect if it is an outlier).

$$MSE_{(i)} = \frac{SSE_{(i)}}{n-p'-1} = \frac{SSE - \frac{e_i^2}{1-v_{ii}}}{n-p'-1} \quad \text{for simple regression } p' = 2$$

**Standardized Residuals** - Residuals divided by their estimated standard errors. They make use of the usual estimate of $\sigma$: $\sqrt{MSE}$. These are like $t$-statistics. Note that if an observation is an outlier, it will tend to inflate $MSE$, thus decreasing the Standardized Residual.

$$V\{e_i\} = \sigma^2(1-v_{ii}) \qquad \hat{SE}\{e_i\} = \sqrt{MSE(1-v_{ii})}$$

$$r_i = \frac{e_i}{\hat{SE}\{e_i\}} = \frac{e_i}{\sqrt{MSE(1-v_{ii})}}$$

**Studentized Residuals** - Residuals divided by their estimated standard error, with their contribution to $SSE$ having been removed (see above). Since residuals have mean 0, the studentized residuals are like $t$-statistics. Since we are simultaneously checking whether $n$ of these are outliers, we conclude any cases are outliers if the absolute value of their studentized residuals exceed $t_{\alpha/2n,n-p'-1}$, where $p'$ is the number of independent variables plus one (for simple regression, $p'$=2).

$$r_i^* = \frac{e_i}{\sqrt{MSE_{(i)}(1-v_{ii})}}$$

**Studentized Deleted Residuals** - Residuals that are obtained by subtracting off the fitted value based on the regression fit on the remaining $n-1$ observations and divided by their corresponding standard errors. For linear regression models, the individual regressions do not need to be re-fit. Outliers are detected in the same manner as Studentized Residuals. These are returned as **rstudent** in R.

$$r_i^{**} = e_i\left[\frac{n-p'-1}{SSE(1-v_{ii}) - e_i^2}\right]^{1/2}$$

**Leverage Values (Hat Values)** - These measure each case's potential to influence the regression due to its $X$ levels. Cases with high leverage values (often denoted $v_{ii}$ or $h_{ii}$) have $X$ levels "away" from the

center of the distribution. The leverage values sum to $p'$ (2 for simple regression), and cases with leverage values greater than $2p'/n$ (twice the average) are considered to be potentially influential due to their $X$-levels.

$$\hat{\mathbf{Y}} = \mathbf{PY} \quad \Rightarrow \quad \hat{Y}_i = \sum_{j=1}^{n-1} v_{ij}Y_j + v_{ii}Y_i + \sum_{j=i+1}^{n} v_{ij}Y_j \text{ subject to } \sum_{j=1}^{n} v_{ij} = 1$$

**DFFITS** - These measure how much an individual case's fitted value shifts when it is included in the regression fit $\left(\hat{Y}_i\right)$, and when it is excluded$\left(\hat{Y}_{i(i)}\right)$. The shift is divided by its standard error, so we are measuring how many standard errors a fitted value shifts, due to its being included in the regression model. Cases with the DFFITS values greater than $2\sqrt{p'/n}$ in absolute value are considered influential on their own fitted values.

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)}v_{ii}}} = e_i\left[\frac{n-p'-1}{SSE\,(1-v_{ii}) - e_i^2}\right]^{1/2}\left(\frac{v_{ii}}{1-v_{ii}}\right)^{1/2} = r_i^{**}\left(\frac{v_{ii}}{1-v_{ii}}\right)^{1/2}$$

**DFBETAS** - One of these is computed for each case, for each regression coefficient (including the intercept). DFBETAS measures how much the estimated regression coefficient shifts when that case is included $\left(\hat{\beta}_j\right)$ and excluded $\left(\hat{\beta}_{j(i)}\right)$ from the model, in units of standard errors. Cases with DFBETAS values larger than $2/\sqrt{n}$ in absolute value are considered to be influential on the estimated regression coefficient.

$$DFBETAS_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{MSE_{(i)}c_{j+1,j+1}}} \text{ where } c_{j+1,j+1} \text{ is the } (j+1)^{th} \text{ diagonal element of } \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

**Cook's D** - A single measure that represents each case's aggregate influence on all regression coefficients, and all cases' fitted values. Cases with Cook's D larger than $F_{.50,p',n-p'}$ are considered influential, although various textbooks and software packages have varying criteria.

$$D_i = \frac{\left(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)}\right)'\left(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)}\right)}{p'MSE} = \frac{\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}\right)'\left(\mathbf{X}'\mathbf{X}\right)\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}\right)}{p'MSE} = \frac{r_i^2}{p'}\left[\frac{v_{ii}}{1-v_{ii}}\right]$$

**COVRATIO** - This measures each case's influence on the estimated standard errors of the regression coefficients (inflating or deflating them). It represents the ratio of the determinants of the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ without and with the $i^{th}$ observation. Cases with COVRATIO outside of $1 \pm 3p'/n$ are considered influential. The matrix $\mathbf{X}_{(\mathbf{i})}$ is the $\mathbf{X}$ matrix with the $i^{th}$ row removed.

$$\text{COVRATIO}_i = \frac{\left|MSE_{(i)}\left(\mathbf{X}'_{(\mathbf{i})}\mathbf{X}_{(\mathbf{i})}\right)^{-1}\right|}{\left|MSE\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right|}$$

Note that R uses different criteria for "flagging" cases. For instance, for it uses the following rules (others are the same).

Leverage/Hat Values: $v_{ii} > \frac{3p'}{n}$ DFFITS: $|DFFITS_i| > 3\sqrt{\frac{p'}{n}}$ DFBETAS: $|DFBETAS_{j(i)}| > 1$

**Example: Total Phenolic Content and DPPH Radical Scavenging Activity in Lager Beers**

Zhao, Li, Sun, Yang, and Zhao (2013) report the results of a study relating antioxidant activity to phenolic content in $n = 40$ lager beers. The response is DPPH Radical Scavenging Activity ($Y$), and the predictor is Total Phenolic Content ($X$). The data are given in Table 5.4 and plotted in Figure 5.7 along with the OLS simple linear regression line. The "cut-off values" for the various diagnostic measures are computed below, with $p'$=1+1=2. The R program and output are given below the "cut-off values" computations.

Studentized Deleted Residuals: $|r_i^*| > t_{\alpha/(2n),n-p'-1} = t_{.05/(2(40)),40-2-1} = t_{.000625,37} = 3.495$

Leverage/Hat Values: $v_{ii} > \frac{2p'}{n} = \frac{2(2)}{40} = 0.10$

DFFITS: $|DFFITS_i| > 2\sqrt{\frac{p'}{n}} = 2\sqrt{\frac{2}{40}} = 0.4472$

DFBETAS: $|DFBETAS_{j(i)}| > \frac{2}{\sqrt{n}} = \frac{2}{\sqrt{40}} = 0.3162$

Cook's D: $D_i > F_{.50,p',n-p'} = F_{.50,2,40-2} = F_{.50,2,38} = 0.7059$

COVRATIO: $\text{COVRATIO}_i < 1 - \frac{3p'}{n} = 1 - \frac{3(2)}{40} = 0.85$ or $> 1 - \frac{3p'}{n} = 1 + \frac{3(2)}{40} = 1.15$

**R Program**

```
lager <- read.csv("http://www.stat.ufl.edu/~winner/data/lager_antioxidant_reg.csv",
      header=T)
attach(lager); names(lager)

lager.mod1 <- lm(dsa ~ tpc)
summary(lager.mod1)
anova(lager.mod1)
rstudent(lager.mod1)
influence.measures(lager.mod1)
e.mod1 <- resid(lager.mod1)
yhat.mod1 <- predict(lager.mod1)

par(mfrow=c(2,1))
plot(tpc,dsa)
abline(lager.mod1)
plot(yhat.mod1,e.mod1)
abline(h=0)
par(mfrow=c(1,1))
```

**R Output**

```
> summary(lager.mod1)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0343018  0.0639781   0.536    0.595
tpc         0.0034132  0.0003694   9.240 2.93e-11 ***

Residual standard error: 0.09629 on 38 degrees of freedom
Multiple R-squared:  0.692,     Adjusted R-squared:  0.6839
F-statistic: 85.38 on 1 and 38 DF,  p-value: 2.926e-11

> anova(lager.mod1)
Analysis of Variance Table
Response: dsa
          Df  Sum Sq Mean Sq F value    Pr(>F)
tpc        1 0.79175 0.79175  85.385 2.926e-11 ***
Residuals 38 0.35236 0.00927

> rstudent(lager.mod1)
          1           2           3           4           5           6
 1.27362241  0.50306312  0.04206598  1.67326277  1.13669877 -0.30507059
          7           8           9          10          11          12
-0.32569557  0.31113199  0.13536449  1.76837157 -0.86879480 -1.00698654
         13          14          15          16          17          18
-1.10301595 -0.33802210  0.12742060  0.20092759 -0.39670755 -1.31488291
         19          20          21          22          23          24
-0.48375729 -0.07785662 -0.20536135 -0.32680824 -0.87171742 -0.08034019
         25          26          27          28          29          30
-1.41906053 -1.20990613  0.18547066 -0.89654216 -0.92331109  1.35037405
         31          32          33          34          35          36
-0.11708938  0.98369936 -0.19073790  4.93877695 -0.32025202  0.20370294
         37          38          39          40
-0.11876585  0.42727476 -0.94183250 -0.41188552
> influence.measures(lager.mod1)
Influence measures of
         lm(formula = dsa ~ tpc) :
     dfb.1_   dfb.tpc    dffit cov.r   cook.d    hat inf
1   0.14508 -0.099246  0.22737 0.999 2.54e-02 0.0309
2   0.03408 -0.015345  0.08204 1.068 3.43e-03 0.0259
3   0.00126  0.000357  0.00675 1.082 2.34e-05 0.0251
4  -0.19384  0.266053  0.37999 0.959 6.89e-02 0.0490
5   0.13904 -0.098390  0.20751 1.018 2.14e-02 0.0323
6   0.00215 -0.014192 -0.05092 1.079 1.33e-03 0.0271
7  -0.01139 -0.001052 -0.05216 1.076 1.39e-03 0.0250
8   0.07091 -0.060569  0.07902 1.117 3.20e-03 0.0606
9   0.00988 -0.004861  0.02223 1.082 2.54e-04 0.0263
10  0.16902 -0.104522  0.30230 0.923 4.33e-02 0.0284
11 -0.02891 -0.004326 -0.13919 1.039 9.75e-03 0.0250
12 -0.26217  0.229408 -0.28280 1.078 4.00e-02 0.0731
13 -0.14562  0.106452 -0.20693 1.023 2.13e-02 0.0340
14  0.01104 -0.024663 -0.05961 1.081 1.82e-03 0.0302
15  0.03399 -0.029858  0.03648 1.140 6.83e-04 0.0757
16 -0.02430  0.033003  0.04639 1.109 1.10e-03 0.0506
17 -0.14147  0.129280 -0.14552 1.187 1.08e-02 0.1186   *
18  0.26542 -0.326426 -0.39194 1.048 7.54e-02 0.0816
19 -0.04639  0.028744 -0.08275 1.072 3.49e-03 0.0284
```

| Lager ID | X | Y | Lager ID | X | Y | Lager ID | X | Y | Lager ID | X | Y |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 148.23 | 0.66 | 11 | 169.51 | 0.53 | 21 | 159.81 | 0.56 | 31 | 177.83 | 0.63 |
| 2 | 160.38 | 0.63 | 12 | 111.05 | 0.32 | 22 | 163.23 | 0.56 | 32 | 150.11 | 0.64 |
| 3 | 170.41 | 0.62 | 13 | 143.50 | 0.42 | 23 | 169.59 | 0.53 | 33 | 135.92 | 0.48 |
| 4 | 208.65 | 0.90 | 14 | 186.96 | 0.64 | 24 | 135.76 | 0.49 | 34 | 162.99 | 0.96 |
| 5 | 146.03 | 0.64 | 15 | 109.50 | 0.42 | 25 | 198.62 | 0.58 | 35 | 183.54 | 0.63 |
| 6 | 180.19 | 0.62 | 16 | 209.95 | 0.77 | 26 | 221.94 | 0.68 | 36 | 236.37 | 0.86 |
| 7 | 169.06 | 0.58 | 17 | 88.47 | 0.30 | 27 | 148.80 | 0.56 | 37 | 163.23 | 0.58 |
| 8 | 119.04 | 0.47 | 18 | 230.25 | 0.70 | 28 | 120.02 | 0.36 | 38 | 212.48 | 0.80 |
| 9 | 158.99 | 0.59 | 19 | 152.96 | 0.51 | 29 | 84.64 | 0.24 | 39 | 235.06 | 0.75 |
| 10 | 153.04 | 0.72 | 20 | 147.42 | 0.53 | 30 | 238.33 | 0.97 | 40 | 267.27 | 0.91 |

Table 5.4: Total Phenolic Content ($X$) and DPPH Radical Scavenging Activity ($Y$) for 40 lager beers

```
20 -0.00911  0.006314 -0.01401 1.089 1.01e-04 0.0314
21 -0.01436  0.006720 -0.03358 1.081 5.78e-04 0.0260
22 -0.01862  0.006347 -0.05272 1.076 1.42e-03 0.0254
23 -0.02874 -0.004612 -0.13966 1.039 9.82e-03 0.0250
24 -0.01301  0.010215 -0.01651 1.099 1.40e-04 0.0405
25  0.10942 -0.168722 -0.28431 0.987 3.94e-02 0.0386
26  0.20358 -0.258140 -0.32539 1.047 5.23e-02 0.0675
27  0.02072 -0.014038  0.03293 1.086 5.56e-04 0.0306
28 -0.20080  0.170928 -0.22489 1.074 2.54e-02 0.0592
29 -0.34509  0.316998 -0.35345 1.155 6.27e-02 0.1278
30 -0.31773  0.382193  0.44337 1.061 9.62e-02 0.0973
31 -0.00022 -0.004371 -0.01926 1.083 1.91e-04 0.0264
32  0.10500 -0.069410  0.17249 1.033 1.49e-02 0.0298
33 -0.03076  0.024130 -0.03912 1.097 7.85e-04 0.0404
34  0.28588 -0.100522  0.79736 0.393 1.97e-01 0.0254   *
35  0.00631 -0.019083 -0.05480 1.080 1.54e-03 0.0284
36 -0.04626  0.055918  0.06535 1.161 2.19e-03 0.0933   *
37 -0.00677  0.002307 -0.01916 1.081 1.88e-04 0.0254
38 -0.05589  0.074562  0.10190 1.104 5.31e-03 0.0538
39  0.20877 -0.253208 -0.29749 1.106 4.44e-02 0.0907
40  0.14975 -0.171693 -0.18597 1.258 1.77e-02 0.1693   *
```

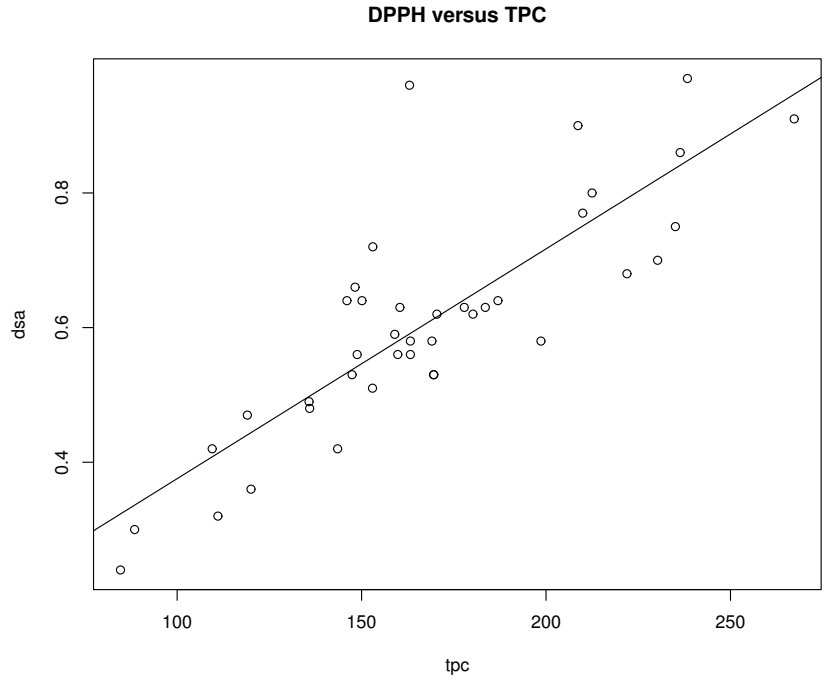


Figure 5.7: Plot of DPPH Radical Scavenging Activity versus Total Phenolic Content for 40 lager beers

# Chapter 6

# Multiple Linear Regression

When there are more than one predictor variables, the model generalizes to multiple linear regression. The calculations become more complex, but conceptually, the ideas remain the same. We will use the notation of $p$ as the number of predictors, and $p' = p + 1$ as the number of parameters in the model (including the intercept). The model can be written as follows.

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \qquad \epsilon \sim N(0, \sigma^2) \text{ independent}$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \sigma^2\mathbf{I}\right) \qquad \mathbf{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

We then obtain least squares (and maximum likelihood) estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ that minimize the error sum of squares.

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} \qquad \hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{P}\mathbf{Y} \qquad \boldsymbol{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \left(\mathbf{I} - \mathbf{P}\right)\mathbf{Y}$$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots \hat{\beta}_p X_{ip} \qquad e_i = Y_i - \hat{Y}_i \qquad SSE = \sum_{i=1}^{n} e_i^2 = \boldsymbol{e}'\boldsymbol{e} = \mathbf{Y}'\left(\mathbf{I} - \mathbf{P}\right)\mathbf{Y}$$

The degrees of freedom for error are now $n - p' = n - (p + 1)$, as we have now estimated $p' = p + 1$ parameters. This results from the following fact.

$$\text{trace}\left(\mathbf{I} - \mathbf{P}\right) = \text{trace}\left(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right) = n - \text{trace}\left(\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right) = n - p'$$

In the multiple linear regression model, $\beta_j$ represents the change in $E\{Y\}$ when $X_j$ increases by 1 unit, with all other predictor variables being held constant. It is thus often referred to as the **partial regression coefficient**.

## 6.1 Testing and Estimation for Partial Regression Coefficients

Once we fit the model, obtaining the estimated regression coefficients, we also obtain standard errors for each coefficient (actually, we obtain an estimated variance-covariance matrix for the vector of coefficients).

$$V\left\{\hat{\boldsymbol{\beta}}\right\} = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1} \qquad \hat{V}\left\{\hat{\boldsymbol{\beta}}\right\} = MSE\left(\mathbf{X}'\mathbf{X}\right)^{-1} \qquad MSE = \frac{SSE}{n-p'}$$

If we wish to test whether $Y$ is associated with $X_j$, after controlling for the remaining $p-1$ predictors, we are testing whether $\beta_j = 0$. This is equivalent to the $t$-test from simple regression (in general, we can test whether a regression coefficient is any specific number, although software packages are testing whether it is 0).

$$H_0 : \beta_j = \beta_{j0} \qquad H_A : \beta_j \neq \beta_{j0} \quad TS : t_{obs} = \frac{\hat{\beta}_j - \beta_{j0}}{\hat{SE}\{\hat{\beta}_j\}} \qquad RR : |t_{obs}| \geq t_{\alpha/2,n-p'} \qquad P\text{-value} : 2P(t_{n-p'} \geq |t_{obs}|)$$

One-sided tests make the same adjustments as in simple linear regression.

$$H_A^+ : \beta_j > \beta_{j0} \qquad RR : t_{obs} \geq t_{\alpha,n-p'} \qquad P\text{-value} : P(t_{n-p'} \geq t_{obs})$$

$$H_A^- : \beta_j < \beta_{j0} \qquad RR : t_{obs} \leq -t_{\alpha,n-p'} \qquad P\text{-value} : P(t_{n-p'} \leq t_{obs})$$

A $(1-\alpha)100\%$ Confidence Interval for $\beta_j$ is obtained as:

$$\hat{\beta}_j \pm t_{\alpha/2,n-p'}\hat{SE}\{\hat{\beta}_j\}$$

Note that the Confidence Interval represents the values of $\beta_{j0}$ that the two-sided test: $H_0 : \beta_j = \beta_{j0} \quad H_A : \beta_j \neq \beta_{j0}$ fails to reject the null hypothesis.

## 6.2 Analysis of Variance

When there is no association between $Y$ and $X_1, \ldots, X_p$ ($\beta_1 = \cdots = \beta_p = 0$), the best predictor of each observation is $\overline{Y} = \hat{\beta}_0$ (in terms of minimizing the sum of squares of prediction errors). In this case, the total variation can be denoted as $TSS = \sum_{i=1}^n (Y_i - \overline{Y})^2$, the **Total Sum of Squares**, just as with simple regression.

When there is an association between $Y$ and at least one of $X_1, \ldots, X_p$ (not all $\beta_i = 0$), the best predictor of each observation is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}$ (in terms of minimizing the sum of squares of prediction errors). In this case, the error variation can be denoted as $SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$, the **Error Sum of Squares**.

The difference between $TSS$ and $SSE$ is the variation "explained" by the regression of $Y$ on $X_1, \ldots, X_p$ (as opposed to having ignored $X_1, \ldots, X_p$). It represents the difference between the fitted values and the mean: $SSR = \sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2$ the **Regression Sum of Squares**.

$$TSS = SSE + SSR \qquad \sum_{i=1}^n (Y_i - \overline{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2$$

$$\mathbf{Y}'\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y} \quad = \quad \mathbf{Y}'(\mathbf{I} - \mathbf{P})\mathbf{Y} \quad + \quad \mathbf{Y}'\left(\mathbf{P} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y}$$

Each sum of squares has a **degrees of freedom** associated with it. The **Total Degrees of Freedom** is $df_{\text{Total}} = n - 1$. The **Error Degrees of Freedom** is $df_{\text{Error}} = n - p'$. The **Regression Degrees of Freedom** is $df_{\text{Regression}} = p$. Note that when we have $p = 1$ predictor, this generalizes to simple regression.

$$df_{\text{Total}} = \text{trace}\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right) = n - 1 \qquad \text{trace}(\mathbf{P}) = \text{trace}(\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \text{trace}(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{I}_{p'} = p'$$

$$\Rightarrow \qquad df_{\text{Error}} = n - p' \qquad df_{\text{Regression}} = p' - 1 = p$$

$$df_{\text{Total}} = df_{\text{Error}} + df_{\text{Regression}} \qquad n - 1 = n - p' + p$$

The Error and Regression sums of squares have **Mean Squares**, which are the sums of squares divided by its corresponding degrees of freedom: $MSE = SSE/(n - p')$ and $MSR = SSR/p$. It can be shown (as was done for simple regression) that these mean squares have the following **Expected Values**, average values in repeated sampling at the same observed $X$ levels.

$$E\{MSE\} = \sigma^2 \qquad E\{MSR\} = \sigma^2 + \boldsymbol{\beta}'\mathbf{X}'\left(\mathbf{I} - \left(\frac{1}{n}\right)\mathbf{J}\right)\mathbf{X}\boldsymbol{\beta}$$

where $\boldsymbol{\beta}$ and $\mathbf{X}$ are vector/matrix extensions of the simple linear regression model (see below). Note that when $\beta_1 = \cdots \beta_p = 0$, then $E\{MSR\} = E\{MSE\}$, otherwise $E\{MSR\} > E\{MSE\}$. A way of testing whether $\beta_1 = \cdots \beta_p = 0$ is by the $F$-test:

$$H_0 : \beta_1 = \cdots \beta_p = 0 \qquad H_A : \text{ Not all } \beta_j = 0$$

$$TS : F_{obs} = \frac{MSR}{MSE} \qquad RR : F_{obs} \geq F_{\alpha,p,n-p'} \qquad P\text{-value} : P(F_{p,n-p'} \geq F_{obs})$$

The Analysis of Variance is typically set up as in Table 6.1.

A measure often reported from a regression analysis is the **Coefficient of Determination** or $R^2$. This represents the variation in $Y$ "explained" by $X_1, \ldots, X_p$, divided by the total variation in $Y$.

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2} = \frac{SSR}{TSS} = 1 - \frac{SSE}{TSS} \qquad 0 \leq R^2 \leq 1$$

| Source | $df$ | $SS$ | $MS$ | $F_{obs}$ | $P$-value |
|---|---|---|---|---|---|
| Regression (Model) | $p$ | $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$ | $MSR = \frac{SSR}{p}$ | $F_{obs} = \frac{MSR}{MSE}$ | $P(F_{p,n-p'} \geq F_{obs})$ |
| Error (Residual) | $n-p'$ | $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ | $MSE = \frac{SSE}{n-p'}$ | | |
| Total (Corrected) | $n-1$ | $TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$ | | | |

Table 6.1: Analysis of Variance Table for Multiple Linear Regression

The interpretation of $R^2$ is the proportion of variation in $Y$ that is "explained" by $X_1, \ldots, X_p$, and is often reported as a percentage ($100R^2$).

## 6.3 Testing a Subset of $\beta^s = 0$

The $F$-test from the Analysis of Variance and the $t$-tests represent extremes as far as model testing (all variables simultaneously versus one-at-a-time). Often we wish to test whether a group of predictors do not improve prediction, after controlling for the remaining predictors.

Suppose that after controlling for $g$ predictors, we wish to test whether the remaining $p-g$ predictors are associated with $Y$. That is, we wish to test:

$$H_0 : \beta_{g+1} = \cdots \beta_p = 0 \qquad H_A : \text{ Not all of } \beta_{g+1}, \ldots, \beta_p = 0$$

Note that, the $t$-tests control for all other predictors, while here, we want to control for only $X_1, \ldots, X_g$. To do this, we fit two models: the **Complete** or **Full Model** with all $p$ predictors, and the **Reduced Model** with only the $g$ "control" variables. For each model, we obtain the Regression and Error sums of squares, as well as $R^2$. This leads to the following test statistic and rejection region.

$$TS : F_{obs} = \frac{\left[\frac{SSE(R)-SSE(F)}{(n-g')-(n-p')}\right]}{\left[\frac{SSE(F)}{n-p'}\right]} = \frac{\left[\frac{SSR(F)-SSR(R)}{p-g}\right]}{\left[\frac{SSE(F)}{n-p'}\right]} = \frac{\left[\frac{R_F^2-R_R^2}{p-g}\right]}{\left[\frac{1-R_F^2}{n-p'}\right]}$$

$$RR : F_{obs} \geq F_{\alpha,p-g,n-p'} \qquad P\text{-value} : P(F_{p-g,n-p'} \geq F_{obs})$$

## 6.4 Tests Based on the Matrix Form of Multiple Regression Model

The matrix form is virtually identical (at least symbolically) for multiple regression as simple regression. The primary difference is the dimension of the various matrices and vectors. Now, $\mathbf{X}$ still has $n$ rows, but it has $p+1$ columns (one for the intercept, and one each for the $p$ predictors). The vectors $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$ each have $p' = p+1$ rows.

We still have that the estimated variance of $\hat{\boldsymbol{\beta}} = s^2(\mathbf{X}'\mathbf{X})^{-1}$ which is how the estimated standard errors for the partial regression coefficients used in $t$-tests and confidence intervals are obtained, in the case of the model with normal, independent errors with constant variance.

The **general linear test** can be used to test any set of up to $p+1$ linear hypotheses among the $\beta^s$, that are linearly independent. The tests described above are special cases. Here we wish to test:

$$H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m} \quad \Rightarrow \quad \mathbf{K}'\boldsymbol{\beta} - \mathbf{m} = \mathbf{0}$$

where $\mathbf{K}'$ is a $q \times (p+1)$ matrix of constants defining the the hypotheses among the $\boldsymbol{\beta}$ elements and $\mathbf{m}$ is the $q \times 1$ vector of hypothesized values for the $q$ linear functions. Some special cases are given below, assuming $p = 3$ (three predictor variables):

$$H_{01} : \beta_1 = \beta_2 = \beta_3 = 0 \qquad \mathbf{K}_1' = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad \mathbf{m_1} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$H_{02} : \beta_1 = \beta_2 = \beta_3 \qquad \mathbf{K}_2' = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix} \qquad \mathbf{m_2} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$H_{03} : \beta_0 = 100, \beta_1 = 10, \beta_2 = \beta_3 \qquad \mathbf{K}_3' = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \qquad \mathbf{m_3} = \begin{bmatrix} 100 \\ 10 \\ 0 \end{bmatrix}$$

The estimator $\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m}$ has an estimated variance-covariance matrix of $s^2\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}$ which is $q \times q$. Then, we can form the $F$-statistic (based on assuming normal and independent errors with constant variance).

$$F_{obs} = \frac{Q}{qMSE} = \frac{\left(\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m}\right)' \left[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\right]^{-1} \left(\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m}\right)}{qs^2}$$

which under the null hypothesis is distributed $F_{q,n-p'}$. This holds from the following results.

$$\boldsymbol{\Sigma}_{\hat{\beta}} = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1} \qquad \Rightarrow \quad \boldsymbol{\Sigma}_{K'\hat{\beta}-m} = \sigma^2 \mathbf{K}' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{K}$$

$$Q = \left(\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m}\right)' \left[\mathbf{K}' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{K}\right]^{-1} \left(\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m}\right) \qquad \mathbf{A}_{K'\hat{\beta}-m} = \frac{1}{\sigma^2} \left[\mathbf{K}' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{K}\right]^{-1}$$

$$\mathbf{A}_{K'\hat{\beta}-m} \boldsymbol{\Sigma}_{K'\hat{\beta}-m} = \frac{1}{\sigma^2} \left[\mathbf{K}' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{K}\right]^{-1} \sigma^2 \mathbf{K}' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{K} = \mathbf{I}$$

$$\Rightarrow \quad \mathbf{A}_{K'\hat{\beta}-m} \boldsymbol{\Sigma}_{K'\hat{\beta}-m} \mathbf{A}_{K'\hat{\beta}-m} \boldsymbol{\Sigma}_{K'\hat{\beta}-m} = \mathbf{A}_{K'\hat{\beta}-m} \boldsymbol{\Sigma}_{K'\hat{\beta}-m} \qquad \text{rank}\left(\mathbf{A}_{K'\hat{\beta}-m}\right) = q$$

$$\Omega_Q = \frac{\left(\mathbf{K}'\boldsymbol{\beta} - \mathbf{m}\right)' \mathbf{A}_{K'\hat{\beta}-m} \left(\mathbf{K}'\boldsymbol{\beta} - \mathbf{m}\right)}{2} = \frac{\left(\mathbf{K}'\boldsymbol{\beta} - \mathbf{m}\right)' \left[\mathbf{K}' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{K}\right]^{-1} \left(\mathbf{K}'\boldsymbol{\beta} - \mathbf{m}\right)}{2\sigma^2}$$

That is, $Q/sigma^2$ is distributed $\chi^2_{q,\Omega_Q}$. Now, to see that $Q$ and $SSE = \mathbf{Y}'(\mathbf{I}-\mathbf{P})\mathbf{Y}$ are independent, consider $Q$ in terms of $\mathbf{Y}$.

$$\mathbf{K}'\hat{\boldsymbol{\beta}} = \mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \qquad \left(\mathbf{K}'\hat{\boldsymbol{\beta}}\right)' = \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}$$

$$\Rightarrow \quad Q = \mathbf{Y}'\mathbf{A_1}\mathbf{Y} - 2\mathbf{A_2}\mathbf{Y} + \mathbf{A_3} \text{ where:}$$

$$\mathbf{Y}'\mathbf{A_1}\mathbf{Y} = \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\left[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\right]^{-1}\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

$$\mathbf{A_2}\mathbf{Y} = \mathbf{m}'\left[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\right]^{-1}\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \qquad \mathbf{A_3} = \mathbf{m}'\left[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\right]^{-1}\mathbf{m}$$

Clearly $SSE$ and $\mathbf{A_3}$ are independent since $\mathbf{A_3}$ is a scalar constant. Recall that $\mathbf{\Sigma_Y} = \sigma^2\mathbf{I}$. Then we just need to show that $\mathbf{A_1}(\mathbf{I}-\mathbf{P}) = \mathbf{A_2}(\mathbf{I}-\mathbf{P}) = 0$.

$$\mathbf{A_1}(\mathbf{I}-\mathbf{P}) = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\left[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\right]^{-1}\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I}-\mathbf{P}) = 0 \qquad \text{since} \quad \mathbf{X}'\mathbf{P} = \mathbf{X}'$$

$$\mathbf{A_2}(\mathbf{I}-\mathbf{P}) = \mathbf{m}'\left[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\right]^{-1}\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I}-\mathbf{P}) = 0$$

Thus, $Q/\sigma^2$ and $SSE/\sigma^2$ are both distributed chi-square and are independent.

$$\Rightarrow \quad F_{obs} = \frac{Q/q}{MSE} \sim F_{q,n-p',\Omega_Q} \qquad \Omega_Q = \frac{(\mathbf{K}'\boldsymbol{\beta}-\mathbf{m})'\left[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\right]^{-1}(\mathbf{K}'\boldsymbol{\beta}-\mathbf{m})}{2\sigma^2}$$

Under the null hypothesis $H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$, the noncentrality parameter $\Omega_Q = 0$, and the $F$-statistic is distributed $F_{q,n-p'}$. We reject the null hypothesis if $F_{obs} \geq F_{\alpha,q,n-p'}$.

Note that even if the data are not normally distributed, the quantity $qF_{obs}$ is asymptotically distributed as $\chi^2_q$, so the test can be conducted in this manner in large samples. Note that in this large-sample case, the tests are identical as $F_{\alpha,q,\infty} = \frac{\chi^2_{\alpha,q}}{q}$.

### 6.4.1 Equivalence of Complete/Reduced Model and Matrix Based Test

When the restriction $\mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$ is applied, Lagrange Multipliers can be used to obtain the constrained least squares estimator. A brief description of the process is given here to demonstrate that $Q = SSE(R) -$

$SSE(F)$, and the equivalence of the two forms of the $F$-test. Here $\mathbf{K}'$ is $q \times p'$, with linearly independent rows ($q \leq p'$)), and $\lambda$ is $q \times 1$.

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \text{subject to} \qquad \mathbf{K}'\boldsymbol{\beta} = \mathbf{m} \quad \Rightarrow \quad \mathbf{K}'\boldsymbol{\beta} - \mathbf{m} = \mathbf{0}$$

Minimize with respect to $\boldsymbol{\beta}^*$ and $\lambda$: $Q^* = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^*)'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^*) + 2\lambda'(\mathbf{K}'\boldsymbol{\beta}^* - \mathbf{m}) =$

$$\mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\beta}^{*'}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}^* + 2\lambda'\mathbf{K}'\boldsymbol{\beta}^* - 2\lambda'\mathbf{m}$$

$$\frac{\partial Q^*}{\partial \boldsymbol{\beta}^*} = -2\mathbf{X}'\mathbf{Y} + \mathbf{X}'\mathbf{X}\boldsymbol{\beta}^* + \mathbf{K}\lambda \quad \Rightarrow \quad \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}^* + \mathbf{K}\lambda = \mathbf{X}'\mathbf{Y}$$

$$\frac{\partial Q^*}{\partial \lambda} = 2(\mathbf{K}'\boldsymbol{\beta}^* - \mathbf{m}) \quad \Rightarrow \quad \mathbf{K}'\hat{\boldsymbol{\beta}}^* = \mathbf{m}$$

$$\Rightarrow \quad \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{K} \\ \mathbf{K}' & \mathbf{0} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}}^* \\ \lambda \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{Y} \\ \mathbf{m} \end{bmatrix}$$

Now, make use of the following matrix result for partitioned, square full rank matrices. To confirm that its an inverse, show that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$.

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \quad \Rightarrow \quad \mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}_{11}^{-1}\left(\mathbf{I} + \mathbf{A}_{12} + \mathbf{F}_2\mathbf{A}_{12}\mathbf{A}_{11}^{-1}\right) & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{F}_2 \\ -\mathbf{F}_2\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{F}_2 \end{bmatrix}$$

$$\text{where:} \quad \mathbf{F}_2 = \left(\mathbf{A}_{22}^{-1} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right)^{-1}$$

$$\Rightarrow \quad \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{K} \\ \mathbf{K}' & \mathbf{0} \end{bmatrix}^{-1} = \begin{bmatrix} (\mathbf{X}'\mathbf{X})^{-1}\left(\mathbf{I} - \mathbf{K}\left[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\right]^{-1}\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\right) & (\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\left[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\right]^{-1} \\ \left[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\right]^{-1}\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1} & -\left[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\right]^{-1} \end{bmatrix}$$

$$\text{Note that:} \quad \mathbf{F}_2 = \left(\mathbf{0} - \mathbf{K}'(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{K}\right)^{-1} = -\left[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\right]^{-1}$$

$$\Rightarrow \quad \begin{bmatrix} \hat{\boldsymbol{\beta}}^* \\ \lambda \end{bmatrix} =$$

$$\begin{bmatrix} (\mathbf{X}'\mathbf{X})^{-1}\left(\mathbf{I} - \mathbf{K}\left[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\right]^{-1}\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\right) & (\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\left[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\right]^{-1} \\ \left[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\right]^{-1}\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1} & -\left[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\right]^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{X}'\mathbf{Y} \\ \mathbf{m} \end{bmatrix}$$

$$\Rightarrow \quad \hat{\boldsymbol{\beta}}^* = (\mathbf{X}'\mathbf{X})^{-1}\left(\mathbf{I} - \mathbf{K}\left[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\right]^{-1}\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\right)\mathbf{X}'\mathbf{Y} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\left[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\right]^{-1}\mathbf{m} =$$

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\left[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\right]^{-1}\left(\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{Y} - \mathbf{m}\right) =$$

$$\hat{\boldsymbol{\beta}} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\left[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\right]^{-1}\left(\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{Y} - \mathbf{m}\right)$$

The Error Sum of Squares for the constrained regression model are given below, leading to the equivalence of $Q$ and $SSE(R) - SSE(F)$.

$$\mathbf{e}^* = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}^* = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\left(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}\right) = \mathbf{e} - \mathbf{X}\left(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}\right)$$

$$\text{where:} \quad \mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} \quad \text{and} \quad SSE(F) = \mathbf{e}'\mathbf{e}$$

$$\Rightarrow \quad SSE(R) = \mathbf{e}^{*'}\mathbf{e}^* = \left(\mathbf{e} - \mathbf{X}\left(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}\right)\right)'\left(\mathbf{e} - \mathbf{X}\left(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}\right)\right) =$$

$$\mathbf{e}'\mathbf{e} - 2\left(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}\right)\mathbf{X}'\mathbf{e} + \left(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}\right)'\mathbf{X}'\mathbf{X}\left(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}\right) =$$

$$\mathbf{e}'\mathbf{e} + \left(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}\right)'\mathbf{X}'\mathbf{X}\left(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}\right) \qquad \text{since} \quad \mathbf{X}'\mathbf{e} = \mathbf{0}$$

$$\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}} = -(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\left[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\right]^{-1}\left(\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{Y} - \mathbf{m}\right) \quad \Rightarrow \quad SSE(R) - SSE(F) =$$

$$\left(\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m}\right)'\left[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\right]^{-1}\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{K}\left[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\right]^{-1}\left(\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m}\right) =$$

$$\left(\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m}\right)'\left[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\right]^{-1}\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{K}\left[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\right]^{-1}\left(\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m}\right) =$$

$$\left(\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m}\right)'\left[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\right]^{-1}\left(\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m}\right) = Q$$

**Example: Cobb-Douglas Production Function**

Cobb and Douglas (1928) proposed a multiplicative production function, where the dependent variable is the output: Quantity Produced ($Y$), and the independent variables are the inputs: Capital ($X_1$) and Labor ($X_2$). The function is nonlinear, but can be linearized by logarithmic transformation.

$$Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2}\epsilon \quad E\{\epsilon\} = 1 \quad \Rightarrow \quad \ln(Y) = \ln\beta_0 + \beta_1 \ln X_1 + \beta_2 \ln X_2 + \ln\epsilon$$

$$\Rightarrow \quad Y^* = \beta_0^* + \beta_1 X_1^* + \beta_2 X_2^* + \epsilon^* \quad E\{\epsilon\} = 0$$

When we ignore the multiplicative error term in the original model, we obtain the following **elasticities** of Quantity produced with Capital and Labor.

$$\zeta_1 = \frac{\partial E\{Y\}/E\{Y\}}{\partial X_1/X_1} = \left(\frac{\partial E\{Y\}}{\partial X_1}\right)\left(\frac{X_1}{E\{Y\}}\right) = \beta_0\beta_1 X_1^{\beta_1 - 1}X_2^{\beta_2}\frac{X_1}{\beta_0 X_1^{\beta_1}X_2^{\beta_2}} = \beta_1$$

$$\zeta_2 = \frac{\partial E\{Y\}/E\{Y\}}{\partial X_2/X_2} = \left(\frac{\partial E\{Y\}}{\partial X_2}\right)\left(\frac{X_2}{E\{Y\}}\right) = \beta_0 X_1^{\beta_1}\beta_2 X_2^{\beta_2 - 1}\frac{X_2}{\beta_0 X_1^{\beta_1}X_2^{\beta_2}} = \beta_2$$

| Year | Quantity ($Y$) | Capital ($X_1$) | Labor ($X_2$) |
|---|---|---|---|
| 1899 | 100 | 100 | 100 |
| 1900 | 101 | 107 | 105 |
| 1901 | 112 | 114 | 110 |
| 1902 | 122 | 122 | 118 |
| 1903 | 124 | 131 | 123 |
| 1904 | 122 | 138 | 116 |
| 1905 | 143 | 149 | 125 |
| 1906 | 152 | 163 | 133 |
| 1907 | 151 | 176 | 138 |
| 1908 | 126 | 185 | 121 |
| 1909 | 155 | 198 | 140 |
| 1910 | 159 | 208 | 144 |
| 1911 | 153 | 216 | 145 |
| 1912 | 177 | 226 | 152 |
| 1913 | 184 | 236 | 154 |
| 1914 | 169 | 244 | 149 |
| 1915 | 189 | 266 | 154 |
| 1916 | 225 | 298 | 182 |
| 1917 | 227 | 335 | 196 |
| 1918 | 223 | 366 | 200 |
| 1919 | 218 | 387 | 193 |
| 1920 | 231 | 407 | 193 |
| 1921 | 179 | 417 | 147 |
| 1922 | 240 | 431 | 161 |

Table 6.2: U.S. Production Data: 1899-1922

In this economic model, the elasticity of scale is defined as the sum of the two elasticities, $\zeta = \zeta_1 + \zeta_2 = \beta_1 + \beta_2$. If the elasticity of scales is 1, which implies that a 1% increase in all inputs leads to a 1% increase in output. This is referred to as "Constant Returns to Scale."

In their seminal paper, Cobb and Douglas applied this model to the U.S. economy with annual numbers for years 1899-1922. The data were indexed, so that the 1899 values were set at 100. The data are given in Table 6.2. The hypothesis we wish to test is $H_0 : \zeta = \beta_1 + \beta_2 = 1$. This involves a single restriction, so that $q = 1$. In matrix form, based on the transformed (linear) model, we the following. Recall that the regression makes use of $Y^* = \ln Y$, $X_1^* = \ln X_1$, and $X_2^* = \ln X_2$.

$$H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0^* \\ \beta_1 \\ \beta_2 \end{bmatrix} \qquad \mathbf{K}' = \begin{bmatrix} 0 & 1 & 1 \end{bmatrix} \qquad \mathbf{m} = [1]$$

Ordinary Least Squares estimates, and the Error Sum of Squares for the unconstrained (Full) model are obtained as follow (rounded to 4 decimal places except when necessary, all computations were done in EXCEL with more decimal places).

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 24.0000 & 128.5556 & 119.1054 \\ 128.5556 & 693.4555 & 639.9174 \\ 119.1054 & 639.9174 & 592.0168 \end{bmatrix} \qquad \mathbf{X}'\mathbf{Y} = \begin{bmatrix} 121.8561 \\ 655.4095 \\ 605.9387 \end{bmatrix} \qquad \mathbf{Y}'\mathbf{Y} = 620.3713$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 55.8006 & 5.9123 & -17.6170 \\ 5.9123 & 1.1941 & -2.4802 \\ -17.6170 & -2.4802 & 6.2268 \end{bmatrix} \qquad \hat{\boldsymbol{\beta}} = \begin{bmatrix} -0.1773 \\ 0.2331 \\ 0.8073 \end{bmatrix}$$

$$\mathbf{Y}'\mathbf{P}\mathbf{Y} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} = -0.1773(121.8561) + 0.2331(655.4095) + 0.8073(605.9387) = 620.3003$$

$$SSE(F) = \mathbf{Y}'(\mathbf{I} - \mathbf{P})\mathbf{Y} = \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} = 620.3713 - 620.3003 = 0.0710$$

$$\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m} = (0 + 0.2331 + 0.8073) - 1 = 0.0403$$

$$\left[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\right]^{-1} = (1.1941 - 2.4802 - 2.4802 + 6.2268)^{-1} = \frac{1}{2.4605} = 0.4064$$

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K} = \begin{bmatrix} 5.9123 - 17.6170 \\ 1.1941 - 2.4802 \\ -2.4802 + 6.2268 \end{bmatrix} = \begin{bmatrix} -11.7047 \\ -1.2861 \\ 3.7467 \end{bmatrix}$$

$$\hat{\boldsymbol{\beta}}^* = \begin{bmatrix} -0.1773 \\ 0.2331 \\ 0.8073 \end{bmatrix} - \begin{bmatrix} -11.7047 \\ -1.2861 \\ 3.7467 \end{bmatrix}(0.4064)(0.0403) = \begin{bmatrix} -0.1773 \\ 0.2331 \\ 0.8073 \end{bmatrix} - \begin{bmatrix} -0.1919 \\ -0.0211 \\ 0.0614 \end{bmatrix} = \begin{bmatrix} 0.0145 \\ 0.2541 \\ 0.7459 \end{bmatrix}$$

$$\hat{\boldsymbol{\beta}}^*\mathbf{X}'\mathbf{Y} = 0.0145(121.8561) + 0.2541(655.4095) + 0.7459(605.9387) = 620.2833 \qquad \hat{\boldsymbol{\beta}}^{*'}\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}^* = 620.2669$$

$$SSE(R) = \left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}^*\right)'\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}^*\right) = 620.3713 - 2(620.2833) + 620.2669 = 0.0716$$

$$Q = SSE(R) - SSE(F) = 0.0716 - 0.0710 = 0.000661 \quad \text{showing more decimal places}$$

$$s^2 = MSE(F) = \frac{0.0710}{24 - 3} = 0.00338$$

$$F_{obs} = \frac{Q}{qs^2} = \frac{0.000661}{1(0.00338)} = 0.1956 \qquad F_{.05,1,21} = 4.325 \qquad P\left(F_{.05,1,21} \geq 0.1956\right) = .6628$$

This was a single hypothesis, and could also have been conducted as a $t$-test as follows.

$$\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m} = (0 + 0.2331 + 0.8073) - 1 = 0.0403 \qquad V\left\{\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m}\right\} = \sigma^2(2.6405)$$

$$\hat{SE}\left\{\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m}\right\} = \sqrt{0.00338(2.6405)} = 0.0912$$

$$t_{obs} = \frac{0.0403}{0.0912} = 0.4422 \qquad t_{.025,21} = 2.0796 \qquad P\left(t_{.025,21} \geq |0.4422|\right) = .6628$$

$$95\% \text{ CI for } \mathbf{K}'\boldsymbol{\beta}: \quad 1.0403 \pm 2.0796(0.0912) \quad \equiv \quad 1.0403 \pm 0.1897 \quad \equiv \quad (0.8506, 1.2300)$$

The confidence interval contains 1. Based on all evidence, there is no reason to reject the hypothesis of Constant Returns to Scale. An R program and its output are given below.

**R Program**

```
cobbdoug <- read.table("http://www.stat.ufl.edu/~winner/data/cobbdoug1.dat",header=F,
      col.names=c("year","Q.indx","K.indx","L.indx"))
attach(cobbdoug)

log.Q <- log(Q.indx); log.K <- log(K.indx); log.L <- log(L.indx)
log.K_L <- log.K - log.L

cobbdoug.mod1 <- lm(log.Q ~ log.K + log.L)
summary(cobbdoug.mod1)
anova(cobbdoug.mod1)

### Reduced Model:
## E(log.Q) = b0 + b1*log.K + (1-b1)*log.L = b0 + b1*(log.K - log.L) + log.L
cobbdoug.mod2 <- lm(log.Q ~ log.K_L, offset=log.L)
summary(cobbdoug.mod2)
anova(cobbdoug.mod2)

anova(cobbdoug.mod2,cobbdoug.mod1)

#### Matrix Form
n <- length(log.Q)
Y <- log.Q
X0 <- rep(1,n)
X <- cbind(X0,log.K,log.L)
Kp <- matrix(c(0,1,1),ncol=3)
m <- 1
```

```
XPXI <- solve(t(X) %*% X)
beta.ols <- XPXI  %*% t(X) %*% Y
Yhat.ols <- X %*% beta.ols
e.ols <- Y - Yhat.ols
(SSE.ols <- sum(e.ols^2))
beta.diff <- XPXI %*% t(Kp) %*% solve(Kp %*% XPXI %*% t(Kp)) %*% (Kp %*% beta.ols - m)
beta.const <- beta.ols - beta.diff
Yhat.const <- X %*% beta.const
e.const <- Y - Yhat.const
(SSE.const <- sum(e.const^2))
(Q.1 <- SSE.const - SSE.ols)
(Q.2 <- t(Kp %*% beta.ols - m) %*% solve(Kp %*% XPXI %*% t(Kp)) %*% (Kp %*% beta.ols - m))
(s2 <- SSE.ols / (n-ncol(X)))
(F_obs <- Q.1/(1*s2))
(p.F_obs <- 1 - pf(F_obs,1,n-ncol(X)))
```

### R Output

```
> summary(cobbdoug.mod1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.17731    0.43429  -0.408  0.68721
log.K        0.23305    0.06353   3.668  0.00143 **
log.L        0.80728    0.14508   5.565  1.6e-05 ***

Residual standard error: 0.05814 on 21 degrees of freedom
Multiple R-squared:  0.9574,    Adjusted R-squared:  0.9534
F-statistic: 236.1 on 2 and 21 DF,  p-value: 4.038e-15

> anova(cobbdoug.mod1)
Analysis of Variance Table
Response: log.Q
          Df  Sum Sq Mean Sq F value    Pr(>F)
log.K      1 1.49156 1.49156 441.280 1.402e-15 ***
log.L      1 0.10466 0.10466  30.964 1.601e-05 ***
Residuals 21 0.07098 0.00338

> summary(cobbdoug.mod2)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.01454    0.01998   0.728    0.474
log.K_L      0.25413    0.04122   6.165 3.32e-06 ***

Residual standard error: 0.05707 on 22 degrees of freedom
Multiple R-squared:  0.9562,    Adjusted R-squared:  0.9542
F-statistic: 479.9 on 1 and 22 DF,  p-value: < 2.2e-16

> anova(cobbdoug.mod2)
Analysis of Variance Table
Response: log.Q
          Df   Sum Sq  Mean Sq F value    Pr(>F)
log.K_L    1 0.123761 0.123761  38.005 3.324e-06 ***
Residuals 22 0.071643 0.003256

> anova(cobbdoug.mod2,cobbdoug.mod1)
Analysis of Variance Table

Model 1: log.Q ~ log.K_L
Model 2: log.Q ~ log.K + log.L
  Res.Df      RSS Df  Sum of Sq      F Pr(>F)
```

```
1     22 0.071643
2     21 0.070982  1 0.00066109 0.1956 0.6628

> #### Matrix Form
> (SSE.const <- sum(e.const^2))
[1] 0.07164273
> (Q.1 <- SSE.const - SSE.ols)
[1] 0.0006610878
> (Q.2 <- t(Kp %*% beta.ols - m) %*% solve(Kp %*% XPXI %*% t(Kp)) %*% (Kp %*% beta.ols - m))
[1,] 0.0006610878
> (s2 <- SSE.ols / (n-ncol(X)))
[1] 0.003380078
> (F_obs <- Q.1/(1*s2))
[1] 0.1955836
> (p.F_obs <- 1 - pf(F_obs,1,n-ncol(X)))
[1] 0.6628307
```

$\nabla$

### 6.4.2 $R$-Notation for Sums of Squares

$R$ notation is helpful in labeling the Regression Sums of Squares for various models, depending on which predictors are included. The **Model** Sum of Squares for the full set of $p$ predictors is labeled $R(\beta_0, \beta_1, \ldots, \beta_p)$. Note that $R(\beta_0, \beta_1, \ldots, \beta_p) = \mathbf{Y'PY}$. It includes $SS\mu = n\overline{Y}^2$. The **Regression** Sum of Squares for the full set of predictors is $R(\beta_1, \ldots, \beta_p|\beta_0)$. This is computed as follows.

$$R(\beta_1, \ldots, \beta_p|\beta_0) = R(\beta_0, \beta_1, \ldots, \beta_p) - R(\beta_0)$$

$$\text{where:} \quad R(\beta_0, \beta_1, \ldots, \beta_p) = \mathbf{Y'P}_{01\ldots p}\mathbf{Y} \qquad R(\beta_0) = \mathbf{Y'P}_0\mathbf{Y}$$

In these cases, $\mathbf{P}$ depends on which predictors and/or the intercept are included in a model. Consider $R(\beta_0)$:

$$\mathbf{X}_0 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \qquad \mathbf{X}_0'\mathbf{X}_0 = n \qquad (\mathbf{X}_0'\mathbf{X}_0)^{-1} = \frac{1}{n} \qquad \mathbf{P}_0 = \mathbf{X}_0(\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_0' = \frac{1}{n}\mathbf{J}.$$

$$R(\beta_0) = \mathbf{Y'P}_0\mathbf{Y} = \frac{\left(\sum_{i=1}^n Y_i\right)^2}{n} = n\overline{Y}^2 = SS\mu$$

Note that $Q$ for many linear hypotheses (that some set of $\beta_j = 0$) can be constructed from the difference between two $R$ values. Two important sets that are computed by some statistical software packages are the

**Sequential** and the **Partial** Sums of Squares. In SAS, these are labeled Type I and Type III, respectively. The Sequential Sums of Squares represent the impact of each variable being added sequentially on the Regression Sum of Squares.

$$R(\beta_1|\beta_0) = \mathbf{Y}'\mathbf{P}_{01}\mathbf{Y} - \mathbf{Y}'\mathbf{P}_0\mathbf{Y} \qquad R(\beta_2|\beta_0,\beta_1) = \mathbf{Y}'\mathbf{P}_{012}\mathbf{Y} - \mathbf{Y}'\mathbf{P}_{01}\mathbf{Y} \qquad \cdots$$
$$R(\beta_p|\beta_0,\ldots,\beta_{p-1}) = \mathbf{Y}'\mathbf{P}_{01\ldots p}\mathbf{Y} - \mathbf{Y}'\mathbf{P}_{01\ldots p-1}\mathbf{Y} = \mathbf{Y}'\left(\mathbf{P}_{01\ldots p} - \mathbf{P}_{01\ldots p-1}\right)\mathbf{Y}$$

The Partial Sums of Squares give the impact of each variable above and beyond the $p-1$ other predictors on the Regression Sum of Squares. The $F$-test based on the Partial Sums of Squares are comparable to the $t$-tests used for the individual predictors.

$$R(\beta_j|\beta_0,\ldots,\beta_{j-1},\beta_{j+1},\ldots,\beta_p) = \mathbf{Y}'\mathbf{P}_{01\ldots p}\mathbf{Y} - \mathbf{Y}'\mathbf{P}_{01\ldots j-1,j+1,\ldots p}\mathbf{Y} = \mathbf{Y}'\left(\mathbf{P}_{01\ldots p} - \mathbf{P}_{01\ldots j-1,j+1,\ldots p}\right)\mathbf{Y}$$

**Example: Association Between Climate Factors and Phenolic Measurements in Bordeaux Wines**

A study over $n = 16$ years looked at the relation between $p = 4$ climate variables and sugar content in Cabernet Sauvignon wine (Jones and Storchmann (2001)). The predictors were various functional forms of precipitation levels during stages of the growing season, and the number of days of one stage over 25C degrees. The response variable for this model is Sugar content of Cabernet Sauvignon ($Y$, in grams/liter). The predictor variables that are in the authors' final model are given below. The dataset is given in Table 6.3.

- Reciprocal of the Precipitation During Bud break stage (BPREC): $X_1 = 1/\text{BPREC}$
- Precipitation During Floraison stage (FPREC): $X_2 = \text{FPREC}$
- Reciprocal of the number of days over 25°C in the Floraison stage (FTEMP25): $X_3 = 1/\text{FTEMP25}$
- Logarithm of the Precipitation During Veraison stage (VPREC): $X_4 = \ln(\text{VPREC})$

The Sequential and Partial Sums of Squares are computed by using the $R$-notation below.

```
wpb <- read.csv("http://www.stat.ufl.edu/~winner/data/wineprice_bordeaux.csv",
     header=T)
attach(wpb); names(wpb)

y <- CABSUG
x1 <- 1/BPREC
x2 <- FPREC
x3 <- 1/FTEMP25
x4 <- log(VPREC)
x0 <- rep(1,16)


X0 <- x0
X01 <- cbind(x0,x1)
X012 <- cbind(X01,x2)
```

| YEAR | BPREC | X1 | X2 | FTEMP25 | X3 | VPREC | X4 | Y=CABSUG |
|---|---|---|---|---|---|---|---|---|
| 1980 | 216 | 0.00463 | 120 | 30 | 0.0333 | 128 | 4.852 | 179 |
| 1981 | 204 | 0.00490 | 81 | 27 | 0.0370 | 116 | 4.754 | 186 |
| 1982 | 129 | 0.00775 | 159 | 33 | 0.0303 | 55 | 4.007 | 200 |
| 1983 | 220 | 0.00455 | 114 | 45 | 0.0222 | 92 | 4.522 | 195 |
| 1984 | 195 | 0.00513 | 75 | 38 | 0.0263 | 261 | 5.565 | 185 |
| 1985 | 230 | 0.00435 | 123 | 26 | 0.0385 | 14 | 2.639 | 200 |
| 1986 | 103 | 0.00971 | 31 | 41 | 0.0244 | 146 | 4.984 | 199 |
| 1987 | 111 | 0.00901 | 117 | 24 | 0.0417 | 153 | 5.030 | 176 |
| 1988 | 237 | 0.00422 | 118 | 34 | 0.0294 | 53 | 3.970 | 191 |
| 1989 | 198 | 0.00505 | 79 | 49 | 0.0204 | 60 | 4.094 | 205 |
| 1990 | 128 | 0.00781 | 98 | 36 | 0.0278 | 74 | 4.304 | 199 |
| 1991 | 222 | 0.00450 | 115 | 41 | 0.0244 | 179 | 5.187 | 183 |
| 1992 | 258 | 0.00388 | 324 | 34 | 0.0294 | 167 | 5.118 | 168 |
| 1993 | 218 | 0.00459 | 170 | 36 | 0.0278 | 242 | 5.489 | 175 |
| 1994 | 312 | 0.00321 | 111 | 43 | 0.0233 | 166 | 5.112 | 193 |
| 1995 | 276 | 0.00362 | 80 | 48 | 0.0208 | 117 | 4.762 | 194 |

Table 6.3: Phenolic and Climate Data for Bordeaux Wines 1980-1995

```
X0123 <- cbind(X012,x3)
X0124 <- cbind(X012,x4)
X0134 <- cbind(X01,x3,x4)
X0234 <- cbind(x0,x2,x3,x4)
X01234 <- cbind(X0123,x4)

P01234 <- X01234 %*% solve(t(X01234) %*% X01234) %*% t(X01234)
P0234 <- X0234 %*% solve(t(X0234) %*% X0234) %*% t(X0234)
P0134 <- X0134 %*% solve(t(X0134) %*% X0134) %*% t(X0134)
P0124 <- X0124 %*% solve(t(X0124) %*% X0124) %*% t(X0124)
P0123 <- X0123 %*% solve(t(X0123) %*% X0123) %*% t(X0123)
P012 <- X012 %*% solve(t(X012) %*% X012) %*% t(X012)
P01 <- X01 %*% solve(t(X01) %*% X01) %*% t(X01)
P0 <- X0 %*% solve(t(X0) %*% X0) %*% t(X0)

> ### Total Corrected Sum of Squares
> (TSS <- t(y) %*% (diag(16) - P0) %*% y)
[1,] 1745
> ### Regression Sum of Squares
> (SSR <- t(y) %*% (P01234 - P0) %*% y)
[1,] 1625.105
> ### Error Sum of Squares
> (SSE <- t(y) %*% (diag(16) - P01234) %*% y)
[1,] 119.8947
> ### Sequential Sums of Squares
> (R1_0 <- t(y) %*% (P01 - P0) %*% y)
[1,] 66.82349
> (R2_01 <- t(y) %*% (P012 - P01) %*% y)
[1,] 562.973
> (R3_012 <- t(y) %*% (P0123 - P012) %*% y)
[1,] 129.6306
> (R4_0123 <- t(y) %*% (P01234 - P0123) %*% y)
[1,] 865.6782
> ### Partial Sums of Squares
> (R1_0234 <- t(y) %*% (P01234 - P0234) %*% y)
[1,] 74.31272
> (R2_0134 <- t(y) %*% (P01234 - P0134) %*% y)
[1,] 203.6555
> (R3_0124 <- t(y) %*% (P01234 - P0124) %*% y)
```

```
[1,] 401.1116
> (R4_0123 <- t(y) %*% (P01234 - P0123) %*% y)
[1,] 865.6782
```

$\nabla$

### 6.4.3 Coefficients of Partial Determination

The Coefficient of Determination, $R^2$, represents the Regression Sum of Squares, $SSR$, as a proportion of the Total Sum of Squares, $TSS$. It represents the proportion of the variation in $Y$ that is "explained" by the set of predictors $X_1, \ldots, X_p$. It is often useful to consider how much of the variation "not explained" by previously entered predictors that is "explained" by subsequently entered predictor(s). These are **Coefficients of Partial Determination**.

Suppose a given response variable has a (corrected) Total Sum of Squares of $TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$. The first predictor, $X_1$ is used to predict $Y$, and we obtain $SSR(X_1) = R(\beta_1|\beta_0)$. Then the Coefficient of Determination for using $X_1$ to predict $Y$ and the remaining "unexplained" variation are

$$R^2(X_1) = \frac{SSR(X_1)}{TSS} = \frac{R(\beta_1|\beta_0)}{TSS} \qquad SSE(X_1) = TSS - SSR(X_1) = TSS - R(\beta_1|\beta_0).$$

Now, variable $X_2$ is added to the model, and we obtain

$$SSR(X_1, X_2) = R(\beta_1|\beta_0) + R(\beta_2|\beta_0, \beta_1).$$

Then the proportion of variation "explained" by $X_2$, that was not "explained" by $X_1$ is

$$R^2(X_2|X_1) = \frac{SSR(X_1, X_2) - SSR(X_1)}{TSS - SSR(X_1)} = \frac{R(\beta_2|\beta_0, \beta_1)}{TSS - R(\beta_1|\beta_0)}.$$

This continues for all predictors entered in the model. In general, we have:

$$R^2(X_j|X_1, \ldots, X_{j-1}) = \frac{R(\beta_j|\beta_0, \beta_1, \ldots, \beta_{j-1})}{TSS - R(\beta_1, \ldots, \beta_{j-1}|\beta_0)}.$$

**Example: Association Between Climate Factors and Phenolic Measurements in Bordeaux Wines**

For this dataset, we obtained the following Sequential Sums of Squares previously, and make use of them to compute the Coefficients of Partial Determination.

$$TSS = 1745 \qquad R(\beta_1|\beta_0) = 66.8 \quad \Rightarrow \quad R^2(X_1) = \frac{66.8}{1745} = 0.0382$$

$$R(\beta_2|\beta_0,\beta_1) = 563.0 \quad \Rightarrow \quad R^2(X_2|X_1) = \frac{563.0}{1745-66.8} = 0.3355$$

$$R(\beta_1,\beta_2|\beta_0) = 66.8 + 563.0 = 629.8$$

$$R(\beta_3|\beta_0,\beta_1,\beta_2) = 129.6 \quad \Rightarrow \quad R^2(X_3|X_1,X_2) = \frac{129.6}{1745-629.8} = 0.1162$$

$$R(\beta_1,\beta_2,\beta_3|\beta_0) = 629.8 + 129.6 = 759.4$$

$$R(\beta_4|\beta_0,\beta_1,\beta_2,\beta_3) = 865.7 \quad \Rightarrow \quad R^2(X_4|X_1,X_2,X_3) = \frac{865.7}{1745-759.4} = 0.8783$$

$$R(\beta_1,\beta_2,\beta_3,\beta_4|\beta_0) = 759.4 + 865.7 = 1625.1 \quad \Rightarrow \quad R^2(X_1,X_2,X_3,X_4) = \frac{1625.1}{1745} = 0.9313$$

$\nabla$

## 6.5 Models With Categorical (Qualitative) Predictors

Often, one or more categorical variables are included in a model. If we have a categorical variable with $m$ levels, we will need to create $m-1$ **dummy** or **indicator variables**. The variable will take on 1 if the $i^{th}$ observation corresponds to that level of the variable, 0 otherwise. Note that one level of the variable will have $0^s$ for all $m-1$ dummy variables, making it the reference group or category. The $\beta^s$ for the other groups (levels of the qualitative variable) reflect the difference in the mean for that group with the reference group, controlling for all other predictors.

Note that if the qualitative variable has 2 levels, there will be a single dummy variable, and we can test for differences in the effects of the 2 levels with a $t$-test, controlling for all other predictors. If there are $m-1>2$ dummy variables, we can use the $F$-test to test whether all $m-1$ $\beta^s$ are 0, meaning there are no differences in group means, controlling for all other predictors.

**Example: Relationship Between Weight and Height Among NBA, NHL, and EPL Athletes**

Samples of male athletes from the National Basketball Association, National Hockey League, and English Premier (Football) League are obtained, and the relationship between players' Weight ($Y$) and Height ($X_1$) is measured. There are $m=3$ sports, so we create $m-1=2$ dummy variables. Let $X_2=1$ if the player is from the NBA (0 otherwise) and $X_3=1$ if the player is from the NHL (0 otherwise). This makes the EPL the "reference" category. The data will be based on random samples of 12 athletes per league. The model is given below. The data are in Table 6.4.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i \qquad i = 1,\ldots,36$$

$$\text{NBA:} \quad E\{Y_i\} = \beta_0 + \beta_1 X_{i1} + \beta_2(1) + \beta_3(0) = (\beta_0+\beta_2) + \beta_1 X_{i1}$$

| Player | Height | Weight | NBA | NHL |
|---|---|---|---|---|
| Juwan Howard | 81 | 250 | 1 | 0 |
| Luc Mbah a Moute | 80 | 230 | 1 | 0 |
| Jeff Withey | 84 | 235 | 1 | 0 |
| Brook Lopez | 84 | 265 | 1 | 0 |
| Cory Joseph | 75 | 185 | 1 | 0 |
| Quincy Acy | 79 | 233 | 1 | 0 |
| Giannis Antetokou | 81 | 205 | 1 | 0 |
| Chase Budinger | 79 | 218 | 1 | 0 |
| Darren Collison | 72 | 160 | 1 | 0 |
| Carlos Boozer | 81 | 266 | 1 | 0 |
| Pero Antic | 82 | 260 | 1 | 0 |
| Jamaal Tinsley | 75 | 185 | 1 | 0 |
| Matt Niskanen | 72 | 209 | 0 | 1 |
| Andrew Ference | 71 | 184 | 0 | 1 |
| Brooks Orpik | 74 | 219 | 0 | 1 |
| Elias Lindholm | 73 | 192 | 0 | 1 |
| Martin Brodeur | 74 | 220 | 0 | 1 |
| Sam Bennett | 73 | 178 | 0 | 1 |
| Rob Klinkhammer | 75 | 214 | 0 | 1 |
| Nick Bonino | 73 | 196 | 0 | 1 |
| B.J. Crombeen | 74 | 209 | 0 | 1 |
| Michael Raffl | 72 | 195 | 0 | 1 |
| John-Michael Lile | 70 | 185 | 0 | 1 |
| Jonathan Ericsson | 76 | 220 | 0 | 1 |
| Kieran Trippier | 70 | 157 | 0 | 0 |
| Moussa Dembele | 73 | 181 | 0 | 0 |
| Marouane Chamakh | 73 | 154 | 0 | 0 |
| Martin Kelly | 75 | 170 | 0 | 0 |
| Jesse Lingard | 69 | 128 | 0 | 0 |
| Darren Randolph | 74 | 216 | 0 | 0 |
| James Ward-Prowse | 68 | 146 | 0 | 0 |
| Mame Diouf | 73 | 168 | 0 | 0 |
| Nemanja Matic | 76 | 185 | 0 | 0 |
| Danny Rose | 68 | 159 | 0 | 0 |
| Mathieu Flamini | 70 | 148 | 0 | 0 |
| Ikechi Anya | 65 | 159 | 0 | 0 |

Table 6.4: Heights and Weights of Samples of 12 NBA, NHL, EPL Athletes

$$\text{NHL:} \quad E\{Y_i\} = \beta_0 + \beta_1 X_{i1} + \beta_2(0) + \beta_3(1) = (\beta_0 + \beta_3) + \beta_1 X_{i1}$$
$$\text{EPL:} \quad E\{Y_i\} = \beta_0 + \beta_1 X_{i1} + \beta_2(0) + \beta_3(0) = \beta_0 + \beta_1 X_{i1}$$

Two models are fit, the first with the Height and the 2 League dummy variables, the second with only Height, with the following results.

$$\text{Model 1:} \quad \hat{Y} = -290.2045 + 6.3858X_1 + 7.4007X_2 + 25.2606X_3$$
$$SSE_1 = 8925.0 \quad df_1 = 36 - 4 = 32 \quad R(\beta_1, \beta_2, \beta_3|\beta_0) = 34481.2 \quad R_1^2 = 0.7944$$

$$\text{Model 2:} \quad \hat{Y} = -277.8713 + 6.3664X_1$$
$$SSE_2 = 12972 \quad df_2 = 36 - 2 = 34 \quad R(\beta_1|\beta_0) = 30434.1 \quad R_2^2 = 0.7011$$

The R program and output are given below.

**R Program**

```
set.seed(1357)
nba <- read.csv("http://stat.ufl.edu/~winner/data/nba_ht_wt.csv",header=T)
attach(nba); names(nba)
nba.samp <- sample(1:length(Height),12,replace=F)
nba.sample <- nba[nba.samp,c(1,3,4)]
nba.sample
detach(nba)

nhl <- read.csv("http://stat.ufl.edu/~winner/data/nhl_ht_wt.csv",header=T)
attach(nhl); names(nhl)
nhl.samp <- sample(1:length(Height),12,replace=F)
nhl.sample <- nhl[nhl.samp,c(2,4,5)]
nhl.sample
detach(nhl)

epl <- read.csv("http://stat.ufl.edu/~winner/data/epl_2015_ht_wt.csv",header=T)
attach(epl); names(epl)
epl.samp <- sample(1:length(Height),12,replace=F)
epl.sample <- epl[epl.samp,c(2,6,7)]
epl.sample
detach(epl)

all.sample <- rbind(nba.sample,nhl.sample,epl.sample)
all.sample
plot(Height,Weight)

NBA <- c(rep(1,12),rep(0,24))
NHL <- c(rep(0,12),rep(1,12),rep(0,12))
League <- c(rep(1,12),rep(2,12),rep(3,12))
League <- factor(League)

all.sample1 <- data.frame(all.sample,NBA,NHL,League)

wtht.1 <- lm(Weight ~ Height + NBA + NHL, data=all.sample1)
summary(wtht.1)
```

```
anova(wtht.1)
drop1(wtht.1, test="F")

wtht.2 <- lm(Weight ~ Height, data=all.sample1)
summary(wtht.2)
anova(wtht.2)

anova(wtht.2,wtht.1)

ht.x <- seq(62,82,0.1)
yhat.nba <- coef(wtht.1)[1] + ht.x*coef(wtht.1)[2] + coef(wtht.1)[3]
yhat.nhl <- coef(wtht.1)[1] + ht.x*coef(wtht.1)[2] + coef(wtht.1)[4]
yhat.epl <- coef(wtht.1)[1] + ht.x*coef(wtht.1)[2]

plot(Height,Weight,pch=as.numeric(League),ylim=c(120,240))
lines(ht.x,yhat.nba,lty=1)
lines(ht.x,yhat.nhl,lty=2)
lines(ht.x,yhat.epl,lty=5)
legend(65,240,c("NBA","NHL","EPL"),pch=c(1,2,3),lty=c(1,2,5))

wtht.3 <- lm(Weight ~ Height + NBA + NHL + I(Height*NBA) + I(Height*NHL),
data=all.sample1)
summary(wtht.3)
anova(wtht.3)
drop1(wtht.3,test="F")

anova(wtht.1,wtht.3)
ht.x <- seq(62,82,0.1)
yhat.nba3 <- coef(wtht.3)[1] + ht.x*coef(wtht.3)[2] + coef(wtht.3)[3] +
   ht.x*coef(wtht.3)[5]
yhat.nhl3 <- coef(wtht.3)[1] + ht.x*coef(wtht.3)[2] + coef(wtht.3)[4] +
   ht.x*coef(wtht.3)[6]
yhat.epl3 <- coef(wtht.3)[1] + ht.x*coef(wtht.3)[2]

plot(Height,Weight,pch=as.numeric(League),ylim=c(120,240))
lines(ht.x,yhat.nba3,lty=1)
lines(ht.x,yhat.nhl3,lty=2)
lines(ht.x,yhat.epl3,lty=5)
legend(65,240,c("NBA","NHL","EPL"),pch=c(1,2,3),lty=c(1,2,5))

wtht.4 <- lm(Weight[League==3] ~ Height[League==3],data=all.sample1)
anova(wtht.4)
```

### R Output

```
Model 1:
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -290.2045    68.3924  -4.243 0.000176 ***
Height         6.3858     0.9586   6.661 1.62e-07 ***
NBA            7.4007    10.4418   0.709 0.483609
NHL           25.2606     7.0612   3.577 0.001129 **

Residual standard error: 16.7 on 32 degrees of freedom
Multiple R-squared:  0.7944,    Adjusted R-squared:  0.7751
F-statistic: 41.21 on 3 and 32 DF,  p-value: 4.233e-11

Sequential Sums of Squares:
Analysis of Variance Table
Response: Weight
          Df  Sum Sq Mean Sq  F value    Pr(>F)
```

```
Height     1 30434.1 30434.1 109.1196 7.712e-12 ***
NBA        1   477.7   477.7   1.7129  0.199933
NHL        1  3569.4  3569.4  12.7977  0.001129 **
Residuals 32  8925.0   278.9

Partial Sums of Squares:
> drop1(wtht.1,test="F")
Single term deletions
Model:
Weight ~ Height + NBA + NHL
       Df Sum of Sq     RSS    AIC F value    Pr(>F)
<none>               8925.0 206.47
Height  1   12376.2 21301.2 235.79 44.3739 1.623e-07 ***
NBA     1     140.1  9065.1 205.03  0.5023  0.483609
NHL     1    3569.4 12494.4 216.58 12.7977  0.001129 **

Model 2:
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -277.8713    53.2440  -5.219 8.93e-06 ***
Height         6.3664     0.7128   8.931 1.94e-10 ***

Analysis of Variance Table
Response: Weight
          Df Sum Sq Mean Sq F value   Pr(>F)
Height     1  30434 30434.1  79.768 1.94e-10 ***
Residuals 34  12972   381.5
```

To test whether there are league effects, after controlling for Height, we test $H_0 : \beta_2 = \beta_3 = 0$. This can be constructed in many ways, In matrix form, we are testing $H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$ where we have 2 restrictions (rows) in $\mathbf{K}'$.

$$\mathbf{K}' = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \qquad \mathbf{m} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Compute $Q$ as either $SSE(R) - SSE(F) = SSE_2 - SSE_1$ or $R(\beta_2, \beta_3 | \beta_0, \beta_1)$.

$$Q = 12972 - 8925 = 34481 - 30434 = 4047 \qquad s^2 = MSE(F) = \frac{8925}{32} = 278.9$$

$$F_{obs} = \frac{4047}{2(278.9)} = 7.255 \qquad F_{0.05,2,32} = 3.295 \qquad P\left(Y_{2,32} \geq 7.255\right) = .0025$$

There is evidence of league differences in Weight, controlling for Height. A plot of the data and regression lines are shown in Figure 6.1.

$\nabla$

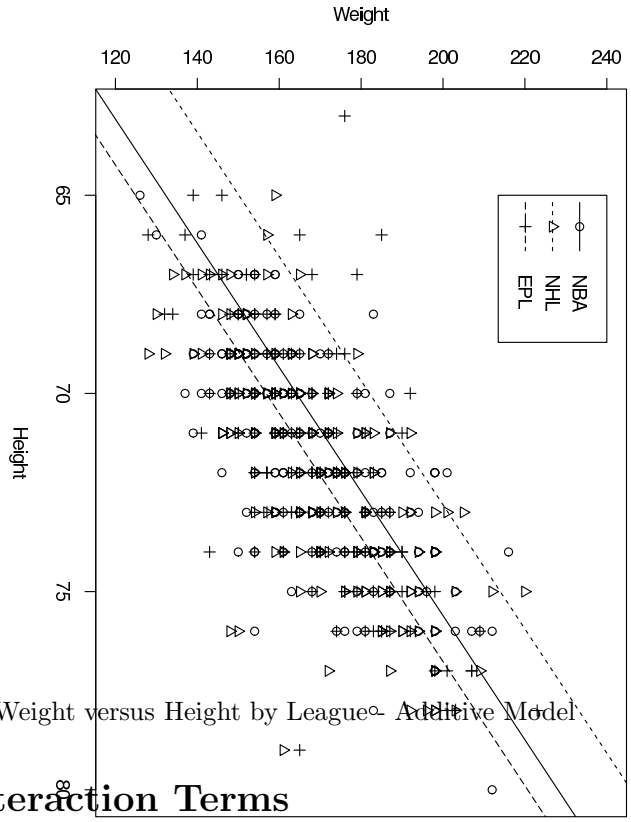


Figure 6.1: Weight versus Height by League - Additive Model

## 6.6 Models With Interaction Terms

When the effect of one predictor depends on the level of another predictor (and vice versa), the predictors are said to **interact**. The way we can model interaction(s) is to create a new variable that is the product of the 2 predictors. Suppose we have $Y$, and 2 numeric predictors: $X_1$ and $X_2$. We create a new predictor $X_3 = X_1X_2$. Now, consider the model:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 = \beta_0 + \beta_2 X_2 + (\beta_1 + \beta_3 X_2) X_1$$

Thus, the slope with respect to $X_1$ depends on the level of $X_2$, unless $\beta_3 = 0$, which we can test with a $t$-test. This logic extends to qualitative variables as well. We create cross-product terms between numeric (or other categorical) predictors with the $m - 1$ dummy variables representing the qualitative predictor. Then a $t$-test ($m - 1 = 1$) or a $F$-test ($m - 1 > 2$) can be conducted to test for interactions among predictors.

**Example: Relationship Between Weight and Height Among NBA, NHL, and EPL Athletes**

Continuing the Athletes' Weight and Height regression model, we can include 2 interaction terms, one for the NBA, one for the NHL. These allow for the slopes to differ among the leagues, as well as the intercepts. We extend the model below.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} | \beta_4 X_{i1} X_{i2} + \beta_5 X_{i1} X_{i3} + \epsilon_i \qquad i = 1, \ldots, 36$$

$$\text{NBA:} \quad E\{Y_i\} = \beta_0 + \beta_1 X_{i1} + \beta_2(1) + \beta_3(0) + \beta_4 X_{i1}(1) + \beta_5 X_{i5}(0) = (\beta_0 + \beta_2) + (\beta_1 + \beta_4) X_{i1}$$

$$\text{NHL:} \quad E\{Y_i\} = \beta_0 + \beta_1 X_{i1} + \beta_2(0) + \beta_3(1) = +\beta_4 X_{i1}(0) + \beta_5 X_{i5}(1) (\beta_0 + \beta_3) + (\beta_1 + \beta_5) X_{i1}$$

$$\text{EPL:} \quad E\{Y_i\} = \beta_0 + \beta_1 X_{i1} + \beta_2(0) + \beta_3(0) + \beta_4 X_{i1}(0) + \beta_5 X_{i5}(0) = \beta_0 + \beta_1 X_{i1}$$

$$\text{Model 3:} \qquad \hat{Y} = -128.508 + 4.114 X_1 - 294.595 X_2 - 159.656 X_3 + 4.039 X_1 X_2 + 2.590 X_1 X_3$$

$$SSE_3 = 7822.8 \quad df_3 = 36 - 6 = 30 \quad R(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5 | \beta_0) = 35583.4 \quad R_3^2 = 0.8198$$

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -128.508    104.291  -1.232  0.22744
Height            4.114      1.464   2.810  0.00864 **
NBA            -294.595    147.634  -1.995  0.05515 .
NHL            -159.656    236.532  -0.675  0.50485
I(Height * NBA)    4.039      1.968   2.053  0.04891 *
I(Height * NHL)    2.590      3.252   0.796  0.43212

Residual standard error: 16.15 on 30 degrees of freedom
Multiple R-squared:  0.8198,    Adjusted R-squared:  0.7897
F-statistic: 27.29 on 5 and 30 DF,  p-value: 2.565e-10


Sequential Sums of Squares:
Analysis of Variance Table

Response: Weight
               Df  Sum Sq Mean Sq  F value    Pr(>F)
Height          1 30434.1 30434.1 116.7134 7.341e-12 ***
NBA             1   477.7   477.7   1.8321 0.1859924
NHL             1  3569.4  3569.4  13.6883 0.0008651 ***
I(Height * NBA) 1   936.9   936.9   3.5928 0.0677003 .
I(Height * NHL) 1   165.3   165.3   0.6341 0.4321155
Residuals      30  7822.8   260.8

Partial Sums of Squares:
> drop1(wtht.3, test="F")
Single term deletions
Model:
Weight ~ Height + NBA + NHL + I(Height * NBA) + I(Height * NHL)
               Df Sum of Sq    RSS    AIC F value  Pr(>F)
<none>                      7822.8 205.73
Height          1   2058.91 9881.7 212.14  7.8958 0.00864 **
NBA             1   1038.29 8861.1 208.21  3.9818 0.05515 .
NHL             1    118.80 7941.6 204.27  0.4556 0.50485
I(Height * NBA) 1   1098.74 8921.5 208.46  4.2136 0.04891 *
I(Height * NHL) 1    165.35 7988.1 204.48  0.6341 0.43212
```

To test whether there are league by Height interactions, we test $H_0 : \beta_4 = \beta_5 = 0$. This can be constructed in many ways, In matrix form, we are testing $H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$ where we have 2 restrictions (rows) in $\mathbf{K}'$.

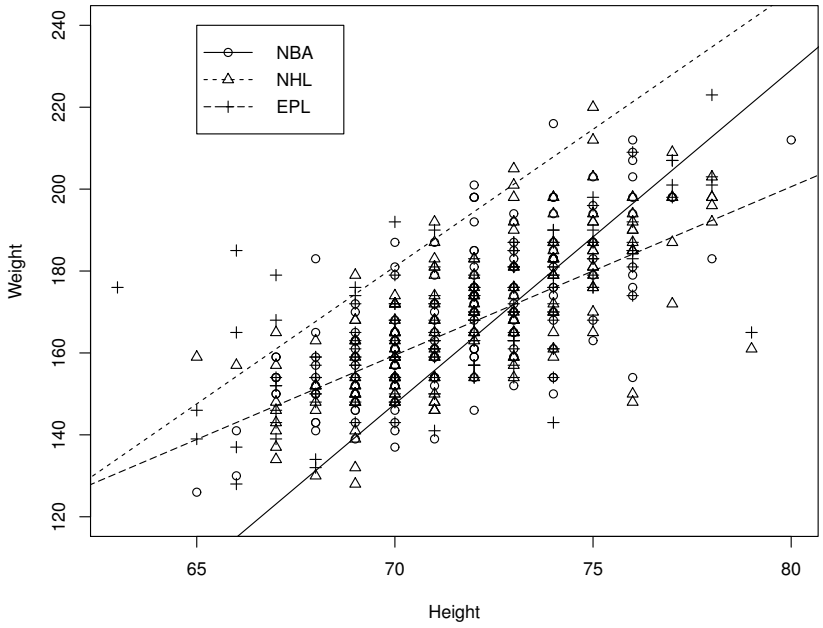


Figure 6.2: Weight versus Height by League - Interaction Model

$$\mathbf{K}' = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} \qquad \mathbf{m} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Compute $Q$ as either $SSE(R) - SSE(F) = SSE_1 - SSE_3$ or $R(\beta_4, \beta_5 | \beta_0, \beta_1, \beta_2, \beta_3)$.

$$Q = 8925 - 7823 = 35583 - 34481 = 1102 \qquad s^2 = MSE(F) = \frac{7823}{30} = 260.8$$

$$F_{obs} = \frac{1102}{2(260.8)} = 2.113 \qquad F_{0.05,2,30} = 3.316 \qquad P(Y_{2,30} \geq 2.113) = .1385$$

While the test is not significant, we include a plot of the fitted equations in Figure 6.2.

$\nabla$

**Test for Equal Variances**

A general test for common variances across groups is **Bartlett's Test**. It can be used in many settings, such as in a 1-Way ANOVA model comparing several groups. In the case of "grouped" regression models, we may want to test whether the error variance when $Y$ is regressed on $X$ separately for the $m$ groups. This could be extended to include several numeric predictors as well.

We then fit the individual regressions for each group separately, and obtain $s_j^2 = MSE_j \quad j = 1, ..., m$. When there is a single numeric predictor, $MSE_j$ has $\nu_j = n_j - 2$ degrees of freedom. This clearly generalizes to $n_j - p'$ when there are $p$ predictor in each regression.

$$\nu = \sum_{j=1}^{m} \nu_j \qquad s^2 = MSE = \frac{\sum_{j=1}^{m} \nu_j s_j^2}{\nu}$$

$$X_B^2 = \left[1 + \frac{\sum_{j=1}^{m} \nu_j^{-1} - \nu^{-1}}{3\,(m-1)}\right]^{-1} \left[\nu \ln s^2 - \sum_{i=1}^{m} \nu_j \ln s_j^2\right]$$

Conclude that the error variances among the regressions differ if $X_B^2 \geq \chi^2\alpha, m-1$.

**Example: Relationship Between Weight and Height Among NBA, NHL, and EPL Athletes**

When we fit the $m = 3$ regression models individually by league, we get the same regression coefficients as we did in the single interaction model with all of the players. Fitting the individual models gives the following variance estimates. Each league had 12 players in the sample.

$$\nu_1 = \nu_2 = \nu_3 = 12 - 2 = 10 \qquad \Rightarrow \qquad \nu = 10 + 10 + 10 = 30$$

$$s_1^2 = 315.8 \quad s_2^2 = 116.3 \quad s_3^2 = 350.3 \quad \Rightarrow \quad s^2 = \frac{10(315.8) + 10(116.3) + 10(350.3)}{30} = 260.8 = MSE$$

$$X_B^2 = \left[1 + \frac{3(10)^{-1} - (30)^{-1}}{3\,(3-1)}\right]^{-1} [30\ln(260.8) - 10\,(\ln(315.8) + \ln(116.3) + \ln(350.3))] =$$

$$\frac{166.9126 - (57.5511 + 47.5609 + 58.5879)}{1 + \frac{3(.10) - .0333}{3(2)}} = \frac{3.2128}{1.0444} = 3.0760 \qquad \chi^2_{.05,2} = 5.991 \qquad P\left(\chi_2^2 \geq 3.0760\right) = .2148$$

There is no evidence of unequal error variance among the 3 Leagues.

$\nabla$

## 6.7 Models With Curvature

When a plot of $Y$ versus one or more of the predictors displays curvature, we can include polynomial terms to "bend" the regression line. Often, to avoid multicollinearity, we work with centered predictor(s), by subtracting off their mean(s). If the data show $k$ bends, we will include $k+1$ polynomial terms. Suppose we have a single predictor variable, with 2 "bends" appearing in a scatterplot. Then, we will include terms up to the a third order term. Note that even if lower order terms are not significant, when a higher order term is significant, we keep the lower order terms (unless there is some physical reason not to). We can now fit the model:

$$E\{Y\} = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3.$$

If we wish to test whether the fit is linear, as opposed to "not linear," we could test $H_0 : \beta_2 = \beta_3 = 0$. In many instances it is preferable to center the data (subtract off the mean) or to center and scale the data (divide centered values by a scale constant) for ease of interpretation and to reduce collinearity among the predictors. Extrapolating outside the observed $X$-levels can lead to highly erroneous predictions. Consider the un-centered and centered models for a quadratic model.

$$\text{Un-centered:} \quad E\{Y\} = \beta_0 + \beta_1 X + \beta_2 X^2$$

$$\text{Centered:} \quad E\{Y\} = \gamma_0 + \gamma_1\left(X - \overline{X}\right) + \gamma_2\left(X - \overline{X}\right)^2 = \gamma_0 + \gamma_1 X - \gamma_1\overline{X} + \gamma_2 X^2 - 2\gamma_2 X\overline{X} + \gamma_2\overline{X}^2$$

$$\Rightarrow \quad \beta_0 = \gamma_0 - \gamma_1\overline{X} + \gamma_2\overline{X}^2 \qquad \beta_1 = \gamma_1 - 2\gamma_2\overline{X} \qquad \beta_2 = \gamma_2$$

The fitted values are the same for the 2 models, but the parameters (except the coefficient of the highest order polynomial) are different. For plotting purposes, fit the model in the original units.

**Example: Relationship Between Container Ship Speed and Fuel Consumption**

Wang and Meng (2012) studied the relationship between Container Ship speed ($X$, in knots) and fuel consumption ($Y$, in tons/day). They studied 5 Ship Type/Voyage Leg combinations. This data is from the third combination: 5000 TEU Hong Kong/Singapore (TEU = 20-foot Equivalent unit). A plot of the data is given in Figure 6.3. There appear to be 2 bends in the data, so try a cubic regression model. We fit both a centered and an un-centered model. The data are given in Table 6.5.

$$\text{Centered:} \quad \hat{Y} = 58.7020 + 13.3245\left(X - \overline{X}\right) + 0.7779\left(X - \overline{X}\right)^2 - 1.1479\left(X - \overline{X}\right)^3$$

$$\text{Centered:} \quad \hat{Y} = 6328.5818 - 1081.3793X + 61.4035X^2 - 1.1479X^3$$

The R program and output are given below.

**R Program**

| speed | fuel | centered speed |
|---|---|---|
| 16.1 | 45 | -1.505 |
| 16.4 | 46 | -1.205 |
| 16.5 | 46 | -1.105 |
| 16.8 | 48 | -0.805 |
| 16.9 | 49 | -0.705 |
| 17 | 51 | -0.605 |
| 17 | 51 | -0.605 |
| 17 | 52 | -0.605 |
| 17.1 | 52 | -0.505 |
| 17.1 | 53 | -0.505 |
| 17.4 | 56 | -0.205 |
| 17.6 | 59 | -0.005 |
| 17.9 | 64 | 0.295 |
| 18.2 | 67 | 0.595 |
| 18.4 | 69 | 0.795 |
| 18.5 | 70 | 0.895 |
| 18.5 | 71 | 0.895 |
| 18.9 | 74 | 1.295 |
| 19 | 75 | 1.395 |
| 19.8 | 80 | 2.195 |

Table 6.5: Speed and Fuel Consumption for 5000 TEU Hong Kong/Singapore Container Ship

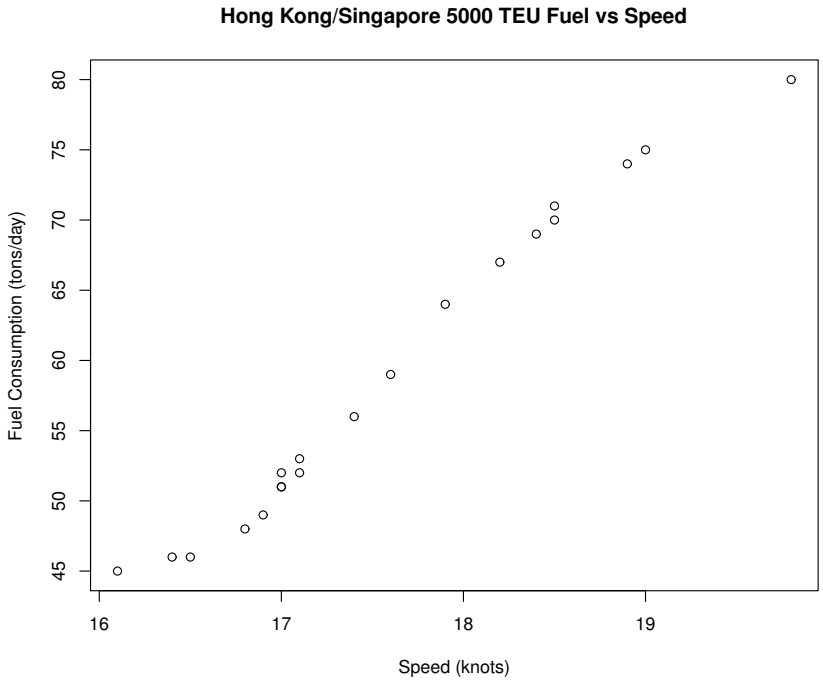


Figure 6.3: Fuel Consumption vs Speed Hong Kong/Singapore 5000 TEU Container Ship

```
spdfuel <- read.csv("http://www.stat.ufl.edu/~winner/data/ship_speed_fuel.csv")
attach(spdfuel); names(spdfuel)

speed_leg3 <- speed[ship_leg==3]
fuel_leg3 <- fuel[ship_leg==3]

plot(speed_leg3,fuel_leg3,xlab="Speed (knots)",ylab="Fuel Consumption (tons/day)",
main = "Hong Kong/Singapore 5000 TEU Fuel vs Speed")

speed_leg3c <- speed_leg3 - mean(speed_leg3)

spdfuel.1 <- lm(fuel_leg3 ~ speed_leg3c + I(speed_leg3c^2) + I(speed_leg3c^3))
summary(spdfuel.1)
anova(spdfuel.1)
drop1(spdfuel.1,test="F")

spdfuel.2 <- lm(fuel_leg3 ~ speed_leg3c)
summary(spdfuel.2)
anova(spdfuel.2)

spdfuel.3 <- lm(fuel_leg3 ~ speed_leg3 + I(speed_leg3^2) + I(speed_leg3^3))
summary(spdfuel.3)


xs <- seq(16,20,.01)
plot(speed_leg3,fuel_leg3,xlab="Speed (knots)",xlim=c(16,20),ylab="Fuel Consumption (tons/day)",
main = "Hong Kong/Singapore 5000 TEU Fuel vs Speed")
lines(xs,predict(spdfuel.3,list(speed_leg3=xs)),lty=1)
```

### R Output

```
> summary(spdfuel.1)
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     58.7020     0.2324 252.566  < 2e-16 ***
speed_leg3c     13.3245     0.2993  44.518  < 2e-16 ***
I(speed_leg3c^2)  0.7779     0.2152   3.616  0.00232 **
I(speed_leg3c^3) -1.1479     0.1384  -8.294 3.46e-07 ***

Residual standard error: 0.7171 on 16 degrees of freedom
Multiple R-squared:  0.9966,    Adjusted R-squared:  0.9959
F-statistic:  1551 on 3 and 16 DF,  p-value: < 2.2e-16

> anova(spdfuel.1)
Analysis of Variance Table
Response: fuel_leg3
               Df  Sum Sq Mean Sq   F value    Pr(>F)
speed_leg3c     1 2355.43 2355.43 4580.2738 < 2.2e-16 ***
I(speed_leg3c^2)  1    2.77    2.77    5.3784   0.03394 *
I(speed_leg3c^3)  1   35.37   35.37   68.7881 3.462e-07 ***
Residuals      16    8.23    0.51

> drop1(spdfuel.1,test="F")
Single term deletions
Model:
fuel_leg3 ~ speed_leg3c + I(speed_leg3c^2) + I(speed_leg3c^3)
               Df Sum of Sq     RSS    AIC  F value    Pr(>F)
<none>                       8.23 -9.764
speed_leg3c     1   1019.18 1027.41 84.781 1981.860 < 2.2e-16 ***
I(speed_leg3c^2)  1      6.72   14.95  0.181   13.073  0.002321 **
I(speed_leg3c^3)  1     35.37   43.60 21.588   68.788 3.462e-07 ***
```

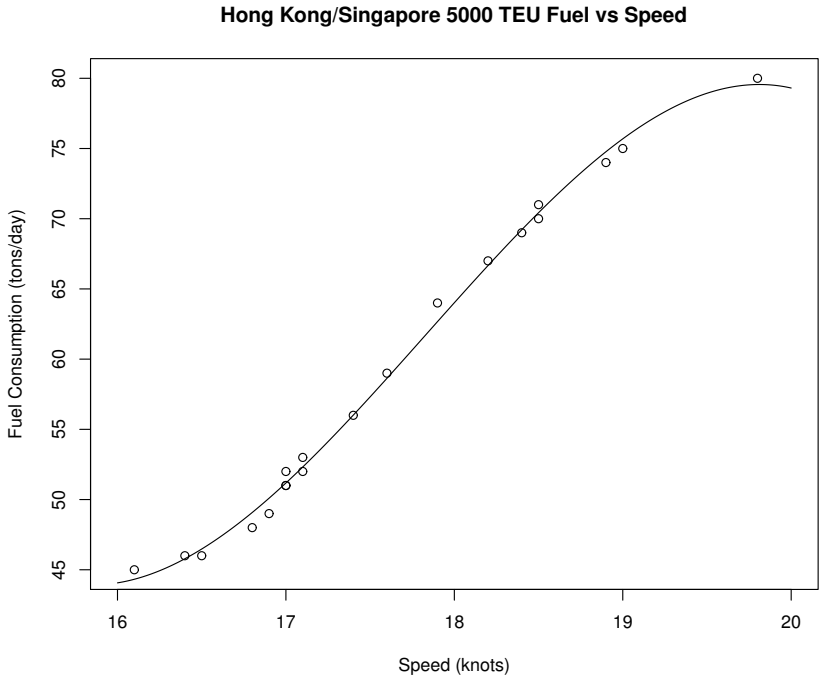


Figure 6.4: Data and Fitted Cubic Regression Equation - Container Ship Data

```
> summary(spdfuel.3)
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      6328.5818   795.9240   7.951 6.00e-07 ***
speed_leg3      -1081.3793   133.4939  -8.101 4.71e-07 ***
I(speed_leg3^2)    61.4035     7.4511   8.241 3.77e-07 ***
I(speed_leg3^3)    -1.1479     0.1384  -8.294 3.46e-07 ***

Residual standard error: 0.7171 on 16 degrees of freedom
Multiple R-squared:  0.9966,    Adjusted R-squared:  0.9959
F-statistic:  1551 on 3 and 16 DF,  p-value: < 2.2e-16
```

Clearly the model fits very well. A plot of the data and the fitted equation are given in Figure 6.4.

$\nabla$

## 6.8 Response Surfaces

Response surfaces are often fit when we have 2 or more predictors, and include "linear effects", "quadratic effects", and "interaction effects." In the case of 3 predictors, a full model would be of the form:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{33} X_3^2 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{23} X_2 X_3.$$

This is a "second-order" model, in some (apparently rare) circumstances, experimenters fit "third-order" models which include cubic terms, and interactions between main effects and quadratic terms. We typically wish to simplify the model, to make it more parsimonious, when possible. Response surfaces are typically used to optimize a process in terms of choosing the input $X$ values that maximize or minimize the process. Consider the second model with $k$ input (predictor) variables.

$$\mathbf{x} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix} \qquad \boldsymbol{\beta}_1 = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ bh_k \end{bmatrix} \qquad \boldsymbol{\beta}_2 = \begin{bmatrix} \hat{\beta}_{11} & \hat{\beta}_{12}/2 & \cdots & \hat{\beta}_{1k}/2 \\ \hat{\beta}_{12}/2 & \hat{\beta}_{22} & \cdots & \hat{\beta}_{2k}/2 \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\beta}_{1k}/2 & \hat{\beta}_{2k}/2 & \cdots & \hat{\beta}_{kk} \end{bmatrix}$$

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j X_j + \sum_{j=1}^{k-1} \sum_{j'=j+1}^{k} \hat{\beta}_{jj'} X_j X_{j'} \sum_{j=1}^{k} \hat{\beta}_{jj} X_j^2 = \hat{\beta}_0 + \mathbf{x}'\hat{\boldsymbol{\beta}}_1 + \mathbf{x}'\hat{\boldsymbol{\beta}}_2\mathbf{x}$$

Taking the derivative of $\hat{Y}$ with respect to $\mathbf{x}$ and setting equal to 0 leads to the optimal settings of $\mathbf{x}$ to maximize or minimize the response surface (assuming it is at an interior point).

$$\frac{\partial \hat{Y}}{\partial \mathbf{x}} = \hat{\boldsymbol{\beta}}_1 + 2\hat{\boldsymbol{\beta}}_2\mathbf{x} \quad \Rightarrow \quad 2\hat{\boldsymbol{\beta}}_2\mathbf{x}^* = -\hat{\boldsymbol{\beta}}_1 \quad \Rightarrow \quad \mathbf{x}^* = -\frac{1}{2}\hat{\boldsymbol{\beta}}_2^{-1}\hat{\boldsymbol{\beta}}_1$$

Most practitioners work with coded values (centered at 0) when fitting the model, then "convert back" to original units for plots and interpretations. The **rsm** package in R uses the original scale for the predictors.

**Example: Factors Affecting Mango Wine**

Kumar, Praksam, and Reddy (2009) reported results from a 3-factor study relating Ethanol percentage ($Y$) to 3 fermenting factors: Temperature ($X_1$, Celsius), pH ($X_2$), and Inoculum Size ($X_3$, percent) in Mango Wine production. They fit a second-order response surface, as described above. The data are given in Table 6.6. The data are coded in the following manner such that the center points (coded values = 0) for Temperature, pH, and Inoculum are 24, 3.8, and 10, respectively. The coded inner lower (-1) values are 18, 3.4, and 5 respectively. The coded inner upper (+1) values are 30, 4.2, and 15, respectively. The extreme, or axial, (+/-1.682) values are (13.908, 34.902) for temperature, (3.1272, 4.4728) for pH, and (1.59, 18.41) for inoculum. Note that there are 6 runs where each factor is at its center value. This permits a goodness-of-fit test.

We begin by fitting the second-order model using a regression package, then we fit it with a specialized package (R package **rsm**). In the original units, we get the following fitted equation (rounded to 3 decimal places).

| Run | CodedTemp | Coded pH | Coded Inoculum | Ethanol | Temp | pH | Inoculum |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | -1.682 | 4.8 | 24 | 3.8 | 1.59 |
| 2 | 0 | 0 | 0 | 9.6 | 24 | 3.8 | 10 |
| 3 | 0 | 0 | 0 | 10.2 | 24 | 3.8 | 10 |
| 4 | 0 | 0 | 1.682 | 8.5 | 24 | 3.8 | 18.41 |
| 5 | 0 | -1.682 | 0 | 7.3 | 24 | 3.1272 | 10 |
| 6 | 1 | 1 | -1 | 4.8 | 30 | 4.2 | 5 |
| 7 | -1 | -1 | 1 | 7.9 | 18 | 3.4 | 15 |
| 8 | 1 | -1 | -1 | 6.7 | 30 | 3.4 | 5 |
| 9 | 0 | 0 | 0 | 9.8 | 24 | 3.8 | 10 |
| 10 | 0 | 0 | 0 | 10.2 | 24 | 3.8 | 10 |
| 11 | 0 | 0 | 0 | 9.8 | 24 | 3.8 | 10 |
| 12 | 0 | 0 | 0 | 10.2 | 24 | 3.8 | 10 |
| 13 | -1 | 1 | 1 | 8.2 | 18 | 4.2 | 15 |
| 14 | -1 | -1 | -1 | 7.1 | 18 | 3.4 | 5 |
| 15 | -1.682 | 0 | 0 | 8.2 | 13.908 | 3.8 | 10 |
| 16 | -1 | 1 | -1 | 5.6 | 18 | 4.2 | 5 |
| 17 | 1 | 1 | 1 | 7.5 | 30 | 4.2 | 15 |
| 18 | 0 | 1.682 | 0 | 6.7 | 24 | 4.4728 | 10 |
| 19 | 1 | -1 | 1 | 5.5 | 30 | 3.4 | 15 |
| 20 | 1.682 | 0 | 0 | 6.5 | 34.092 | 3.8 | 10 |

Table 6.6: Ethanol Percentage, Temperature, pH, Inoculum Content in Mango Wine Experiment

$$\hat{Y} = -88.088 + 1.012X_1 + 45.822X_2 - 0.036X_3 + 0.067X_1X_2 - 0.008X_1X_3 + 0.356X_2X_3 - 0.027X_1^2 - 6.763X_2^2 - 0.048X_3^2$$

$$\mathbf{x} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \qquad \hat{\boldsymbol{\beta}}_1 = \begin{bmatrix} 1.012243 \\ 45.8220 \\ -0.03629 \end{bmatrix} \qquad \hat{\boldsymbol{\beta}}_2 = \begin{bmatrix} -0.02662 & 0.067708/2 & -0.00792/2 \\ 0.067708/2 & -6.76258 & 0.35625/2 \\ -0.00792/2 & 0.35625/2 & -0.04823 \end{bmatrix}$$

$$\Rightarrow \quad \mathbf{x}^* = -\frac{1}{2}\hat{\boldsymbol{\beta}}_2^{-1}\hat{\boldsymbol{\beta}}_1 = \begin{bmatrix} 22.09 \\ 3.81 \\ 11.89 \end{bmatrix}$$

The R program and output are given below.

**R Program**

```
mango <- read.table("http://www.stat.ufl.edu/~winner/data/mangowine.dat",header=F,
   col.names=c("runnum","c_temp","c_pH","c_inoc","ethanol","glycerol","acidity",
   "one","temperature","pH","inoculum"))
attach(mango)

mango.1 <- lm(ethanol ~ temperature + pH + inoculum + I(temperature*pH) +
  I(temperature*inoculum) + I(pH*inoculum) + I(temperature^2) + I(pH^2) + I(inoculum^2))
summary(mango.1)
anova(mango.1)
```

```
library(rsm)

mango.2 <- rsm(ethanol ~ SO(temperature,pH,inoculum))
summary(mango.2)

par(mfrow=c(1,3))
contour(mango.2, ~ temperature + pH + inoculum, at=summary(mango.2)$canonical$xs)
}

{\bf R Output - lm Function}

{\footnotesize
\begin{verbatim}

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               -88.088035  12.223256  -7.207 2.90e-05 ***
temperature                 1.012243   0.283907   3.565 0.005135 **
pH                         45.821999   5.543945   8.265 8.84e-06 ***
inoculum                   -0.036288   0.320992  -0.113 0.912228
I(temperature * pH)         0.067708   0.062167   1.089 0.301644
I(temperature * inoculum)  -0.007917   0.004973  -1.592 0.142510
I(pH * inoculum)            0.356250   0.074600   4.775 0.000751 ***
I(temperature^2)           -0.026619   0.003087  -8.622 6.07e-06 ***
I(pH^2)                    -6.762575   0.694648  -9.735 2.03e-06 ***
I(inoculum^2)              -0.048229   0.004446 -10.848 7.50e-07 ***

Residual standard error: 0.422 on 10 degrees of freedom
Multiple R-squared:  0.9714,    Adjusted R-squared:  0.9457
F-statistic:  37.8 on 9 and 10 DF,  p-value: 1.491e-06

Analysis of Variance Table
Response: ethanol
                         Df  Sum Sq Mean Sq  F value    Pr(>F)
temperature               1  3.7528  3.7528  21.0730 0.0009944 ***
pH                        1  0.3257  0.3257   1.8290 0.2060372
inoculum                  1  9.0590  9.0590  50.8684 3.171e-05 ***
I(temperature * pH)       1  0.2113  0.2113   1.1862 0.3016437
I(temperature * inoculum) 1  0.4513  0.4513   2.5339 0.1425104
I(pH * inoculum)          1  4.0613  4.0613  22.8049 0.0007509 ***
I(temperature^2)          1  8.2896  8.2896  46.5483 4.611e-05 ***
I(pH^2)                   1 13.4792 13.4792  75.6892 5.607e-06 ***
I(inoculum^2)             1 20.9585 20.9585 117.6868 7.500e-07 ***
Residuals                10  1.7809  0.1781
```

Note that the sum of squares for the interaction terms given the main effects is obtained by adding the 3 interaction sequential sums of squares. Further, the sum of squares for the quadratic terms given main effects and interactions can also be obtained this way.

$$R\left(\beta_{12}, \beta_{13}, \beta_{23} | \beta_0, \beta_1, \beta_2, \beta_3\right) = 0.2113 + 0.4513 + 4.0613 = 4.7239$$

$$R\left(\beta_{11}, \beta_{22}, \beta_{33} | \beta_0, \beta_1, \beta_2, \beta_3, \beta_{12}, \beta_{13}, \beta_{23}\right) = 8.2896 + 13.4792 + 20.9585 = 42.7273$$

| $Y$ | $X_1$ | $X_2$ | $X_3$ | $\overline{Y}_{\text{Grp}}$ | $\hat{Y}$ | $Y - \overline{Y}_{\text{Grp}}$ | $\overline{Y}_{\text{Grp}} - \hat{Y}$ |
|---|---|---|---|---|---|---|---|
| 4.8 | 24 | 3.8 | 1.59 | 4.8 | 5.190776 | 0 | -0.39078 |
| 9.6 | 24 | 3.8 | 10 | 9.966667 | 9.971758 | -0.36667 | -0.00509 |
| 10.2 | 24 | 3.8 | 10 | 9.966667 | 9.971758 | 0.233333 | -0.00509 |
| 8.5 | 24 | 3.8 | 18.41 | 8.5 | 7.930448 | 0 | 0.569552 |
| 7.3 | 24 | 3.1272 | 10 | 7.3 | 7.170358 | 0 | 0.129642 |
| 4.8 | 30 | 4.2 | 5 | 4.8 | 4.920205 | 0 | -0.12021 |
| 7.9 | 18 | 3.4 | 15 | 7.9 | 7.90624 | 0 | -0.00624 |
| 6.7 | 30 | 3.4 | 5 | 6.7 | 6.329059 | 0 | 0.370941 |
| 9.8 | 24 | 3.8 | 10 | 9.966667 | 9.971758 | -0.16667 | -0.00509 |
| 10.2 | 24 | 3.8 | 10 | 9.966667 | 9.971758 | 0.233333 | -0.00509 |
| 9.8 | 24 | 3.8 | 10 | 9.966667 | 9.971758 | -0.16667 | -0.00509 |
| 10.2 | 24 | 3.8 | 10 | 9.966667 | 9.971758 | 0.233333 | -0.00509 |
| 8.2 | 18 | 4.2 | 15 | 8.2 | 8.697386 | 0 | -0.49739 |
| 7.1 | 18 | 3.4 | 5 | 7.1 | 7.227422 | 0 | -0.12742 |
| 8.2 | 13.908 | 3.8 | 10 | 8.2 | 8.142285 | 0 | 0.057715 |
| 5.6 | 18 | 4.2 | 5 | 5.6 | 5.168568 | 0 | 0.431432 |
| 7.5 | 30 | 4.2 | 15 | 7.5 | 7.499023 | 0 | 0.000977 |
| 6.7 | 24 | 4.4728 | 10 | 6.7 | 6.650866 | 0 | 0.049134 |
| 5.5 | 30 | 3.4 | 15 | 5.5 | 6.057877 | 0 | -0.55788 |
| 6.5 | 34.092 | 3.8 | 10 | 6.5 | 6.378939 | 0 | 0.121061 |
| | | | | | Sum of Squares | 0.353333 | 1.427535 |

Table 6.7: Mango Wine Experiment - Lack-of-Fit Test Calculations

The Goodness-of-Fit test can be computed as follows. The error degrees of freedom is $n-p' = 20-10 = 10$. There are $c = 15$ distinct groupings (6 at the center point, 14 individual combinations of factor levels). That leaves $n - c = 20 - 15 = 5$ degrees of freedom for Pure Error, and $c - p' = 15 - 10 = 5$ degrees of freedom for Lack of Fit. The "group means" for the 14 individual points is the observation itself, the group mean at the center value is the average of those 6 cases. Computations are given in Table 6.7.

$$F_{\text{LOF}} = \frac{MSLF}{MSPE} = \frac{1.427535/5}{0.353333/5} = 4.040 \qquad F_{.05,5,5} = 5.050 \qquad P\left(F_{5,5} \geq 4.040\right) = .0758$$

We fail to conclude that the model does not fit the data.

The output from the R package **rsm** are given below. Figure 6.5 gives the contours of the response surface for all three pairs of factors.

**R Output - rsm package**

```
Call:
rsm(formula = ethanol ~ SO(temperature, pH, inoculum))
                      Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)         -88.0880353  12.2232556  -7.2066 2.903e-05 ***
temperature           1.0122431   0.2839067   3.5654 0.0051348 **
pH                   45.8219993   5.5439454   8.2652 8.839e-06 ***
inoculum             -0.0362881   0.3209922  -0.1130 0.9122284
temperature:pH        0.0677083   0.0621670   1.0891 0.3016437
temperature:inoculum -0.0079167   0.0049734  -1.5918 0.1425104
```

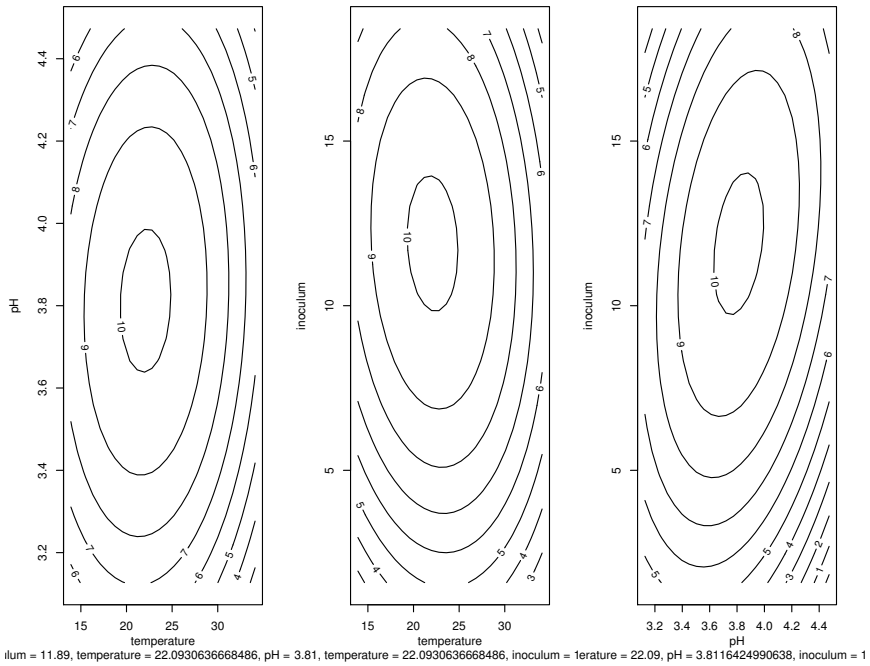


Figure 6.5: Contour Plots of Ethanol for all Pairs of predictors

```
pH:inoculum             0.3562500   0.0746003   4.7754 0.0007509 ***
temperature^2          -0.0266194   0.0030873  -8.6222 6.075e-06 ***
pH^2                   -6.7625745   0.6946475  -9.7353 2.032e-06 ***
inoculum^2             -0.0482290   0.0044457 -10.8484 7.500e-07 ***

Multiple R-squared:  0.9714,    Adjusted R-squared:  0.9457
F-statistic:  37.8 on 9 and 10 DF,  p-value: 1.491e-06

Analysis of Variance Table
Response: ethanol
                             Df Sum Sq Mean Sq F value    Pr(>F)
FO(temperature, pH, inoculum)   3 13.138  4.3792 24.5901 6.227e-05
TWI(temperature, pH, inoculum)  3  4.724  1.5746  8.8417  0.003657
PQ(temperature, pH, inoculum)   3 42.727 14.2424 79.9748 2.730e-07
Residuals                      10  1.781  0.1781
Lack of fit                     5  1.428  0.2855  4.0402  0.075809
Pure error                      5  0.353  0.0707

Stationary point of response surface:
temperature          pH    inoculum
  22.093064    3.811642   11.888138
```

$\nabla$

## 6.9 Trigonometric Models

Many biological, economic, and meteorological time series of measurements display cyclic patterns, along with possibly trend. Use of regression models with sine and cosine components, along with a trend term, can be used to model the process. If there are an even number of $k$ measurements per period, we can include up to $k/2$ cosine and $(k/2)-1$ sine terms before they start repeating themselves. For monthly ($k=12$), we have the following model, with centered time.

$$\text{Monthly: } Y_t = \beta_0 + \sum_{j=1}^{5}\left[\beta_j^C \cos\left(\frac{2\pi jt}{12}\right) + \beta_j^S \sin\left(\frac{2\pi jt}{12}\right)\right] + \beta_6^C \cos(\pi t) + \beta_7\left(t-\overline{t}\right) + \epsilon_t \qquad t = 1,\ldots,n$$

**Example: Minneapolis/St.Paul Mean Monthly Temperature - 1/1900-12/2014**

This model is fit to monthl mean temperature for Minneapolis/St. Paul, Minnesota over the $n = 1380$ months from 1/1900-12/2014. After fitting the full model, we see that we can drop the two highest order cosine and the highest order sine term. We then include a plot of the data and the fitted curve for 1/1900-12/1904 in Figure 6.6. A plot pf the full series is so long that you cannot see the pattern. The R program and (partial) output are given below.

```
### Program
msw <- read.csv("E:\\coursenotes\\minn_stp_weather.csv",header=T)
attach(msw); names(msw)

X1 <- cos(2*pi*1*MonthS/12); X2 <- sin(2*pi*1*MonthS/12)
X3 <- cos(2*pi*2*MonthS/12); X4 <- sin(2*pi*2*MonthS/12)
X5 <- cos(2*pi*3*MonthS/12); X6 <- sin(2*pi*3*MonthS/12)
X7 <- cos(2*pi*4*MonthS/12); X8 <- sin(2*pi*4*MonthS/12)
X9 <- cos(2*pi*5*MonthS/12); X10 <- sin(2*pi*5*MonthS/12)
X11 <- cos(2*pi*6*MonthS/12)
X12 <- MonthS - mean(MonthS)

msw.mod1 <- lm(meanTemp~X1+X2+X3+X4+X5+X6+X7+X8+X9+X10+X11+X12)
summary(msw.mod1)
anova(msw.mod1)

msw.mod2 <- lm(meanTemp~X1+X2+X3+X4+X5+X6+X7+X12)
summary(msw.mod2)
anova(msw.mod2)
plot(msw.mod2)
anova(msw.mod2,msw.mod1)

plot(MonthS[1:60],meanTemp[1:60],pch=16,ylim=c(0,80),
main="Minneapolis-St.Paul Mean Temp1/1900-12/1904")
lines(1:60,predict(msw.mod2)[1:60])

### Output
> summary(msw.mod1)
Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept)  45.269467   0.118994  380.436  < 2e-16 ***
X1          -24.507707   0.168283 -145.634  < 2e-16 ***
X2          -16.183906   0.168283  -96.171  < 2e-16 ***
```

```
X3           -1.272425   0.168343   -7.559 7.41e-14 ***
X4           -1.715098   0.168221  -10.195  < 2e-16 ***
X5            0.162471   0.168283    0.965   0.3345
X6           -0.675480   0.168283   -4.014 6.29e-05 ***
X7           -0.278256   0.168343   -1.653   0.0986 .
X8            0.338942   0.168222    2.015   0.0441 *
X9           -0.008014   0.168282   -0.048   0.9620
X10           0.115859   0.168282    0.688   0.4913
X11           0.078679   0.118994    0.661   0.5086
X12           0.001248   0.000297    4.202 2.81e-05 ***

Residual standard error: 4.433 on 1375 degrees of freedom
Multiple R-squared:  0.9571,    Adjusted R-squared:  0.9568
F-statistic:  2558 on 12 and 1375 DF,  p-value: < 2.2e-16

> summary(msw.mod2)
Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept)  4.527e+01  1.190e-01  380.304  < 2e-16 ***
X1          -2.451e+01  1.683e-01 -145.583  < 2e-16 ***
X2          -1.618e+01  1.683e-01  -96.137  < 2e-16 ***
X3          -1.272e+00  1.684e-01   -7.554 7.63e-14 ***
X4          -1.715e+00  1.683e-01  -10.192  < 2e-16 ***
X5           1.630e-01  1.683e-01    0.969    0.333
X6          -6.758e-01  1.683e-01   -4.015 6.28e-05 ***
X7          -2.780e-01  1.684e-01   -1.651    0.099 .
X12          1.248e-03  2.971e-04    4.200 2.84e-05 ***

Residual standard error: 4.435 on 1379 degrees of freedom
Multiple R-squared:  0.957,     Adjusted R-squared:  0.9567
F-statistic:  3834 on 8 and 1379 DF,  p-value: < 2.2e-16

> anova(msw.mod2,msw.mod1)
Analysis of Variance Table

Model 1: meanTemp ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X12
Model 2: meanTemp ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 +
    X11 + X12
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1   1379 27120
2   1375 27023  4     97.52 1.2405 0.2918
```

The trend coefficient is significant ($\hat{\beta} = .001248$, $P < .0001$). Above and beyond the cyclic behavior of the series, temperature has increased by (1380-1)(.001248)=1.72 degrees Fahrenheit on average over the period.

$\nabla$

## 6.10 Model Building

When we have many predictors, we may wish to use an algorithm to determine which variables to include in the model. These variables can be main effects, interactions, and polynomial terms. Note that there are two common approaches. One method involves testing variables based on $t$-tests, or equivalently $F$-tests for partial regression coefficients. An alternative method involves comparing models based on model based

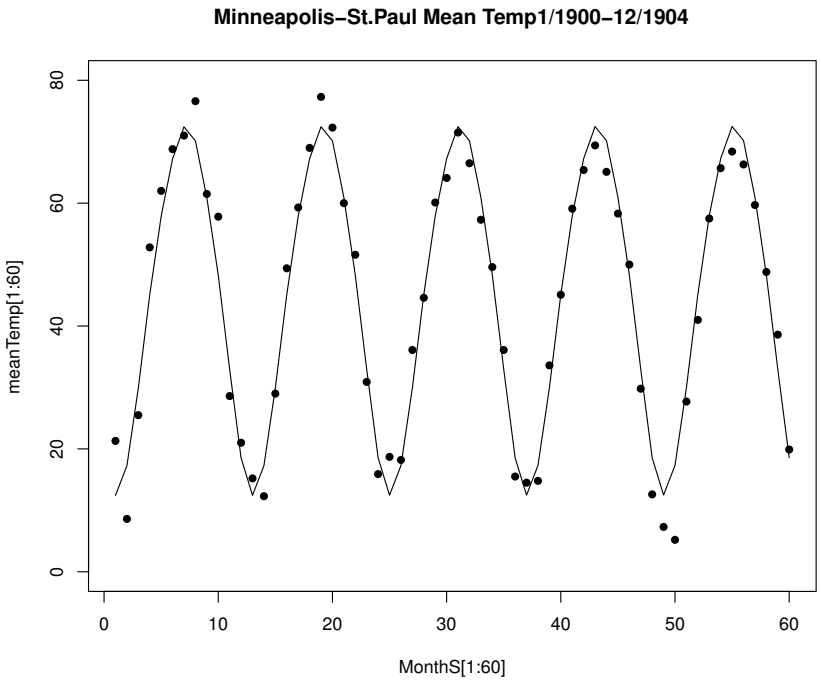


Figure 6.6: Minneapolis/St. Paul Temperatures and Fitted values 1/1900-12/1904

measures, such as Akaike Information Criterion ($AIC$), or Schwartz Bayesian Information criterion ($BIC$ or $SBC$). These measures can be written as follows (note that different software packages print different versions, as some parts are constant for all potential models). The goal is to minimize the measures.

$$AIC(\text{Model}) = n\ln(SSE(\text{Model}))+2p'-n\ln(n) \qquad BIC(\text{Model}) = n\ln(SSE(\text{Model}))+[\ln(n)]p'-n\ln(n)$$

Note that $SSE$(Model) depends on the variables included in the current model. The measures put a penalty on excess predictor variables, with BIC placing a higher penalty when $\ln(n) > 2$. Note that $p'$ is the number of parameters in the model (including the intercept), and $n$ is the sample size. Be aware that different computer packages can use different formulations of $AIC$ and $BIC$, unless clearly stated, the goal is still to minimize the values.

**Example: Time-to-Incapacitation for Animals Exposed to Burning Aircraft Materials**

Spurgeon (1978) reports results from experiments of animals being exposed to $n = 71$ burning aircraft materials. The outcome measured was the time-to-incapacitation. Due to theoretical reasons the response $Y$ is the reciprocal of time-to-incapacitation, multiplied by 1000. For each burned material, levels of $p = 7$ gases were measured: CO, HCN, H2S, HCL, HBr, NO2, and SO2. These were the potential predictors of $Y$. The R program and output for the full model, containing all predictors main effects (no interactions or quadratic terms are included) are given below.

### R Program

```
toxic <- read.table("http://www.stat.ufl.edu/~winner/data/air_int_incap.dat",
header=F,col.names=c("matcat","matid","timeinc","Y","CO","HCN","H2S",
    "HCL","HBr","NO2","SO2"))
attach(toxic)

library(car)

toxic.1 <- lm(Y ~ CO + HCN + H2S + HCL + HBr + NO2 + SO2)
summary(toxic.1)
anova(toxic.1)
drop1(toxic.1,test="F")
vif(toxic.1)

######### Perform Backward Elimination, Forward Selection, and Stepwise Regression
######### Based on Model AIC (not individual regression coefficients)
######### fit1 and fit2 represent "extreme" models
library(MASS)
fit1 <- lm(Y ~ CO + HCN + H2S + HCL + HBr + NO2 + SO2)
fit2 <- lm(Y ~ 1)
stepAIC(fit1,direction="backward")
stepAIC(fit2,direction="forward",scope=list(upper=fit1,lower=fit2))
stepAIC(fit2,direction="both",scope=list(upper=fit1,lower=fit2))

########## Perform all possible regressions (aka all subset regressions)
########## Prints out best 4 models of each # of predictors
install.packages("leaps")
library(leaps)

alltoxic <- regsubsets(Y ~ CO + HCN + H2S + HCL + HBr + NO2 + SO2,
 nbest=4,data=toxic)
aprout <- summary(alltoxic)
n <- length(toxic$Y)
pprime <- apply(aprout$which, 1, sum)
aprout$aic <- aprout$bic - log(n) * pprime + 2 * pprime
with(aprout,round(cbind(which,rsq,adjr2,cp,bic,aic),3))     ## Prints "readable" results
plot(alltoxic,scale="bic")
plot(alltoxic,scale="adjr2")
```

### R Output - Full Model

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  75.44679   15.16710   4.974 5.34e-06 ***
CO            0.64036    0.09717   6.590 1.03e-08 ***
HCN          11.77871    0.90006  13.087  < 2e-16 ***
H2S         -10.90998    2.99044  -3.648 0.000537 ***
HCL          -0.10200    0.07284  -1.400 0.166337
HBr          -0.67569    0.88126  -0.767 0.446105
NO2          44.81628   21.06666   2.127 0.037312 *
SO2           6.61409    2.72454   2.428 0.018069 *

Residual standard error: 55.75 on 63 degrees of freedom
Multiple R-squared:  0.8345,    Adjusted R-squared:  0.8161
F-statistic: 45.38 on 7 and 63 DF,  p-value: < 2.2e-16

> anova(toxic.1)
Analysis of Variance Table (Sequential SS)
Response: Y
```

```
          Df Sum Sq Mean Sq  F value    Pr(>F)
CO         1  81119   81119  26.0966 3.243e-06 ***
HCN        1 787782  787782 253.4359 < 2.2e-16 ***
H2S        1  68561   68561  22.0567 1.478e-05 ***
HCL        1   5208    5208   1.6755  0.200254
HBr        1   2298    2298   0.7392  0.393174
NO2        1  24125   24125   7.7611  0.007044 **
SO2        1  18319   18319   5.8932  0.018069 *
Residuals 63 195830    3108

> drop1(toxic.1,test="F")
Single term deletions
Model:
Y ~ CO + HCN + H2S + HCL + HBr + NO2 + SO2
       Df Sum of Sq    RSS    AIC  F value    Pr(>F)
<none>              195830 578.48
CO      1    135010 330840 613.72  43.4338 1.028e-08 ***
HCN     1    532345 728174 669.73 171.2595 < 2.2e-16 ***
H2S     1     41373 237203 590.09  13.3100 0.0005374 ***
HCL     1      6095 201925 578.66   1.9608 0.1663369
HBr     1      1827 197657 577.14   0.5879 0.4461054
NO2     1     14068 209897 581.41   4.5256 0.0373116 *
SO2     1     18319 214148 582.83   5.8932 0.0180690 *
```

Full Model: $SSE = 195830 \quad p' = 7 + 1 = 8$ Null Model: $TSS = 1183241 \quad p' = 0 + 1 = 1$

Full Model: $AIC = 71\ln(195830) + 2(8) - 71\ln(71) = 865.135 + 16 - 302.650 = 578.485$

Full Model: $BIC = 71\ln(195830) + [\ln(71)](8) - 71\ln(71) = 865.135 + 34.101 - 302.650 = 596.586$

Null Model: $AIC = 71\ln(1183241) + 2(1) - 71\ln(71) = 992.848 + 2 - 302.650 = 692.198$

Null Model: $BIC = 71\ln(1183241) + [\ln(71)](1) - 71\ln(71) = 992.848 + 4.263 - 302.650 = 694.461$

$\nabla$

### 6.10.1 Backward Elimination

This is a "top-down" method, which begins with a "Complete" Model, with all potential predictors. The analyst then chooses a significance level to stay in the model (SLS). The model is fit, and the predictor with the lowest $t$-statistic in absolute value (largest $P$-value) is identified. If the $P$-value is larger than SLS, the variable is dropped from the model. Then the model is re-fit with all other predictors (this will change all regression coefficients, standard errors, and $P$-values). The process continues until all variables have $P$-values below SLS.

The model based approach fits the full model, with all predictors and computes $AIC$ (or $BIC$). Then, each variable is dropped one-at-a-time, and $AIC$ (or $BIC$) is obtained for each model. If none of the models with one dropped variable has $AIC$ (or $BIC$) below that for the full model, the full model is kept, otherwise the model with the lowest $AIC$ (or $BIC$) is kept as the new full model. The process continues until no variables should be dropped (none of the "drop one variable models" has a lower $AIC$ (or $BIC$) than the "full model."

**Example: Time-to-Incapacitation for Animals Exposed to Burning Aircraft Materials**

We make use of the **stepAIC** command in the R package **MASS** to use $AIC$ to obtain the Backward Elimination model selection. Note that in the output, $RSS$ is the Residual (Error) Sum of Squares.

**R Output - Backward Elimination**

```
> library(MASS)
> fit1 <- lm(Y ~ CO + HCN + H2S + HCL + HBr + NO2 + SO2)
> fit2 <- lm(Y ~ 1)
> stepAIC(fit1,direction="backward")
Start:  AIC=578.48
Y ~ CO + HCN + H2S + HCL + HBr + NO2 + SO2
       Df Sum of Sq    RSS    AIC
- HBr   1      1827 197657 577.14
<none>              195830 578.48
- HCL   1      6095 201925 578.66
- NO2   1     14068 209897 581.41
- SO2   1     18319 214148 582.83
- H2S   1     41373 237203 590.09
- CO    1    135010 330840 613.72
- HCN   1    532345 728174 669.73

Step:  AIC=577.14
Y ~ CO + HCN + H2S + HCL + NO2 + SO2
       Df Sum of Sq    RSS    AIC
- HCL   1      5281 202938 577.02
<none>              197657 577.14
- NO2   1     15112 212769 580.38
- SO2   1     17751 215408 581.25
- H2S   1     42033 239690 588.83
- CO    1    154961 352618 616.24
- HCN   1    531912 729569 667.86

Step:  AIC=577.02
Y ~ CO + HCN + H2S + NO2 + SO2
       Df Sum of Sq    RSS    AIC
<none>              202938 577.02
- SO2   1     13934 216872 579.73
- NO2   1     21378 224316 582.13
- H2S   1     38630 241568 587.39
- CO    1    166472 369410 617.55
- HCN   1    529032 731971 666.10

Call: lm(formula = Y ~ CO + HCN + H2S + NO2 + SO2)
Coefficients:
(Intercept)           CO          HCN          H2S          NO2          SO2
    68.4756       0.6211      11.7355     -10.4239      53.2167       5.5480
```

In the first step, if HBr is removed, $SSE = 197657$, $p' = 6 + 1 = 7$ and $AIC$ drops to 577.14. When the model drops each of the remaining 6 predictors, one-at-a-time, $AIC$ increases. In the second step, we begin with the 6 predictor model, and drop each variable, one-at-a-time. The model has $AIC = 577.14$. When HCL is dropped from the model, $SSE = 202938$, $p' = 5 + 1 = 6$ and $AIC$ drops to 577.02. In the third step, we begin with the 5 predictor model, and drop each variable, one-at-a-time. In each case $AIC$ decreases, and we select the 5 predictor model: CO, HCN, H2S, NO2, SO2.

$\nabla$

### 6.10.2 Forward Selection

This is a "bottom-up" method, which begins with all "Simple" Models, each with one predictor. The analyst then chooses a significance level to enter into the model (SLE). Each model is fit, and the predictor with the highest $t$-statistic in absolute value (smallest $P$-value) is identified. If the $P$-value is smaller than SLE, the variable is entered into the model. Then all two variable models including the best predictor in the first round, with each of the other predictors. The best second variable is identified, and its $P$-value is compared with SLE. If its $P$-value is below SLE, the variable is added to the model. The process continues until no potential added variables have $P$-values below SLE.

The model based approach fits each simple model, with one predictor and computes $AIC$ (or $BIC$). The best variable is identified (assuming its $AIC$ (or $BIC$) is smaller than that for the null model, with no predictors). Then, each potential variable is added one-at-a-time, and $AIC$ (or $BIC$) is obtained for each model. If none of the models with one added variable has $AIC$ (or $BIC$) below that for the best simple model, the simple model is kept, otherwise the model with the lowest $AIC$ (or $BIC$) is kept as the new full model. The process continues until no variables should be added (none of the "add one variable models" has a lower $AIC$ (or $BIC$) than the "reduced model."

**Example: Time-to-Incapacitation for Animals Exposed to Burning Aircraft Materials**

We make use of the **stepAIC** command in the R package **MASS** to use $AIC$ to obtain the Forward Selection model choice. Note that in the output, $RSS$ is the Residual (Error) Sum of Squares for that model.

**R Output - Forward Selection**

```
> library(MASS)
> fit1 <- lm(Y ~ CO + HCN + H2S + HCL + HBr + NO2 + SO2)
> fit2 <- lm(Y ~ 1)
> stepAIC(fit2,direction="forward",scope=list(upper=fit1,lower=fit2))
Start:  AIC=692.2
Y ~ 1
       Df Sum of Sq     RSS    AIC
+ HCN   1    677833  505408 633.80
+ NO2   1    212831  970410 680.12
+ H2S   1    102071 1081170 687.79
+ CO    1     81119 1102122 689.15
+ HCL   1     51156 1132086 691.06
<none>              1183241 692.20
+ HBr   1     14339 1168902 693.33
+ SO2   1     13344 1169897 693.39

Step:  AIC=633.8
Y ~ HCN
       Df Sum of Sq    RSS    AIC
+ CO    1    191068 314340 602.08
+ NO2   1    109591 395817 618.45
+ H2S   1     73868 431541 624.58
+ HCL   1     28118 477291 631.74
+ HBr   1     25890 479518 632.07
<none>              505408 633.80
+ SO2   1       803 504606 635.69

Step:  AIC=602.08
```

```
Y ~ HCN + CO
       Df Sum of Sq    RSS    AIC
+ H2S   1     68561 245779 586.62
+ NO2   1     65161 249179 587.59
+ SO2   1     14129 300211 600.82
<none>              314340 602.08
+ HCL   1      4387 309953 603.09
+ HBr   1      3701 310639 603.24

Step:  AIC=586.62
Y ~ HCN + CO + H2S
       Df Sum of Sq    RSS    AIC
+ NO2   1   28906.7 216872 579.73
+ SO2   1   21462.7 224316 582.13
<none>              245779 586.62
+ HCL   1    5208.0 240571 587.09
+ HBr   1    1509.8 244269 588.18

Step:  AIC=579.73
Y ~ HCN + CO + H2S + NO2
       Df Sum of Sq    RSS    AIC
+ SO2   1   13933.9 202938 577.02
<none>              216872 579.73
+ HCL   1    1463.9 215408 581.25
+ HBr   1     901.2 215971 581.44

Step:  AIC=577.02
Y ~ HCN + CO + H2S + NO2 + SO2
       Df Sum of Sq    RSS    AIC
<none>              202938 577.02
+ HCL   1    5280.9 197657 577.14
+ HBr   1    1013.4 201925 578.66

Call: lm(formula = Y ~ HCN + CO + H2S + NO2 + SO2)

Coefficients:
(Intercept)          HCN           CO          H2S          NO2          SO2
    68.4756      11.7355       0.6211     -10.4239      53.2167       5.5480
```

In the first stage, each one variable model is compared to the null model. The best single variable is HCN, with $AIC = 633.80$, compared to the null model with $AIC = 692.20$. In the second stage, each two variable model containing HCN is fit, and CO drops $AIC$ to 602.02. The best three variable model containing HCN and CO adds H2S, with $AIC = 586.62$. In the fourth stage, NO2 is added, with $AIC = 579.93$, in the fifth stage, SO2 is added with $AIC = 577.02$. When either HCL or HBr is added, $AIC$ increases.

$\nabla$

### 6.10.3 Stepwise Regression

This approach is a hybrid of forward selection and backward elimination. It begins like forward selection, but then applies backward elimination at each step. In forward selection, once a variable is entered, it stays in the model. In stepwise regression, once a new variable is entered, all previously entered variables are tested, to confirm they should stay in the model, after controlling for the new entrant, as well as the other previous entrant.

**Example: Time-to-Incapacitation for Animals Exposed to Burning Aircraft Materials**

The Stepwise Regression leads to the same model as Backward Elimination and Forward selection for this data. The output is virtually the same as Forward selection, with the exception that it tries to remove previously entered predictors at each step.

$\nabla$

### 6.10.4 All Possible Regressions

We can fit all possible regression models, and use model based measures to choose the "best" model. Commonly used measures are: Adjusted-$R^2$ (equivalently $MSE$), Mallow's $C_p$ statistic, $AIC$, and $BIC$. The formulas, and decision criteria are given below (where $p'$ is the number of parameters in the "current" model being fit:

**Adjusted-$R^2$** - $1 - \left(\frac{n-1}{n-p'}\right)\frac{SSE}{TSS}$ - Goal is to maximize

**Mallow's $C_p$** - $C_p = \frac{SSE(\text{Model})}{MSE(\text{Complete})} + 2p' - n$ - Goal is to have $C_p \leq p'$

**Akaike Information Criterion** - $AIC(\text{Model}) = n\ln(SSE(\text{Model})) + 2p' - n\ln(n)$ - Goal is to minimize

**Bayesian Information Criterion** - $BIC(\text{Model}) = n\ln(SSE(\text{Model})) + [\ln(n)]p' - n\ln(n)$ - Goal is to minimize

**Example: Time-to-Incapacitation for Animals Exposed to Burning Aircraft Materials**

The R package **leaps** will perform all possible regressions. It does not include $AIC$, however i can be computed from BIC. The following output gives the best four $p = 1, 2, 3, 4, 5$, and 6 variable models, as well as the 7 variable model.

```
> alltoxic <- regsubsets(Y ~ CO + HCN + H2S + HCL + HBr + NO2 + SO2,
+  nbest=4,data=toxic)
> aprout <- summary(alltoxic)
> n <- length(toxic$Y)
> pprime <- apply(aprout$which, 1, sum)
> aprout$aic <- aprout$bic - log(n) * pprime + 2 * pprime
> with(aprout,round(cbind(which,rsq,adjr2,cp,bic,aic),3))     ## Prints "readable" results
  (Intercept) CO HCN H2S HCL HBr NO2 SO2   rsq adjr2      cp     bic      aic
1           1  0   1   0   0   0   0   0 0.573 0.567  95.594 -51.871  -56.396
1           1  0   0   0   0   0   1   0 0.180 0.168 245.189  -5.554  -10.079
1           1  0   0   1   0   0   0   0 0.086 0.073 280.821   2.120   -2.405
1           1  1   0   0   0   0   0   0 0.069 0.055 287.562   3.483   -1.042
2           1  1   1   0   0   0   0   0 0.734 0.727  36.126 -81.325  -88.113
2           1  0   1   0   0   0   1   0 0.665 0.656  62.338 -64.961  -71.749
2           1  0   1   1   0   0   0   0 0.635 0.625  73.830 -58.826  -65.614
2           1  0   1   0   1   0   0   0 0.597 0.585  88.548 -51.672  -58.460
```

```
3          1  1   1   1   0   0   0   0 0.792 0.783  16.069 -94.532 -103.582
3          1  1   1   0   0   0   1   0 0.789 0.780  17.163 -93.556 -102.607
3          1  1   1   0   0   0   0   1 0.746 0.735  33.580 -80.328  -89.378
3          1  1   1   0   1   0   0   0 0.738 0.726  36.714 -78.060  -87.111
4          1  1   1   1   0   0   1   0 0.817 0.806   8.769 -99.153 -110.466
4          1  1   1   1   0   0   0   1 0.810 0.799  11.164 -96.757 -108.070
4          1  1   1   1   1   0   0   0 0.797 0.784  16.394 -91.790 -103.103
4          1  1   1   0   0   0   1   1 0.796 0.783  16.714 -91.496 -102.809
5          1  1   1   1   0   0   1   1 0.828 0.815   6.287 -99.605 -113.181
5          1  1   1   1   1   0   0   1 0.820 0.806   9.450 -96.246 -109.822
5          1  1   1   1   1   0   1   0 0.818 0.804  10.298 -95.371 -108.947
5          1  1   1   1   0   1   1   0 0.817 0.803  10.480 -95.186 -108.762
6          1  1   1   1   1   0   1   1 0.833 0.817   6.588 -97.214 -113.053
6          1  1   1   1   0   1   1   1 0.829 0.813   7.961 -95.698 -111.536
6          1  1   1   1   1   1   0   1 0.823 0.806  10.526 -92.948 -108.787
6          1  1   1   1   1   1   1   0 0.819 0.802  11.893 -91.525 -107.363
7          1  1   1   1   1   1   1   1 0.834 0.816   8.000 -93.611 -111.712
```

$\nabla$

## 6.11 Issues of Collinearity

When the predictor variables are highly correlated among themselves, the regression coefficients become unstable, with increased standard errors. This leads to smaller $t$-statistics for tests regarding the partial regression coefficients and wider confidence intervals. At its most extreme case, the sign of a regression coefficient can change when a new predictor variable is included. One widely reported measure of collinearity is the **Variance Inflation Factor (VIF)**. This is computed for each predictor variable, by regressing it on the remaining $p-1$ predictors. Then $VIF_J = \frac{1}{1-R_j^2}$ where $R_j^2$ is the coefficent of determination of the regression of $X_j$ on the remaining predictors. Values of $VIF_j$ greater than 10 are considered problematic. Collinearity is not problematic when the primary goal of the model is for prediction.

Various remedies exist. One is determining which variable(s) make the most sense theoretically for the model, and removing other variables, which are correlated with the other more meaningful predictors. A second method involves generating uncorrelated predictor variables from the original set of predictors. While this method based on **principal components** removes the collinearity problem, the new variables may lose their meaning, thus making it harder to describe the process. A third method, **ridge regression**, introduces a bias factor into the regression that reduces the inflated variance due to collinearity, and through that reduces the Mean Square Error of the regression coefficients. Unfortunately, there is no simple rule on choosing the bias factor.

**Example: Time-to-Incapacitation for Animals Exposed to Burning Aircraft Materials**

The R package **car** has a **vif** function that computes the Variance Inflation Factors for each predictor. We apply it to the regression fit on all $p = 7$ predictors, and find that the gases in their functional forms are not highly correlated among themselves.

**R Output - Variance Inflation Factors**

```
library(car)
toxic.1 <- lm(Y ~ CO + HCN + H2S + HCL + HBr + NO2 + SO2)
vif(toxic.1)
```

```
> vif(toxic.1)
      CO      HCN      H2S      HCL      HBr      NO2      SO2
1.483027 2.043812 2.072767 1.223480 1.359955 1.374012 1.189694
```

$\nabla$

### 6.11.1 Principal Components Regression

For **principal components regression**, if we have $p$ predictors: $X_1, \ldots, X_p$, we can generate $p$ linearly independent predictors that are linear functions of $X_1, \ldots, X_p$. When the new variables with small eigenvalues are removed, the estimate of $\boldsymbol{\beta}$ obtained from the new regression is biased. The amount of bias depends on the relative size of the eigenvalues of the removed principal components, however the collinearity problem will be removed and the variance of the estimator will have been reduced. The process is conducted as follows (see e.g. Rawlings, Pantula, and Dickey (1998), Section 13.2.2).

1. Create $Z_1, \ldots Z_p$ from the original variables $X_1, \ldots, X_p$ by subtracting the mean and dividing by a multiple of the standard deviation.

$$Z_{ij} = \frac{X_{ij} - \overline{X}_j}{\sqrt{n-1}s_j} \quad i = 1, ..., n; j = 1, ..., p \qquad \overline{X}_j = \frac{\sum_{i=1}^n X_{ij}}{n} \qquad s_j = \sqrt{\frac{\sum_{i=1}^n \left(X_{ij} - \overline{X}_j\right)^2}{n-1}}$$

2. Obtain the eigenvalues $\lambda_1, \ldots, \lambda_p$ (and place in a diagonal matrix $\mathbf{L}$) and eigenvectors (as columns in matrix $\mathbf{V}$) of the $p \times p$ matrix $\mathbf{R}_{XX} = \mathbf{Z}'\mathbf{Z}$, where $\mathbf{R}_{XX}$ is the correlation matrix among the predictor variables $X_1, \ldots, X_p$. These can be obtained in any matrix computer package ($\mathbf{Z}$ does not contain a column for an intercept).

$$\mathbf{R} = \mathbf{Z}'\mathbf{Z} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{12} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & \cdots & 1 \end{bmatrix} \qquad \mathbf{Z}'\mathbf{Z} = \mathbf{V}\mathbf{L}\mathbf{V}' = \sum_{i=1}^p \lambda_i \left(\mathbf{v_i}\mathbf{v_i'}\right)$$

$$\text{where:} \quad \mathbf{Z}_{ij} = \frac{X_{ij} - \overline{X}_j}{\sqrt{n-1}s_j} \qquad \overline{X}_j = \frac{\sum_{i=1}^n X_{ij}}{n} \qquad s_j^2 = \frac{\sum_{i=1}^n \left(X_{ij} - \overline{X}_j\right)}{n-1}$$

3. The Condition Index for the $j^{th}$ principal component is $\sqrt{\frac{\lambda \max}{\lambda_j}}$. Values between 30 and 100 imply moderate to large depencies among the predictors. Values over 100 are signs of serious collinearity problems.

4. Create the matrix of principal components $\mathbf{W} = \mathbf{ZV}$ and augment it with a column of $1^s$ for the intercept: $\mathbf{W}^* = [\mathbf{1}|\mathbf{W}]$.

5. Fit the regression $\mathbf{Y} = \mathbf{W}^*\boldsymbol{\gamma}$ and obtain $SSR(\hat{\gamma}_j)$, the partial sum of squares for each generated predictor variable (principal component):

$$\hat{\boldsymbol{\gamma}} = (\mathbf{W}^{*'}\mathbf{W}^*)^{-1}\mathbf{W}^{*'}\mathbf{Y} \qquad \hat{V}\{\hat{\boldsymbol{\gamma}}\} = s^2(\mathbf{W}^{*'}\mathbf{W}^*)^{-1}$$

6. For each generated predictor, test $H_0 : \gamma_j = 0$, based on the $t$-test or $F$-test. Eliminate any principal components with high $VIF$ and do not have significant coefficients.

7. Let $\hat{\boldsymbol{\gamma}}_{(g)}$ be the vector of retained coefficients from previous part. Then $SSR_{PC} = \sum SSR(\hat{\gamma}_j)$, with $g$ degrees of freedom (the number of retained principal components (generated predictors)).

8. Scaling back to the original variables (in their standardized (mean=0, standard deviation=1) format), we get: $\hat{\boldsymbol{\beta}}_g^{PC} = \mathbf{V}_{(g)}\hat{\boldsymbol{\gamma}}_{(g)}$where $\mathbf{V}_{(g)}$ is the $p \times g$ portion of the eigenvector matrix (columns) corresponding to the retained principal components. This ignores the intercept in

9. The estimated variance-covariance matrix of $\hat{\boldsymbol{\beta}}_g^{PC}$ is:

$$\hat{V}\left\{\hat{\boldsymbol{\beta}}_g^{PC}\right\} = s^2\mathbf{V}_{(g)}\mathbf{L}_{(g)}^{-1}\mathbf{V}'_{(g)} \qquad \mathbf{L}_{(g)} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_g \end{bmatrix}$$

10. The fitted regression equation can be written with respect to the original (standardized) variables, or the principal components:

$$\hat{\mathbf{Y}}_{(g)} = \mathbf{W}^*\hat{\boldsymbol{\gamma}}_{(g)}$$

Note that the bias in the estimation comes when we reduce the principal components regression from $p$ to $g$ components. If the model reflects no need to remove any principal components, the estimator is unbiased. The bias comes from the fact that we have:

$$\hat{\boldsymbol{\beta}}_g^{PC} = \mathbf{V}_{(g)}\mathbf{V}'_{(g)}\hat{\boldsymbol{\beta}} \quad \Rightarrow \quad E\left\{\hat{\boldsymbol{\beta}}_g^{PC}\right\} = \mathbf{V}_{(g)}\mathbf{V}'_{(g)}\boldsymbol{\beta} \qquad \text{Note: } \mathbf{V}_{(p)}\mathbf{V}'_{(p)} = \mathbf{I}.$$

**Example: Standing Heights of Female Police Applicants**

In a paper describing Principal Components Regression, Lafi and Kaneene (1992) report data from $n = 33$ female police applicants, data are in cms, with possibly the exception of Foot Length. The response if Standing Height ($Y$), and the $p = 9$ predictors are: Sitting Height ($X_1$), Upper Arm Length, ($X_2$), Forearm Length ($X_3$), Hand Length ($X_4$), Upper Leg Length ($X_5$), Lower Leg Length ($X_6$), Foot Length ($X_7$), BRACH ($X_8 = 100X_3/X_2$), and TIBIO ($X_9 = 100X_6/X_5$). The data are given in Table 6.8. A typo in the original paper's table has been corrected, so results differ from authors'.

The correlation matrix among the predictor variables, and its inverse are given below. The diagonal elements of the inverse are the Variance Inflation Factors.

| ID | Y | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 165.8 | 88.7 | 31.8 | 28.1 | 18.7 | 40.3 | 38.9 | 6.7 | 88.4 | 96.5 |
| 2 | 169.8 | 90.0 | 32.4 | 29.1 | 18.3 | 43.3 | 42.7 | 6.4 | 89.8 | 98.6 |
| 3 | 170.7 | 87.7 | 33.6 | 29.5 | 20.7 | 43.7 | 41.1 | 7.2 | 87.8 | 94.1 |
| 4 | 170.9 | 87.1 | 31.0 | 28.2 | 18.6 | 43.7 | 40.6 | 6.7 | 91.0 | 92.9 |
| 5 | 157.5 | 81.3 | 32.1 | 27.3 | 17.5 | 38.1 | 39.6 | 6.6 | 85.0 | 103.9 |
| 6 | 165.9 | 88.2 | 31.8 | 29.0 | 18.6 | 42.0 | 40.6 | 6.5 | 91.2 | 96.7 |
| 7 | 158.7 | 86.1 | 30.6 | 27.8 | 18.4 | 40.0 | 37.0 | 5.9 | 90.8 | 92.5 |
| 8 | 166.0 | 88.7 | 30.2 | 26.9 | 17.5 | 41.6 | 39.0 | 5.9 | 89.1 | 93.8 |
| 9 | 158.7 | 83.7 | 31.1 | 27.1 | 18.1 | 38.9 | 37.5 | 6.1 | 87.1 | 96.4 |
| 10 | 161.5 | 81.2 | 32.3 | 27.8 | 19.1 | 42.8 | 40.1 | 6.2 | 86.1 | 93.7 |
| 11 | 167.3 | 88.6 | 34.8 | 27.3 | 18.3 | 43.1 | 41.8 | 7.3 | 78.4 | 97 |
| 12 | 167.4 | 83.2 | 34.3 | 30.1 | 19.2 | 43.4 | 42.2 | 6.8 | 87.8 | 97.2 |
| 13 | 159.2 | 81.5 | 31.0 | 27.3 | 17.5 | 39.8 | 39.6 | 4.9 | 88.1 | 99.5 |
| 14 | 170.0 | 87.9 | 34.2 | 30.9 | 19.4 | 43.1 | 43.7 | 6.3 | 90.4 | 101.4 |
| 15 | 166.3 | 88.3 | 30.6 | 28.8 | 18.3 | 41.8 | 41.0 | 5.9 | 94.1 | 98.1 |
| 16 | 169.0 | 85.6 | 32.6 | 28.8 | 19.1 | 42.7 | 42.0 | 6.0 | 88.3 | 98.4 |
| 17 | 156.2 | 81.6 | 31.0 | 25.6 | 17.0 | 44.2 | 39.0 | 5.1 | 82.6 | 88.2 |
| 18 | 159.6 | 86.6 | 32.7 | 25.4 | 17.7 | 42 | 37.5 | 5.0 | 77.7 | 89.3 |
| 19 | 155.0 | 82.0 | 30.3 | 26.6 | 17.3 | 37.9 | 36.1 | 5.2 | 87.8 | 95.3 |
| 20 | 161.1 | 84.1 | 29.5 | 26.6 | 17.8 | 38.6 | 38.2 | 5.9 | 90.2 | 99 |
| 21 | 170.3 | 88.1 | 34.0 | 29.3 | 18.2 | 43.2 | 41.4 | 5.9 | 86.2 | 95.8 |
| 22 | 167.8 | 83.9 | 32.5 | 28.6 | 20.2 | 43.3 | 42.9 | 7.2 | 88.0 | 99.1 |
| 23 | 163.1 | 88.1 | 31.7 | 26.9 | 18.1 | 40.1 | 39.0 | 5.9 | 84.9 | 97.3 |
| 24 | 165.8 | 87.0 | 33.2 | 26.3 | 19.5 | 43.2 | 40.7 | 5.9 | 79.2 | 94.2 |
| 25 | 175.4 | 89.6 | 35.2 | 30.1 | 19.1 | 45.1 | 44.5 | 6.3 | 85.5 | 98.7 |
| 26 | 159.8 | 85.6 | 31.5 | 27.1 | 19.2 | 42.3 | 39.0 | 5.7 | 86.0 | 92.2 |
| 27 | 166.0 | 84.9 | 30.5 | 28.1 | 17.8 | 41.2 | 43.0 | 6.1 | 92.1 | 104.4 |
| 28 | 161.2 | 84.1 | 32.8 | 29.2 | 18.4 | 42.6 | 41.1 | 5.9 | 89.0 | 96.5 |
| 29 | 160.4 | 84.3 | 30.5 | 27.8 | 16.8 | 41.0 | 39.8 | 6.0 | 91.1 | 97.1 |
| 30 | 164.3 | 85.0 | 35.0 | 27.8 | 19.0 | 47.2 | 42.4 | 5.0 | 79.4 | 89.8 |
| 31 | 165.5 | 82.6 | 36.2 | 28.6 | 20.2 | 45.0 | 42.3 | 5.6 | 79.0 | 94.0 |
| 32 | 167.2 | 85.0 | 33.6 | 27.1 | 19.8 | 46.0 | 41.6 | 5.6 | 80.7 | 90.4 |
| 33 | 167.2 | 83.4 | 33.5 | 29.7 | 19.4 | 45.2 | 44.0 | 5.2 | 88.7 | 97.3 |

Table 6.8: Female Police Applicant Stature Data

$$\mathbf{Z} = \begin{bmatrix} 0.2101 & -0.0596 & 0.0102 & 0.0242 & -0.1584 & -0.1437 & 0.1868 & 0.0688 & 0.0215 \\ 0.2973 & 0.0035 & 0.1443 & -0.0502 & 0.0774 & 0.1769 & 0.1035 & 0.1258 & 0.1196 \\ 0.1431 & 0.1298 & 0.1980 & 0.3963 & 0.1089 & 0.0419 & 0.3257 & 0.0443 & -0.0906 \\ 0.1028 & -0.1439 & 0.0236 & 0.0056 & 0.1089 & -0.0003 & 0.1868 & 0.1747 & -0.1466 \\ -0.2861 & -0.0281 & -0.0972 & -0.1990 & -0.3314 & -0.0846 & 0.1591 & -0.0698 & 0.3670 \\ 0.1766 & -0.0596 & 0.1309 & 0.0056 & -0.0248 & -0.0003 & 0.1313 & 0.1828 & 0.0308 \\ 0.0358 & -0.1860 & -0.0301 & -0.0316 & -0.1820 & -0.3040 & -0.0353 & 0.1665 & -0.1653 \\ 0.2101 & -0.2281 & -0.1508 & -0.1990 & -0.0562 & -0.1353 & -0.0353 & 0.0973 & -0.1046 \\ -0.1252 & -0.1333 & -0.1240 & -0.0874 & -0.2685 & -0.2618 & 0.0202 & 0.0158 & 0.0168 \\ -0.2928 & -0.0070 & -0.0301 & 0.0987 & 0.0381 & -0.0424 & 0.0480 & -0.0249 & -0.1092 \\ 0.2034 & 0.2561 & -0.0972 & -0.0502 & 0.0617 & 0.1010 & 0.3535 & -0.3386 & 0.0449 \\ -0.1587 & 0.2035 & 0.2785 & 0.1173 & 0.0853 & 0.1348 & 0.2146 & 0.0443 & 0.0542 \\ -0.2727 & -0.1439 & -0.0972 & -0.1990 & -0.1977 & -0.0846 & -0.3131 & 0.0565 & 0.1616 \\ 0.1565 & 0.1930 & 0.3858 & 0.1545 & 0.0617 & 0.2613 & 0.0757 & 0.1502 & 0.2503 \\ 0.1833 & -0.1860 & 0.1041 & -0.0502 & -0.0405 & 0.0335 & -0.0353 & 0.3010 & 0.0962 \\ 0.0022 & 0.0246 & 0.1041 & 0.0987 & 0.0303 & 0.1179 & -0.0076 & 0.0647 & 0.1102 \\ -0.2660 & -0.1439 & -0.3252 & -0.2920 & 0.1482 & -0.1353 & -0.2575 & -0.1675 & -0.3660 \\ 0.0693 & 0.0351 & -0.3521 & -0.1618 & -0.0248 & -0.2618 & -0.2853 & -0.3671 & -0.3147 \\ -0.2392 & -0.2175 & -0.1911 & -0.2362 & -0.3471 & -0.3800 & -0.2297 & 0.0443 & -0.0345 \\ -0.0984 & -0.3017 & -0.1911 & -0.1432 & -0.2920 & -0.2028 & -0.0353 & 0.1421 & 0.1382 \\ 0.1699 & 0.1719 & 0.1712 & -0.0688 & 0.0696 & 0.0672 & -0.0353 & -0.0209 & -0.0112 \\ -0.1118 & 0.0140 & 0.0772 & 0.3033 & 0.0774 & 0.1938 & 0.3257 & 0.0525 & 0.1429 \\ 0.1699 & -0.0702 & -0.1508 & -0.0874 & -0.1741 & -0.1353 & -0.0353 & -0.0738 & 0.0589 \\ 0.0961 & 0.0877 & -0.2313 & 0.1731 & 0.0696 & 0.0082 & -0.0353 & -0.3060 & -0.0859 \end{bmatrix}$$

The $\mathbf{V}$ matrix containing the eigenvectors of $\mathbf{R}_{XX} = \mathbf{Z}'\mathbf{Z}$ is obtained using the **eigen** function in R.

$$\mathbf{V} = \begin{bmatrix} -0.1854 & -0.1530 & 0.8020 & -0.2796 & 0.3690 & 0.2330 & -0.1741 & -0.0003 & 0.0094 \\ -0.4414 & 0.2347 & -0.0974 & 0.2322 & 0.2543 & 0.3188 & 0.3976 & 0.5776 & -0.1693 \\ -0.3933 & -0.3338 & -0.1661 & -0.2331 & -0.1218 & 0.3164 & 0.4962 & -0.5132 & 0.1657 \\ -0.4183 & 0.0807 & 0.0275 & 0.2041 & -0.5771 & 0.3718 & -0.5521 & 0.0000 & 0.0034 \\ -0.4127 & 0.2997 & -0.0129 & -0.3507 & -0.0544 & -0.4663 & -0.0272 & 0.1786 & 0.6031 \\ -0.4645 & -0.1012 & -0.2522 & -0.1634 & 0.2718 & -0.3795 & -0.2793 & -0.1831 & -0.5952 \\ -0.2140 & -0.3578 & 0.3798 & 0.5839 & -0.2177 & -0.4819 & 0.2477 & 0.0018 & -0.0025 \\ 0.0851 & -0.5461 & -0.0531 & -0.4556 & -0.3661 & -0.0366 & 0.0417 & 0.5659 & -0.1634 \\ -0.0462 & -0.5265 & -0.3288 & 0.2714 & 0.4393 & 0.1038 & -0.3447 & 0.1315 & 0.4461 \end{bmatrix}$$

The vector of eigenvalues of $\mathbf{R}_{XX} = \mathbf{Z}'\mathbf{Z}$ is also obtained using the **eigen** function in R. The eigenvalues $\lambda$ and the condition indices are given below.

$$\text{Eigenvalues:} \quad \begin{bmatrix} 3.6299 & 2.4432 & 1.0129 & 0.7666 & 0.6111 & 0.3025 & 0.2326 & 0.0008 & 0.0004 \end{bmatrix}$$

$$\text{Condition Indices} \begin{bmatrix} 1.000 & 1.219 & 1.893 & 2.176 & 2.437 & 3.464 & 3.951 & 67.174 & 92.266 \end{bmatrix}$$

Clearly the eighth and ninth components have very high condition indices.

The matrix of principal components is obtained as $\mathbf{W} = \mathbf{ZV}$. When fitting the principal component regression, we will use $\mathbf{W}^* = [\mathbf{1}|\mathbf{W}]$.

$$\mathbf{W} = \begin{bmatrix}
0.0702 & -0.1962 & 0.2718 & 0.0926 & -0.0397 & 0.0803 & 0.0175 & 0.0004 & 0.0017 \\
-0.2235 & -0.2603 & 0.1604 & -0.1467 & 0.1498 & -0.0479 & -0.0129 & -0.0036 & -0.0002 \\
-0.4536 & -0.0902 & 0.2193 & 0.1253 & -0.2885 & 0.0501 & 0.0052 & -0.0012 & 0.0058 \\
-0.0303 & -0.1093 & 0.2012 & -0.1149 & -0.1798 & -0.1746 & 0.0347 & 0.0042 & 0.0007 \\
0.3060 & -0.2493 & -0.2470 & 0.4098 & 0.1611 & -0.0297 & 0.0430 & -0.0009 & 0.0106 \\
-0.0639 & -0.2547 & 0.1562 & -0.0821 & -0.0499 & 0.0105 & 0.0376 & 0.0017 & 0.0022 \\
0.3462 & -0.0567 & 0.1620 & -0.0805 & -0.2108 & 0.1218 & 0.0673 & 0.0036 & -0.0029 \\
0.3110 & -0.0400 & 0.2609 & -0.1687 & 0.0452 & -0.0653 & -0.0216 & 0.0016 & -0.0009 \\
0.3960 & -0.0565 & 0.0016 & 0.1612 & -0.0739 & 0.0726 & 0.0358 & -0.0022 & -0.0006 \\
0.0246 & 0.1308 & -0.1608 & 0.1107 & -0.2260 & -0.0785 & 0.0381 & -0.0023 & -0.0027 \\
-0.2705 & 0.1005 & 0.2642 & 0.3497 & 0.2718 & -0.1407 & 0.0740 & 0.0051 & -0.0062 \\
-0.3614 & -0.1290 & -0.1639 & 0.1186 & -0.1156 & -0.0307 & 0.1784 & -0.0022 & -0.0018 \\
0.4207 & -0.0303 & -0.3452 & -0.0567 & 0.0957 & 0.0757 & -0.0499 & -0.0003 & -0.0002 \\
-0.4923 & -0.3438 & -0.0813 & -0.0780 & 0.0768 & 0.1335 & 0.0197 & -0.0053 & 0.0018 \\
0.0581 & -0.3284 & 0.0775 & -0.2519 & -0.0123 & 0.0199 & -0.0641 & 0.0087 & -0.0003 \\
-0.1587 & -0.1149 & -0.0878 & -0.0329 & -0.0058 & 0.0319 & -0.0644 & -0.0043 & 0.0001 \\
0.4223 & 0.5264 & -0.0897 & -0.1461 & -0.0148 & -0.2449 & 0.0781 & -0.0083 & 0.0016 \\
0.3540 & 0.5894 & 0.1871 & 0.0046 & 0.1592 & 0.0849 & 0.0128 & -0.0052 & -0.0029 \\
0.6885 & 0.0409 & -0.1233 & 0.0328 & -0.0498 & 0.1382 & 0.0632 & 0.0002 & -0.0031 \\
0.5144 & -0.2083 & -0.0332 & 0.0604 & -0.0299 & 0.0065 & -0.1045 & 0.0073 & 0.0011 \\
-0.1995 & -0.0043 & 0.0627 & -0.1111 & 0.1502 & 0.0817 & 0.1352 & -0.0018 & 0.0010 \\
-0.3365 & -0.1978 & -0.0714 & 0.2246 & -0.2010 & -0.1380 & -0.1267 & -0.0041 & -0.0039 \\
0.2286 & -0.0157 & 0.1732 & 0.0657 & 0.1469 & 0.0953 & -0.0737 & -0.0036 & 0.0021 \\
-0.0850 & 0.3421 & 0.1398 & 0.1525 & 0.0665 & 0.0253 & -0.1883 & -0.0043 & -0.0028 \\
-0.6018 & -0.0896 & 0.0491 & -0.0812 & 0.2184 & 0.0342 & 0.0313 & -0.0033 & -0.0019 \\
0.1281 & 0.1852 & 0.0946 & -0.0350 & -0.1600 & 0.0539 & -0.0871 & -0.0046 & 0.0001 \\
0.0895 & -0.4336 & -0.2046 & -0.0493 & 0.1614 & -0.1362 & -0.1607 & 0.0042 & -0.0010 \\
-0.0649 & -0.0766 & -0.1467 & -0.0771 & -0.0325 & 0.0193 & 0.1064 & -0.0027 & 0.0004 \\
0.3407 & -0.1948 & -0.0633 & -0.1072 & 0.0575 & -0.1382 & 0.1137 & 0.0036 & -0.0027 \\
-0.3167 & 0.6053 & -0.0902 & -0.1710 & 0.0774 & -0.0187 & 0.0223 & 0.0090 & 0.0087 \\
-0.4490 & 0.4368 & -0.2392 & 0.1438 & -0.0288 & 0.1256 & 0.0106 & 0.0144 & -0.0052 \\
-0.2391 & 0.4893 & 0.0133 & -0.0281 & -0.0911 & -0.0498 & -0.1099 & 0.0015 & 0.0054 \\
-0.3521 & 0.0336 & -0.3474 & -0.2340 & -0.0280 & 0.0321 & -0.0610 & -0.0052 & -0.0041
\end{bmatrix}$$

The principal component regression coefficients, standard errors, and $t$-statistics are given in Table 6.10. The estimate of the error variance is $s^2 = MSE = 3.5737$.

As the standard errors are very large, and the $t$-statistics are very small for $W8$ and $W9$, remove them, and keep the first $g = 7$ principal components. We create $\mathbf{W}^*_{(7)} = \left[\mathbf{1}|\mathbf{W}_{(7)}\right]$ and compute $\hat{\boldsymbol{\gamma}}_{(7)} = (\mathbf{W}^{*'}_{(7)}\mathbf{W}^*_{(7)})^{-1}\mathbf{W}^*_{(7)}\mathbf{Y}$, and its variance-covariance matrix $s^2(\mathbf{W}^{*'}_{(7)}\mathbf{W}^*_{(7)})^{-1}$.

The principal component regression coefficients, standard errors, and $t$-statistics are given in Table 6.11. The estimate of the error variance is $s^2 = MSE = 3.2969$.

The Principal Component estimates have much smaller standard errors than the OLS estimates. The predicted values (not shown here) are very similar for the two methods. The primary difference is the

| Parameter | Estimate | Std. Error | t-statistic |
|---|---|---|---|
| Intercept | 164.5636 | 0.3275 | 502.4518 |
| W1 | -12.1261 | 0.9875 | -12.2792 |
| W2 | -4.5428 | 1.2037 | -3.7740 |
| W3 | 7.6001 | 1.8695 | 4.0654 |
| W4 | -4.9758 | 2.1488 | -2.3156 |
| W5 | 3.5712 | 2.4069 | 1.4837 |
| W6 | -3.2902 | 3.4210 | -0.9618 |
| W7 | -6.8250 | 3.9014 | -1.7494 |
| W8 | 22.6248 | 66.3364 | 0.3411 |
| W9 | -37.2527 | 91.1147 | -0.4089 |

Table 6.9: Female Police Applicant Stature Principal Components Regression - Full Model

| Parameter | Estimate | Std. Error | t-statistic |
|---|---|---|---|
| Intercept | 164.5636 | 0.3161 | 520.6434 |
| W1 | -12.1261 | 0.9530 | -12.7238 |
| W2 | -4.5428 | 1.1616 | -3.9107 |
| W3 | 7.6001 | 1.8041 | 4.2126 |
| W4 | -4.9758 | 2.0737 | -2.3994 |
| W5 | 3.5712 | 2.3228 | 1.5375 |
| W6 | -3.2902 | 3.3015 | -0.9966 |
| W7 | -6.8250 | 3.7651 | -1.8127 |

Table 6.10: Female Police Applicant Stature Principal Components Regression - Reduced Model

| | Principal Components | | | Ordinary Least Squares | | |
|---|---|---|---|---|---|---|
| Parameter | Estimate | Std. Error | t-statistic | Estimate | Std. Error | t-statistic |
| Intercept | N/A | N/A | N/A | 164.5636 | 0.3275 | 502.452 |
| Z1 | 12.1690 | 2.0612 | 5.9038 | 11.8120 | 2.3013 | 5.133 |
| Z2 | -0.4643 | 2.0524 | -0.2262 | 18.9116 | 41.3585 | 0.457 |
| Z3 | 1.3199 | 2.2969 | 0.5746 | -16.4646 | 37.3195 | -0.441 |
| Z4 | 4.3827 | 2.8237 | 1.5521 | 4.2577 | 2.9421 | 1.447 |
| Z5 | 6.8153 | 1.7893 | 3.8089 | -11.6107 | 56.2438 | -0.206 |
| Z6 | 9.1153 | 1.8988 | 4.7995 | 27.1454 | 55.6130 | 0.488 |
| Z7 | 3.3191 | 2.4096 | 1.3774 | 3.4503 | 2.5096 | 1.375 |
| Z8 | 1.8407 | 1.4398 | 1.2784 | 20.7340 | 40.4162 | 0.513 |
| Z9 | 2.6830 | 1.9716 | 1.3608 | -10.9605 | 41.6235 | -0.263 |

Table 6.11: Female Police Applicant Stature Principal Components Regression - Back-Transformed to $Z$-scale, and OLS on Z-scale

| ID | Y | Yhat(W) | Yhat(Z) | ID | Y | Yhat(W) | Yhat(Z) | ID | Y | Yhat(W) | Yhat(Z) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 165.8 | 165.7 | 165.7 | 12 | 167.4 | 166.2 | 166.2 | 23 | 163.1 | 163.6 | 163.6 |
| 2 | 169.8 | 171.2 | 171.2 | 13 | 159.2 | 157.7 | 157.7 | 24 | 165.8 | 165.8 | 165.8 |
| 3 | 170.7 | 170.3 | 170.3 | 14 | 170.0 | 171.6 | 171.6 | 25 | 175.4 | 173.5 | 173.5 |
| 4 | 170.9 | 167.2 | 167.2 | 15 | 166.3 | 167.5 | 167.5 | 26 | 159.8 | 162.9 | 162.9 |
| 5 | 157.5 | 158.4 | 158.4 | 16 | 169.0 | 166.8 | 166.8 | 27 | 166.0 | 166.3 | 166.3 |
| 6 | 165.9 | 167.6 | 167.6 | 17 | 156.2 | 157.3 | 157.3 | 28 | 161.2 | 164.1 | 164.1 |
| 7 | 158.7 | 160.6 | 160.6 | 18 | 159.6 | 159.2 | 159.2 | 29 | 160.4 | 161.3 | 161.3 |
| 8 | 166.0 | 164.3 | 164.3 | 19 | 155.0 | 153.9 | 153.9 | 30 | 164.3 | 166.0 | 166.0 |
| 9 | 158.7 | 158.5 | 158.5 | 20 | 161.1 | 159.3 | 159.3 | 31 | 165.5 | 164.9 | 164.9 |
| 10 | 161.5 | 161.1 | 161.1 | 21 | 170.3 | 167.4 | 167.4 | 32 | 167.2 | 166.1 | 166.1 |
| 11 | 167.3 | 168.6 | 168.6 | 22 | 167.8 | 168.5 | 168.5 | 33 | 167.2 | 167.4 | 167.4 |

Table 6.12: Female Police Applicant Stature Principal Components Regression - Observed and Fitted Values

stability of the regression coefficients.

The observed standing heights, and predicted heights from both of the fitted equations (based on $\mathbf{W}_{(7)}$ and $\mathbf{Z}$ are given in Table 6.12.

$\nabla$

### 6.11.2 Ridge Regression

In **Ridge Regression**, a biased estimator is directly induced that reduces its variance and mean square error (variance + squared bias). Unfortunately, the bias-inducing constant varies among applications, so it must be selected comparing results over various possible levels. We begin with a **standardized regression model** with no bias, based on the $p \times p$ correlation matrix among the predictors $\mathbf{R}_{XX}$ and the $p \times 1$ vector of correlations $\mathbf{R}_{XY}$ between the predictors and response variable.

$$\mathbf{R}_{XX} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{12} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & \cdots & 1 \end{bmatrix} \qquad \mathbf{R}_{XY} = \begin{bmatrix} r_{Y1} \\ r_{Y2} \\ \vdots \\ r_{Yp} \end{bmatrix}$$

The estimated standardized regression coefficients are obtained as:

$$\mathbf{R}_{XX}\hat{\boldsymbol{\beta}}^* = \mathbf{R}_{XY} \qquad \Rightarrow \qquad \hat{\boldsymbol{\beta}}^* = \mathbf{R}_{XX}^{-1}\mathbf{R}_{XY}$$

The standardized regression coefficients $\hat{\boldsymbol{\beta}}^*$ measure the change in $Y$ in standard deviation units as each predictor increases by 1 standard deviation, thus removing the effects of scaling each predictor. It can also be obtained by transforming each $X$ and $Y$ by the following transformations.

$$X_{ij}^* = \frac{X_{ij} - \overline{X}_j}{\sqrt{n-1}s_j} \qquad Y_i^* = \frac{Y_i - \overline{Y}}{\sqrt{n-1}s_Y}$$

In matrix form, we have:

$$\mathbf{X}^* = \begin{bmatrix} X_{11}^* & X_{12}^* & \cdots & X_{1p}^* \\ X_{21}^* & X_{22}^* & \cdots & X_{2p}^* \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1}^* & X_{n2}^* & \cdots & X_{np}^* \end{bmatrix} \qquad \mathbf{Y}^* = \begin{bmatrix} Y_1^* \\ Y_2^* \\ \vdots \\ Y_n^* \end{bmatrix} \qquad \hat{\boldsymbol{\beta}}^* = \left(\mathbf{X}^{*'}\mathbf{X}^*\right)^{-1}\mathbf{X}^{*'}\mathbf{Y}^*$$

Note that $\mathbf{R}_{XX} = \mathbf{X}^{*'}\mathbf{X}^*$ and $\mathbf{R}_{XY} = \mathbf{X}^{*'}\mathbf{Y}^*$. The standardized ridge estimator is obtained as follows (see e.g. Kutner, Nachtsheim, Neter, and Li (2005), Section 11.2).

$$\hat{\boldsymbol{\beta}}_{OLS}^* = (\mathbf{R}_{XX})^{-1}\mathbf{R}_{XY} = \left(\mathbf{X}^{*'}\mathbf{X}^*\right)^{-1}\mathbf{X}^{*'}\mathbf{Y}^*$$

$$V\left\{\hat{\boldsymbol{\beta}}_{OLS}^*\right\} = \sigma^2\left(\mathbf{X}^{*'}\mathbf{X}^*\right)^{-1}$$

$$\hat{\boldsymbol{\beta}}_{RR}^* = (\mathbf{R}_{XX} + c\mathbf{I})^{-1}\mathbf{R}_{XY} = \left(\mathbf{X}^{*'}\mathbf{X}^* + c\mathbf{I}\right)^{-1}\mathbf{X}^{*'}\mathbf{Y}^*$$

$$V\left\{\hat{\boldsymbol{\beta}}_{RR}^*\right\} = \sigma^2\left(\mathbf{X}^{*'}\mathbf{X}^* + c\mathbf{I}\right)^{-1}\left(\mathbf{X}^{*'}\mathbf{X}^*\right)\left(\mathbf{X}^{*'}\mathbf{X}^* + c\mathbf{I}\right)^{-1}$$

Making use of the following matrix identity, we can write the Ridge Estimator as a linear function of the Ordinary Least Squares Estimator.

$$\left(\mathbf{A}^{-1} + \mathbf{B}^{-1}\right)^{-1} = \mathbf{A}\left(\mathbf{A} + \mathbf{B}\right)^{-1}\mathbf{B} = \mathbf{B}\left(\mathbf{A} + \mathbf{B}\right)^{-1}\mathbf{A}$$

$$\Rightarrow \quad \left(\mathbf{X}^{*'}\mathbf{X}^* + c\mathbf{I}\right)^{-1} = \frac{1}{c}\mathbf{I}\left[\left(\mathbf{X}^{*'}\mathbf{X}^*\right)^{-1} + \frac{1}{c}\mathbf{I}\right]^{-1}\left(\mathbf{X}^{*'}\mathbf{X}^*\right)^{-1} = \left[c\left(\mathbf{X}^{*'}\mathbf{X}^*\right)^{-1} + \mathbf{I}\right]^{-1}\left(\mathbf{X}^{*'}\mathbf{X}^*\right)^{-1}$$

$$\Rightarrow \hat{\boldsymbol{\beta}}_{RR}^* = \left[c\left(\mathbf{X}^{*'}\mathbf{X}^*\right)^{-1} + \mathbf{I}\right]^{-1}\hat{\boldsymbol{\beta}}_{OLS}^*$$

A **ridge trace plot** of the regression coefficients (vertical axis) versus $c$ (horizontal axis) leads to a choice of $c$, where the coefficients stabilize or "flatten out." A second graphical measure involves plotting each of the Variance Inflation Factors versus $c$, and determining when they all get below 10.

The fitted regression equation in transformed scale is:

$$\hat{\mathbf{Y}}^* = \mathbf{X}^*\hat{\boldsymbol{\beta}}^{RR} \qquad \Rightarrow \qquad \hat{Y}_i^* = \hat{\beta}_1^{RR}X_{i1}^* + \cdots + \hat{\beta}_p^{RR}X_{ip}^*.$$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}$$

In terms of the originally scaled response, we have:

$$\hat{\beta}_j = \left(\frac{s_Y}{s_j}\right)\hat{\beta}_j^* \qquad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}_1 - \cdots - \hat{\beta}_p\overline{X}_p.$$

$$VIF \equiv \text{ Diagonal Elements of } \mathbf{R}^{-1} = \left(\mathbf{X}^{*'}\mathbf{X}^*\right)^{-1}$$

$$VIF(c) \equiv \text{ Diagonal Elements of } \left(\mathbf{X}^{*'}\mathbf{X}^{*} + c\mathbf{I}\right)^{-1}\left(\mathbf{X}^{*'}\mathbf{X}^{*}\right)\left(\mathbf{X}^{*'}\mathbf{X}^{*} + c\mathbf{I}\right)^{-1}$$

There are also several numerical based approaches (see e.g. Myers (1990), Section 8.4) including Generalized Cross-Validation (GCV), the "PRESS" Statistic, and the $C_c$ Statistic.

**Example: Population and Carbon Emissions in China**

Zhu and Peng (2012) fit a model relating China Carbon Emissions ($Y$ in millions of tons) to $p = 5$ predictors: Population $\left(X_1, \text{ in } 10^7\right)$, Urbanization Rate ($X_2$, in percent), Percent of Population of Working Age ($X_3$), Average Household Size ($X_4$), Per Capita Expenditures ($X_5$, adjusted to Year=2000), for $n = 31$ years (1978-2008). The theoretical model was multiplicative, and fit as a linear regression after taking logarithms. The data in its original units is given in Table 6.13.

$$\ln Y_t = \beta_0 + \beta_1 \ln X_{t1} + \beta_2 \ln X_{t2} + \beta_3 \ln X_{t3} + \beta_4 \ln X_{t4} + \beta_5 \ln X_{t5} + \epsilon_t \qquad t = 1 \ldots, 31$$

The correlation matrix of the logs of the predictors, $\mathbf{R}_{XX}$, and the vector of correlations of the logs of the predictors and the log of the carbon emissions, $\mathbf{R}_{XY}$, are given in Table 6.14. The table also includes $\mathbf{R}_{XX}^{-1}$, where the Variance Inflation Factors for the predictors are on the main diagonal. Each predictor is highly correlated with emissions, and all of the predictors are highly correlated among themselves. The Variance Inflation Factors are all well over 10, with the largest, $VIF_5 = 213.60$. This implies the regression standard errors will be very large, and the estimates very unstable.

A plot of the Ridge Trace is given in Figure 6.7 and a plot of the VIF is given in Figure 6.8. The $VIF^s$ get below 10 for a very low value of $c$, around 0.02. The Ridge Trace shows it takes higher values of $c$, around 0.15 or so. The authors used $c = 0.20$, based on the Ridge Trace. These plots are obtained by "brute force" computing the Ridge Regression coefficients over a grid of $c$ values. The values were $c = 0$ to 0.5 by 0.0001. All regression coefficients and $VIF^s$ were saved and plotted versus $c$.

Based on using $c = 0.20$, we get the following estimated standardized regression coefficients, standard errors, and $t$-statistics based on Ridge and OLS in Table 6.15.

$\nabla$

## 6.12 Models with Unequal Variances (Heteroskedasticity)

When the data are independent, but with unequal variances, we can use (estimated) **Weighted Least Squares**. In rare occasions, the variances are known, and they will be used directly. One setting where this occurs in practice is when the "data" are averages among a group of units with common $X$ levels. If each individual unit is independent with constant variance $\sigma^2$, the average of the $m_i$ units ($Y_i$ in this setting) has variance $V\{Y_i\} = \sigma^2/m_i$. In this case, we would use the reciprocal of the variance as the weight for each

| Year | $Y$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ |
|---|---|---|---|---|---|---|
| 1978 | 40.77 | 96.26 | 17.92 | 59.5 | 4.66 | 740 |
| 1979 | 41.65 | 97.54 | 18.96 | 60 | 4.65 | 791 |
| 1980 | 40.7 | 98.71 | 19.39 | 60.5 | 4.61 | 862 |
| 1981 | 40.29 | 100.07 | 20.16 | 61 | 4.54 | 934 |
| 1982 | 43.12 | 101.65 | 21.13 | 61.5 | 4.51 | 997 |
| 1983 | 45.47 | 103.01 | 21.62 | 62.37 | 4.46 | 1079 |
| 1984 | 49.43 | 104.36 | 23.01 | 63.24 | 4.41 | 1207 |
| 1985 | 53.59 | 105.85 | 23.71 | 64.12 | 4.33 | 1370 |
| 1986 | 56.35 | 107.51 | 24.52 | 64.99 | 4.24 | 1435 |
| 1987 | 60.12 | 109.3 | 25.32 | 65.86 | 4.15 | 1520 |
| 1988 | 64.45 | 111.03 | 25.81 | 66.15 | 4.05 | 1638 |
| 1989 | 65.47 | 112.7 | 26.21 | 66.45 | 3.97 | 1635 |
| 1990 | 65.86 | 114.33 | 26.41 | 66.74 | 3.93 | 1695 |
| 1991 | 69.15 | 115.82 | 26.94 | 66.3 | 3.89 | 1842 |
| 1992 | 72.14 | 117.17 | 27.46 | 66.2 | 3.85 | 2086 |
| 1993 | 77.02 | 118.52 | 27.99 | 66.7 | 3.81 | 2262 |
| 1994 | 81.81 | 119.85 | 28.51 | 66.6 | 3.78 | 2367 |
| 1995 | 88.47 | 121.12 | 29.04 | 67.2 | 3.74 | 2553 |
| 1996 | 92.6 | 122.39 | 30.48 | 67.2 | 3.72 | 2793 |
| 1997 | 91.49 | 123.63 | 31.91 | 67.5 | 3.64 | 2919 |
| 1998 | 86.61 | 124.76 | 33.35 | 67.6 | 3.63 | 3091 |
| 1999 | 90.5 | 125.79 | 34.78 | 67.7 | 3.58 | 3346 |
| 2000 | 92.89 | 126.74 | 36.22 | 70.15 | 3.44 | 3632 |
| 2001 | 95.14 | 127.63 | 37.66 | 70.4 | 3.42 | 3855 |
| 2002 | 100.96 | 128.45 | 39.09 | 70.3 | 3.39 | 4125 |
| 2003 | 118.72 | 129.23 | 40.53 | 70.4 | 3.38 | 4415 |
| 2004 | 139.07 | 129.99 | 41.76 | 70.92 | 3.31 | 4773 |
| 2005 | 153.42 | 130.76 | 42.99 | 72.04 | 3.24 | 5142 |
| 2006 | 166.46 | 131.45 | 43.9 | 72.32 | 3.17 | 5636 |
| 2007 | 180.17 | 132.13 | 44.94 | 72.53 | 3.17 | 6239 |
| 2008 | 192.27 | 132.8 | 45.68 | 72.8 | 3.16 | 6782 |

Table 6.13: China Carbon Emissions and Population Characteristics 1978-2008

| $\mathbf{R}_{XX}$ | | | | | $\mathbf{R}_{XY}$ | $\mathbf{R}_{XX}^{-1}$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1.0000 | 0.9753 | 0.9712 | -0.9873 | 0.9847 | 0.9525 | 50.62 | 36.87 | -9.85 | 32.98 | -44.13 |
| 0.9753 | 1.0000 | 0.9801 | -0.9907 | 0.9952 | 0.9748 | 36.87 | 147.72 | -27.81 | 21.50 | -134.78 |
| 0.9712 | 0.9801 | 1.0000 | -0.9802 | 0.9773 | 0.9601 | -9.85 | -27.81 | 31.40 | 13.30 | 19.91 |
| -0.9873 | -0.9907 | -0.9802 | 1.0000 | -0.9942 | -0.9786 | 32.98 | 21.50 | 13.30 | 122.38 | 54.80 |
| 0.9847 | 0.9952 | 0.9773 | -0.9942 | 1.0000 | 0.9826 | -44.13 | -134.78 | 19.91 | 54.80 | 213.60 |

Table 6.14: China Carbon Emissions and Population Characteristics - Correlations and Variance Inflation Factors

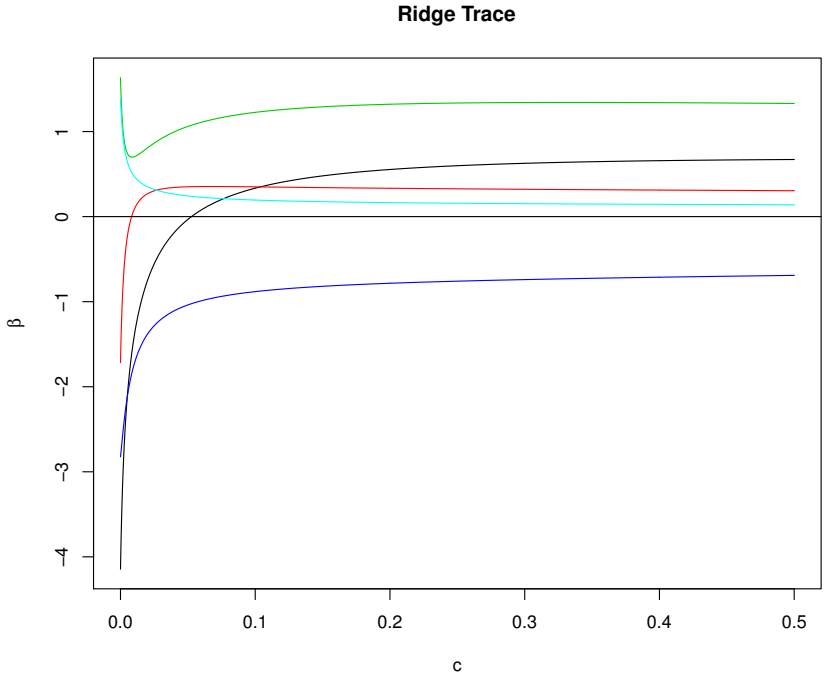


Figure 6.7: Ridge Trace for China Carbon Data

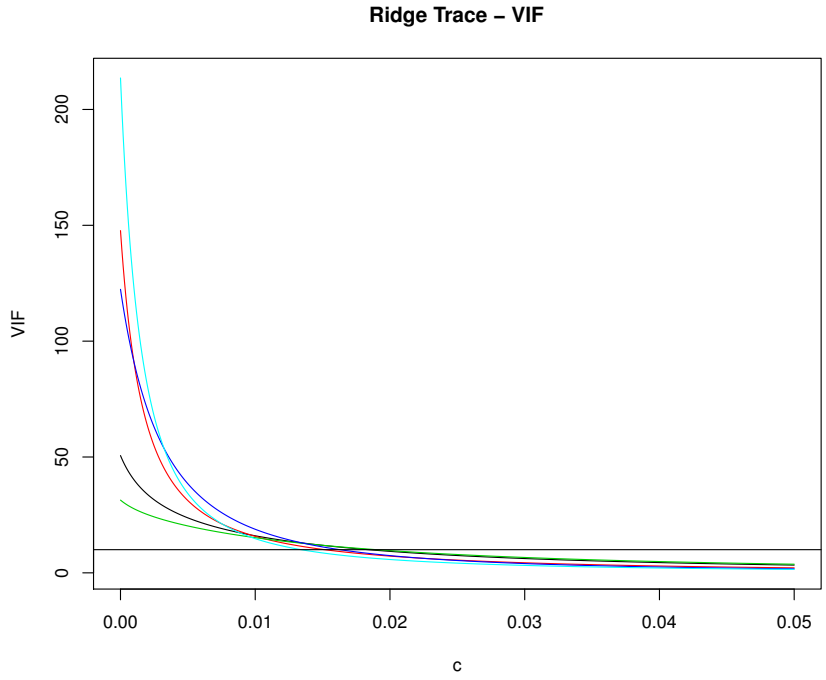


Figure 6.8: VIF versus $c$ for China Carbon Data

| | Ridge Regression | | | Ordinary Least Squares | | |
|---|---|---|---|---|---|---|
| Parameter | Estimate | Std. Error | t-statistic | Estimate | Std. Error | t-statistic |
| $(\ln X_1)^*$ | 0.1245 | 0.0163 | 7.6496 | -0.9311 | 0.1799 | -5.1766 |
| $(\ln X_2)^*$ | 0.2028 | 0.0126 | 16.0901 | -1.0452 | 0.3073 | -3.4016 |
| $(\ln X_3)^*$ | 0.1678 | 0.0175 | 9.5773 | 0.2074 | 0.1417 | 1.4643 |
| $(\ln X_4)^*$ | -0.2149 | 0.0105 | -20.5254 | -0.7746 | 0.2797 | -2.7698 |
| $(\ln X_5)^*$ | 0.2337 | 0.0104 | 22.4318 | 1.9668 | 0.3695 | 5.3232 |

Table 6.15: Standardized Regression Coefficients - China Carbon Emissions

case (observations based on larger sample sizes have smaller variances and larger weights).

$$\mathbf{W} = \boldsymbol{\Sigma}_Y^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_n^2} \end{bmatrix} = \begin{bmatrix} \frac{m_1}{\sigma^2} & 0 & \cdots & 0 \\ 0 & \frac{m_2}{\sigma^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{m_n}{\sigma^2} \end{bmatrix} \qquad \hat{\boldsymbol{\beta}}^W = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y}$$

The variance of the least squares estimator is obtained as follows:

$$V\left\{\hat{\boldsymbol{\beta}}^W\right\} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\boldsymbol{\Sigma}_Y\mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$$

In the case with $V\{Y_i\} = \sigma^2/m_i$, we can estimate $\sigma^2$ based on weighted mean square error:

$$MSE_W = \frac{\sum_{i=1}^n m_i\left(Y_i - \hat{Y}_i\right)^2}{n - p'} \qquad \hat{Y}_i = \hat{\beta}_0^W + \hat{\beta}_1^W X_{i1} + \cdots + \hat{\beta}_p^W X_{ip}.$$

In this case (where data are averages):

$$\hat{V}\left\{\hat{\boldsymbol{\beta}}^W\right\} = \left(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X}\right)^{-1} \qquad \hat{\mathbf{W}} = \begin{bmatrix} \frac{m_1}{MSE_W} & 0 & \cdots & 0 \\ 0 & \frac{m_2}{MSE_W} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{m_n}{MSE_W} \end{bmatrix}$$

Note that weighted least squares can be conducted as ordinary least squares on transformed $\mathbf{X}$ and $\mathbf{Y}$, which makes it possible to conduct using EXCEL, and non-statistical computing packages.

$$\mathbf{X}^* = \mathbf{W}^{1/2}\mathbf{X} \qquad \mathbf{Y}^* = \mathbf{W}^{1/2}\mathbf{Y} \qquad \hat{\boldsymbol{\beta}}^W = \left(\mathbf{X}^{*'}\mathbf{X}^*\right)^{-1}\mathbf{X}^{*'}\mathbf{Y}^*$$

where $\mathbf{W}^{1/2}$ is the (diagonal) matrix with elements equal to the square roots of the elements of $\mathbf{W}$.

**Example: Dose-Response Study of Rosuvastin**

Saito, et al (2003) give results of a Japanese Dose-Response study for Rosuvastin in patients with high cholesterol. There were six doses: 1,2.5,5,10,20,40. The response was percentage change in LDL cholesterol at week 12. The data reported were group means per dose, $\overline{Y}_j$, with varying sample sizes among doses, $m_j$. Assuming equal variances among individual patients, $V\left\{\overline{Y}_j\right\} = \sigma^2/m_j$. The data are given in Table 6.16. Figure 6.9 plots the percentage LDL change ($Y$) versus Dose ($X$) and $\ln(X)$. As is often seen with this type

of data, the relationship is curvilinear when $Y$ is plotted versus $X$, and more linear when it is plotted versus $\ln(X)$. This analysis will fit the following model.

$$\overline{Y}_j = \beta_0 + \beta_1 \ln X_j + \epsilon_j \qquad \epsilon_j \sim N\left(0, \frac{\sigma^2}{m_j}\right) \qquad j = 1, \ldots, 6$$

$$\mathbf{Y} = \begin{bmatrix} -35.8 \\ -45.0 \\ -52.7 \\ -49.7 \\ -58.2 \\ -66.0 \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & 0.0000 \\ 1 & 0.9163 \\ 1 & 1.6094 \\ 1 & 2.3026 \\ 1 & 2.9957 \\ 1 & 3.6889 \end{bmatrix} \qquad \mathbf{W} = \frac{1}{\sigma^2} \begin{bmatrix} 15 & 0 & 0 & 0 & 0 & 0 \\ 0 & 17 & 0 & 0 & 0 & 0 \\ 0 & 0 & 12 & 0 & 0 & 0 \\ 0 & 0 & 0 & 14 & 0 & 0 \\ 0 & 0 & 0 & 0 & 18 & 0 \\ 0 & 0 & 0 & 0 & 0 & 13 \end{bmatrix}$$

Note that for the Weighted Least Squares estimator, $\hat{\boldsymbol{\beta}}^W = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y}$, the unknown constant $\sigma^2$ "cancels out" and does not need to be known. It does need to be estimated to obtain the variance-covariance matrix of $\hat{\boldsymbol{\beta}}^W$.

$$\hat{\boldsymbol{\beta}}^W = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y} = \begin{bmatrix} -36.9588 \\ -7.3753 \end{bmatrix} \qquad \hat{\mathbf{Y}}^W = \mathbf{X}\hat{\boldsymbol{\beta}}^W = \begin{bmatrix} -36.9588 \\ -43.7168 \\ -48.8289 \\ -53.9411 \\ -59.0533 \\ -64.1654 \end{bmatrix} \qquad \boldsymbol{e}^W = \mathbf{Y} - \hat{\mathbf{Y}}^W = \begin{bmatrix} 1.1588 \\ -1.2832 \\ -3.8711 \\ 4.2411 \\ 0.8533 \\ -1.8346 \end{bmatrix}$$

$$SSE_W = 15(1.1588)^2 + \cdots + 13(-1.8346)^2 = 536.635 \qquad s^2_W = MSE_W = \frac{536.635}{6-2} = 134.159$$

$$\hat{V}\left\{\hat{\boldsymbol{\beta}}^W\right\} = \left(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X}\right)^{-1} = \begin{bmatrix} 5.0361 & -1.8581 \\ -1.8581 & 0.9786 \end{bmatrix}$$

$$\hat{SE}\left\{\hat{\beta}_0^W\right\} = \sqrt{5.0361} = 2.2441 \qquad \hat{SE}\left\{\hat{\beta}_1^W\right\} = \sqrt{0.9786} = 0.9892$$

The $t$-tests and Confidence Intervals for $\beta_0$ and $\beta_1$ are given below.

$$\beta_0: \qquad t_{obs} = \frac{-36.9588}{2.2441} = -16.4691 \quad t_{.025,4} = 2.776 \quad 2P\left(t_4 \geq |-16.4691|\right) = .0001$$

$$95\%\text{ CI for } \beta_0\text{: } -36.9588 \pm 2.776(2.2441) \quad \equiv \quad -36.9588 \pm 6.2296 \quad \equiv \quad (-43.1884, -30.7292)$$

$$\beta_1: \qquad t_{obs} = \frac{-7.3753}{0.9892} = -7.4556 \quad t_{.025,4} = 2.776 \quad 2P\left(t_4 \geq |-7.4556|\right) = .0017$$

$$95\%\text{ CI for } \beta_1\text{: } -7.3753 \pm 2.776(0.9892) \quad \equiv \quad -7.3753 \pm 2.7460 \quad \equiv \quad (-10.1213, -4.6293)$$

A plot of the data and fitted equation is given in Figure 6.10. It also includes the Ordinary Least Squares (OLS) regression line, the two are very similar. The R program using the **lm** function and **weights** option is given below.

| Dose Group ($j$) | Change in LDL ($\overline{Y}_j$) | Dose ($X_j$) | ln(Dose) ($\ln(X_j)$) | Group Size ($m_j$) |
|---|---|---|---|---|
| 1 | -35.8 | 1 | 0.0000 | 15 |
| 2 | -45.0 | 2.5 | 0.9163 | 17 |
| 3 | -52.7 | 5 | 1.6094 | 12 |
| 4 | -49.7 | 10 | 2.3026 | 14 |
| 5 | -58.2 | 20 | 2.9957 | 18 |
| 6 | -66.0 | 40 | 3.6889 | 13 |

Table 6.16: Rovustatin Data

The calculations of the Mean, and the Total, Error, and Regression Sums of Squares are obtained below.

$$\overline{Y} = \frac{\sum_{j=1}^{n} m_j \overline{Y}_j}{\sum_{j=1}^{n} m_j} = \frac{15(-35.8) + \cdots + 13(-66.0)}{15 + \cdots + 13} = \frac{-4535.8}{89} = -50.9640$$

$$TSS_W = \sum_{j=1}^{n} m_j \left(Y_j - \overline{Y}\right)^2 = \left[15(-35.8 - -50.9640)^2 + \cdots + 13(-66.0 - -50.9640)^2\right] = 7993.945$$

$$SSE_W = \sum_{j=1}^{n} m_j \left(Y_j - \hat{Y}_j\right)^2 = \left[15(-35.8 - -36.9588)^2 + \cdots + 13(-66.0 - -64.1654)^2\right] = 536.635$$

$$SSR_W = \sum_{j=1}^{n} m_j \left(\hat{Y}_j - \overline{Y}\right)^2 = \left[15(-36.9588 - -50.9640)^2 + \cdots + 13(-64.1654 - -50.9640)^2\right] = 7457.310$$

The $F$-test for testing $H_0 : \beta_1 = 0$ is as follows.

$$F_{obs} = \frac{MSR_W}{MSE_W} = \frac{7457.310/1}{536.635/(6-2)} = 55.584 \qquad F_{.05,1,4} = 7.709 \qquad P\left(F_{1,4} \geq 55.584\right) = .0017$$

The R program and output are given below.

```
### Program
dLDL <- c(-35.8,-45.0,-52.7,-49.7,-58.2,-66.0)
DOSE <- c(1,2.5,5,10,20,40)
r <- c(15,17,12,14,18,13)
lnDOSE <- log(DOSE)

cholest <- data.frame(dLDL,lnDOSE,m)
attach(cholest)
cholest.wls <- lm(dLDL ~ lnDOSE, weights=r)
summary(cholest.wls)
confint(cholest.wls)

cholest.ols <- lm(dLDL ~ lnDOSE)
summary(cholest.ols)
confint(cholest.ols)

### Output
```

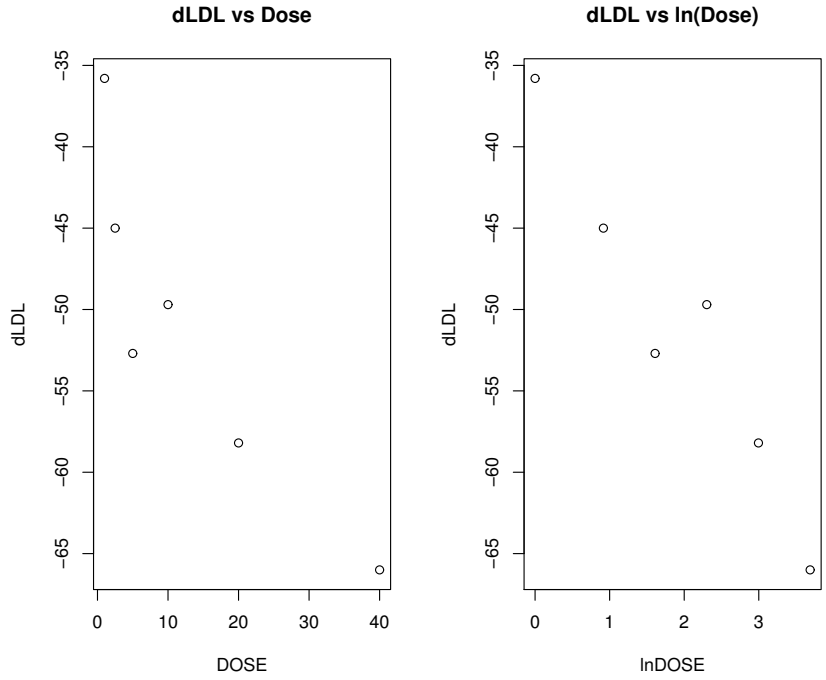


Figure 6.9: Plots of LDL Change vs Dose and ln(Dose)

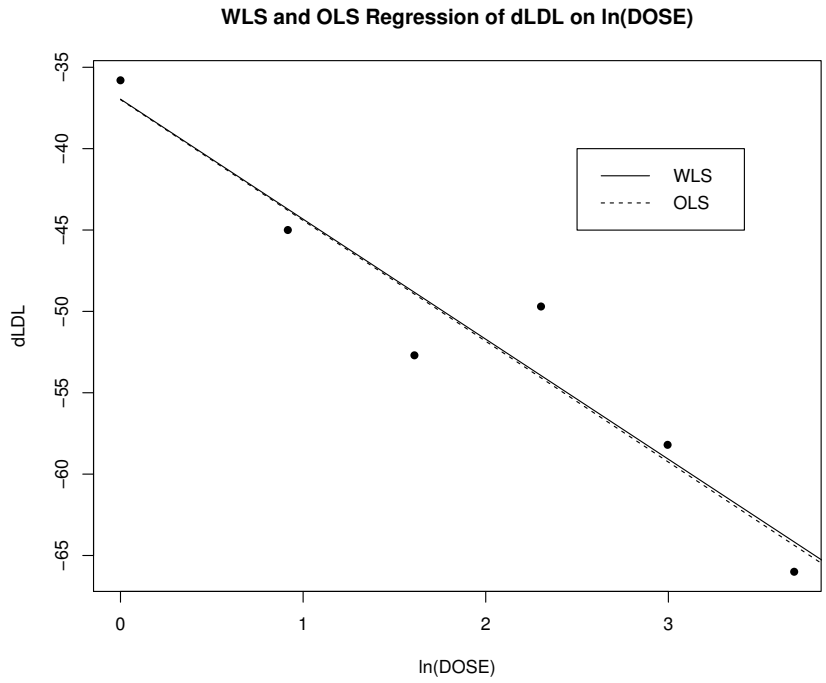


Figure 6.10: Plot of LDL Change versus ln(Dose) with WLS and OLS Fitted Equations

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -36.9588     2.2441 -16.469 7.96e-05 ***
lnDOSE       -7.3753     0.9892  -7.456  0.00173 **
Residual standard error: 11.58 on 4 degrees of freedom
Multiple R-squared:  0.9329,    Adjusted R-squared:  0.9161
F-statistic: 55.59 on 1 and 4 DF,  p-value: 0.001729

Analysis of Variance Table
Response: dLDL
          Df Sum Sq Mean Sq F value   Pr(>F)
lnDOSE     1 7457.3  7457.3  55.586 0.001729 **
Residuals  4  536.6   134.2
> confint(cholest.wls)
                2.5 %     97.5 %
(Intercept) -43.18954 -30.728138
lnDOSE      -10.12186  -4.628755

###ORDINARY LEAST SQUARES:
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -36.990      2.377 -15.560 9.96e-05 ***
lnDOSE        -7.423      1.041  -7.134  0.00204 **
Residual standard error: 3.16 on 4 degrees of freedom
Multiple R-squared:  0.9271,    Adjusted R-squared:  0.9089
F-statistic: 50.89 on 1 and 4 DF,  p-value: 0.002042

Analysis of Variance Table
Response: dLDL
          Df Sum Sq Mean Sq F value   Pr(>F)
lnDOSE     1 508.19  508.19   50.89 0.002042 **
Residuals  4  39.94    9.99

> confint(cholest.ols)
                2.5 %    97.5 %
(Intercept) -43.58992 -30.38978
lnDOSE      -10.31209  -4.53399
```

$\nabla$

### 6.12.1 Estimated Weighted Least Squares

In most cases, the variances are unknown, and must be estimated. In this case, the squared residuals (variance) or absolute residuals (standard deviation) are regressed against one or more of the predictor variables or the mean (fitted values). The process is iterative. We begin by fitting ordinary least squares, obtaining the residuals, then regressing the squared or absolute residuals on the the predictor variables or fitted values, leading to (assuming all $p$ predictors are used in the residual regression):

Variance Case: $\hat{v}_i = \hat{\delta}_0 + \hat{\delta}_1 X_{i1} + \cdots \hat{\delta}_p X_{ip}$ Standard Deviation Case: $\hat{s}_i = \hat{\delta}_0 + \hat{\delta}_1 X_{i1} + \cdots \hat{\delta}_p X_{ip}$.

Variance Case: $\hat{v}_i = \hat{\delta}_0 + \hat{\delta}_1 \hat{Y}_i$ Standard Deviation Case: $\hat{s}_i = \hat{\delta}_0 + \hat{\delta}_1 \hat{Y}_i$.

Once the estimated variance (standard deviation) is obtained for each case, we get the estimated weights:

$$\text{Variance Case:} \quad \hat{w}_i = \frac{1}{\hat{v}_i} \qquad \text{Standard Deviation Case:} \quad \hat{w}_i = \frac{1}{\hat{s}_i^2}$$

Then we compute the estimated weighted least squares estimator as:

$$\hat{\boldsymbol{\beta}}^{\hat{W}} = \left(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{W}}\mathbf{Y} \qquad \hat{\mathbf{W}} = \begin{bmatrix} \hat{w}_1 & 0 & \cdots & 0 \\ 0 & \hat{w}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{w}_n \end{bmatrix}.$$

The process is continued until the estimated regression coefficients are stable from iteration to iteration. The estimated variance is:

$$\hat{V}\left\{\hat{\boldsymbol{\beta}}^{\hat{W}}\right\} = MSE_{\hat{W}}\left(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X}\right)^{-1} \qquad MSE_{\hat{W}} = \frac{1}{n-p'}\left(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}^{\hat{W}}\right)'\left(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}^{\hat{W}}\right).$$

It is also possible to estimate the power relation between the standard deviation and the mean $\sigma^2 \propto \mu^\delta$. The Variance and Standard deviation cases considered above correspond to $\delta = 1$ and 2, respectively.

$$\sigma^2 = c\mu^\delta \quad \Rightarrow \quad \ln\left(\sigma^2\right) = \ln c + \delta \ln \mu$$

At each stage in the iteration process, fit a regression of $\ln\left(e^2\right)$ on $\ln \overline{Y}$ to estimate $\delta$.

**Example: Construction Plant Maintenance Costs**

Edwards, Holt, and Harris (2000), studied the relationship between Maintenance Costs ($Y$) and $p = 4$ predictors: Machine Weight ($X_1$), and indicators for Industry Type ($X_2 = 1$=opencast coal, 0 if slate), Machine Type ($X_3 = 1$ if front shovel, 0 if backacter), and Company attitude to Oil Analysis ($X_4 = 1$, if regular use, 0 if not). The data, fitted values, and residuals are given in Table 6.17, based on a regression of $n = 33$ maintenance plants.

The program and output for the regression analysis is given below, as well as the Breusch-Pagan test for constant error variance. Plots of the residuals and absolute residuals versus fitted values are given in Figure 6.11. The correlation between the absolute residuals and fitted values is 0.6589, while the correlation between the squared residuals and fitted values is 0.5992. We will fit Estimated Weighted Least Squares, based on the Standard deviation case, regressing the absolute residuals on the predicted values.

**R Program**

```
cmc <- read.table("http://www.stat.ufl.edu/~winner/data/const_maint.dat",
     header=F,col.names=c("mach_id","mach_cost","coal","front_shov","use_oil","mach_wt"))
attach(cmc)

mach.mod1 <- lm(mach_cost ~ mach_wt + coal + front_shov + use_oil)
summary(mach.mod1); anova(mach.mod1)
yhat.1 <- predict(mach.mod1); e.1 <- resid(mach.mod1)

par(mfrow=c(2,1))
plot(yhat.1, e.1)
abline(h=0)
```

| Plant ID | $Y$ | $X_2$ | $X_3$ | $X_4$ | $X_1$ | $\hat{Y}_{OLS}$ | $e_{OLS}$ |
|---|---|---|---|---|---|---|---|
| 1 | 6.068 | 1 | 0 | 0 | 16.6 | 5.028 | 1.040 |
| 2 | 4.602 | 1 | 0 | 0 | 20.37 | 5.844 | -1.242 |
| 3 | 3.282 | 1 | 0 | 0 | 20.37 | 5.844 | -2.562 |
| 4 | 2.192 | 1 | 1 | 0 | 20.37 | 1.686 | 0.506 |
| 5 | 2.572 | 1 | 1 | 0 | 20.37 | 1.686 | 0.886 |
| 6 | 4.142 | 1 | 0 | 0 | 20.37 | 5.844 | -1.702 |
| 7 | 5.321 | 1 | 1 | 1 | 21 | 6.214 | -0.893 |
| 8 | 4.421 | 1 | 1 | 1 | 21 | 6.214 | -1.793 |
| 9 | 5.301 | 1 | 1 | 1 | 21 | 6.214 | -0.913 |
| 10 | 9.679 | 1 | 0 | 0 | 27.01 | 7.283 | 2.396 |
| 11 | 11.997 | 1 | 0 | 1 | 31 | 12.538 | -0.541 |
| 12 | 7.757 | 1 | 0 | 0 | 31 | 8.147 | -0.390 |
| 13 | 12.597 | 1 | 0 | 0 | 31 | 8.147 | 4.450 |
| 14 | 13.045 | 1 | 0 | 1 | 33.2 | 13.014 | 0.031 |
| 15 | 2.471 | 0 | 0 | 0 | 45 | 3.677 | -1.206 |
| 16 | 6.689 | 0 | 0 | 0 | 46.27 | 3.952 | 2.737 |
| 17 | 12.139 | 1 | 0 | 0 | 46.27 | 11.454 | 0.685 |
| 18 | 10.859 | 1 | 0 | 0 | 46.27 | 11.454 | -0.595 |
| 19 | 7.119 | 0 | 1 | 0 | 68.42 | 4.592 | 2.527 |
| 20 | 3.728 | 0 | 1 | 0 | 83.7 | 7.902 | -4.174 |
| 21 | 46.057 | 1 | 1 | 0 | 218 | 44.494 | 1.563 |
| 22 | 49.847 | 1 | 1 | 0 | 218 | 44.494 | 5.353 |
| 23 | 61.671 | 1 | 1 | 0 | 335 | 69.837 | -8.166 |
| 24 | 56.167 | 1 | 0 | 1 | 228 | 55.209 | 0.958 |
| 25 | 9.721 | 1 | 0 | 1 | 22.6 | 10.718 | -0.997 |
| 26 | 8.535 | 1 | 0 | 1 | 21.7 | 10.523 | -1.988 |
| 27 | 66.91 | 1 | 1 | 1 | 229 | 51.268 | 15.642 |
| 28 | 11.101 | 1 | 0 | 1 | 22.6 | 10.718 | 0.383 |
| 29 | 12.511 | 1 | 0 | 1 | 22.6 | 10.718 | 1.793 |
| 30 | 11.397 | 1 | 0 | 1 | 31 | 12.538 | -1.141 |
| 31 | 4.069 | 0 | 0 | 0 | 46.27 | 3.952 | 0.117 |
| 32 | 3.622 | 1 | 0 | 0 | 20.37 | 5.844 | -2.222 |
| 33 | 40.727 | 1 | 1 | 1 | 229 | 51.268 | -10.541 |

Table 6.17: Plant Maintenance Cost Data

```
plot(yhat.1, abs(e.1))
abline(lm(abs(e.1) ~ yhat.1))

library(lmtest)
bptest(mach_cost ~ coal + front_shov + use_oil + mach_wt, studentize=F)

plot(yhat.1, abs(e.1))
cor(yhat.1, abs(e.1))
plot(yhat.1, e.1^2)
cor(yhat.1, e.1^2)
print(cbind(cmc,yhat.1,e.1))
```

**R Output - OLS regression and Breusch-Pagan Test**

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.06990    2.09210  -2.901  0.00716 **
mach_wt      0.21661    0.01041  20.804  < 2e-16 ***
coal         7.50193    2.29401   3.270  0.00285 **
front_shov  -4.15823    1.82931  -2.273  0.03089 *
use_oil      4.39110    1.67713   2.618  0.01410 *

Residual standard error: 4.418 on 28 degrees of freedom
Multiple R-squared:  0.9518,    Adjusted R-squared:  0.9449
F-statistic: 138.3 on 4 and 28 DF,  p-value: < 2.2e-16

Analysis of Variance Table
Response: mach_cost
           Df  Sum Sq Mean Sq  F value    Pr(>F)
mach_wt     1 10164.6 10164.6 520.7042 < 2.2e-16 ***
coal        1   413.0   413.0  21.1570 8.281e-05 ***
front_shov  1    88.0    88.0   4.5071   0.04273 *
use_oil     1   133.8   133.8   6.8550   0.01410 *
Residuals  28   546.6    19.5

        Breusch-Pagan test
data:  mach_cost ~ coal + front_shov + use_oil + mach_wt
BP = 51.742, df = 4, p-value = 1.562e-10
> cor(yhat.1, abs(e.1))
[1] 0.6588788
> cor(yhat.1, e.1^2)
[1] 0.5991547
```

In the first stage of the EWLS algorithm, we regress the absolute residuals from the OLS regression on the corresponding fitted values.

**R Output - Regression of —e— on $\hat{Y}$**

```
> e.reg.ols <- lm(abs.e.ols ~ yhat.ols)
> summary(e.reg.ols)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.64678    0.57616   1.123     0.27
yhat.ols     0.11728    0.02405   4.877 3.06e-05 ***

Residual standard error: 2.499 on 31 degrees of freedom
```

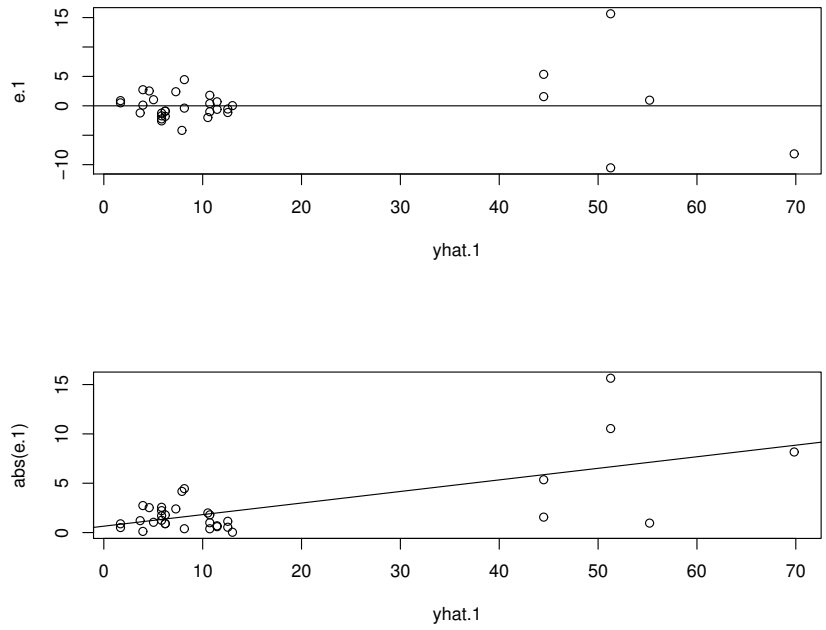


Figure 6.11: Residuals vs Fitted (Top) and —Residuals— vs Fitted (Bottom) - Maintenance Costs Example

```
Multiple R-squared:  0.4341,    Adjusted R-squared:  0.4159
F-statistic: 23.78 on 1 and 31 DF,  p-value: 3.057e-05
```

The predicted (fitted) values from this regression are the estimated standard deviations of the observations, as a function of their means. The weights are the inverse of the square of the estimated standard deviations. The "first round" uses these estimates.

$$\hat{Y}_i = -6.070 + 0.217X_{i1} + 7.502X_{i2} - 4.158X_{i3} + 4.391X_{i4} \qquad \hat{s}_i = 0.647 + 0.117\hat{Y}_i \qquad \hat{w}_i = \frac{1}{\hat{s}_i^2} \qquad \hat{\mathbf{W}}_{(0)} = \text{diag}\{\hat{w}_i\}$$

Next, obtain the first iteration of the EWLS estimator and its new prediction and residual vectors.

$$\hat{\boldsymbol{\beta}}_{(1)} = \left(\mathbf{X}'\hat{\mathbf{W}}_{(0)}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{W}}_{(0)}\mathbf{Y} \qquad \hat{\mathbf{Y}}_{(1)} = \mathbf{X}\hat{\boldsymbol{\beta}}_{(1)} \qquad \boldsymbol{e}_{(1)} = \mathbf{Y} - \hat{\mathbf{Y}}_{(1)}$$

The process continued for 3 iterations until the sum of squared differences of the estimated regression coefficient dropped below 0.0001. The iteration history and the final estimated variance-covariance are given below. The final estimates, standard errors, and $t$-statistics are given in Table 6.18.

| Parameter | Estimate | Std Error | t-stat |
|---|---|---|---|
| Constant | -5.935 | 1.177 | -5.041 |
| X1 | 0.221 | 0.017 | 12.715 |
| X2 | 7.328 | 1.010 | 7.256 |
| X3 | -3.952 | 0.718 | -5.501 |
| X4 | 3.495 | 0.846 | 4.130 |

Table 6.18: Plant Maintenance Cost - Estimated Weighted Least Squares Results

$$\hat{\boldsymbol{\beta}}_{OLS} = \begin{bmatrix} -6.0699 \\ 0.2166 \\ 7.5019 \\ -4.1582 \\ 4.3911 \end{bmatrix} \quad \hat{\boldsymbol{\beta}}_{(1)} = \begin{bmatrix} -5.9342 \\ 0.2202 \\ 7.3465 \\ -3.9819 \\ 3.5340 \end{bmatrix} \quad \hat{\boldsymbol{\beta}}_{(2)} = \begin{bmatrix} -5.9326 \\ 0.2206 \\ 7.3281 \\ -3.9525 \\ 3.4948 \end{bmatrix} \quad \hat{\boldsymbol{\beta}}^{\hat{W}} = \begin{bmatrix} -5.9346 \\ 0.2206 \\ 7.3275 \\ -3.9521 \\ 3.4954 \end{bmatrix}$$

$$\left(\mathbf{X}'\hat{\mathbf{W}}_{(3)}\mathbf{X}\right)^{-1} = \begin{bmatrix} 0.6507 & -0.0074 & -0.4546 & -0.0332 & 0.0162 \\ -0.0074 & 0.0001 & 0.0042 & -0.0007 & -0.0002 \\ -0.4546 & 0.0042 & 0.4788 & -0.0482 & -0.1114 \\ -0.0332 & -0.0007 & -0.0482 & 0.2423 & -0.0215 \\ 0.0162 & -0.0002 & -0.1114 & -0.0215 & 0.3363 \end{bmatrix}$$

$$MSE_{\hat{W}} = \frac{1}{n-p'}\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\hat{W}}\right)'\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\hat{W}}\right) = \left(\frac{1}{33-5}\right)(59.6469) = 2.1302$$

**R Program for EWLS**

```
cmc <- read.table("http://www.stat.ufl.edu/~winner/data/const_maint.dat",
     header=F,col.names=c("mach_id","mach_cost","coal","front_shov","use_oil","mach_wt"))
attach(cmc)

#### Matrix form (using lm for |e|,y-hat regressions) ############

n <- length(mach_cost)
X0 <- rep(1,n)
X <- as.matrix(cbind(X0,mach_wt,coal,front_shov,use_oil))
Y <- as.matrix(mach_cost)
p.star <- ncol(X)

#### Fit original regression, and regress |e| on Y-hat

b.ols <- solve(t(X) %*% X) %*% t(X) %*% Y
yhat.ols <- X %*% b.ols
e.ols <- Y - yhat.ols
abs.e.ols <- abs(e.ols)
e.reg.ols <- lm(abs.e.ols ~ yhat.ols)
summary(e.reg.ols)
s.ols <- predict(e.reg.ols)
```

```
w.ols <- 1/s.ols^2

b.old <- b.ols
wm.old <- as.matrix(diag(w.ols))
b.diff <- 100
num.iter <- 0

while (b.diff > 0.0001) {
num.iter <- num.iter + 1
b.new <- solve(t(X) %*% wm.old %*% X) %*% t(X) %*% wm.old %*% Y
yhat.new <- X %*% b.new
abs.e.new <- abs(Y - yhat.new)
wm.new <- as.matrix(diag(1/predict(lm(abs.e.new~yhat.new))^2))
b.diff <- sum((b.new-b.old)^2)
b.old <- b.new
wm.old <- wm.new
print(b.old)
}

num.iter
b.wls <- b.new
wm.wls <- wm.new
mse.w <- (t(Y-X%*%b.wls) %*% wm.wls %*% (Y-X%*%b.wls))/(n-p.star)
s2.b.wls <- mse.w[1,1]*solve(t(X) %*% wm.wls %*% X)
s.b.wls <- sqrt(diag(s2.b.wls))
t.b.wls <- b.wls/s.b.wls
print(round(cbind(b.wls,s.b.wls,t.b.wls),3))
```

$\nabla$

**Analysis Based on Estimated Variances with Replicated $X$ Values**

When the data collection process is based on a well designed controlled experiment, with multiple cases for each set of $X$ levels, the variance of the errors can be estimated within each distinct group, and used in the estimated weighted least squares equation directly. If we have $g$ groups of observations with distinct $X$ levels, with the $j^{th}$ having $n_j$ observations and sample variance $s_j^2$. Then, the vector of responses, $\mathbf{Y}$, its estimated variance-covariance matrix, and the estimated weight matrix for Estimated Weighted Least Squares (EWLS) are as follow.

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_g \end{bmatrix} \qquad \mathbf{Y}_j = \begin{bmatrix} Y_{1j} \\ Y_{2j} \\ \vdots \\ Y_{n_j j} \end{bmatrix} \qquad \hat{\mathbf{\Sigma}}_Y = \begin{bmatrix} s_1^2\mathbf{I}_{n_1\times n_1} & \mathbf{0}_{n_1\times n_2} & \cdots & \mathbf{0}_{n_1\times n_g} \\ \mathbf{0}_{n_2\times n_1} & s_2^2\mathbf{I}_{n_2\times n_2} & \cdots & \mathbf{0}_{n_2\times n_g} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{n_g\times n_1} & \mathbf{0}_{n_g\times n_2} & \cdots & s_g^2\mathbf{I}_{n_g\times n_g} \end{bmatrix} \qquad \hat{\mathbf{W}} = \hat{\mathbf{\Sigma}}_Y^{-1}$$

The EWLS estimator does not need to be iterated, so this is very simple to implement in any matrix spreadsheet or software package.

$$\hat{\boldsymbol{\beta}}^{\hat{W}} = \left(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{W}}\mathbf{Y} \qquad \hat{V}\left\{\hat{\boldsymbol{\beta}}^{\hat{W}}\right\} = MSE_{\hat{W}}\left(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X}\right)^{-1}$$

| Distance ($j$) | $Y_{1j}$ | $Y_{2j}$ | $Y_{3j}$ | $Y_{4j}$ | $Y_{5j}$ | $Y_{6j}$ | $\overline{Y}_j$ | $s_j$ |
|---|---|---|---|---|---|---|---|---|
| 10 (1) | 3.01 | 3.02 | 3.29 | 3.00 | 3.20 | 3.11 | 3.105 | 0.119 |
| 20 (2) | 5.57 | 5.00 | 5.42 | 5.73 | 5.29 | 5.10 | 5.353 | 0.280 |
| 30 (3) | 8.09 | 6.80 | 7.95 | 8.62 | 8.41 | 8.62 | 8.082 | 0.685 |
| 40 (4) | 10.81 | 10.19 | 13.01 | 11.17 | 11.33 | 9.35 | 10.797 | 1.232 |
| 50 (5) | 16.07 | 14.90 | 17.47 | 14.21 | 13.13 | 11.93 | 14.618 | 1.996 |

Table 6.19: Shotgun Pellet Spread by Distance

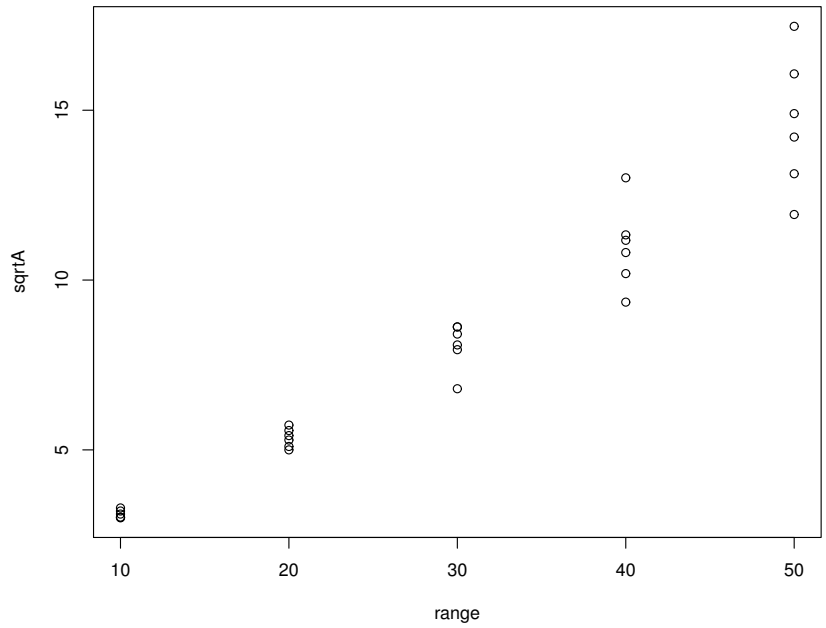


Figure 6.12: Plot of spread versus Range

**Example: Shotgun Pellet Spread by Distance**

Rowe and Hanson (1985) give results from a forensic experiment to relate shotgun pellet spread as function of distance. The data given here are from a Remington 12-gauge shotgun shooting Winchester Super X 12-gauge buckshot cartridges. The response $Y$, was the square root of the area of the smallest rectangle that covered the pellet pattern. There were $g = 5$ ranges ($X$,in feet) with $n_j = 6$ replicates per distance. The data, along with group means and standard deviations are given in Table 6.19. A plot of the data is given in Figure 6.12. Plots of the mean and standard deviation versus range are given in Figure 6.13.

Clearly the standard deviation (and variance) increase with the mean. Further, the mean of $Y$ is not linearly related to the range ($X$), as it bends upward. This leads to fitting a quadratic regression model, with Estimated Weighted Least Squares.

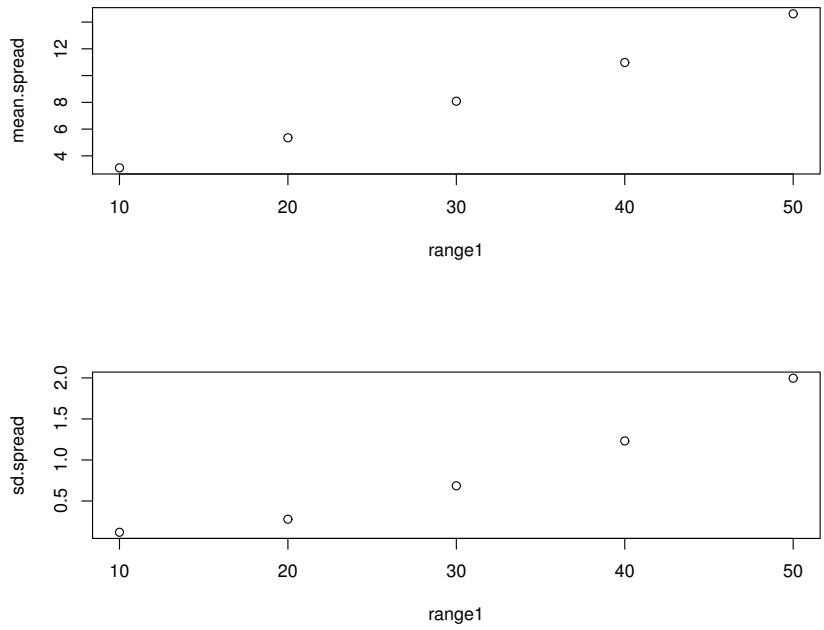


Figure 6.13: Plots of mean and SD of spread vs Range

$$Y_{ij} = \beta_0 + \beta_1 X_j + \beta_2 X_j^2 + \epsilon_{ij} \qquad i = 1, \ldots, 6 \quad j = 1, \ldots 5 \quad X_j = 10j \qquad \text{with weights} \quad \hat{w}_{ij} = \frac{1}{s_j^2}$$

The R program for the matrix form of Ordinary and Estimated Least Squares are given below and results are summarized in Table 6.20.

```
### Program
sg_spread <- read.csv("http://www.stat.ufl.edu/~winner/sta4210/mydata/shotgun_spread.csv",
      header=T)

attach(sg_spread); names(sg_spread)

sg_spread1 <- sg_spread[cartridge==1,]
detach(sg_spread); attach(sg_spread1)

### Matrix Approach

X <- cbind(rep(1,length(sqrtA)),range,range^2)
W <- diag(1/SD_range^2)
Y <- sqrtA

(beta.ols <- solve(t(X) %*% X) %*% t(X) %*% Y)
e.ols <- Y - X %*% beta.ols
(sse.ols <- t(e.ols) %*% e.ols)
(s2.ols <- sse.ols/(length(Y)-ncol(X)))
```

| | Ordinary Least Squares | | | Estimated Weighted Least Squares | | | OLS - Robust SEs | | |
|---|---|---|---|---|---|---|---|---|---|
| Parameter | Estimate | Std Error | $t$-stat | Estimate | Std Error | $t$-stat | Estimate | Std Error | $t$-stat |
| Constant | 1.30867 | 0.92959 | 1.408 | 1.25564 | 0.24319 | 5.163 | 1.30867 | 0.56985 | 2.297 |
| $X$ | 0.15987 | 0.07084 | 2.257 | 0.16460 | 0.02895 | 5.686 | 0.15987 | 0.05921 | 2.700 |
| $X^2$ | 0.00211 | 0.00116 | 1.822 | 0.00203 | 0.00068 | 2.992 | 0.00211 | 0.00118 | 1.795 |

Table 6.20: OLS and EWLS estimates, standard errors, and $t$-tests

```
(v.beta.ols <- s2.ols[1,1] * solve(t(X) %*% X))
se.beta.ols <- sqrt(diag(v.beta.ols))
t.beta.ols <- beta.ols/se.beta.ols

(beta.wls <- solve(t(X) %*% W %*% X) %*% t(X) %*% W %*% Y)
e.wls <- Y - X %*% beta.wls
(sse.wls <- t(e.wls) %*% W %*% e.wls)
(s2.wls <- sse.wls/(length(Y)-ncol(X)))
(v.beta.wls <- s2.wls[1,1] * solve(t(X) %*% W %*% X))
se.beta.wls <- sqrt(diag(v.beta.wls))
t.beta.wls <- beta.wls/se.beta.wls


print(round(cbind(beta.ols,se.beta.ols,t.beta.ols,
beta.wls,se.beta.wls,t.beta.wls),4))

### Output
> (beta.ols <- solve(t(X) %*% X) %*% t(X) %*% Y)
      1.308666667
range 0.159873810
      0.002110714
> (sse.ols <- t(e.ols) %*% e.ols)
[1,] 30.43237
> (s2.ols <- sse.ols/(length(Y)-ncol(X)))
[1,] 1.127125
> (v.beta.ols <- s2.ols[1,1] * solve(t(X) %*% X))
       0.8641289841 -6.199186e-02  9.392706e-04
range -0.0619918619  5.018389e-03 -8.050891e-05
       0.0009392706 -8.050891e-05  1.341815e-06

> (beta.wls <- solve(t(X) %*% W %*% X) %*% t(X) %*% W %*% Y)
      1.255644821
range 0.164596451
      0.002029814
> (sse.wls <- t(e.wls) %*% W %*% e.wls)
[1,] 25.09147
> (s2.wls <- sse.wls/(length(Y)-ncol(X)))
[1,] 0.9293137
> (v.beta.wls <- s2.wls[1,1] * solve(t(X) %*% W %*% X))
       0.059140712 -6.868167e-03  1.494750e-04
range -0.006868167  8.380629e-04 -1.895007e-05
       0.000149475 -1.895007e-05  4.603523e-07
```

The R commands and output for OLS and EWLS (with the **weight** statement are given below, using the **lm** function.

```
### Program
sg_spread <- read.csv("http://www.stat.ufl.edu/~winner/sta4210/mydata/shotgun_spread.csv",
      header=T)
```

```
attach(sg_spread); names(sg_spread)

sg_spread1 <- sg_spread[cartridge==1,]
detach(sg_spread); attach(sg_spread1)

regweight <- 1/(SD_range^2)
### Ordinary Least Squares
sg.mod1 <- lm(sqrtA ~ range + I(range^2))
summary(sg.mod1)

#### Weighted Least Squares
sg.mod2 <- lm(sqrtA ~ range + I(range^2),
  weight=regweight)
summary(sg.mod2)

### Output
> ### Ordinary Least Squares
> sg.mod1 <- lm(sqrtA ~ range + I(range^2))
> summary(sg.mod1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.308667   0.929585   1.408   0.1706
range       0.159874   0.070841   2.257   0.0323 *
I(range^2)  0.002111   0.001158   1.822   0.0795 .

Residual standard error: 1.062 on 27 degrees of freedom
Multiple R-squared:  0.9422,    Adjusted R-squared:  0.9379
F-statistic: 220.2 on 2 and 27 DF,  p-value: < 2.2e-16

> #### Weighted Least Squares
> sg.mod2 <- lm(sqrtA ~ range + I(range^2),
+   weight=regweight)
> summary(sg.mod2)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.2556448  0.2431886   5.163 1.96e-05 ***
range       0.1645965  0.0289493   5.686 4.86e-06 ***
I(range^2)  0.0020298  0.0006785   2.992  0.00586 **

Residual standard error: 0.964 on 27 degrees of freedom
Multiple R-squared:  0.9754,    Adjusted R-squared:  0.9736
F-statistic: 535.4 on 2 and 27 DF,  p-value: < 2.2e-16
```

$\nabla$

### Robust Standard Errors

In general, the variance-covariance matrix of the Ordinary Least Squares estimator is:

$$V\left\{\hat{\boldsymbol{\beta}}\right\} = V\left\{\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}\right\} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{\mathbf{Y}}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

A simple, yet effective method is to obtain robust standard errors of the ordinary least squares (OLS) estimators based on the residuals from the linear regression (using the squared residuals as estimates of the

variances for the individual cases). This method was originally proposed by White (1980). The estimated variance-covariance matrix with resulting **robust to heteroskedasticity standard errors** for $\hat{\boldsymbol{\beta}}$ is:

$$\hat{V}\left\{\hat{\boldsymbol{\beta}}\right\} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{E}}_{\mathbf{2}}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \qquad \hat{\mathbf{E}}_{\mathbf{2}} = \begin{bmatrix} e_1^2 & 0 & \cdots & 0 \\ 0 & e_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e_n^2 \end{bmatrix} = \text{diag}\left(\boldsymbol{e}\boldsymbol{e}'\right).$$

**Example: Shotgun Pellet Spread by Distance**

For the shotgun pellet spread data, we obtain the OLS estimator for $\boldsymbol{\beta}$ and obtain robust standard errors in matrix form. The results for the estimated coefficients, robust standard errors and $t$-statistics are given in Table 6.20.

```
### Robust Standard Errors
(beta.ols <- solve(t(X) %*% X) %*% t(X) %*% Y)
e.ols <- Y - X %*% beta.ols
E2 <- e.ols %*% t(e.ols) * diag(length(Y))
(v.beta.ols.robust<-  solve(t(X) %*% X) %*% t(X) %*% E2 %*%
   X %*% solve(t(X) %*% X))
se.beta.ols.robust <- sqrt(diag(v.beta.ols.robust))
t.beta.ols.robust <- beta.ols/se.beta.ols.robust

print(round(cbind(beta.ols,se.beta.ols.robust,t.beta.ols.robust),5))
```

```
> (v.beta.ols.robust<-  solve(t(X) %*% X) %*% t(X) %*% E2 %*%
+    X %*% solve(t(X) %*% X))
                            range
       0.3247278501 -3.349684e-02  6.465852e-04
range -0.0334968425  3.505749e-03 -6.867561e-05
       0.0006465852 -6.867561e-05  1.383327e-06
```

$\nabla$

## 6.12.2 Bootstrap Methods When Distribution of Errors is Unknown

The bootstrap is widely used in many applications when the distribution of the data is unknown, or when the distribution of the estimator of is unknown. In regression applications, there are various ways of bootstrapping (see e.g. Cameron and Trivedi (2009, Chapter 13) and Kutner, et al (2005, Section 11.5)). All sampling is done with replacement (except the parametric bootstrap).

One possibility is to bootstrap the individual cases from the dataset, and repeatedly re-fit the regression, and saving the regression coefficients, obtaining their standard error. Then we can construct Confidence

| Lager ID | X | Y | Lager ID | X | Y | Lager ID | X | Y | Lager ID | X | Y |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 148.23 | 0.66 | 11 | 169.51 | 0.53 | 21 | 159.81 | 0.56 | 31 | 177.83 | 0.63 |
| 2 | 160.38 | 0.63 | 12 | 111.05 | 0.32 | 22 | 163.23 | 0.56 | 32 | 150.11 | 0.64 |
| 3 | 170.41 | 0.62 | 13 | 143.50 | 0.42 | 23 | 169.59 | 0.53 | 33 | 135.92 | 0.48 |
| 4 | 208.65 | 0.90 | 14 | 186.96 | 0.64 | 24 | 135.76 | 0.49 | 34 | 162.99 | 0.96 |
| 5 | 146.03 | 0.64 | 15 | 109.50 | 0.42 | 25 | 198.62 | 0.58 | 35 | 183.54 | 0.63 |
| 6 | 180.19 | 0.62 | 16 | 209.95 | 0.77 | 26 | 221.94 | 0.68 | 36 | 236.37 | 0.86 |
| 7 | 169.06 | 0.58 | 17 | 88.47 | 0.30 | 27 | 148.80 | 0.56 | 37 | 163.23 | 0.58 |
| 8 | 119.04 | 0.47 | 18 | 230.25 | 0.70 | 28 | 120.02 | 0.36 | 38 | 212.48 | 0.80 |
| 9 | 158.99 | 0.59 | 19 | 152.96 | 0.51 | 29 | 84.64 | 0.24 | 39 | 235.06 | 0.75 |
| 10 | 153.04 | 0.72 | 20 | 147.42 | 0.53 | 30 | 238.33 | 0.97 | 40 | 267.27 | 0.91 |

Table 6.21: Total Phenolic Content ($X$) and DPPH Radical Scavenging Activity ($Y$) for 40 lager beers

Intervals for the regression coefficients by taking the original estimate and adding and subtracting 2 standard errors for approximate 95% Confidence Intervals. This method is widely used when the $X$ levels are random (not set up by the experimenter), and when the errors may not have constant variance. Also, the cut-off values for the middle $(1-\alpha)100\%$ of the bootstrap estimates can be used.

Another possibility that is useful is to retain the fitted values from the original regression $\hat{Y}_1, \ldots, \hat{Y}_n$ and bootstrap the residuals $e_1, \ldots, e_n$. Then the bootstrapped residuals are added to the original fitted values and the regression coefficients are obtained, and their standard error is computed and used as above.

The reflection method (see e.g. Kutner, et al (2005, Section 11.5)) is another possibility. In this case, we obtain the lower $\alpha/2$ percentile of the bootstrapped regression coefficients $\left(\hat{\beta}_j^*(\alpha/2)\right)$ and the upper $1-\alpha/2$ percentile of the regression coefficients $\left(\hat{\beta}_j^*(1-\alpha/2)\right)$ and obtain the interval:

$$\hat{\beta}_j - \hat{\beta}_j^*(\alpha/2) \leq \beta_j \leq \hat{\beta}_j^*(1-\alpha/2) - \hat{\beta}_j \quad j = 0, 1, \ldots, p.$$

In the parametric bootstrap, the residuals are sampled from a specified distribution with parameter(s) estimated from the original sample.

There are various bias-corrected methods applied as well that are computed by statistical software packages.

**Example: Total Phenolic Content and DPPH Radical Scavenging Activity in Lager Beers**

Zhao, Li, Sun, Yang, and Zhao (2013) report the results of a study relating antioxidant activity to phenolic content in $n = 40$ lager beers. The response is DPPH Radical Scavenging Activity ($Y$), and the predictor is Total Phenolic Content ($X$). The data are given in Table 6.21 and plotted in Figure 6.14 along with the OLS simple linear regression line.

The standard output for the regression model is given below.

```
lager <- read.csv("http://www.stat.ufl.edu/~winner/data/lager_antioxidant_reg.csv",
```

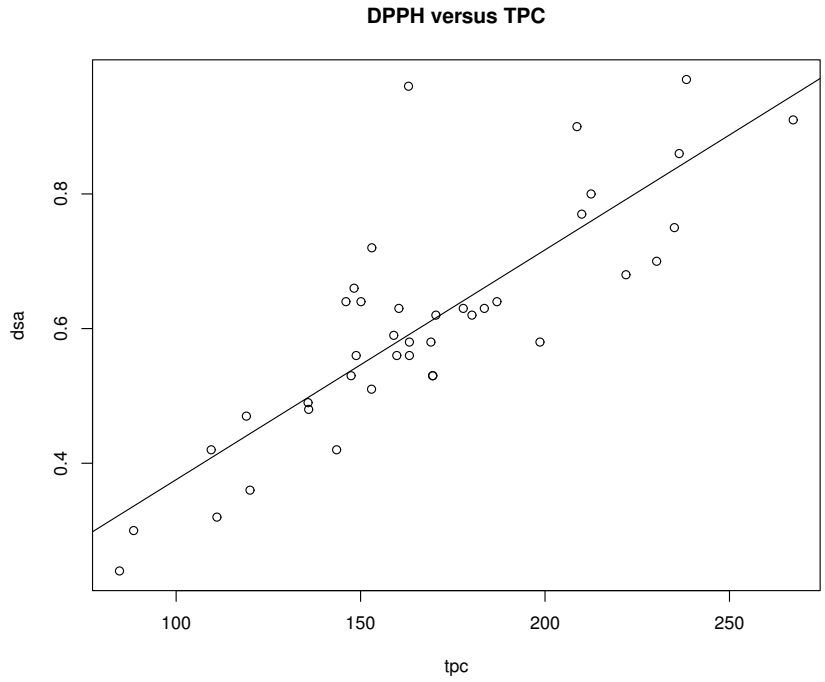


Figure 6.14: Plot of DPPH Radical Scavenging Activity vs Total Phenolic Content and Fitted Equation - Lager Beer Data

```
      header=T)
attach(lager); names(lager)


# print(cbind(tpc,dsa))

lager.mod1 <- lm(dsa ~ tpc)
summary(lager.mod1)
anova(lager.mod1)
confint(lager.mod1)

plot(tpc,dsa, main="DPPH versus TPC")
abline(lager.mod1)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0343018  0.0639781   0.536    0.595
tpc         0.0034132  0.0003694   9.240 2.93e-11 ***

Residual standard error: 0.09629 on 38 degrees of freedom
Multiple R-squared:  0.692,     Adjusted R-squared:  0.6839
F-statistic: 85.38 on 1 and 38 DF,  p-value: 2.926e-11

> anova(lager.mod1)
Analysis of Variance Table
Response: dsa
          Df  Sum Sq Mean Sq F value    Pr(>F)
tpc        1 0.79175 0.79175  85.385 2.926e-11 ***
Residuals 38 0.35236 0.00927

> confint(lager.mod1)
                  2.5 %      97.5 %
(Intercept) -0.09521518 0.163818791
tpc          0.00266544 0.004160979
```

We next apply the two versions of the bootstrap. The first involves saving the fitted values and residuals from the regression of the actual dataset, then bootstrapping the residuals and adding them to the fixed fitted values for the "new data" and regressing them on the original $X$ values. The program below fits the original regression, and saves $\hat{\mathbf{Y}}$ and $\boldsymbol{e}$, and the estimated slope $\hat{\beta}_1$.

```
lager <-
read.csv("http://www.stat.ufl.edu/~winner/sta4210/mydata/lager_antioxidant_reg.csv",
         header=TRUE)
attach(lager); names(lager)

lager.mod1 <- lm(dsa ~ tpc)
summary(lager.mod1)
confint(lager.mod1)
yhat <- predict(lager.mod1)
e <- residuals(lager.mod1)
b1 <- coef(lager.mod1)[2]
```

The following part of the program sets up the "fixed" $\mathbf{X}$ matrix and $\hat{\mathbf{Y}}$ vector. Then it samples $n = 40$ elements of $\boldsymbol{e}$ with replacement, and adds them to $\hat{\mathbf{Y}}$ to obtain $\mathbf{Y}_{\text{boot}}$. We then obtain the least squares estimate $\boldsymbol{\beta}_{\text{boot}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_{\text{boot}}$. We then save the estimated slope $\hat{\beta}_{1,\text{boot}}$. The process is repeated 10000 times. Confidence intervals for $\beta_1$ are obtained directly from the empirical "bootstrap distribution" and from the reflection method. The histogram is given in Figure 6.15.

```
############## Bootstrap w/ Fixed X's. Adds random errors to predicted values
n <- length(dsa)
X <- as.matrix(cbind(rep(1,n),tpc))

set.seed(13579)
num.boot <- 10000
b1.boot <- rep(0,num.boot)

for (i in 1:num.boot) {
e.boot <- as.matrix(sample(e,size=n,replace=TRUE))
Y.boot <- yhat + e.boot

b.boot <- solve(t(X) %*% X) %*% t(X) %*% Y.boot
b1.boot[i] <- b.boot[2,1]
}

hist(b1.boot, breaks=24,main="Bootstrap Method 1")

(b1.boot_025 <- quantile(b1.boot,.025))
(b1.boot_975 <- quantile(b1.boot,.975))

(b1.boot.sd <- sd(b1.boot))

(d1 <- b1-b1.boot_025)
(d2 <- b1.boot_975-b1)

(beta1.95CI <- c(b1-d2,b1+d1))
```

```
> (b1.boot_025 <- quantile(b1.boot,.025))
       2.5%
0.002700025
> (b1.boot_975 <- quantile(b1.boot,.975))
      97.5%
0.004141108
> (b1.boot.sd <- sd(b1.boot))
[1] 0.0003603104
> (d1 <- b1-b1.boot_025)
0.0007131838
> (d2 <- b1.boot_975-b1)
0.0007278994
> (beta1.95CI <- c(b1-d2,b1+d1))
        tpc         tpc
0.002685310 0.004126393
```

The second method involves sampling the observed $(X, Y)$ pairs with replacement for samples of size $n = 40$, and fitting the regression on each sample, saving the estimated slope each time. The program and output are given below, and the histogram is given in Figure 6.16.

```
###### Bootstrap by selecting n (X,Y) pairs with replacement
num.boot <- 10000
set.seed(34567)
b1.boot <- rep(0,num.boot)

for (i in 1:num.boot)  {
boot.sample <- sample(1:n,size=n,replace=TRUE)
dsa.b <- dsa[boot.sample]
tpc.b <- tpc[boot.sample]
X.boot <- as.matrix(cbind(rep(1,n),tpc.b))
Y.boot <- dsa.b
b.boot <- solve(t(X.boot) %*% X.boot) %*% t(X.boot) %*% Y.boot
b1.boot[i] <- b.boot[2,1]
```

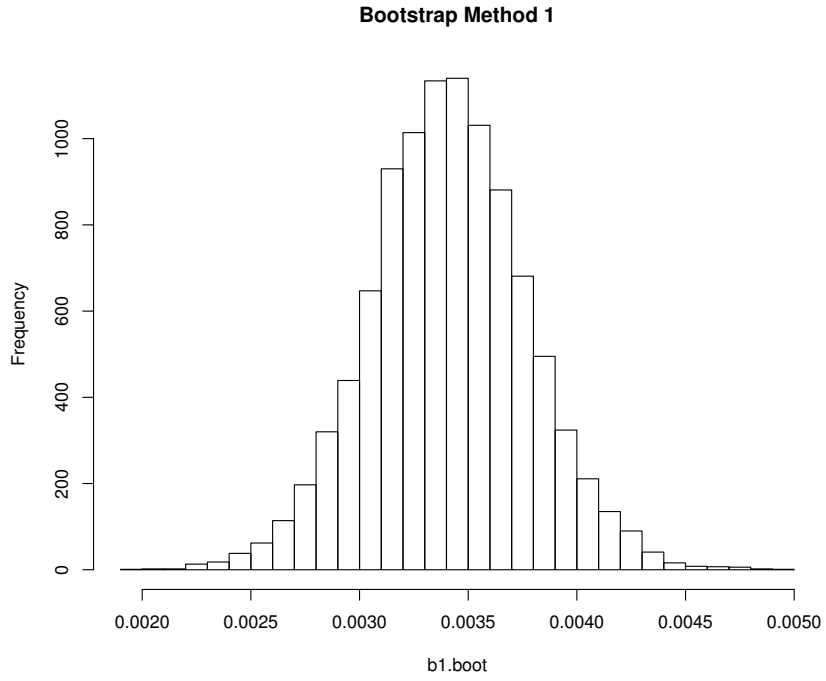


Figure 6.15: Histogram of Bootstrap Estimates of Regression Slope

```
}

hist(b1.boot, breaks=24,main="Bootstrap Method 2")

(b1.boot_025 <- quantile(b1.boot,.025))
(b1.boot_975 <- quantile(b1.boot,.975))

(b1.boot.sd <- sd(b1.boot))

(d1 <- b1-b1.boot_025)
(d2 <- b1.boot_975-b1)

(beta1.95CI <- c(b1-d2,b1+d1))

> (b1.boot_025 <- quantile(b1.boot,.025))
0.002754063
> (b1.boot_975 <- quantile(b1.boot,.975))
0.004018062
> (b1.boot.sd <- sd(b1.boot))
[1] 0.0003181617
> (d1 <- b1-b1.boot_025)
0.0006591461
> (d2 <- b1.boot_975-b1)
0.0006048525
> (beta1.95CI <- c(b1-d2,b1+d1))
0.002808357 0.004072355
```

The original interval, based on the $t$-distribution and the bootstrap intervals are very similar in this case, as the data are very "well-behaved" in terms of their residuals.

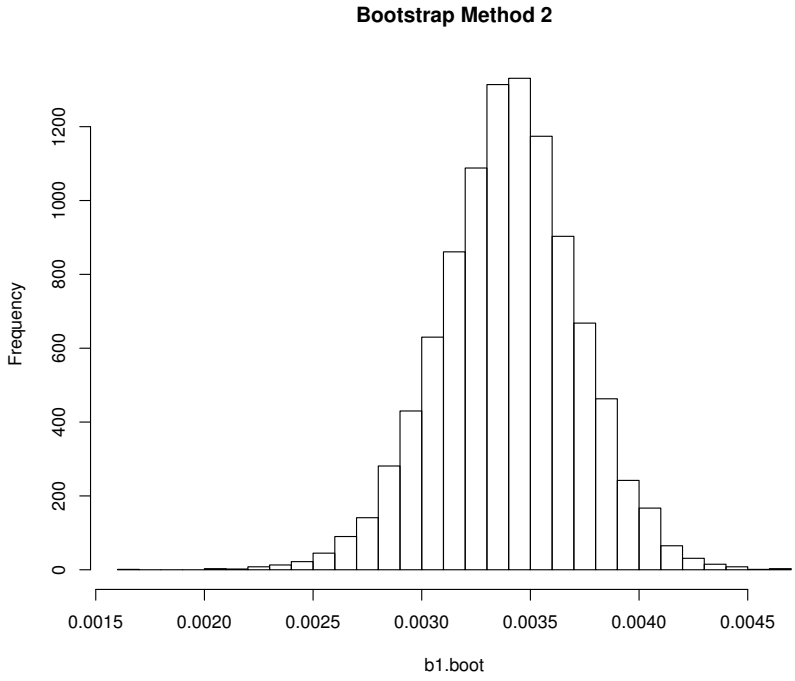


Figure 6.16: Histogram of slope estimates from bootstrap with $(X, Y)$ pairs being re-sampled

$\nabla$

## 6.13 Generalized Least Squares for Correlated Errors

Typically when data are collected over time and/or space, the errors are correlated, with correlation tending to be higher among observations that are closer in time or space. In this case, the variance-covariance matrix of the error terms is written generally:

$$V\{\boldsymbol{\varepsilon}\} = \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1n} & \sigma_{2n} & \cdots & \sigma_n^2 \end{bmatrix}.$$

For the case where the observations are equally spaced in time, and the error terms form an autoregressive process of order 1, we have:

$$\epsilon_t = \nu_t + \rho\epsilon_{t-1} \quad -1 < \rho < 1 \qquad \nu_t \sim iid\left(0, \sigma^2\right) \qquad \{\nu_{t+k}\} \perp \{\epsilon_t\} \quad \forall \quad k > 0$$

Note that this autoregressive process can extend back to $q$ time points, but the covariance structure gets more complicated. If we start with the definition that $E\{\epsilon_1\} = 0$ and $V\{\epsilon_1\} = \frac{\sigma^2}{1-\rho^2}$, we obtain:

$$E\{\epsilon_2\} = E\{\nu_2 + \rho\epsilon_1\} = 0$$

$$V\{\epsilon_2\} = V\{\nu_2 + \rho\epsilon_1\} = V\{\nu_2\} + V\{\rho\epsilon_1\} + 2\text{COV}\{\nu_2, \rho\epsilon_1\} = \sigma^2 + \frac{\rho^2\sigma^2}{1-\rho^2} + 2(0) = \frac{\sigma^2}{1-\rho^2}.$$

$$\text{COV}\{\epsilon_1, \epsilon_2\} = \text{COV}\{\epsilon_1, \nu_2 + \rho\epsilon_1\} = \frac{\rho\sigma^2}{1-\rho^2}$$

In general, this extends to the following results.

$$V\{\epsilon_t\} = \frac{\sigma^2}{1-\rho^2} = \gamma(0) \qquad \text{COV}\{\epsilon_t, \epsilon_{t+k}\} = \frac{\rho^{|k|}\sigma^2}{1-\rho^2} = \gamma(k) \qquad \rho = \frac{\text{COV}\{\epsilon_t, \epsilon_{t+1}\}}{V\{\epsilon_t\}} = \frac{\gamma(1)}{\gamma(0)}$$

$$V\{\boldsymbol{\varepsilon}\} = \boldsymbol{\Sigma} = \frac{\sigma^2}{1-\rho^2}\begin{bmatrix} 1 & \rho & \cdots & \rho^{n-1} \\ \rho & 1 & \cdots & \rho^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \cdots & 1 \end{bmatrix}$$

If $\rho$ were known, then we could use **Generalized Least Squares** to estimate $\boldsymbol{\beta}$. Let $\boldsymbol{\Sigma} = \sigma^2\mathbf{W}$. Then we would have:

$$\hat{\boldsymbol{\beta}}^{\mathbf{GLS}} = \left(\mathbf{X}'\mathbf{W}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}^{-1}\mathbf{Y} \qquad s^2 = \frac{1}{n-p'}\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathbf{GLS}}\right)'\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathbf{GLS}}\right) \qquad \hat{V}\left\{\hat{\boldsymbol{\beta}}^{\mathbf{GLS}}\right\} = s^2\left(\mathbf{X}'\mathbf{W}^{-1}\mathbf{X}\right)^{-1}.$$

In practice, $\rho$ will be unknown, and can be estimated from the data. Further, if we make the following transformations for $AR(1)$ errors, the transformed response will have uncorrelated errors:

$$\mathbf{Y}^* = \mathbf{T}^{-1}\mathbf{Y} \qquad \mathbf{X}^* = \mathbf{T}^{-1}\mathbf{X} \qquad \mathbf{T}^{-1} = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & \cdots & 0 & 0 & 0 \\ -\rho & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -\rho & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\rho & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & -\rho & 1 \end{bmatrix}$$

For this model, the transformed $\mathbf{Y}^*$, has the following model and variance-covariance structure:

$$\mathbf{Y}^* = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{T}^{-1}\boldsymbol{\varepsilon} \qquad V\{\mathbf{Y}^*\} = \mathbf{T}^{-1}\sigma^2\mathbf{W}\mathbf{T}^{-1'} = \sigma^2\mathbf{I}$$

Then for **Estimated Generalized Least Squares (EGLS)**, also known as **Feasible Generalized Least Squares (FGLS)**, we obtain estimates of $\rho$ and $\sigma^2$ based on the residuals from the OLS regression, then re-fit the EGLS model.

$$\hat{\gamma}(0) = \frac{\sum_{t=1}^n e_t^2}{n} \qquad \hat{\gamma}(1) = \frac{\sum_{t=2}^n e_t e_{t-1}}{n} \qquad \hat{\rho} = \frac{\hat{\gamma}(1)}{\hat{\gamma}(0)} \qquad \hat{\sigma}^2 = \hat{\gamma}(0) - \hat{\rho}\hat{\gamma}(1)$$

Next, we obtain the "estimated" transformation matrix, and the transformed $\mathbf{Y}^*$ and $\mathbf{X}^*$:

$$\mathbf{Y}^* = \hat{\mathbf{T}}^{-1}\mathbf{Y} \qquad \mathbf{X}^* = \hat{\mathbf{T}}^{-1}\mathbf{X} \qquad \hat{\mathbf{T}}^{-1} = \begin{bmatrix} \sqrt{1-\hat{\rho}^2} & 0 & 0 & \cdots & 0 & 0 & 0 \\ -\hat{\rho} & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -\hat{\rho} & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\hat{\rho} & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & -\hat{\rho} & 1 \end{bmatrix}$$

Note that the transformed responses have the following pattern (the predictors (and intercept) will have a similar structure).

$$Y_1^* = \sqrt{1-\hat{\rho}^2}Y_1 \qquad Y_t^* = Y_t - \hat{\rho}Y_{t-1} \quad t = 2, \ldots, n$$

The EGLS estimator, its variance-covariance matrix, and the estimator for $\sigma^2$ are obtained as follow.

$$\hat{\boldsymbol{\beta}}^{EGLS} = \left(\mathbf{X}'\hat{\mathbf{T}}^{-1'}\hat{\mathbf{T}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{T}}^{-1'}\hat{\mathbf{T}}^{-1}\mathbf{Y} = \left(\mathbf{X}^{*'}\mathbf{X}^*\right)^{-1}\mathbf{X}^{*'}\mathbf{Y}^*$$

$$\hat{V}\left\{\hat{\boldsymbol{\beta}}^{EGLS}\right\} = \hat{\sigma}_e^2\left(\mathbf{X}'\hat{\mathbf{T}}^{-1'}\hat{\mathbf{T}}^{-1}\mathbf{X}\right)^{-1} = \hat{\sigma}_e^2\left(\mathbf{X}^{*'}\mathbf{X}^*\right)^{-1}$$

$$\hat{\sigma}_e^2 = \frac{\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{EGLS}\right)'\hat{\mathbf{T}}^{-1'}\hat{\mathbf{T}}^{-1}\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{EGLS}\right)}{n-p'-1} = \frac{\left(\mathbf{Y}^* - \mathbf{X}^*\hat{\boldsymbol{\beta}}^{EGLS}\right)'\left(\mathbf{Y}^* - \mathbf{X}^*\hat{\boldsymbol{\beta}}^{EGLS}\right)}{n-p'-1}$$

Tests and Confidence Intervals for regression coefficients are obtained as follow.

$$H_0 : \beta_j = 0 \quad H_A : \beta_j \neq 0 \qquad TS : t_{obs} = \frac{\hat{\beta}_j^{EGLS}}{\hat{SE}\left\{\hat{\beta}_j^{EGLS}\right\}} \qquad RR : |t_{obs}| \geq t\left(n-p'-1, \alpha/2\right)$$

$$(1-\alpha)100\% \text{ Confidence Interval: } \hat{\beta}_j^{EGLS} \pm t\left(n-p'-1, \alpha/2\right)\hat{SE}\left\{\hat{\beta}_j^{EGLS}\right\}$$

A test for the autoregressive parameter $\rho$ is obtained as follows:

$$s^2 = \frac{\hat{\gamma}(0) - \hat{\rho}\hat{\gamma}(1)}{n-p'-1} \qquad \hat{SE}\left\{\hat{\rho}\right\} = \sqrt{\frac{s^2}{\hat{\gamma}(0)}} \qquad t_{obs} = \frac{\hat{\rho}}{\hat{SE}\left\{\hat{\rho}\right\}}$$

This procedure was first described by Gallant and Goebel (1976) that can be applied to linear or nonlinear regression problems. The method is the basis for the **Autoreg** procedure in SAS. R has a function **gls** in the **nlme** (nonlinear mixed effects) package. If you use **method="ML"** as an option, it uses Maximum Likelihood, and will obtain slightly different estimates than this method. The ML method is an iterative procedure. Note that the **gls** function has Restricted Maximum Likelihood (REML) as the default. This gives totally different and inappropriate estimates for regression models.

**Example: U.S. Annual Wine Consumption and Adult Population 1934-2002**

Table 6.22 contains annual U.S. adult population ($X$, in millions) and wine consumption ($Y$, in millions of gallons) for the years 1934-2003. Note that 1934 was the first year after Prohibition was repealed. A plot

| Year | Wine (Y) | Pop(X) | Year | Wine (Y) | Pop(X) | Year | Wine (Y) | Pop(X) |
|---|---|---|---|---|---|---|---|---|
| 1934 | 33 | 91.642 | 1957 | 152 | 119.383 | 1980 | 480 | 175.935 |
| 1935 | 46 | 92.868 | 1958 | 155 | 121.052 | 1981 | 506 | 178.212 |
| 1936 | 60 | 94.068 | 1959 | 156 | 123.091 | 1982 | 514 | 180.334 |
| 1937 | 67 | 95.251 | 1960 | 163 | 124.594 | 1983 | 528 | 182.324 |
| 1938 | 67 | 96.504 | 1961 | 172 | 126.137 | 1984 | 555 | 184.343 |
| 1939 | 77 | 97.760 | 1962 | 168 | 128.634 | 1985 | 580 | 186.389 |
| 1940 | 90 | 99.181 | 1963 | 176 | 130.777 | 1986 | 587 | 188.599 |
| 1941 | 101 | 100.463 | 1964 | 186 | 132.942 | 1987 | 581 | 190.430 |
| 1942 | 113 | 101.734 | 1965 | 190 | 135.052 | 1988 | 551 | 192.047 |
| 1943 | 98 | 103.023 | 1966 | 191 | 137.301 | 1989 | 524 | 193.598 |
| 1944 | 99 | 104.300 | 1967 | 203 | 139.653 | 1990 | 509 | 195.477 |
| 1945 | 94 | 105.350 | 1968 | 214 | 142.022 | 1991 | 466 | 197.735 |
| 1946 | 140 | 106.301 | 1969 | 236 | 144.417 | 1992 | 476 | 200.309 |
| 1947 | 97 | 107.462 | 1970 | 267 | 147.114 | 1993 | 449 | 202.824 |
| 1948 | 122 | 108.623 | 1971 | 305 | 149.927 | 1994 | 459 | 205.324 |
| 1949 | 133 | 109.812 | 1972 | 337 | 152.849 | 1995 | 464 | 208.006 |
| 1950 | 140 | 110.875 | 1973 | 347 | 155.749 | 1996 | 500 | 210.691 |
| 1951 | 127 | 111.981 | 1974 | 349 | 158.651 | 1997 | 520 | 213.560 |
| 1952 | 138 | 113.070 | 1975 | 368 | 161.611 | 1998 | 526 | 216.374 |
| 1953 | 141 | 114.138 | 1976 | 376 | 164.658 | 1999 | 551 | 219.085 |
| 1954 | 142 | 115.336 | 1977 | 401 | 167.642 | 2000 | 558 | 221.937 |
| 1955 | 145 | 116.559 | 1978 | 435 | 170.630 | 2001 | 561 | 224.833 |
| 1956 | 150 | 117.904 | 1979 | 444 | 173.602 | 2002 | 595 | 227.723 |

Table 6.22: U.S. Adult Population and Wine Consumption 1934-2002

of the data and ordinary least squares regression line is given in Figure 6.17, and a plot of residuals versus time is given in Figure 6.18. As population is increasing over time, the plot of wine sales versus population has an inherent "time effect" contained in it, and shows periodic behavior in the top portion. The residuals display clear autocorrelation.

We first give the analysis based on the **lm** (OLS) and **gls** (EGLS) functions, as well as the Durbin-Watson test. The Durbin-Watson statistic is very small, demonstrating very strong evidence of autocorrelation of the error terms. The **lmtest** package has the **dwtest** function, that provides a p-value, based on an assumption of normality, computed by the "pan" algorithm.

```
### Program
winepop <- read.table("http://www.stat.ufl.edu/~winner/data/winepop.dat",header=F,
col.names=c("year","tpop","pop5","pop14","pop24","pop34","pop44","pop54",
            "pop64","pop65p","wine"))

attach(winepop)
adultpop <- (tpop-pop5-pop14)/1000

n.wine <- length(wine)
wine.mod1 <- lm(wine ~ adultpop)
summary(wine.mod1)
plot(adultpop,wine)
abline(wine.mod1)
e <- residuals(wine.mod1)
plot(e, type="o", main="Residuals vs Time")
sse <- deviance(wine.mod1)
dw1 <- 0
gamma1 <- 0
```

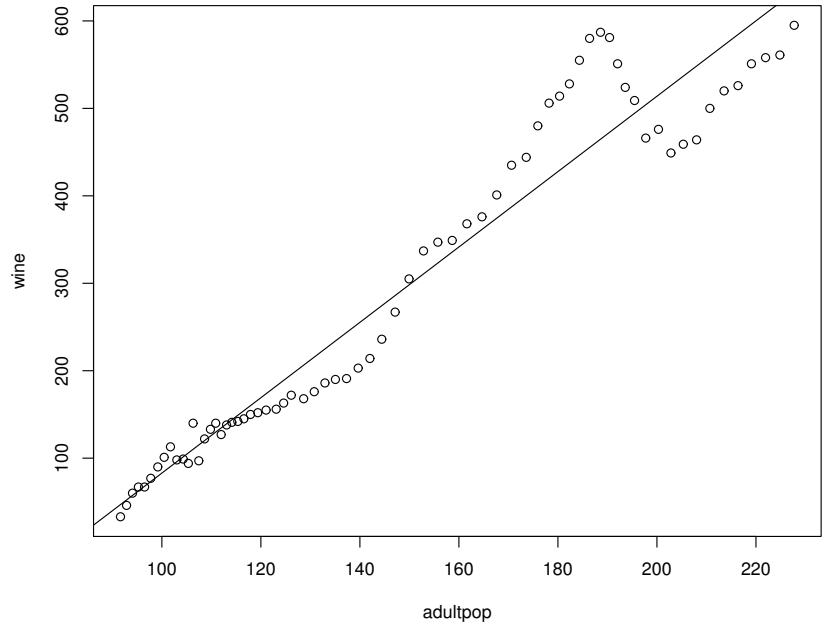


Figure 6.17: Plot of Wine Sales versus Adult Population: US 1934-2002

```
for (i in 2:n.wine) {
dw1 <- dw1 + (e[i]-e[i-1])^2
gamma1 <- gamma1 + (e[i]*e[i-1])
}
(dw <- dw1/sse)
(gamma0 <- sse/n.wine)
(gamma1 <- gamma1/n.wine)
(rho <- gamma1/gamma0)
(sigma2 <- gamma0 - rho*gamma1)

library(lmtest)
dwtest(wine ~ adultpop)

library(nlme)
wine.mod2 <- gls(wine ~ adultpop, correlation=corAR1(form=~year), method='ML')
summary(wine.mod2)

### Output
Ordinary Least Squares
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -347.9736    21.9895  -15.82   <2e-16 ***
adultpop       4.3092     0.1417   30.40   <2e-16 ***

Residual standard error: 48.64 on 67 degrees of freedom
Multiple R-squared:  0.9324,    Adjusted R-squared:  0.9314
F-statistic: 924.2 on 1 and 67 DF,  p-value: < 2.2e-16

> (dw <- dw1/sse)
0.1198987
> (gamma0 <- sse/n.wine)
```

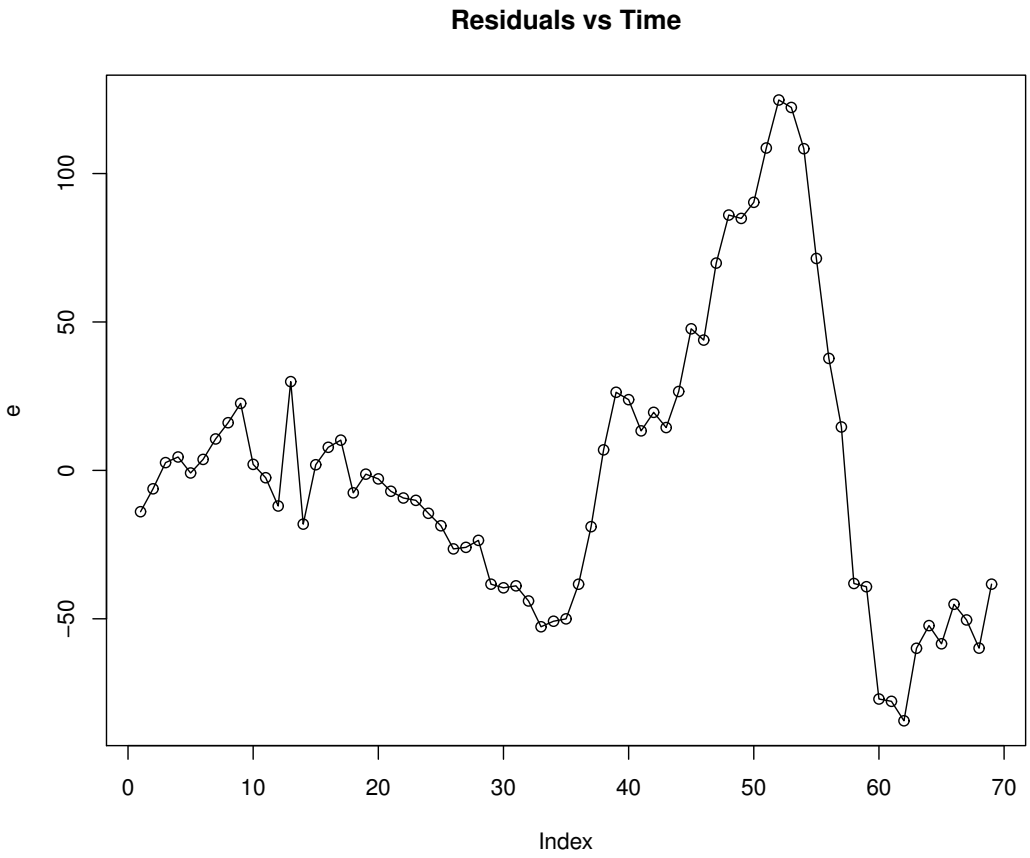


Figure 6.18: Plot of Residuals versus Time - Wine Sales

```
[1] 2297.689
> (gamma1 <- gamma1/n.wine)
2147.894
> (rho <- gamma1/gamma0)
0.9348065
> (sigma2 <- gamma0 - rho*gamma1)
289.823
> dwtest(wine ~ adultpop)
        Durbin-Watson test
data:  wine ~ adultpop
DW = 0.1199, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
> summary(wine.mod2)
Generalized least squares fit by maximum likelihood
Correlation Structure: AR(1)
 Formula: ~year
 Parameter estimate(s):
      Phi
0.9319506

Coefficients:
               Value Std.Error   t-value p-value
(Intercept) -346.7972  71.60350 -4.843299       0
adultpop       4.2528   0.44049  9.654536       0
```

The matrix form is used here to obtain the OLS and EGLS estimators, variance-covariance matrices, standard errors and $t$-statistics. The final results for OLS and EGLS are given in Table 6.23. Note that the estimated standard errors based on EGLS are much larger than those based on OLS. OLS leads to overstating the precision of estimated regression coefficients when errors are autocorrelated.

```
### Matrix form of OLS and EGLS
X <- cbind(rep(1,n.wine),adultpop)
Y <- wine
T.inv <- matrix(rep(0,n.wine**2),ncol=n.wine)
T.inv[1,1] <- sqrt(1-rho**2)
for (i in 2:n.wine) {
T.inv[i,i-1] <- -rho
T.inv[i,i] <- 1
}

(beta.ols <- solve(t(X)%*%X) %*% t(X)%*%Y)
(SSE.ols <- t(Y-X%*%beta.ols) %*% (Y-X%*%beta.ols))
(sigma2.ols <- SSE.ols/(n.wine-2))
(V.beta.ols <- sigma2.ols[1,1] * solve(t(X) %*% X))
SE.beta.ols <- sqrt(diag(V.beta.ols))
t.beta.ols <- beta.ols/SE.beta.ols

(beta.egls <- solve(t(X)%*%t(T.inv)%*%T.inv%*%X) %*% t(X)%*%t(T.inv)%*%T.inv%*%Y)
(SSE.egls <- t(Y-X%*%beta.egls) %*%t(T.inv)%*%T.inv %*% (Y-X%*%beta.egls))
(sigma2.egls <- SSE.egls/(n.wine-3))
(V.beta.egls <- sigma2.egls[1,1] * solve(t(X)%*%t(T.inv)%*%T.inv%*%X))
SE.beta.egls <- sqrt(diag(V.beta.egls))
t.beta.egls <- beta.ols/SE.beta.egls

s2.rho <- (gamma0-rho*gamma1)/(n.wine-2-1)
SE.rho <- sqrt(s2.rho/gamma0)
t.rho <- rho/SE.rho

print(round(cbind(beta.ols,SE.beta.ols,t.beta.ols,beta.egls,SE.beta.egls,t.beta.egls),4))
print(round(cbind(rho,SE.rho,t.rho),4))
```

| | Ordinary Least Squares | | | Estimated Generalized Least Squares | | |
|---|---|---|---|---|---|---|
| Parameter | Estimate | Std. Error | $t$-stat | Estimate | Std. Error | $t$-stat |
| Constant | -347.9736 | 21.9895 | -15.8245 | -347.2297 | 74.0420 | -4.6997 |
| Adult Pop | 4.3092 | 0.1417 | 30.4012 | 4.2540 | 0.4546 | 9.4796 |

Table 6.23: OLS and EGLS Regression Coefficients, Standard Errors, and $t$-statistics

```
####

> (beta.ols <- solve(t(X)%*%X) %*% t(X)%*%Y)
                [,1]
         -347.973637
adultpop    4.309183
> (SSE.ols <- t(Y-X%*%beta.ols) %*% (Y-X%*%beta.ols))
         [,1]
[1,] 158540.5
> (sigma2.ols <- SSE.ols/(n.wine-2))
         [,1]
[1,] 2366.276
> (V.beta.ols <- sigma2.ols[1,1] * solve(t(X) %*% X))
                      adultpop
         483.538924 -3.00432140
adultpop  -3.004321  0.02009137

> (beta.egls <- solve(t(X)%*%t(T.inv)%*%T.inv%*%X) %*% t(X)%*%t(T.inv)%*%T.inv%*%Y)
                [,1]
         -347.229711
adultpop    4.254013
> (SSE.egls <- t(Y-X%*%beta.egls) %*%t(T.inv)%*%T.inv %*% (Y-X%*%beta.egls))
         [,1]
[1,] 18516.16
> (sigma2.egls <- SSE.egls/(n.wine-3))
         [,1]
[1,] 280.5479
> (V.beta.egls <- sigma2.egls[1,1] * solve(t(X)%*%t(T.inv)%*%T.inv%*%X))
                      adultpop
         5482.22471 -31.5150787
adultpop  -31.51508   0.2066391

     rho SE.rho   t.rho
2 0.9348 0.0437 21.3832
```

$\nabla$

# Chapter 7

# Nonlinear Regression

Often theory leads to a relationship between the response and the predictor variable(s) to be nonlinear, based on differential equations. While polynomial regression models allow for bends, these models are linear with respect to the parameters. Many models with multiplicative errors can be transformed to be linear models. For instance:

$$Y = \beta_0 X^{\beta_1}\epsilon \quad E\{\epsilon\} = 1 \quad \Rightarrow \quad \ln(Y) = \ln(\beta_0) + \beta_1 \ln(X) + \ln(\epsilon) \quad \Rightarrow \quad Y^* = \beta_0^* + \beta_1 X^* + \epsilon^*.$$

If the error term had been additive (with mean=0), the linearizing transformation would not have been possible, and a nonlinear regression would need to have been fitted. Consider the relationship:

$$Y_i = g(\mathbf{x_i'}, \boldsymbol{\beta}) + \epsilon_i \qquad \epsilon_i \sim NID\left(0, \sigma^2\right)$$

for some nonlinear function $g$ (noting that linear regression ends up being a special case of this method). In matrix form, we have:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \mathbf{g}(\boldsymbol{\beta}) = \begin{bmatrix} g(\mathbf{x_1'}, \boldsymbol{\beta}) \\ g(\mathbf{x_2'}, \boldsymbol{\beta}) \\ \vdots \\ g(\mathbf{x_n'}, \boldsymbol{\beta}) \end{bmatrix} \qquad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \qquad \mathbf{Y} = \mathbf{g}(\boldsymbol{\beta}) + \epsilon$$

Then by nonlinear least squares (NLS), we wish to estimate $\boldsymbol{\beta}$.

$$Q = \sum_{i=1}^{n} \left[Y_i - g(\mathbf{x_i'}, \boldsymbol{\beta})\right]^2 = (\mathbf{Y} - \mathbf{g}(\boldsymbol{\beta}))'(\mathbf{Y} - \mathbf{g}(\boldsymbol{\beta})) \qquad \frac{\partial Q}{\partial \boldsymbol{\beta}'} = -2\sum_{i=1}^{n} \left[Y_i - g(\mathbf{x_i'}, \boldsymbol{\beta})\right] \left(\frac{\partial g(\mathbf{x_i'}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}\right)$$

Note that for linear regression, $\frac{\partial g(\mathbf{x_i'}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} = \mathbf{x_i'}$. Defining the matrix $\mathbf{G}(\boldsymbol{\beta})$ as follows, we can obtain the NLS iterative algorithm.

$$\mathbf{G}(\boldsymbol{\beta}) = \begin{bmatrix} \frac{\partial g(\mathbf{x_1'}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1} & \frac{\partial g(\mathbf{x_1'}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}_2} & \cdots & \frac{\partial g(\mathbf{x_1'}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}_p} \\ \frac{\partial g(\mathbf{x_2'}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1} & \frac{\partial g(\mathbf{x_2'}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}_2} & \cdots & \frac{\partial g(\mathbf{x_2'}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g(\mathbf{x_n'}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1} & \frac{\partial g(\mathbf{x_n'}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}_2} & \cdots & \frac{\partial g(\mathbf{x_n'}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}_p} \end{bmatrix} \quad \text{where} \quad \mathbf{x_i'} = \begin{bmatrix} x_{i1} & x_{i2} & \cdots & x_{ip} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$$

The Gauss-Newton algorithm can be used to obtain the values $\hat{\boldsymbol{\beta}}$ that minimize $Q$ by setting $\frac{\partial Q}{\partial \hat{\boldsymbol{\beta}}} = \mathbf{0}$:

$$\frac{\partial Q}{\partial \boldsymbol{\beta}'} = -2\sum_{i=1}^{n}\left[Y_i - g\left(\mathbf{x}_\mathbf{i}', \beta\right)\right]\left(\frac{\partial g\left(\mathbf{x}_\mathbf{i}', \beta\right)}{\partial \boldsymbol{\beta}'}\right) = -2\left[\mathbf{Y} - \mathbf{g}\left(\boldsymbol{\beta}\right)\right]'\mathbf{G}\left(\boldsymbol{\beta}\right) = \left[\begin{array}{cccc} 0 & 0 & \cdots & 0 \end{array}\right]$$

The algorithm begins with setting starting values $\boldsymbol{\beta}^0$, and iterating to convergence (which can be difficult with poor starting values):

$$\hat{\boldsymbol{\beta}}^{(1)} = \boldsymbol{\beta}^0 + \left(\mathbf{G}\left(\boldsymbol{\beta}^\mathbf{0}\right)'\mathbf{G}\left(\boldsymbol{\beta}^\mathbf{0}\right)\right)^{-1}\mathbf{G}\left(\boldsymbol{\beta}^\mathbf{0}\right)'\left[\mathbf{Y} - \mathbf{g}\left(\boldsymbol{\beta}^\mathbf{0}\right)\right]$$

At the second round $\boldsymbol{\beta}^0$ is replaced by $\hat{\boldsymbol{\beta}}^{(1)}$, and we obtain $\hat{\boldsymbol{\beta}}^{(2)}$. Then iterate to convergence (hopefully).

All of the distributional arguments given below are based on large sample asymptotics, however simulation results have shown that in small samples, tests generally work well. For more information on nonlinear regression models, see e.g. (Gallant (1987), and Rawlings, Pantula, Dickey (2001, Chapter 15)). When the errors are independent and normally distributed with equal variances $\left(\sigma^2\right)$, the estimator $\hat{\boldsymbol{\beta}}$ is approximately Normal, with:

$$E\left\{\hat{\boldsymbol{\beta}}\right\} = \boldsymbol{\beta} \qquad V\left\{\hat{\boldsymbol{\beta}}\right\} = \sigma^2(\mathbf{G}'\mathbf{G})^{-\mathbf{1}} \qquad \hat{\boldsymbol{\beta}} \overset{\text{approx}}{\sim} N\left(\boldsymbol{\beta}, \sigma^2(\mathbf{G}'\mathbf{G})^{-\mathbf{1}}\right)$$

The estimated variance-covariance matrix for $\hat{\boldsymbol{\beta}}$ is:

$$\hat{V}\left\{\hat{\boldsymbol{\beta}}\right\} = s^2\left(\hat{\mathbf{G}}'\hat{\mathbf{G}}\right)^{-\mathbf{1}} = \hat{\mathbf{S}}\hat{\rho}\hat{\mathbf{S}} \qquad \hat{G} = G\left(\hat{\boldsymbol{\beta}}\right) \qquad s^2 = \frac{\left(\mathbf{Y} - \mathbf{g}\left(\hat{\boldsymbol{\beta}}\right)\right)'\left(\mathbf{Y} - \mathbf{g}\left(\hat{\boldsymbol{\beta}}\right)\right)}{n-p}$$

where $\hat{\mathbf{S}}$ is the diagonal matrix of estimated standard errors of the elements of $\hat{\boldsymbol{\beta}}$, and $\hat{\rho}$ is the matrix of correlations of the elements of $\hat{\boldsymbol{\beta}}$, which are printed out in various software packages. Estimates (Confidence Intervals) and tests for the regression coefficients can be conducted (approximately) based on the $t$-distribution as in linear regression.

$$(1-\alpha)100\% \text{ CI for } \beta_j : \hat{\beta}_j \pm t\left(\alpha/2, n-p\right)\hat{SE}\left\{\hat{\beta}_j\right\} \qquad \frac{\hat{\beta}_j - \beta_{j0}}{\hat{SE}\left\{\hat{\beta}_j\right\}} \overset{\text{approx}}{\sim} t_{n-p} \text{ Under } H_0 : \beta_j = \beta_{j0}$$

**Example: Winning Velocities and Completion Times in the Kentucky Derby**

It has been argued in the academic literature that there are limits to performance in animals (e.g. Denny (2008)). Denny studied historical results involving speed among horses, dogs, and humans with a wide variety of theoretically based nonlinear models relating performance to year. One model considered for Velocity was an "S-shaped" logistic function, of the following form, where $Y_t$ is the winning velocity (meters per second) in year $t$. The data are given in Table 7.1.

$$Y_t = \beta_1 + (\beta_2 - \beta_1))\left[\frac{\exp\left\{\beta_3\left(t - \beta_4\right)\right\}}{1 + \exp\left\{\beta_3\left(t - \beta_4\right)\right\}}\right] + \epsilon_t$$

In this model, $\beta_1$ is the lower asymptote (minimum mean speed), $\beta_2$ is the upper asymptote (maximum mean speed), $\beta_3$ is a shape parameter determining the steepness of the curve between lower and upper

| Year | t | WinTime | Velocity | Year | t | WinTime | Velocity | Year | t | WinTime | Velocity |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1896 | 0 | 127.75 | 15.75 | 1938 | 42 | 124.8 | 16.12 | 1980 | 84 | 122 | 16.49 |
| 1897 | 1 | 132.5 | 15.18 | 1939 | 43 | 123.4 | 16.30 | 1981 | 85 | 122 | 16.49 |
| 1898 | 2 | 129 | 15.59 | 1940 | 44 | 125 | 16.09 | 1982 | 86 | 122.4 | 16.44 |
| 1899 | 3 | 132 | 15.24 | 1941 | 45 | 122.4 | 16.44 | 1983 | 87 | 122.2 | 16.46 |
| 1900 | 4 | 126.25 | 15.93 | 1942 | 46 | 124.4 | 16.17 | 1984 | 88 | 122.4 | 16.44 |
| 1901 | 5 | 127.75 | 15.75 | 1943 | 47 | 124 | 16.22 | 1985 | 89 | 120.2 | 16.74 |
| 1902 | 6 | 128.75 | 15.62 | 1944 | 48 | 124.2 | 16.20 | 1986 | 90 | 122.8 | 16.38 |
| 1903 | 7 | 129 | 15.59 | 1945 | 49 | 127 | 15.84 | 1987 | 91 | 123.4 | 16.30 |
| 1904 | 8 | 128.5 | 15.66 | 1946 | 50 | 126.6 | 15.89 | 1988 | 92 | 122.2 | 16.46 |
| 1905 | 9 | 130.75 | 15.39 | 1947 | 51 | 126.8 | 15.86 | 1989 | 93 | 125 | 16.09 |
| 1906 | 10 | 128.8 | 15.62 | 1948 | 52 | 125.4 | 16.04 | 1990 | 94 | 122 | 16.49 |
| 1907 | 11 | 132.6 | 15.17 | 1949 | 53 | 124.2 | 16.20 | 1991 | 95 | 123 | 16.36 |
| 1908 | 12 | 135.2 | 14.88 | 1950 | 54 | 121.6 | 16.54 | 1992 | 96 | 123 | 16.36 |
| 1909 | 13 | 128.2 | 15.69 | 1951 | 55 | 122.6 | 16.41 | 1993 | 97 | 122.4 | 16.44 |
| 1910 | 14 | 126.4 | 15.92 | 1952 | 56 | 121.6 | 16.54 | 1994 | 98 | 123.6 | 16.28 |
| 1911 | 15 | 125 | 16.09 | 1953 | 57 | 122 | 16.49 | 1995 | 99 | 121.2 | 16.60 |
| 1912 | 16 | 129.4 | 15.55 | 1954 | 58 | 123 | 16.36 | 1996 | 100 | 121 | 16.63 |
| 1913 | 17 | 124.8 | 16.12 | 1955 | 59 | 121.8 | 16.52 | 1997 | 101 | 122.4 | 16.44 |
| 1914 | 18 | 123.4 | 16.30 | 1956 | 60 | 123.4 | 16.30 | 1998 | 102 | 122.2 | 16.46 |
| 1915 | 19 | 125.4 | 16.04 | 1957 | 61 | 122.2 | 16.46 | 1999 | 103 | 123.2 | 16.33 |
| 1916 | 20 | 124 | 16.22 | 1958 | 62 | 125 | 16.09 | 2000 | 104 | 121 | 16.63 |
| 1917 | 21 | 124.6 | 16.15 | 1959 | 63 | 122.2 | 16.46 | 2001 | 105 | 119.97 | 16.77 |
| 1918 | 22 | 130.8 | 15.38 | 1960 | 64 | 122.4 | 16.44 | 2002 | 106 | 121.13 | 16.61 |
| 1919 | 23 | 129.8 | 15.50 | 1961 | 65 | 124 | 16.22 | 2003 | 107 | 121.19 | 16.60 |
| 1920 | 24 | 129 | 15.59 | 1962 | 66 | 120.4 | 16.71 | 2004 | 108 | 124.06 | 16.22 |
| 1921 | 25 | 124.2 | 16.20 | 1963 | 67 | 121.8 | 16.52 | 2005 | 109 | 122.75 | 16.39 |
| 1922 | 26 | 124.6 | 16.15 | 1964 | 68 | 120 | 16.76 | 2006 | 110 | 121.36 | 16.58 |
| 1923 | 27 | 125.4 | 16.04 | 1965 | 69 | 121.2 | 16.60 | 2007 | 111 | 122.17 | 16.47 |
| 1924 | 28 | 125.2 | 16.07 | 1966 | 70 | 122 | 16.49 | 2008 | 112 | 121.82 | 16.51 |
| 1925 | 29 | 127.6 | 15.77 | 1967 | 71 | 120.6 | 16.68 | 2009 | 113 | 122.66 | 16.40 |
| 1926 | 30 | 123.8 | 16.25 | 1968 | 72 | 122.2 | 16.46 | 2010 | 114 | 124.45 | 16.16 |
| 1927 | 31 | 126 | 15.97 | 1969 | 73 | 121.8 | 16.52 | 2011 | 115 | 122.04 | 16.48 |
| 1928 | 32 | 130.4 | 15.43 | 1970 | 74 | 123.4 | 16.30 | 2012 | 116 | 121.83 | 16.51 |
| 1929 | 33 | 130.8 | 15.38 | 1971 | 75 | 123.2 | 16.33 | 2013 | 117 | 122.89 | 16.37 |
| 1930 | 34 | 127.6 | 15.77 | 1972 | 76 | 121.8 | 16.52 | 2014 | 118 | 123.66 | 16.27 |
| 1931 | 35 | 121.8 | 16.52 | 1973 | 77 | 119.4 | 16.85 | 2015 | 119 | 123.02 | 16.35 |
| 1932 | 36 | 125.2 | 16.07 | 1974 | 78 | 124 | 16.22 | 2016 | 120 | 121.31 | 16.58 |
| 1933 | 37 | 126.8 | 15.86 | 1975 | 79 | 122 | 16.49 | | | | |
| 1934 | 38 | 124 | 16.22 | 1976 | 80 | 121.6 | 16.54 | | | | |
| 1935 | 39 | 125 | 16.09 | 1977 | 81 | 122.2 | 16.46 | | | | |
| 1936 | 40 | 123.6 | 16.28 | 1978 | 82 | 121.2 | 16.60 | | | | |
| 1937 | 41 | 123.2 | 16.33 | 1979 | 83 | 122.4 | 16.44 | | | | |

Table 7.1: Kentucky Derby Winning Times (sec) and Velocities (meters/sec) 1896-2016

asymptotes. Finally, $\beta_4$ is the year when the curve is steepest, as well as half way between the lower and upper asymptotes. Here we consider the winning speeds of the Kentucky Derby for years 1896-2016, all years that the horse race was run at a distance of 1.25 miles. The variable $t$ represents Year - 1896, so that the "origin" is the first year the race was 1.25 miles long. For this model, we have:

$$g\left(\mathbf{x}_\mathbf{t}', \boldsymbol{\beta}\right) = \beta_1 + \left(\beta_2 - \beta_1\right)) \left[\frac{\exp\left\{\beta_3\left(t - \beta_4\right)\right\}}{1 + \exp\left\{\beta_3\left(t - \beta_4\right)\right\}}\right]$$

$$\frac{\partial g\left(\mathbf{x}_\mathbf{t}', \boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}'} = \left[\begin{array}{cccc} \frac{\partial g\left(\mathbf{x}_\mathbf{t}', \boldsymbol{\beta}\right)}{\partial \beta_1} & \frac{\partial g\left(\mathbf{x}_\mathbf{t}', \boldsymbol{\beta}\right)}{\partial \beta_2} & \frac{\partial g\left(\mathbf{x}_\mathbf{t}', \boldsymbol{\beta}\right)}{\partial \beta_3} & \frac{\partial g\left(\mathbf{x}_\mathbf{t}', \boldsymbol{\beta}\right)}{\partial \beta_4} \end{array}\right]$$

$$\frac{\partial g\left(\mathbf{x}_\mathbf{t}', \boldsymbol{\beta}\right)}{\partial \beta_1} = 1 - \left[\frac{\exp\left\{\beta_3\left(t - \beta_4\right)\right\}}{1 + \exp\left\{\beta_3\left(t - \beta_4\right)\right\}}\right]$$

$$\frac{\partial g\left(\mathbf{x}_\mathbf{t}', \boldsymbol{\beta}\right)}{\partial \beta_2} = \left[\frac{\exp\left\{\beta_3\left(t - \beta_4\right)\right\}}{1 + \exp\left\{\beta_3\left(t - \beta_4\right)\right\}}\right]$$

$$\frac{\partial g\left(\mathbf{x}_\mathbf{t}', \boldsymbol{\beta}\right)}{\partial \beta_3} = \left(\beta_2 - \beta_1\right)) \left[\frac{\left(1 + \exp\left\{\beta_3\left(t - \beta_4\right)\right\}\right)\left(t - \beta_4\right)\exp\left\{\beta_3\left(t - \beta_4\right)\right\} - \left(t - \beta_4\right)\left(\exp\left\{\beta_3\left(t - \beta_4\right)\right\}\right)^2}{\left(1 + \exp\left\{\beta_3\left(t - \beta_4\right)\right\}\right)^2}\right]$$

$$= \left(\beta_2 - \beta_1\right))\left(t - \beta_4\right)\left[\frac{\exp\left\{\beta_3\left(t - \beta_4\right)\right\}}{\left(1 + \exp\left\{\beta_3\left(t - \beta_4\right)\right\}\right)^2}\right]$$

$$\frac{\partial g\left(\mathbf{x}_\mathbf{t}', \boldsymbol{\beta}\right)}{\partial \beta_4} = \left(\beta_2 - \beta_1\right)) \left[\frac{\left(1 + \exp\left\{\beta_3\left(t - \beta_4\right)\right\}\right)\left(-\beta_3\right)\exp\left\{\beta_3\left(t - \beta_4\right)\right\} - \left(-\beta_3\right)\left(\exp\left\{\beta_3\left(t - \beta_4\right)\right\}\right)^2}{\left(1 + \exp\left\{\beta_3\left(t - \beta_4\right)\right\}\right)^2}\right]$$

$$= -\beta_3\left(\beta_2 - \beta_1\right))\left[\frac{\exp\left\{\beta_3\left(t - \beta_4\right)\right\}}{\left(1 + \exp\left\{\beta_3\left(t - \beta_4\right)\right\}\right)^2}\right]$$

When choosing starting values for the parameters, it is helpful to think of their effects on predicted values in terms of sign and magnitude. For this example, $\beta_1$ and $\beta_2$ represent the lower and upper asymptotes of the speeds. Also they enter the equation in a linear form, so that they do not have large effects on the estimation process (as long as $\beta_2 > \beta_1$). As all races have Velocities between 10 and 20 meters per second, we will start with these values for $\beta_1$ and $\beta_2$, respectively. The parameter $\beta_4$ represents the steepest point

of the curve in terms of the year, we will choose the "middle" year, 60. Finally, $\beta_3$ represents the steepness of the curve. Suppose we hypothesize that the points that correspond to being 25% and 75% between the lower and upper asymptotes are $k$ years below and above $\beta_4$, respectively.

$$\Rightarrow \qquad .25 = \frac{\exp\{\beta_3(t_{.25} - \beta_4)\}}{1 + \exp\{\beta_3(t_{.25} - \beta_4)\}} = \frac{\exp\{\beta_3(-k)\}}{1 + \exp\{\beta_3(-k)\}} \qquad \Rightarrow \qquad \beta_3 = \frac{-\ln((1/0.25) - 1)}{-k} = \frac{-\ln(3)}{-k}$$

If we choose $k = 20$ as a reasonable possibility, it leads to a starting value for $\beta_3 = 0.05$. We used 0.10 in the following program. Note that when using built-in nonlinear regression software procedures in R, SAS, SPSS, and STATA, it is not necessary to be so precise for starting values. However, in many exponential growth and decay models, there are multiple parameterizations of the same function, and if a sign is wrong, or orders of magnitude are incorrect, these programs will not converge.

A set of R commands in matrix form are given below, along with the first two and last two iterated estimates of $\boldsymbol{\beta}$ and all inferences regarding $\boldsymbol{\beta}$. The variables Year125,...,Speed125 are all based on the years when the race was 1.25 miles (1896-2016).

```
### Program
kd <- read.csv("http://www.stat.ufl.edu/~winner/data/kentuckyderby.csv",
               header=TRUE)
attach(kd); names(kd)
Year125 <- Year[Length==1.25]
Time125 <- Time[Length==1.25]
Length125 <- Length[Length==1.25]
Year125.0 <- Year125-min(Year125)

Speed125 <- 1609.34*Length125/Time125
summary(Speed125)

### Matrix form for logistic model

Y <- Speed125
X.t <- Year125.0
beta.old <- c(10,20,0.1,60)
diff.beta <- 1000

while (diff.beta > .00001)  {
exp.t <- exp(beta.old[3] * (X.t - beta.old[4]))
G1 <- 1 - (exp.t/(1+exp.t))
G2 <- exp.t/(1+exp.t)
G3 <- (beta.old[2]-beta.old[1]) * (X.t-beta.old[4]) * (exp.t/(1+exp.t)^2)
G4 <- (beta.old[2]-beta.old[1]) * (-beta.old[3]) * (exp.t/(1+exp.t)^2)
G <- cbind(G1,G2,G3,G4)
Yhat <- beta.old[1] + (beta.old[2]-beta.old[1]) * (exp.t/(1+exp.t))
e <- Y - Yhat
beta.new <- beta.old + solve(t(G) %*% G) %*% t(G) %*% e
print(beta.new)
diff.beta <- sum((beta.new-beta.old)^2)
beta.old <- beta.new
}
exp.t <- exp(beta.old[3] * (X.t - beta.old[4]))
G1 <- 1 - (exp.t/(1+exp.t))
G2 <- exp.t/(1+exp.t)
G3 <- (beta.old[2]-beta.old[1]) * (X.t-beta.old[4]) * (exp.t/(1+exp.t)^2)
G4 <- (beta.old[2]-beta.old[1]) * (-beta.old[3]) * (exp.t/(1+exp.t)^2)
G <- cbind(G1,G2,G3,G4)
Yhat <- beta.old[1] + (beta.old[2]-beta.old[1]) * (exp.t/(1+exp.t))
```

```
e <- Y - Yhat
(SSE <- t(e) %*% e)
(MSE <- SSE/(length(Y) - ncol(G)))
V.beta <- MSE[1,1] * solve(t(G) %*% G)
SE.beta <- sqrt(diag(V.beta))
t.beta <- beta.new/SE.beta
P.beta <- 2*(1-pt(abs(t.beta),length(Y)-ncol(G)))
print(round(cbind(beta.new,SE.beta,t.beta,P.beta),4))
(CI.beta1 <- beta.new[1] + qt(c(.025,.975),length(Y)-ncol(G))*SE.beta[1])
(CI.beta2 <- beta.new[2] + qt(c(.025,.975),length(Y)-ncol(G))*SE.beta[2])
(CI.beta3 <- beta.new[3] + qt(c(.025,.975),length(Y)-ncol(G))*SE.beta[3])
(CI.beta4 <- beta.new[4] + qt(c(.025,.975),length(Y)-ncol(G))*SE.beta[4])

### Output
> Speed125 <- 1609.34*Length125/Time125
> summary(Speed125)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  14.88   16.04   16.30   16.20   16.49   16.85

          [,1]
G1 15.69133064
G2 16.47338070
G3  0.09592648
G4 58.53688771
          [,1]
G1 15.67313699
G2 16.47186534
G3  0.05280251
G4 39.73971423

...

          [,1]
G1 15.36095708
G2 16.48585093
G3  0.06381906
G4 27.58451474
         [,1]
G1 15.3609265
G2 16.4858482
G3  0.0638184
G4 27.5832162

> (SSE <- t(e) %*% e)
         [,1]
[1,] 6.607435
> (MSE <- SSE/(length(Y) - ncol(G)))
          [,1]
[1,] 0.0564738
> print(round(cbind(beta.new,SE.beta,t.beta,P.beta),4))
          SE.beta
G1 15.3609  0.2823  54.4092 0.0000
G2 16.4858  0.0464 355.3788 0.0000
G3  0.0638  0.0230   2.7781 0.0064
G4 27.5832  9.2703   2.9755 0.0036
> (CI.beta1 <- beta.new[1] + qt(c(.025,.975),length(Y)-ncol(G))*SE.beta[1])
[1] 14.80180 15.92005
> (CI.beta2 <- beta.new[2] + qt(c(.025,.975),length(Y)-ncol(G))*SE.beta[2])
[1] 16.39398 16.57772
> (CI.beta3 <- beta.new[3] + qt(c(.025,.975),length(Y)-ncol(G))*SE.beta[3])
[1] 0.01832384 0.10931295
> (CI.beta4 <- beta.new[4] + qt(c(.025,.975),length(Y)-ncol(G))*SE.beta[4])
[1]  9.223954 45.942479
```

The analysis based on the **nls** function is given below. Note that we don't need to be so precise on starting values, due to the algorithm used in computations. A plot of the data and the fitted regression equation is given in Figure 7.1. A plot of Residuals versus Year does demonstrate higher variance in early years then in later years, but the difference is not very drastic. It could very well be due to less accurate timing equipment. The residual plot is given in Figure 7.2.

### R Program

```
kd <- read.csv("http://www.stat.ufl.edu/~winner/data/kentuckyderby.csv",
               header=TRUE)
attach(kd); names(kd)

Year125 <- Year[Length==1.25]
Time125 <- Time[Length==1.25]
Length125 <- Length[Length==1.25]
Year125.0 <- Year125-min(Year125)

Speed125 <- 1609.34*Length125/Time125
summary(Speed125)

Speed125.2008 <- Speed125[Year125<=2008]
Year125.2008 <- Year125[Year125<=2008]

kd.mod1 <- nls(Speed125 ~ b0 + (b1-b0)*exp(b2*(Year125.0-b3))/
    (1+exp(b2*(Year125.0-b3))), start=c(b0=1,b1=20,b2=1,b3=60))
summary(kd.mod1)
AIC(kd.mod1)
confint(kd.mod1)

plot(Year125.0,Speed125)
lines(Year125.0,predict(kd.mod1,Year125))
```

### R Output

```
Formula: Speed125 ~ b0 + (b1 - b0) * exp(b2 * (Year125.0 - b3))/(1 + exp(b2 *
    (Year125.0 - b3)))

Parameters:
   Estimate Std. Error t value Pr(>|t|)
b0 15.36090    0.28235  54.404  < 2e-16 ***
b1 16.48585    0.04639 355.369  < 2e-16 ***
b2  0.06382    0.02297   2.778  0.00637 **
b3 27.58248    9.27100   2.975  0.00356 **

Residual standard error: 0.2376 on 117 degrees of freedom

Number of iterations to convergence: 17
Achieved convergence tolerance: 8.259e-06

> AIC(kd.mod1)
[1] 1.564131
> confint(kd.mod1)
Waiting for profiling to be done...
Error in prof$getProfile() : number of iterations exceeded maximum of 50
```

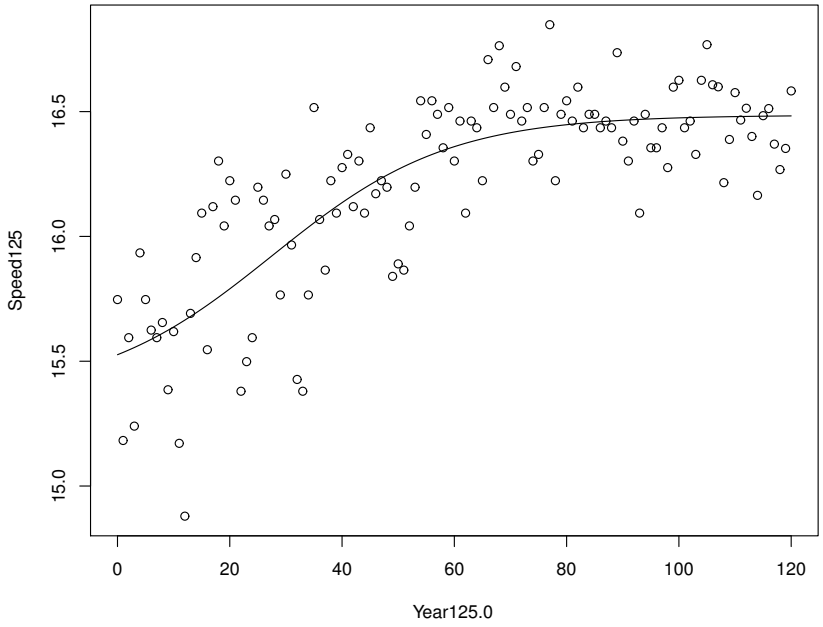


Figure 7.1: Kentucky Derby Winning Velocities and Fitted Logistic Function - 1896-2016

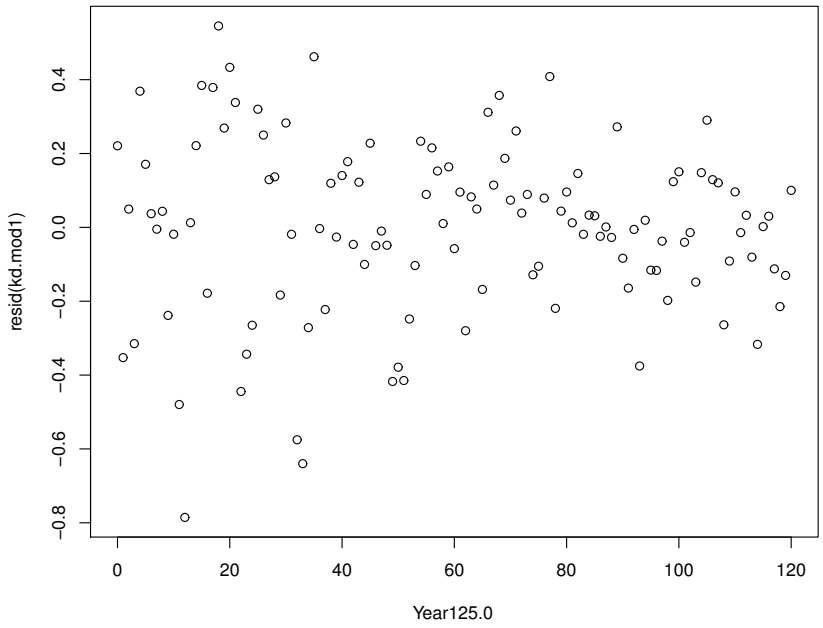


Figure 7.2: Kentucky Derby Residuals versus Year - 1896-2016

For **linear** functions of $\boldsymbol{\beta}$ of the form $\mathbf{K}'\boldsymbol{\beta}$, we then have approximate normality of the estimator $\mathbf{K}'\hat{\boldsymbol{\beta}}$:

$$\mathbf{K}'\hat{\boldsymbol{\beta}} \overset{\text{approx}}{\sim} N\left(\mathbf{K}'\boldsymbol{\beta}, \sigma^2\mathbf{K}'\left(\mathbf{G}'\mathbf{G}\right)^{-1}\mathbf{K}\right)$$

Thus, to test $H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$, where $\mathbf{K}'$ has $q$ linearly independent rows (restrictions on the regression coefficients), we have the following test statistic and rejection region.

$$TS : F_{obs} = \frac{\left(\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m}\right)'\left[\mathbf{K}'\left(\hat{\mathbf{G}}'\hat{\mathbf{G}}\right)^{-1}\mathbf{K}\right]^{-1}\left(\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m}\right)}{qs^2} \qquad RR : F_{obs} \geq F_{q,n-p}$$

with an approximate $P$-value as the area above the test statistic $F_{obs}$ for the $F_{q,n-p}$ distribution.

**Example: Beer Foam Heights Over Time for 3 Beer Brands**

Leike (2002) reported results of an experiment measuring beer foam height over a 6 minute period for 3 brands of beer (Erdinger Weissbier, Augustinerbrau Munchen, and Budweiser). The data are given in Table 7.2. There are a total of $n = 3(15) = 45$ observations when we "stack" the data for 3 brands. An exponential decay model with additive errors is fit, allowing for different curves for the 3 brands, with $t$ representing time, and dummy variables: $X_{i1} = 1$ if Erdinger, 0 otherwise; $X_{i2} = 1$ if Augustinerbrau, 0 otherwise; and $X_{i3} = 1$ if Budweiser, 0 otherwise.

$$Y_i = \beta_{01}X_{i1}\exp\{-\beta_{11}X_{i1}t_i\} + \beta_{02}X_{i2}\exp\{-\beta_{12}X_{i2}t_i\} + \beta_{03}X_{i3}\exp\{-\beta_{13}X_{i3}t_i\} + \epsilon_i \qquad i = 1, \ldots, 45$$

The R program and output are given below. Note that the algorithm fails when $t_i = 0$, so replace it with $t_i = 0.0001$. A plot of the data and the fitted curves are given in Figure 7.3.

**R Program**

```
beerfoam <- read.table("http://www.stat.ufl.edu/~winner/data/beer_foam2.dat",
   header=F,col.names=c("foam.time","brand1","brand2","brand3"))
attach(beerfoam)

foam.time
for (i in 1:length(foam.time)) {
if (foam.time[i] == 0) foam.time[i] <- 0.0001
}
foam.time
Y.foam <- c(brand1,brand2,brand3)
X1 <- c(rep(1,15),rep(0,15),rep(0,15))
X2 <- c(rep(0,15),rep(1,15),rep(0,15))
X3 <- c(rep(0,15),rep(0,15),rep(1,15))
t.foam <- c(foam.time,foam.time,foam.time)
brand <- rep(1:3,each=15)

foam.mod1 <- nls(Y.foam ~ b01*X1*exp(-b11*X1*t.foam) +
   b02*X2*exp(-b12*X2*t.foam) + b03*X3*exp(-b13*X3*t.foam),
   start=c(b01=10,b02=10,b03=10,b11=0.01,b12=0.01,b13=0.01))
summary(foam.mod1)

time.x <- 0:360
```

| Time (sec) | Erdinger | Augustinerbrau | Budweiser |
|---|---|---|---|
| 0 | 17.0 | 14.0 | 14.0 |
| 15 | 16.1 | 11.8 | 12.1 |
| 30 | 14.9 | 10.5 | 10.9 |
| 45 | 14.0 | 9.3 | 10.0 |
| 60 | 13.2 | 8.5 | 9.3 |
| 75 | 12.5 | 7.7 | 8.6 |
| 90 | 11.9 | 7.1 | 8.0 |
| 105 | 11.2 | 6.5 | 7.5 |
| 120 | 10.7 | 6.0 | 7.0 |
| 150 | 9.7 | 5.3 | 6.2 |
| 180 | 8.9 | 4.4 | 5.5 |
| 210 | 8.3 | 3.5 | 4.5 |
| 240 | 7.5 | 2.9 | 3.5 |
| 300 | 6.3 | 1.3 | 2.0 |
| 360 | 5.2 | 0.7 | 0.9 |

Table 7.2: Beer Foam Heights for 3 Brands of Beer over Time

```
yhat.b1 <- coef(foam.mod1)[1] * exp(-coef(foam.mod1)[4]*time.x)
yhat.b2 <- coef(foam.mod1)[2] * exp(-coef(foam.mod1)[5]*time.x)
yhat.b3 <- coef(foam.mod1)[3] * exp(-coef(foam.mod1)[6]*time.x)

plot(t.foam,Y.foam,pch=brand)
lines(time.x,yhat.b1,lty=1)
lines(time.x,yhat.b2,lty=2)
lines(time.x,yhat.b3,lty=5)
legend(240,16,c("Erd","Aug","Bud"),pch=c(1,2,3),lty=c(1,2,5))
```

**R Output**

```
> summary(foam.mod1)
Formula: Y.foam ~ b01 * X1 * exp(-b11 * X1 * t.foam) + b02 * X2 * exp(-b12 *
    X2 * t.foam) + b03 * X3 * exp(-b13 * X3 * t.foam)

Parameters:
     Estimate Std. Error t value Pr(>|t|)
b01 1.650e+01  2.080e-01   79.32   <2e-16 ***
b02 1.323e+01  2.469e-01   53.61   <2e-16 ***
b03 1.337e+01  2.346e-01   57.02   <2e-16 ***
b11 3.396e-03  1.172e-04   28.98   <2e-16 ***
b12 6.758e-03  2.534e-04   26.67   <2e-16 ***
b13 5.625e-03  2.117e-04   26.57   <2e-16 ***

Residual standard error: 0.3931 on 39 degrees of freedom

Number of iterations to convergence: 6
Achieved convergence tolerance: 4.779e-07
```

We test whether the curves for Augustinerbrau and Budweiser are the same. This can be conducted based on the matrix form described above, as well as by fitting a Reduced Model and comparing it with the complete model. What is being tested is as follows.

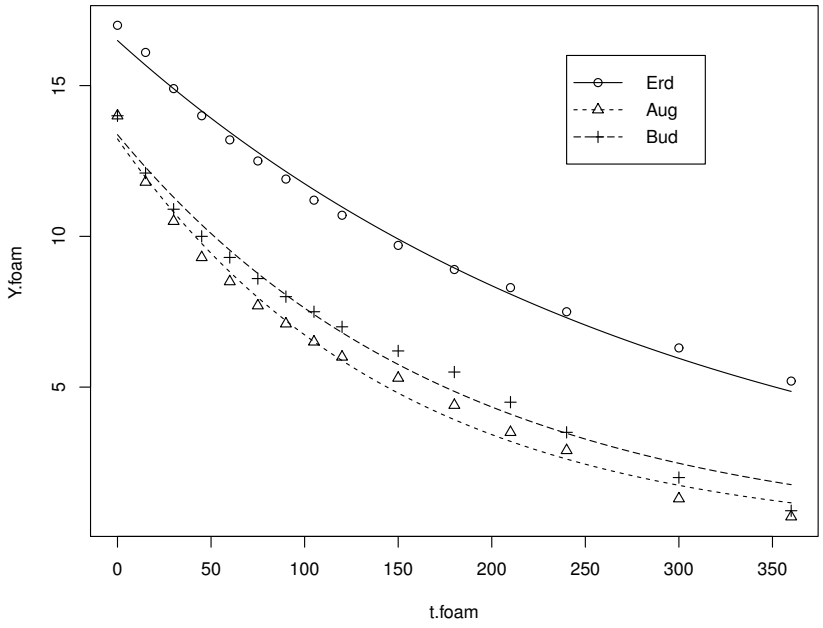


Figure 7.3: Plot of Beer Foam versus Time and Brand Specific Fitted Equations

$$H_0 : \beta_{02} = \beta_{03}, \beta_{12} = \beta_{13} \qquad \Rightarrow \qquad \mathbf{K}'\boldsymbol{\beta} = \mathbf{0} \qquad \mathbf{K}' = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_{01} \\ \beta_{02} \\ \beta_{03} \\ \beta_{11} \\ \beta_{12} \\ \beta_{13} \end{bmatrix}$$

In R, the object produced by many regression functions including **lm** and **nls** contains many items that can be "selected" from the object. Relevant to this test, we can obtain: $\hat{\boldsymbol{\beta}}$ as **coef(foam.mod1)**, **SSE** as **deviance(foam.mod1)**, and $\hat{V}\{\hat{\boldsymbol{\beta}}\} = s^2\left(\hat{G}'\hat{G}\right)^{-1}$ as **vcov(foam.mod1)**. Then we can isolate $\left(\hat{G}'\hat{G}\right)^{-1}$ by dividing the estimated variance-covariance of $\hat{\boldsymbol{\beta}}$ by $s^2 = MSE = SSE/(n-p)$. The following R commands, run after fitting **foam.mod1** above, will conduct the $F$-test.

```
### Commands
(beta.foam1 <- coef(foam.mod1))
vcov.foam1 <- vcov(foam.mod1)
(mse <- deviance(foam.mod1)/(45-6))
GpGinv <- vcov.foam1/mse

Kp <- matrix(c(0,1,-1,0,0,0,0,0,0,0,1,-1),byrow=T,ncol=6)
(F.foam1a <- t(Kp %*% beta.foam1) %*% solve(Kp %*% GpGinv %*% t(Kp)) %*%
   (Kp %*% beta.foam1))
```

```
(F.foam1b <- nrow(Kp)*mse)
(F.foam1 <- F.foam1a / F.foam1b)
(crit.F <- qf(.95,2,39))
(P.F <- 1-pf(F.foam1,2,39))

### Output
> (beta.foam1 <- coef(foam.mod1))
         b01          b02          b03          b11          b12          b13
16.495318508 13.233329203 13.373355980  0.003396205  0.006757962  0.005625264
> (mse <- deviance(foam.mod1)/(45-6))
[1] 0.1545567
> Kp <- matrix(c(0,1,-1,0,0,0,0,0,0,0,1,-1),byrow=T,ncol=6)
> (F.foam1a <- t(Kp %*% beta.foam1) %*% solve(Kp %*% GpGinv %*% t(Kp)) %*%
+    (Kp %*% beta.foam1))
[1,] 4.186677
> (F.foam1b <- nrow(Kp)*mse)
[1] 0.3091133
> (F.foam1 <- F.foam1a / F.foam1b)
[1,] 13.54415
> (crit.F <- qf(.95,2,39))
[1] 3.238096
> (P.F <- 1-pf(F.foam1,2,39))
[1,] 3.414526e-05
```

We strongly reject the hypothesis that the curves are the same for Augustinerbrau and Budweiser. The second approach involves "forcing" $\beta_{02} = \beta_{03}$ and $\beta_{12} = \beta_{13}$ and fitting the ensuing model. This is done as follows.

$$X_{23} = X_2 + X_3 \qquad Y_i = \beta_{01} X_{i1} \exp\{-\beta_{11} X_{i1} t_i\} + \beta_{023} X_{i23} \exp\{-\beta_{123} X_{i23} t_i\} + \epsilon_i$$

The R Program and Output are given below.

**R Program**

```
### Fit Reduced Model

X23 <- X2 + X3

foam.mod2 <- nls(Y.foam ~ b01*X1*exp(-b11*X1*t.foam) +
   b023*X23*exp(-b123*X23*t.foam),
   start=c(b01=10,b023=10,b11=0.01,b123=0.01))
summary(foam.mod2)

### Compare Reduced and Complete Models
anova(foam.mod2,foam.mod1)
```

**R Output**

```
> summary(foam.mod2)
Formula: Y.foam ~ b01 * X1 * exp(-b11 * X1 * t.foam) + b023 * X23 * exp(-b123 *
    X23 * t.foam)

Parameters:
```

```
      Estimate Std. Error t value Pr(>|t|)
b01  1.650e+01  2.652e-01   62.20   <2e-16 ***
b023 1.329e+01  2.167e-01   61.30   <2e-16 ***
b11  3.396e-03  1.494e-04   22.73   <2e-16 ***
b123 6.148e-03  2.082e-04   29.53   <2e-16 ***

Residual standard error: 0.5014 on 41 degrees of freedom

Number of iterations to convergence: 6
Achieved convergence tolerance: 3.585e-07

>
> anova(foam.mod2,foam.mod1)
Analysis of Variance Table

Model 1: Y.foam ~ b01 * X1 * exp(-b11 * X1 * t.foam) + b023 * X23 * exp(-b123 * X23 * t.foam)
Model 2: Y.foam ~ b01 * X1 * exp(-b11 * X1 * t.foam) + b02 * X2 * exp(-b12 * X2 * t.foam) +
                  b03 * X3 * exp(-b13 * X3 * t.foam)
  Res.Df Res.Sum Sq Df Sum Sq F value    Pr(>F)
1     41    10.3060
2     39     6.0277  2 4.2783  13.841 2.869e-05 ***
```

Note that the matrix form and the Complete versus Reduced form give slightly different $F$-statistics. The equivalency derived in the case of linear regression does not hold in the case of nonlinear regression.

$\nabla$

By the nature of nonlinear models, we may also be interested in **nonlinear** functions of the parameters, say $h(\boldsymbol{\beta})$, which cannot be written in the form $\mathbf{K}'\boldsymbol{\beta}$. In this case, the estimator $h\left(\hat{\boldsymbol{\beta}}\right)$ is approximately normally distributed:

$$h\left(\hat{\boldsymbol{\beta}}\right) \overset{\text{approx}}{\sim} N\left(h(\boldsymbol{\beta}), \sigma^2\left(\mathbf{H}\left(\mathbf{G}'\mathbf{G}\right)^{-1}\mathbf{H}'\right)\right)$$

where

$$\mathbf{H}(\boldsymbol{\beta}) = \left[\begin{array}{cccc} \frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1} & \frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_2} & \cdots & \frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_p} \end{array}\right].$$

The estimated variance of $h\left(\hat{\boldsymbol{\beta}}\right)$ replaces both $\mathbf{H}$ and $\mathbf{G}$ with their estimates, replacing $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}$. Estimates (Confidence Intervals) and tests concerning $h(\boldsymbol{\beta})$ can be obtained as follow.

$(1-\alpha)100\%$ CI for $h(\boldsymbol{\beta}) : h\left(\hat{\boldsymbol{\beta}}\right) \pm t(\alpha/2, n-p)\, \hat{SE}\left\{h\left(\hat{\boldsymbol{\beta}}\right)\right\}$ $\qquad$ $\dfrac{h\left(\hat{\boldsymbol{\beta}}\right) - h_0}{\hat{SE}\left\{h\left(\hat{\boldsymbol{\beta}}\right)\right\}} \overset{\text{approx}}{\sim} t_{n-p}$ Under $H_0 : h(\boldsymbol{\beta}) = h_0$

where:

$$\hat{SE}\left\{h\left(\hat{\boldsymbol{\beta}}\right)\right\} = \sqrt{s^2\hat{\mathbf{H}}\left(\hat{\mathbf{G}}'\hat{\mathbf{G}}\right)^{-1}\hat{\mathbf{H}}'}.$$

When there are several $(q)$ nonlinear functions, an approximate Wald test of $\mathbf{h}(\boldsymbol{\beta}) = \mathbf{h_0}$ is:

$$TS : F_{obs} = \frac{\left(\mathbf{h}\left(\hat{\boldsymbol{\beta}}\right) - \mathbf{h_0}\right)'\left[\hat{\mathbf{H}}\left(\hat{\mathbf{G}}'\hat{\mathbf{G}}\right)^{-1}\hat{\mathbf{H}}'\right]^{-1}\left(\mathbf{h}\left(\hat{\boldsymbol{\beta}}\right) - \mathbf{h_0}\right)}{qs^2} \qquad RR : W \geq F_{q,n-p}$$

with $P$-value being the upper-tail area above $F_{obs}$ for the $F_{q,n-p}$ distribution. Here, we define:

$$\mathbf{h}(\boldsymbol{\beta}) = \begin{bmatrix} h_1(\boldsymbol{\beta}) \\ h_2(\boldsymbol{\beta}) \\ \vdots \\ h_q(\boldsymbol{\beta}) \end{bmatrix} \qquad \mathbf{H}(\boldsymbol{\beta}) = \begin{bmatrix} \frac{\partial h_1(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1} & \frac{\partial h_1(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_2} & \cdots & \frac{\partial h_1(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_p} \\ \frac{\partial h_2(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1} & \frac{\partial h_2(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_2} & \cdots & \frac{\partial h_2(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial h_q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1} & \frac{\partial h_q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_2} & \cdots & \frac{\partial h_q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_p} \end{bmatrix}$$

**Example: Winning Velocities and Completion Times in the Kentucky Derby**

Suppose we are interested in the "percentage change" in mean velocity in horses, based on the Kentucky Derby data. Recall that $\beta_1$ is the lower asymptote and $\beta_2$ is the upper asymptote. Then, we are interested in the following nonlinear function of the coefficient vector $\boldsymbol{\beta}$.

$$h(\boldsymbol{\beta}) = 100\left(\frac{\beta_2 - \beta_1}{\beta_1}\right) \qquad \Rightarrow \qquad \frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1} = \frac{-100\beta_2}{\beta_1^2} \qquad \frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_2} = \frac{100}{\beta_1}$$

From the model fit previously, we have the following estimates.

$$\hat{\beta}_1 = 15.36090 \qquad \hat{\beta}_2 = 16.48586 \qquad s = 0.2376 \qquad s^2 = 0.056454$$

$$\Rightarrow \qquad h\left(\hat{\boldsymbol{\beta}}\right) = 100\left(\frac{16.48586 - 15.36090}{15.36090}\right) = 7.32353 \qquad H\left(\hat{\boldsymbol{\beta}}\right) = \begin{bmatrix} -6.98680 & 6.51004 & 0 & 0 \end{bmatrix}$$

The R Commands and Output to compute a 95% Confidence Interval for $h(\boldsymbol{\beta})$ are given below.

```
### Commands
### Matrix form for logistic model

Y <- Speed125
X.t <- Year125.0
beta.old <- c(10,20,0.1,60)
diff.beta <- 1000

while (diff.beta > .00001)  {
exp.t <- exp(beta.old[3] * (X.t - beta.old[4]))
G1 <- 1 - (exp.t/(1+exp.t))
G2 <- exp.t/(1+exp.t)
G3 <- (beta.old[2]-beta.old[1]) * (X.t-beta.old[4]) * (exp.t/(1+exp.t)^2)
G4 <- (beta.old[2]-beta.old[1]) * (-beta.old[3]) * (exp.t/(1+exp.t)^2)
G <- cbind(G1,G2,G3,G4)
Yhat <- beta.old[1] + (beta.old[2]-beta.old[1]) * (exp.t/(1+exp.t))
e <- Y - Yhat
beta.new <- beta.old + solve(t(G) %*% G) %*% t(G) %*% e
```

```
print(beta.new)
diff.beta <- sum((beta.new-beta.old)^2)
beta.old <- beta.new
}
exp.t <- exp(beta.old[3] * (X.t - beta.old[4]))
G1 <- 1 - (exp.t/(1+exp.t))
G2 <- exp.t/(1+exp.t)
G3 <- (beta.old[2]-beta.old[1]) * (X.t-beta.old[4]) * (exp.t/(1+exp.t)^2)
G4 <- (beta.old[2]-beta.old[1]) * (-beta.old[3]) * (exp.t/(1+exp.t)^2)
G <- cbind(G1,G2,G3,G4)
Yhat <- beta.old[1] + (beta.old[2]-beta.old[1]) * (exp.t/(1+exp.t))
e <- Y - Yhat
(SSE <- t(e) %*% e)
(MSE <- SSE/(length(Y) - ncol(G)))

(h.beta <- 100*(beta.old[2]-beta.old[1]) / beta.old[1])
(H.beta1 <- -100*beta.old[2]/(beta.old[1]^2))
(H.beta2 <- 100/beta.old[1])
H.beta3  <- 0; H.beta4 <- 0
H.beta <- cbind(H.beta1,H.beta2,H.beta3,H.beta4)
(SE.h.beta <- sqrt(MSE[1,1] * (H.beta %*% solve(t(G) %*% G) %*% t(H.beta))))
(CI.h.beta <- h.beta +  qt(c(.025,.975),length(Y)-ncol(G))*SE.h.beta)

### Output
> (SSE <- t(e) %*% e)
[1,] 6.607435
> (MSE <- SSE/(length(Y) - ncol(G)))
[1,] 0.0564738
> (h.beta <- 100*(beta.old[2]-beta.old[1]) / beta.old[1])
[1] 7.323268
> (H.beta1 <- -100*beta.old[2]/(beta.old[1]^2))
[1] -6.98677
> (H.beta2 <- 100/beta.old[1])
[1] 6.510024
> H.beta3  <- 0; H.beta4 <- 0
> H.beta <- cbind(H.beta1,H.beta2,H.beta3,H.beta4)
> (SE.h.beta <- sqrt(MSE[1,1] * (H.beta %*% solve(t(G) %*% G) %*% t(H.beta))))
[1,] 2.136145
> (CI.h.beta <- h.beta +  qt(c(.025,.975),length(Y)-ncol(G))*SE.h.beta)
[1]  3.092744 11.553791
```

$\nabla$

When the error variance is not constant, we can fit estimated weighted NLS. The weights would be the inverse of the estimated variances, as in the case of Linear Regression described previously. The variances may be related to the mean and/or the levels of one or more predictor variables. This will necessarily be an iterative process. The function we want to minimize is:

$$Q_W = \sum_{i=1}^{n} \left[v\left(g\left(\mathbf{x_i'};\boldsymbol{\beta}\right)\right)\right]^{-1} \left[Y_i - g\left(\mathbf{x_i'};\boldsymbol{\beta}\right)\right]^2 \qquad \text{where} \qquad \sigma_i^2 = v\left(g\left(\mathbf{x_i'};\boldsymbol{\beta}\right)\right).$$

When there are replicates at distinct $X$ levels, we can use the estimated variances of the replicates as the weights, in a manner like that used for the Shotgun spread example previously.

**Example: Experiments in Salmonella Growth**

Alvord, et al (1990) report growth of salmonella as a nonlinear function of mutagen dose over two experimental days. We consider the data from the second experimental day, with 3 replicates at each of 7 doses, with $n = 21$. The data are given in Table 7.3, along with the dose specific sample standard deviations. A plot of the data and the subsequent fitted equation is given in Figure 7.4. The authors fit the following nonlinear model, where $Y$ is the number of colonies observed, and $X$ is the dose. The weights are the inverse variance for the various doses.

$$Y_i = [\beta_0 + \exp\{\beta_1 + \beta_2 \ln X_i\}] \exp\{-\beta_3 X_i\} + \epsilon_i \qquad i = 1, \ldots, 21$$

The R Program and Output are given below. The data and fitted equation is given in Figure 7.4.

### R Program

```
salmonella <- read.table("http://www.stat.ufl.edu/~winner/data/salmonella.dat",header=F,
             col.names=c("expt","dose","colonies"))
attach(salmonella)

for (i in 1:length(dose)) {if(dose[i] == 0) dose[i] <- 0.000001}
expt01 <- expt-1
(vardoseexp <- aggregate(colonies,by=list(dose,expt),FUN=var))
(ndoseexp <- aggregate(colonies,by=list(dose,expt),FUN=length))
(varwt <- rep(1/vardoseexp[[3]],each=ndoseexp[[3]]))

salmonella <- data.frame(salmonella,varwt)
expt1 <- subset(salmonella,expt==1)
expt2 <- subset(salmonella,expt==2)

plot(dose,colonies,data=expt2)

salm.mod2 <- nls(colonies ~ (b0+exp(b1+b2*log(dose)))*exp(-b3*dose),
  start=c(b0=20,b1=10,b2=1,b3=1),weight=varwt,data=expt2)
summary(salm.mod2)
deviance(salm.mod2)

plot(dose,colonies,data=expt2)
dosev <- seq(.001,3.200,.001)
yhatexp2 <- predict(salm.mod2,list(dose=dosev))
lines(dosev,yhatexp2,lty=1)
```

### R Output

```
> summary(salm.mod2)

Formula: colonies ~ (b0 + exp(b1 + b2 * log(dose))) * exp(-b3 * dose)

Parameters:
   Estimate Std. Error t value Pr(>|t|)
b0 21.19355    2.52609   8.390 1.90e-07 ***
b1  7.26329    0.04514 160.891  < 2e-16 ***
b2  1.20024    0.05033  23.846 1.66e-14 ***
b3  0.30620    0.03571   8.574 1.40e-07 ***

Residual standard error: 0.9569 on 17 degrees of freedom
```

| Experiment | Dose | Colonies | SD.Dose | Experiment | Dose | Colonies | SD.Dose |
|---|---|---|---|---|---|---|---|
| 2 | 0 | 17 | 4.582576 | 1 | 0 | 25 | 4.358899 |
| 2 | 0 | 26 | 4.582576 | 1 | 0 | 18 | 4.358899 |
| 2 | 0 | 20 | 4.582576 | 1 | 0 | 26 | 4.358899 |
| 2 | 0.1 | 147 | 28.05352 | 1 | 0.1 | 156 | 7.505553 |
| 2 | 0.1 | 91 | 28.05352 | 1 | 0.1 | 149 | 7.505553 |
| 2 | 0.1 | 116 | 28.05352 | 1 | 0.1 | 164 | 7.505553 |
| 2 | 0.2 | 223 | 21.37756 | 1 | 0.2 | 294 | 29.56913 |
| 2 | 0.2 | 192 | 21.37756 | 1 | 0.2 | 268 | 29.56913 |
| 2 | 0.2 | 233 | 21.37756 | 1 | 0.2 | 327 | 29.56913 |
| 2 | 0.4 | 373 | 46.04708 | 1 | 0.4 | 511 | 19.55335 |
| 2 | 0.4 | 462 | 46.04708 | 1 | 0.4 | 473 | 19.55335 |
| 2 | 0.4 | 438 | 46.04708 | 1 | 0.4 | 500 | 19.55335 |
| 2 | 0.8 | 848 | 19.2873 | 1 | 0.8 | 1017 | 46.43634 |
| 2 | 0.8 | 878 | 19.2873 | 1 | 0.8 | 925 | 46.43634 |
| 2 | 0.8 | 884 | 19.2873 | 1 | 0.8 | 960 | 46.43634 |
| 2 | 1.6 | 1796 | 147.6347 | 1 | 1.6 | 1432 | 60.22458 |
| 2 | 1.6 | 1552 | 147.6347 | 1 | 1.6 | 1363 | 60.22458 |
| 2 | 1.6 | 1530 | 147.6347 | 1 | 1.6 | 1483 | 60.22458 |
| 2 | 3.2 | 2187 | 126.4608 | 1 | 3.2 | 1890 | 48.60384 |
| 2 | 3.2 | 2020 | 126.4608 | 1 | 3.2 | 1868 | 48.60384 |
| 2 | 3.2 | 2268 | 126.4608 | 1 | 3.2 | 1961 | 48.60384 |

Table 7.3: Salmonella Colonies Observed and Mutagen Dose for Experiments 2 and 1

```
Number of iterations to convergence: 8
Achieved convergence tolerance: 6.668e-07

> deviance(salm.mod2)
[1] 15.56472
```

A series of models were fit for both experimental days. Let $X_{i1}$ be the dose, and $X_{i2}$ be 1 for experiment 2, and 0 for experiment 1. Now there are $n = 42$ total measurements. The models are as follow.

M3: $Y_i = X_{i2}\left[\beta_0 + \exp\{\beta_1 + \beta_2 \ln X_{i1}\}\right]\exp\{-\beta_3 X_{i1}\} + (1 - X_{i2})\left[\beta_0 + \exp\{\beta_1 + \beta_2 \ln X_{i1}\}\right]\exp\{-\beta_3 X_{i1}\} + \epsilon_i$

M4: $Y_i = X_{i2}\left[\beta_{02} + \exp\{\beta_{12} + \beta_{22} \ln X_{i1}\}\right]\exp\{-\beta_{32} X_{i1}\} + (1 - X_{i2})\left[\beta_{01} + \exp\{\beta_{11} + \beta_{21} \ln X_{i1}\}\right]\exp\{-\beta_{31} X_{i1}\} + \epsilon_i$

M5: $Y_i = X_{i2}\left[\beta_{02} + \exp\{\beta_1 + \beta_{22} \ln X_{i1}\}\right]\exp\{-\beta_3 X_{i1}\} + (1 - X_{i2})\left[\beta_{01} + \exp\{\beta_1 + \beta_{21} \ln X_{i1}\}\right]\exp\{-\beta_3 X_{i1}\} + \epsilon_i$

M6: $Y_i = X_{i2}\left[\beta_0 + \exp\{\beta_1 + \beta_{22} \ln X_{i1}\}\right]\exp\{-\beta_3 X_{i1}\} + (1 - X_{i2})\left[\beta_0 + \exp\{\beta_1 + \beta_{21} \ln X_{i1}\}\right]\exp\{-\beta_3 X_{i1}\} + \epsilon_i$

M7: $Y_i = X_{i2}\left[\beta_{02} + \exp\{\beta_1 + \beta_2 \ln X_{i1}\}\right]\exp\{-\beta_3 X_{i1}\} + (1 - X_{i2})\left[\beta_{01} + \exp\{\beta_1 + \beta_2 \ln X_{i1}\}\right]\exp\{-\beta_3 X_{i1}\} + \epsilon_i$

M8: $Y_i = X_{i2}\left[\beta_0 + \exp\{\beta_{12} + \beta_2 \ln X_{i1}\}\right]\exp\{-\beta_3 X_{i1}\} + (1 - X_{i2})\left[\beta_0 + \exp\{\beta_{11} + \beta_2 \ln X_{i1}\}\right]\exp\{-\beta_3 X_{i1}\} + \epsilon_i$

M9: $Y_i = X_{i2}\left[\beta_{02} + \exp\{\beta_1 + \beta_2 \ln X_{i1}\}\right]\exp\{-\beta_{32} X_{i1}\} + (1 - X_{i2})\left[\beta_{01} + \exp\{\beta_1 + \beta_2 \ln X_{i1}\}\right]\exp\{-\beta_{31} X_{i1}\} + \epsilon_i$

A portion of the R Program and Output are given below, and summaries of the models are given in Table 7.4. A plot of the data and the regression based on Model 4 is given in Figure 7.5. The Residual Standard Error is based on the weighted residuals, not the "true" residuals. Note that the R functions **deviance**, **logLik**, and **AIC** do not work correctly for the **nls** function when the **weight** option is used.

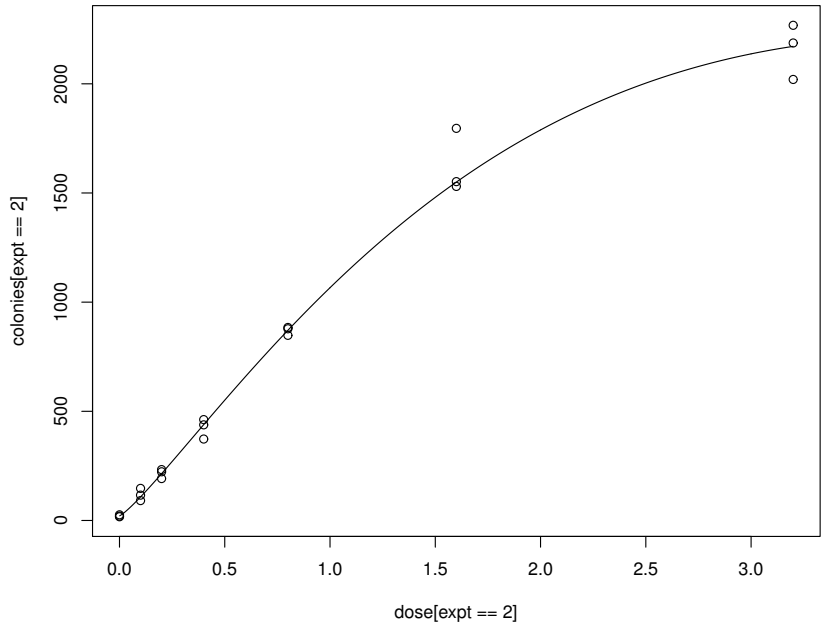


Figure 7.4: Plot of Colonies and Fitted Equation for Experiment 2

We computed $SSE$, $AIC$, and $BIC$ directly. $AIC$ and $BIC$ are computed as a multiplier of the number of etimated parameters (including $\sigma^2$) - twice the log-likelihood evaluated at the ML estimates for $\mu$ and $\sigma^2$.

$$SSE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

$$AIC = 2(p+1) + n\left[\ln(2\pi) + \ln\left(\frac{SSE}{n}\right) + 1\right]$$

$$BIC = \ln(n)(p+1) + n\left[\ln(2\pi) + \ln\left(\frac{SSE}{n}\right) + 1\right]$$

```
salm.mod4 <- nls(colonies ~ expt01*(b02+exp(b12+b22*log(dose)))*exp(-b32*dose)
 + (1-expt01)*(b01+exp(b11+b21*log(dose)))*exp(-b31*dose),
  start=c(b01=20,b11=10,b21=1,b31=1,b02=20,b12=10,b22=1,b32=1),
  weight=varwt,data=salmonella)
summary(salm.mod4)
deviance(salm.mod4)
AIC(salm.mod4)

plot(dose,colonies,pch=expt)
dosev <- seq(.001,3.200,.001)
expt1v <- rep(0,length(seq(.001,3.200,.001)))
expt2v <- rep(1,length(seq(.001,3.200,.001)))
yhatexp1 <- predict(salm.mod6,list(dose=dosev, expt01=expt1v))
yhatexp2 <- predict(salm.mod6,list(dose=dosev, expt01=expt2v))
```

| Model | $SSE$ | $AIC$ | $BIC$ | Parameters |
|---|---|---|---|---|
| 3 | 399778.5 | 513.9527 | 522.6410 | 4+1=5 |
| 4 | 136627.4 | 476.8593 | 492.4983 | 8+1=9 |
| 5 | 155795.1 | 478.3732 | 490.5369 | 6+1=7 |
| 6 | 160459.2 | 477.6121 | 488.0381 | 5+1=6 |
| 7 | 394878.6 | 515.4347 | 525.8607 | 5+1=6 |
| 8 | 490021.1 | 524.5013 | 534.9273 | 5+1=6 |
| 9 | 219897.4 | 492.8472 | 505.0109 | 6+1=7 |

Table 7.4: Model Fit Statistic (Note that # of Parameters includes $\sigma^2$)

```
lines(dosev,yhatexp1,lty=1)
lines(dosev,yhatexp2,lty=2)
legend("topleft",c("Experiment1","Experiment2"),pch=c(1,2),lty=c(1,2))

### Output
> summary(salm.mod4)

Formula: colonies ~ expt01 * (b02 + exp(b12 + b22 * log(dose))) * exp(-b32 *
    dose) + (1 - expt01) * (b01 + exp(b11 + b21 * log(dose))) *
    exp(-b31 * dose)

Parameters:
    Estimate Std. Error t value Pr(>|t|)
b01 23.15862    2.67842   8.646 4.21e-10 ***
b11  7.21858    0.03948 182.857  < 2e-16 ***
b21  1.00008    0.02683  37.274  < 2e-16 ***
b31  0.26099    0.02292  11.389 3.77e-13 ***
b02 21.19355    2.82742   7.496 1.06e-08 ***
b12  7.26329    0.05053 143.744  < 2e-16 ***
b22  1.20024    0.05634  21.305  < 2e-16 ***
b32  0.30620    0.03997   7.660 6.64e-09 ***

Residual standard error: 1.071 on 34 degrees of freedom

Number of iterations to convergence: 8
Achieved convergence tolerance: 4.08e-07

> deviance(salm.mod4)
[1] 38.99907
> AIC(salm.mod4)
[1] 11570.38
```

$\nabla$

If the errors are correlated with a known correlation structure, such as AR(1), the autoregressive parameter(s) can be estimated and plugged into the variance-covariance matrix, and we can fit estimated generalized NLS. Here we want to minimize:

$$\left[Y_i - g\left(\mathbf{x_i'}; \boldsymbol{\beta}\right)\right]' \mathbf{V^{-1}} \left[Y_i - g\left(\mathbf{x_i'}; \boldsymbol{\beta}\right)\right]$$

where the elements of $\mathbf{V}$ are functions of unknown parameters which are estimated from the residuals. See the AR(1) description for Linear Regression.

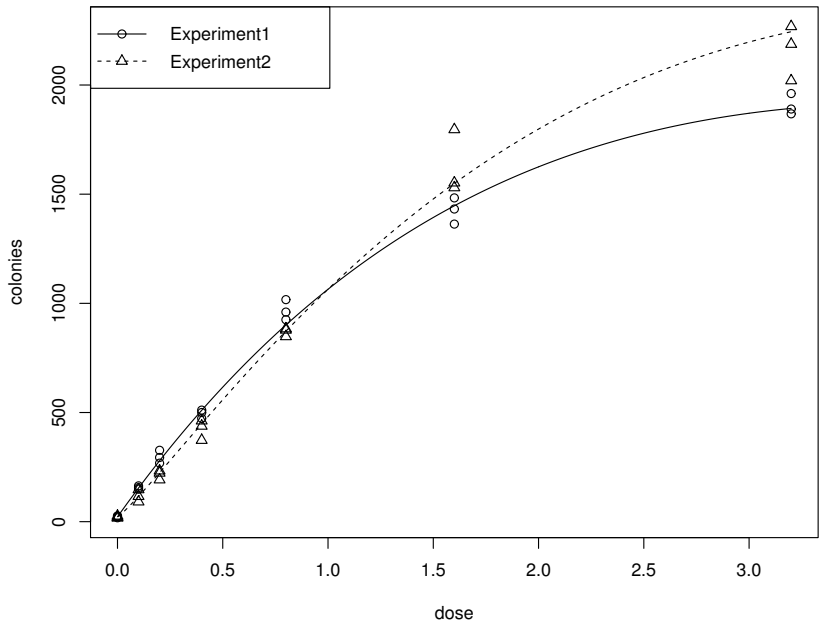


Figure 7.5: Dose Response Curves Based on Model 6 - Salmonella Experiments

# Chapter 8

# Random Coefficient Regression Models

**Random coefficient regression (RCR)** models are typically longitudinal in nature, with a sample of $n$ units being observed on multiple occasions, with $Y$ and $X_1$,...,$X_p$ being observed on units at the various occasions. These models are widely used in various disciplines, and are also known as **Hierarchical Linear Models (HLMs)** and **Multilevel Models**. These methods are described in various books and chapters (e.g. Bryk and Raudenbush (1992), Goldstein (1987), Rawlings, Pantula, and Dickey (1998, Sections 18.3-18.4), and Littell, Milliken, Stroup, and Wolfinger (1996, Chapter 7). Note that nonlinear regression models can also have random coefficients. Here we consider a simple, balanced model, then consider more general cases.

## 8.1 Balanced Model with 1 Predictor

We consider a case where $n$ experimental units are each observed at $t$ occasion, with observed pairs $(X_{ij}, Y_{ij})$ $i = 1, \cdots, n; j = 1, ..., t$. The model allows for different intercepts and slopes among te units. Note that in many growth curve applications among units, the predictor variable is simply the time point.

$$Y_{ij} = \alpha_i + \beta_i X_{ij} + \epsilon_{ij} \qquad i = 1, ..., n; \quad j = 1, ..., t \qquad \epsilon_{ij} \sim N\left(0, \sigma^2\right) \text{ independent}$$

$$\begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} \sim N\left(\begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \begin{bmatrix} \sigma^2_\alpha & \sigma_{\alpha\beta} \\ \sigma_{\alpha\beta} & \sigma^2_\beta \end{bmatrix}\right) \qquad \{\epsilon\} \perp \left\{\begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix}\right\}$$

An alternative, widely used way to re-write this model is as follows.

$$Y_{ij} = (\alpha + \beta X_{ij}) + [(\alpha_i - \alpha) + (\beta_i - \beta) X_{ij} + \epsilon_{ij}]$$

$$\Rightarrow \qquad E\{Y_{ij}\} = \alpha + \beta X_{ij} \qquad\qquad V\{Y_{ij}\} = \sigma^2_\alpha + X^2_{ij}\sigma^2_\beta + \sigma^2 + 2X_{ij}\sigma_{\alpha\beta}$$

$$j \neq j': \quad \text{COV}\{Y_{ij}, Y_{ij'}\} = \sigma^2_\alpha + X_{ij}X_{ij'}\sigma^2_\beta + (X_{ij} + X_{ij'})\sigma_{\alpha\beta}$$

## 8.2 General Model with $p$ Predictors

Gumpertz and Pantula (1989) considered a general (balanced) model, and obtained unbiased estimators of the model parameters. When the number of measurements per subject are equal (e.g. balanced model), these "simple" estimators coincide with those obtained from mixed model statistical software procedures that perform EGLS estimation of the model parameters. The scalar and matrix forms are given below, along with the mean and variance/covariance structure.

$$Y_{ij} = \beta_{i0} + \beta_{i1}X_{ij1} + \cdots + \beta_{ip}X_{ijp} + \epsilon_{ij} \qquad i = 1, \ldots, n; \quad j = 1, \ldots, t$$

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i \qquad \mathbf{Y}_i = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{it} \end{bmatrix} \qquad \mathbf{X}_i = \begin{bmatrix} 1 & X_{i11} & \cdots & X_{i1p} \\ 1 & X_{i21} & \cdots & X_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{it1} & \cdots & X_{itp} \end{bmatrix} \qquad \boldsymbol{\beta}_i = \begin{bmatrix} \beta_{i0} \\ \beta_{i1} \\ \vdots \\ \beta_{ip} \end{bmatrix} \qquad \boldsymbol{\varepsilon}_i = \begin{bmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{it} \end{bmatrix}$$

$$\boldsymbol{\beta}_i \sim NID\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}_\beta\right) \qquad \boldsymbol{\varepsilon}_i \sim NID\left(\mathbf{0}, \sigma^2\mathbf{I}\right) \qquad \{\boldsymbol{\beta}_i\} \perp \{\boldsymbol{\varepsilon}_i\}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \qquad \boldsymbol{\Sigma}_\beta = \begin{bmatrix} V\{\beta_{i0}\} & \text{COV}\{\beta_{i0}, \beta_{i1}\} & \cdots & \text{COV}\{\beta_{i0}, \beta_{ip}\} \\ \text{COV}\{\beta_{i0}, \beta_{i1}\} & V\{\beta_{i1}\} & \cdots & \text{COV}\{\beta_{i1}, \beta_{ip}\} \\ \vdots & \vdots & \ddots & \vdots \\ \text{COV}\{\beta_{i0}, \beta_{ip}\} & \text{COV}\{\beta_{i1}, \beta_{ip}\} & \cdots & V\{\beta_{ip}\} \end{bmatrix}$$

$$E\{\mathbf{Y}_i\} = E\{\mathbf{X}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i\} = \mathbf{X}_i\boldsymbol{\beta} \qquad V\{\mathbf{Y}_i\} = V\{\mathbf{X}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i\} = \mathbf{X}_i\boldsymbol{\Sigma}_\beta\mathbf{X}_i' + \sigma^2\mathbf{I} \qquad \text{COV}\{\mathbf{Y}_i, \mathbf{Y}_{i'}\} = 0 \quad i \neq i'$$

$$\mathbf{Y}_i \sim N\left(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{X}_i\boldsymbol{\Sigma}_\beta\mathbf{X}_i' + \sigma^2\mathbf{I}\right)$$

The predictor of the $i^{th}$ unit's regression coefficient vector can be obtained by fitting the regression for that individual.

$$\hat{\boldsymbol{\beta}}_i = \left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\mathbf{X}_i'\mathbf{Y}_i \qquad E\left\{\hat{\boldsymbol{\beta}}_i\right\} = \left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\mathbf{X}_i'\mathbf{X}_i\boldsymbol{\beta} = \boldsymbol{\beta}$$

$$V\left\{\hat{\boldsymbol{\beta}}_i\right\} = \left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\mathbf{X}_i'\left[\mathbf{X}_i\boldsymbol{\Sigma}_\beta\mathbf{X}_i' + \sigma^2\mathbf{I}\right]\mathbf{X}_i\left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1} = \boldsymbol{\Sigma}_\beta + \sigma^2\left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}$$

Note that each individual's $\hat{\boldsymbol{\beta}}_i$ is an unbiased estimator of the population mean $\boldsymbol{\beta}$. That leads to making use of a simple average of the $\hat{\boldsymbol{\beta}}_i$ as an unbiased estimator of the mean vector.

$$\hat{\boldsymbol{\beta}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\beta}}_i \qquad \Rightarrow \qquad E\left\{\hat{\boldsymbol{\beta}}\right\} = \frac{1}{n}n\boldsymbol{\beta} = \boldsymbol{\beta}$$

$$V\left\{\hat{\boldsymbol{\beta}}\right\}=\frac{1}{n^2}\left[n\boldsymbol{\Sigma}_\beta+\sigma^2\sum_{i=1}^{n}\left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\right]\qquad \text{COV}\left\{\hat{\boldsymbol{\beta}}_i,\hat{\boldsymbol{\beta}}\right\}=\frac{1}{n}\left[\boldsymbol{\Sigma}_\beta+\sigma^2\left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\right]$$

Now, making use of the distributional properties of the regression model, we can estimate $\sigma^2$ and $\boldsymbol{\Sigma}$.

$$\left(\mathbf{Y}_i-\mathbf{X}_i\hat{\boldsymbol{\beta}}_i\right)'\left(\mathbf{Y}_i-\mathbf{X}_i\hat{\boldsymbol{\beta}}_i\right)=\mathbf{Y}_i'\left(\mathbf{I}-\mathbf{P}_i\right)\mathbf{Y}_i\qquad \mathbf{P}_i=\mathbf{X}_i\left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\mathbf{X}_i'$$

$$E\left\{\mathbf{Y}_i'\left(\mathbf{I}-\mathbf{P}_i\right)\mathbf{Y}_i\right\}=\sigma^2\text{trace}\left(\mathbf{I}-\mathbf{P}_i\right)+\boldsymbol{\beta}'\mathbf{X}_i'\left(\mathbf{I}-\mathbf{P}_i\right)\mathbf{X}_i\boldsymbol{\beta}\quad=\quad\sigma^2\left(t-p'\right)+0$$

$$\Rightarrow\qquad E\left\{\sum_{i=1}^{n}\left(\mathbf{Y}_i-\mathbf{X}_i\hat{\boldsymbol{\beta}}_i\right)'\left(\mathbf{Y}_i-\mathbf{X}_i\hat{\boldsymbol{\beta}}_i\right)\right\}=n\left(t-p'\right)\sigma^2$$

$$\Rightarrow\qquad s^2=\frac{1}{n\left(t-p'\right)}\sum_{i=1}^{n}\left(\mathbf{Y}_i-\mathbf{X}_i\hat{\boldsymbol{\beta}}_i\right)'\left(\mathbf{Y}_i-\mathbf{X}_i\hat{\boldsymbol{\beta}}_i\right)=\frac{1}{n\left(t-p'\right)}\sum_{i=1}^{n}SSE_i=\frac{1}{n}\sum_{i=1}^{n}MSE_i$$

$$V\left\{\hat{\boldsymbol{\beta}}_i-\hat{\boldsymbol{\beta}}\right\}=E\left\{\left(\hat{\boldsymbol{\beta}}_i-\hat{\boldsymbol{\beta}}\right)\left(\hat{\boldsymbol{\beta}}_i-\hat{\boldsymbol{\beta}}\right)'\right\}-E\left\{\left(\hat{\boldsymbol{\beta}}_i-\hat{\boldsymbol{\beta}}\right)\right\}E\left\{\left(\hat{\boldsymbol{\beta}}_i-\hat{\boldsymbol{\beta}}\right)'\right\}\qquad E\left\{\left(\hat{\boldsymbol{\beta}}_i-\hat{\boldsymbol{\beta}}\right)\right\}=\boldsymbol{\beta}-\boldsymbol{\beta}=\mathbf{0}$$

$$\Rightarrow\qquad E\left\{\left(\hat{\boldsymbol{\beta}}_i-\hat{\boldsymbol{\beta}}\right)\left(\hat{\boldsymbol{\beta}}_i-\hat{\boldsymbol{\beta}}\right)'\right\}=V\left\{\hat{\boldsymbol{\beta}}_i-\hat{\boldsymbol{\beta}}\right\}=V\left\{\hat{\boldsymbol{\beta}}_i\right\}+V\left\{\hat{\boldsymbol{\beta}}\right\}-2\text{COV}\left\{\hat{\boldsymbol{\beta}}_i,\hat{\boldsymbol{\beta}}\right\}=$$

$$\boldsymbol{\Sigma}_\beta+\sigma^2\left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}+\frac{1}{n^2}\left[n\boldsymbol{\Sigma}_\beta+\sigma^2\sum_{i=1}^{n}\left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\right]-2\frac{1}{n}\left[\boldsymbol{\Sigma}_\beta+\sigma^2\left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\right]=$$

$$\frac{n-1}{n}\boldsymbol{\Sigma}_\beta+\sigma^2\left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\left[1+\frac{1}{n^2}-\frac{2}{n}\right]+\frac{\sigma^2}{n^2}\sum_{\substack{j=1\\j\neq i}}^{n}\left(\mathbf{X}_j'\mathbf{X}_j\right)^{-1}$$

$$\Rightarrow\qquad E\left\{\sum_{i=1}^{n}\left(\hat{\boldsymbol{\beta}}_i-\hat{\boldsymbol{\beta}}\right)\left(\hat{\boldsymbol{\beta}}_i-\hat{\boldsymbol{\beta}}\right)'\right\}=n\left(\frac{n-1}{n}\right)\boldsymbol{\Sigma}_\beta+\sigma^2\left[1+\frac{1}{n^2}-\frac{2}{n}\right]\sum_{i=1}^{n}\left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}+\frac{\sigma^2}{n^2}\sum_{i=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n}\left(\mathbf{X}_j'\mathbf{X}_j\right)^{-1}$$

Note that in the last double summation, each unit $(i)$ appears $n-1$ times.

$$\Rightarrow\qquad E\left\{\sum_{i=1}^{n}\left(\hat{\boldsymbol{\beta}}_i-\hat{\boldsymbol{\beta}}\right)\left(\hat{\boldsymbol{\beta}}_i-\hat{\boldsymbol{\beta}}\right)'\right\}=\left(n-1\right)\boldsymbol{\Sigma}_\beta+\sigma^2\left[1+\frac{1}{n^2}-\frac{2}{n}+\frac{n-1}{n^2}\right]\sum_{i=1}^{n}\left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}$$

$$=\left(n-1\right)\boldsymbol{\Sigma}_\beta+\sigma^2\left(\frac{n-1}{n}\right)\sum_{i=1}^{n}\left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}$$

This leads to the following unbiased estimator for $\boldsymbol{\Sigma}_\beta$.

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}\right)\left(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}\right)' - \frac{s^2}{n}\sum_{i=1}^{n}\left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}$$

**Example: Airline Annual Revenues for 10 Large Markets 1996/7-2000/1**

Data in Table 8.1 are annual revenues for a random sample of $n = 10$ large airline markets over $t = 5$ fiscal years. The years are labeled as 0, 1, 2, 3, 4 in the regression model. As all markets have the same years, the $\mathbf{X}_i$ matrix is the same for all markets. We will fit a linear regression for each market, relating ln(Revenue) to year.

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i \quad i = 1,\ldots,n \qquad \boldsymbol{\beta}_i \sim N\left(\beta = \begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix}, \Sigma_{\boldsymbol{\beta}} = \begin{bmatrix}\sigma^2_{\beta_0} & \sigma_{\beta_0,\beta_1}\\ \sigma_{\beta_0,\beta_1} & \sigma^2_{\beta_1}\end{bmatrix}\right) \qquad \boldsymbol{\varepsilon}_i \sim N\left(\mathbf{0}_5, \sigma^2\mathbf{I}_5\right)$$

$$\mathbf{X}_i = \begin{bmatrix}1 & 0\\ 1 & 1\\ 1 & 2\\ 1 & 3\\ 1 & 4\end{bmatrix} \quad \Rightarrow \quad \mathbf{X}_i'\mathbf{X}_i = \begin{bmatrix}5 & 10\\ 10 & 30\end{bmatrix} \quad \Rightarrow \quad \left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1} = \begin{bmatrix}0.60 & -0.20\\ -0.20 & 0.10\end{bmatrix}$$

$$\left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\mathbf{X}_i' = \begin{bmatrix}0.6 & 0.4 & 0.2 & 0 & -0.2\\ -0.2 & -0.1 & 0.0 & 0.1 & 0.2\end{bmatrix} \quad \Rightarrow \quad \hat{\boldsymbol{\beta}}_i = \left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\mathbf{X}_i'\mathbf{Y}_i \qquad i = 1,\ldots,10$$

The results for the $n = 10$ individual regressions are given in Table 8.2.

$$\hat{\boldsymbol{\beta}} = \frac{1}{10}\left(\begin{bmatrix}7.1821\\ 0.0337\end{bmatrix} + \cdots + \begin{bmatrix}6.3676\\ 0.0325\end{bmatrix}\right) = \begin{bmatrix}7.095741\\ 0.065320\end{bmatrix}$$

$$s^2 = \frac{1}{10(5-2)}\left(0.0188 + \cdots + 0.0132\right) = \frac{0.093505}{30} = 0.003117$$

$$\sum_{i=1}^{n}\left(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}\right)\left(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}\right)' = \begin{bmatrix}0.0864\\ -0.0316\end{bmatrix}\begin{bmatrix}0.0864 & -0.0316\end{bmatrix} + \cdots + \begin{bmatrix}-0.7281\\ -0.0328\end{bmatrix}\begin{bmatrix}-0.7281 & -0.0328\end{bmatrix} =$$

$$\begin{bmatrix}0.007465 & -0.002730\\ -0.002730 & 0.000999\end{bmatrix} + \cdots + \begin{bmatrix}0.530130 & 0.023882\\ 0.023882 & 0.001076\end{bmatrix} = \begin{bmatrix}4.336061 & -0.035867\\ -0.035867 & 0.009864\end{bmatrix}$$

$$\Rightarrow \qquad \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} = \frac{1}{10-1}\begin{bmatrix} 4.336061 & -0.035867 \\ -0.035867 & 0.009864 \end{bmatrix} - \frac{0.003117}{10}(10)\begin{bmatrix} 0.60 & -0.20 \\ -0.20 & 0.10 \end{bmatrix} = \begin{bmatrix} 0.479915 & -0.003362 \\ -0.003362 & 0.000784 \end{bmatrix}$$

The R program and output are given below. They make use of the **lmerTest** package and the **lmer** function. The graphs are given in Figure 8.1 (all lines on one plot) and Figure 8.2 (trellis graphs with data and simple linear regressions).

```
### Program
big20 <- big20 <- read.fwf("http://www.stat.ufl.edu/~winner/sta6208/reg_ex/big20_air_samp.prn",
width=c(3,8,13,8,8), col.names=c("city1", "city2", "market", "revenue", "year"))

attach(big20)
market <- factor(market)
library(lmerTest)

air1 <- lmer(log(revenue) ~ year +  (year|market))
summary(air1)
ranef(air1)
coef(air1)
fixef(air1)

yearplot <- 0:4
Rev_all <- fixef(air1)[1] + fixef(air1)[2]*yearplot
Rev_each <- matrix(rep(0,50),ncol=10)
for (i in 1:10) {
Rev_each[,i] <- coef(air1)$market[i,1] + coef(air1)$market[i,2]*yearplot
}

Rev_each
ylimlo <- 0.9*min(Rev_each); ylimhi <- 1.1*max(Rev_each)

plot(yearplot,Rev_all,type="l",xlab="Year",ylab="ln(Rev)",lwd=4,
    ylim=c(ylimlo,ylimhi),main="Ln(Revenues) by Year")
for (i in 1:10) {
lines(yearplot,Rev_each[,i],lty=i)
}

library(lattice)
xyplot(log(revenue) ~ year | market,
        panel = function(x,y) {
        panel.xyplot(x,y,pch=16)
        panel.abline(lm(y~x))
})

# Text Output
> summary(air1)
Linear mixed model fit by REML t-tests use Satterthwaite approximations to
  degrees of freedom [merModLmerTest]
Formula: log(revenue) ~ year + (year | market)

REML criterion at convergence: -61.4

Random effects:
 Groups   Name        Variance  Std.Dev. Corr
 market   (Intercept) 0.4799143 0.69276
          year        0.0007843 0.02801  -0.17
 Residual             0.0031168 0.05583
Number of obs: 50, groups:  market, 10
```

```
Fixed effects:
            Estimate Std. Error      df t value Pr(>|t|)
(Intercept)  7.09574    0.21950 9.00000  32.327 1.27e-10 ***
year         0.06532    0.01047 9.00000   6.239 0.000152 ***

Correlation of Fixed Effects:
     (Intr)
year -0.173
> ranef(air1)
$market
    (Intercept)          year
2    0.06839075 -0.0225839637
3    0.17257281  0.0091886257
6    1.10490018 -0.0132366004
8    0.25520157  0.0311881120
19  -1.07023650 -0.0003365672
25  -0.14995496 -0.0137260742
31  -0.01721360  0.0378828442
42  -0.57770515  0.0212227803
95   0.96152897 -0.0270340094
108 -0.74748408 -0.0225651472

> coef(air1)
$market
    (Intercept)       year
2      7.164132 0.04273622
3      7.268314 0.07450881
6      8.200641 0.05208359
8      7.350943 0.09650830
19     6.025505 0.06498362
25     6.945786 0.05159411
31     7.078528 0.10320303
42     6.518036 0.08654297
95     8.057270 0.03828618
108    6.348257 0.04275504

attr(,"class")
[1] "coef.mer"
> fixef(air1)
(Intercept)        year
 7.09574112  0.06532019
```

$\nabla$

### 8.2.1 Unequal Sample Sizes Within Subjects

When the data are "stacked" so that the measurements from unit 1 are followed by those for units 2 through $n$, we have the following structure.

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_n \end{bmatrix} \qquad E\{\mathbf{Y}\} = \begin{bmatrix} \mathbf{X}_1\boldsymbol{\beta} \\ \mathbf{X}_2\boldsymbol{\beta} \\ \vdots \\ \mathbf{X}_n\boldsymbol{\beta} \end{bmatrix} = \mathbf{X}\boldsymbol{\beta} \qquad \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{bmatrix} \qquad \mathbf{X}_i = \begin{bmatrix} 1 & X_{i11} & \cdots & X_{i1p} \\ 1 & X_{i21} & \cdots & X_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{it_i1} & \cdots & X_{it_ip} \end{bmatrix}$$

| city1 | city2 | market | revenue | year | $Y = \ln(\text{rev})$ |
|---|---|---|---|---|---|
| ATL | BOS | 2 | 1274.6 | 0 | 7.150388 |
| ATL | BOS | 2 | 1506.1 | 1 | 7.317279 |
| ATL | BOS | 2 | 1307.7 | 2 | 7.176025 |
| ATL | BOS | 2 | 1412.4 | 3 | 7.253046 |
| ATL | BOS | 2 | 1557.9 | 4 | 7.351094 |
| ATL | DFW | 3 | 1511.1 | 0 | 7.320593 |
| ATL | DFW | 3 | 1489.4 | 1 | 7.306129 |
| ATL | DFW | 3 | 1591.2 | 2 | 7.372244 |
| ATL | DFW | 3 | 1713.1 | 3 | 7.44606 |
| ATL | DFW | 3 | 2086.1 | 4 | 7.643052 |
| ATL | LGA | 6 | 3626.3 | 0 | 8.195968 |
| ATL | LGA | 6 | 3957.3 | 1 | 8.283317 |
| ATL | LGA | 6 | 3977.9 | 2 | 8.288509 |
| ATL | LGA | 6 | 4302.7 | 3 | 8.366998 |
| ATL | LGA | 6 | 4430.5 | 4 | 8.396268 |
| ATL | ORD | 8 | 1487.9 | 0 | 7.305121 |
| ATL | ORD | 8 | 1738.1 | 1 | 7.460548 |
| ATL | ORD | 8 | 1924.4 | 2 | 7.56237 |
| ATL | ORD | 8 | 2044.6 | 3 | 7.622957 |
| ATL | ORD | 8 | 2371.8 | 4 | 7.771404 |
| BUR | LAS | 19 | 413.5 | 0 | 6.024658 |
| BUR | LAS | 19 | 444.5 | 1 | 6.09695 |
| BUR | LAS | 19 | 467.6 | 2 | 6.147613 |
| BUR | LAS | 19 | 504.5 | 3 | 6.223568 |
| BUR | LAS | 19 | 532.4 | 4 | 6.277395 |
| BWI | ORD | 25 | 1062.3 | 0 | 6.968192 |
| BWI | ORD | 25 | 1079.4 | 1 | 6.984161 |
| BWI | ORD | 25 | 1163 | 2 | 7.058758 |
| BWI | ORD | 25 | 1195.1 | 3 | 7.085985 |
| BWI | ORD | 25 | 1269.5 | 4 | 7.146378 |
| DEN | SFO | 31 | 1116.2 | 0 | 7.017685 |
| DEN | SFO | 31 | 1269.4 | 1 | 7.1463 |
| DEN | SFO | 31 | 1515.8 | 2 | 7.323699 |
| DEN | SFO | 31 | 1825.2 | 3 | 7.509445 |
| DEN | SFO | 31 | 1683.3 | 4 | 7.428511 |
| DTW | MCO | 42 | 714.8 | 0 | 6.572003 |
| DTW | MCO | 42 | 681.2 | 1 | 6.523856 |
| DTW | MCO | 42 | 753.2 | 2 | 6.624331 |
| DTW | MCO | 42 | 947.8 | 3 | 6.854144 |
| DTW | MCO | 42 | 970.7 | 4 | 6.878017 |
| ORD | LAX | 95 | 3190.1 | 0 | 8.067808 |
| ORD | LAX | 95 | 3304.5 | 1 | 8.10304 |
| ORD | LAX | 95 | 3467.8 | 2 | 8.151276 |
| ORD | LAX | 95 | 3563.4 | 3 | 8.17847 |
| ORD | LAX | 95 | 3548.5 | 4 | 8.17428 |
| PHX | SFO | 108 | 548.3 | 0 | 6.306823 |
| PHX | SFO | 108 | 634.9 | 1 | 6.453468 |
| PHX | SFO | 108 | 631.5 | 2 | 6.448098 |
| PHX | SFO | 108 | 676.9 | 3 | 6.517524 |
| PHX | SFO | 108 | 624.7 | 4 | 6.437272 |

Table 8.1: Airline Market Annual Revenues 1996/7-2000/1

| $\mathbf{Y}_1$ | $\mathbf{Y}_2$ | $\mathbf{Y}_3$ | $\mathbf{Y}_4$ | $\mathbf{Y}_5$ | $\mathbf{Y}_6$ | $\mathbf{Y}_7$ | $\mathbf{Y}_8$ | $\mathbf{Y}_9$ | $\mathbf{Y}_{10}$ |
|---|---|---|---|---|---|---|---|---|---|
| 7.1504 | 7.3206 | 8.1960 | 7.3051 | 6.0247 | 6.9682 | 7.0177 | 6.5720 | 8.0678 | 6.3068 |
| 7.3173 | 7.3061 | 8.2833 | 7.4605 | 6.0970 | 6.9842 | 7.1463 | 6.5239 | 8.1030 | 6.4535 |
| 7.1760 | 7.3722 | 8.2885 | 7.5624 | 6.1476 | 7.0588 | 7.3237 | 6.6243 | 8.1513 | 6.4481 |
| 7.2530 | 7.4461 | 8.3670 | 7.6230 | 6.2236 | 7.0860 | 7.5094 | 6.8541 | 8.1785 | 6.5175 |
| 7.3511 | 7.6431 | 8.3963 | 7.7714 | 6.2774 | 7.1464 | 7.4285 | 6.8780 | 8.1743 | 6.4373 |
| $\hat{\boldsymbol{\beta}}_1$ | $\hat{\boldsymbol{\beta}}_2$ | $\hat{\boldsymbol{\beta}}_3$ | $\hat{\boldsymbol{\beta}}_4$ | $\hat{\boldsymbol{\beta}}_5$ | $\hat{\boldsymbol{\beta}}_6$ | $\hat{\boldsymbol{\beta}}_7$ | $\hat{\boldsymbol{\beta}}_8$ | $\hat{\boldsymbol{\beta}}_9$ | $\hat{\boldsymbol{\beta}}_{10}$ |
| 7.1821 | 7.2606 | 8.2094 | 7.3255 | 6.0276 | 6.9571 | 7.0482 | 6.5020 | 8.0773 | 6.3676 |
| 0.0337 | 0.0785 | 0.0484 | 0.1095 | 0.0632 | 0.0458 | 0.1185 | 0.0942 | 0.0288 | 0.0325 |
| $\hat{\mathbf{Y}}_1$ | $\hat{\mathbf{Y}}_2$ | $\hat{\mathbf{Y}}_3$ | $\hat{\mathbf{Y}}_4$ | $\hat{\mathbf{Y}}_5$ | $\hat{\mathbf{Y}}_6$ | $\hat{\mathbf{Y}}_7$ | $\hat{\mathbf{Y}}_8$ | $\hat{\mathbf{Y}}_9$ | $\hat{\mathbf{Y}}_{10}$ |
| 7.1821 | 7.2606 | 8.2094 | 7.3255 | 6.0276 | 6.9571 | 7.0482 | 6.5020 | 8.0773 | 6.3676 |
| 7.2158 | 7.3391 | 8.2578 | 7.4350 | 6.0908 | 7.0029 | 7.1666 | 6.5962 | 8.1061 | 6.4001 |
| 7.2496 | 7.4176 | 8.3062 | 7.5445 | 6.1540 | 7.0487 | 7.2851 | 6.6905 | 8.1350 | 6.4326 |
| 7.2833 | 7.4961 | 8.3546 | 7.6540 | 6.2172 | 7.0945 | 7.4036 | 6.7847 | 8.1638 | 6.4651 |
| 7.3170 | 7.5746 | 8.4031 | 7.7635 | 6.2805 | 7.1403 | 7.5221 | 6.8789 | 8.1926 | 6.4976 |
| $SSE_1$ | $SSE_2$ | $SSE_3$ | $SSE_4$ | $SSE_5$ | $SSE_6$ | $SSE_7$ | $SSE_8$ | $SSE_9$ | $SSE_{10}$ |
| 0.0188 | 0.0139 | 0.0013 | 0.0024 | 0.0001 | 0.0007 | 0.0228 | 0.0193 | 0.0009 | 0.0132 |
| $\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}$ | $\hat{\boldsymbol{\beta}}_2 - \hat{\boldsymbol{\beta}}$ | $\hat{\boldsymbol{\beta}}_3 - \hat{\boldsymbol{\beta}}$ | $\hat{\boldsymbol{\beta}}_4 - \hat{\boldsymbol{\beta}}$ | $\hat{\boldsymbol{\beta}}_5 - \hat{\boldsymbol{\beta}}$ | $\hat{\boldsymbol{\beta}}_6 - \hat{\boldsymbol{\beta}}$ | $\hat{\boldsymbol{\beta}}_7 - \hat{\boldsymbol{\beta}}$ | $\hat{\boldsymbol{\beta}}_8 - \hat{\boldsymbol{\beta}}$ | $\hat{\boldsymbol{\beta}}_9 - \hat{\boldsymbol{\beta}}$ | $\hat{\boldsymbol{\beta}}_{10} - \hat{\boldsymbol{\beta}}$ |
| 0.0864 | 0.1649 | 1.1136 | 0.2297 | -1.0681 | -0.1387 | -0.0476 | -0.5937 | 0.9816 | -0.7281 |
| -0.0316 | 0.0132 | -0.0169 | 0.0442 | -0.0021 | -0.0195 | 0.0532 | 0.0289 | -0.0365 | -0.0328 |

Table 8.2: Market Specific Revenues, Regression Coefficients, Fitted Values, $SSE$, and Coefficient Deviations

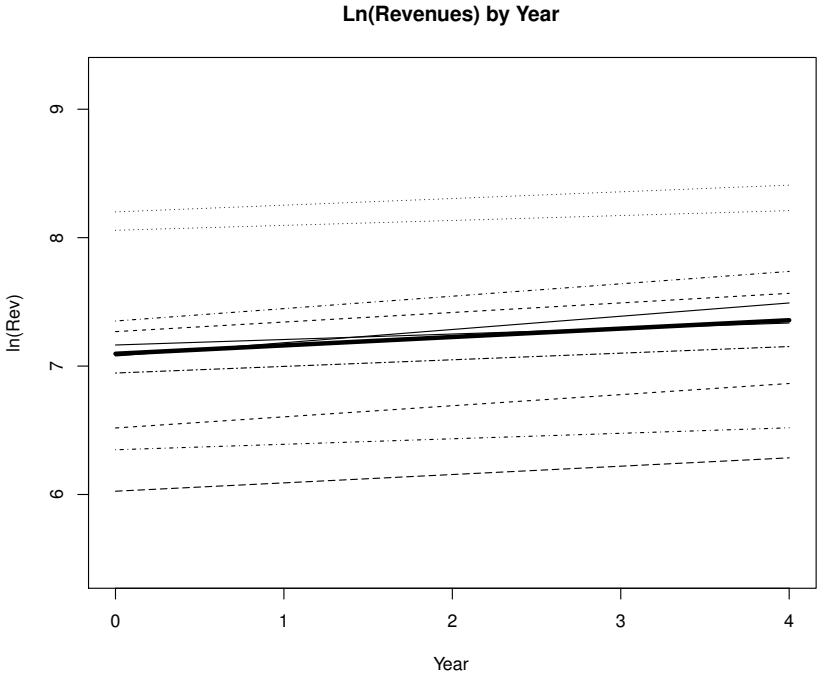


Figure 8.1: Estimated Mean Line (Bold) and Individual Market Lines

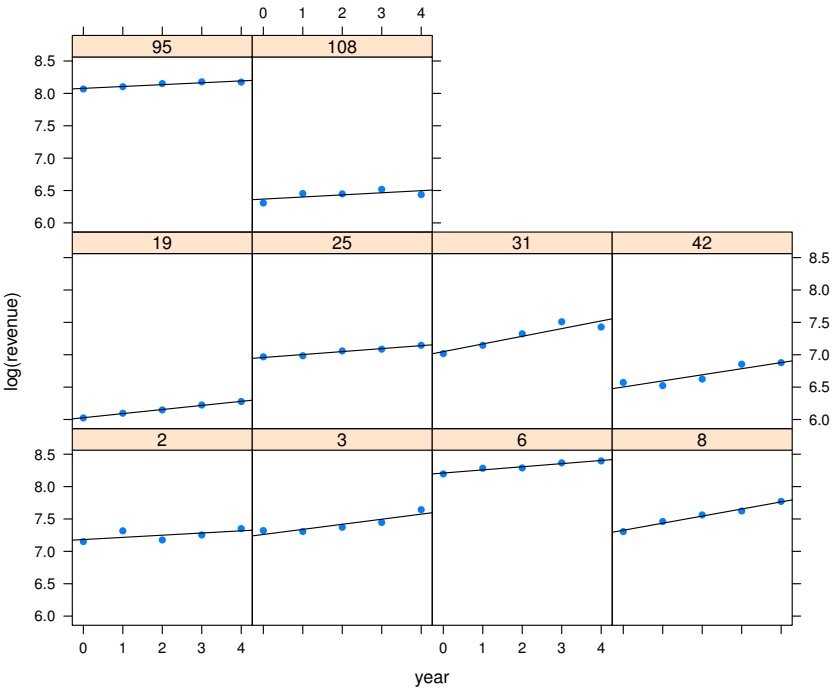


Figure 8.2: Trellis Graph for Individual Markets with RCR Lines

$$V\{\mathbf{Y}\} = \begin{bmatrix} \mathbf{X}_1\boldsymbol{\Sigma}_\beta\mathbf{X}_1' + \sigma^2\mathbf{I}_{t_1} & \mathbf{0}_{t_2} & \cdots & \mathbf{0}_{t_n} \\ \mathbf{0}_{t_1} & \mathbf{X}_2\boldsymbol{\Sigma}_\beta\mathbf{X}_2' + \sigma^2\mathbf{I}_{t_2} & \cdots & \mathbf{0}_{t_n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{t_1} & \mathbf{0}_{t_2} & \cdots & \mathbf{X}_n\boldsymbol{\Sigma}_\beta\mathbf{X}_n' + \sigma^2\mathbf{I}_{t_n} \end{bmatrix}$$

Statistical software packages can be used to estimate the elements of $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}_\beta$ and the error variance $\sigma^2$. These can be used when the number of measurements per unit differ, where unit $i$ has $t_i$ observations. In this case, the overall sample size is $N = \sum_{i=1}^{n} t_i$. The goal is to minimize the following function (this is another application of Generalized Least Squares).

$$(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' (V\{\mathbf{Y}\})^{-1} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

Estimates of the unknown parameters in $V\{\mathbf{Y}\}$ must be obtained. Two methods are maximum likelihood (ML) and restricted maximum likelihood (REML). It has been shown that REML provides less biased estimates of the elements of $V\{\mathbf{Y}\}$, which we will denote as $\boldsymbol{\theta}$. When comparing various models in terms of the fixed effects (elements of $\boldsymbol{\beta}$ in this case), ML should be used. The methods maximize the normal log-likelihood. Maximum likelihood does it "directly," REML transforms $\mathbf{Y}$ to have mean zero, then maximizes the likelihood for the "transformed" response vector.

$$\text{ML: } \ln\lambda(\theta) = -\frac{1}{2}\left[\ln|V\{\mathbf{Y}\}| + N\ln\left(\boldsymbol{\varepsilon}'(V\{\mathbf{Y}\})^{-1}\boldsymbol{\varepsilon}\right)\right]$$

$$\text{REML: } \ln\lambda_R(\theta) = -\frac{1}{2}\left[\ln|V\{\mathbf{Y}\}| + \ln\left|\mathbf{X}'(V\{\mathbf{Y}\})^{-1}\mathbf{X}\right| + (N-p')\ln\left(\boldsymbol{\varepsilon}'(V\{\mathbf{Y}\})^{-1}\boldsymbol{\varepsilon}\right)\right]$$

$$\text{where: } \boldsymbol{\varepsilon} = \left(\mathbf{I} - \mathbf{X}\left(\mathbf{X}'(V\{\mathbf{Y}\})^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'(V\{\mathbf{Y}\})^{-1}\right)\mathbf{Y}$$

The two resulting estimators are labeled as $\hat{\boldsymbol{\theta}}_{\text{ML}}$ and $\hat{\boldsymbol{\theta}}_{\text{RML}}$, with estimated variances for $\mathbf{Y}$ being $\hat{V}_{\text{ML}}\{\mathbf{Y}\}$ and $\hat{V}_{\text{RML}}\{\mathbf{Y}\}$. These lead to two estimators for $\boldsymbol{\beta}$ and their corresponding estimated variance-covariance matrices.

$$\hat{\boldsymbol{\beta}}_{\text{ML}} = \left(\mathbf{X}'\left(\hat{V}_{\text{ML}}\{\mathbf{Y}\}\right)^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\left(\hat{V}_{\text{ML}}\{\mathbf{Y}\}\right)^{-1}\mathbf{Y} \qquad \hat{V}\left\{\hat{\boldsymbol{\beta}}_{\text{ML}}\right\} = \left(\mathbf{X}'\left(\hat{V}_{\text{ML}}\{\mathbf{Y}\}\right)^{-1}\mathbf{X}\right)^{-1}$$

$$\hat{\boldsymbol{\beta}}_{\text{RML}} = \left(\mathbf{X}'\left(\hat{V}_{\text{RML}}\{\mathbf{Y}\}\right)^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\left(\hat{V}_{\text{RML}}\{\mathbf{Y}\}\right)^{-1}\mathbf{Y} \qquad \hat{V}\left\{\hat{\boldsymbol{\beta}}_{\text{RML}}\right\} = \left(\mathbf{X}'\left(\hat{V}_{\text{RML}}\{\mathbf{Y}\}\right)^{-1}\mathbf{X}\right)^{-1}$$

**Example: Women's NBA Players' Points per Game**

In the 2014 Women's National Basketball Association (WNBA) season, there were 117 players who average 10 or more minutes per game they played in. The regular season, there are 34 games (due to injuries, not all players play every game). Games are made up of four 10 minute quarters, some games have overtime. In this example, we take a random sample of $n = 20$ players, and fit a regression relating the player's points ($Y_{it}$) in an individual game to: minutes played ($X_{it1}$), an indicator of whether the game was a home game ($X_{it2}$), and the opponents (centered) average points per game ($X_{it3}$).

$$Y_{ij} = \beta_{i0} + \beta_{i1}X_{ij1} + \beta_{i2}X_{ij2} + \beta_{i3}X_{ij3} + \epsilon_{ij} \quad =$$

$$\beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij3} + (\beta_{i0} - \beta_0) + (\beta_{i1} - \beta_1) X_{ij1} + (\beta_{i2} - \beta_2) X_{ij2} + (\beta_{i3} - \beta_3) X_{ij3} + \epsilon_{ij}$$

A plot of the individual players' points versus minutes with simple linear regression lines in given in Figure 8.3. The R program using the **lmer** function within the **lme4** package is given below along with the output. The program also includes the **lme** function within the **nlme** package, that output is not included. The random effects $\left(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}\right)$ and the random coefficients $\left(\hat{\boldsymbol{\beta}}_i\right)$ are given in Table 8.3.

```
### Program
wnba1 <- read.csv("http://www.stat.ufl.edu/~winner/data/wnba2014a1.csv",
   header=T)
attach(wnba1); names(wnba1)
set.seed(54321)
max.id <- max(id2min)
id2min <- factor(id2min)
wnba1[1:300,]
wnba.samp <- wnba1[is.element(id2min,sample(levels(id2min),20)),]
detach(wnba1); attach(wnba.samp)
wnba.samp


library(nlme)
wnba.rcr1 <- lme(points ~ minutes + c_opp_av + home,
random = ~  minutes|player_id,
control=lmeControl(opt="optim"))
summary(wnba.rcr1)
random.effects(wnba.rcr1)

library(lme4)
library(lattice)
require(optimx)
```

```
xyplot(points ~ minutes | Player,
        panel = function(x,y) {
        panel.xyplot(x,y,pch=16)
        panel.abline(lm(y~x))
})


wnba.rcr2 <- lmer(points~ minutes + c_opp_av + home +
(minutes + home + c_opp_av|player_id),REML=TRUE,
control=lmerControl(opt="optimx",optCtrl  = list(method="nlminb")))

summary(wnba.rcr2)
logLik(wnba.rcr2)

random.effects(wnba.rcr2)
coef(wnba.rcr2)

### Output
> summary(wnba.rcr2)
Linear mixed model fit by REML ['lmerMod']
Formula: points ~ minutes + c_opp_av + home + (minutes + home + c_opp_av |      player_id)
Control: lmerControl(opt = "optimx", optCtrl = list(method = "nlminb"))

REML criterion at convergence: 3685.8

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.5791 -0.6337 -0.0404  0.5914  4.8761

Random effects:
 Groups    Name        Variance Std.Dev. Corr
 player_id (Intercept)  2.61510 1.6171
           minutes      0.01789 0.1337   -0.98
           home         1.06872 1.0338   -0.58  0.73
           c_opp_av     0.01635 0.1279   -0.54  0.36 -0.37
 Residual              18.30500 4.2784
Number of obs: 631, groups:  player_id, 20

Fixed effects:
            Estimate Std. Error t value
(Intercept) -3.25295    0.72662  -4.477
minutes      0.49971    0.03900  12.814
c_opp_av    -0.06108    0.05285  -1.156
home         0.57415    0.41355   1.388

Correlation of Fixed Effects:
         (Intr) minuts c_pp_v
minutes  -0.883
c_opp_av -0.164  0.151
home     -0.355  0.317 -0.095
> logLik(wnba.rcr2)
'log Lik.' -1842.915 (df=15)
```

The (REML) estimates variance-covariance matrix for $\boldsymbol{\beta}_i$ is obtained from the variances, standard deviations, and correlations of the estimated random effects.

$$\hat{V}\{\beta_j\} = \hat{\sigma}^2_{\beta_j} \qquad \hat{\{}\mathrm{COV}\}\{\beta_j, \beta_{j'}\} = \hat{\sigma}_{\beta_j,\beta_{j'}} = \hat{\rho}\hat{\sigma}_j\hat{\sigma}_{j'}$$

| | Random Effects | | | | Random Coefficients | | | |
|---|---|---|---|---|---|---|---|---|
| Player ID | $\hat{\beta}_{i0} - \hat{\beta}_0$ | $\hat{\beta}_{i1} - \hat{\beta}_1$ | $\hat{\beta}_{i2} - \hat{\beta}_2$ | $\hat{\beta}_{i3} - \hat{\beta}_3$ | $\hat{\beta}_{i0}$ | $\hat{\beta}_{i1}$ | $\hat{\beta}_{i2}$ | $\hat{\beta}_{i3}$ |
| 4 | -2.1804 | 0.2096 | 1.8547 | -0.0405 | -5.4334 | 0.7093 | -0.1016 | 2.4288 |
| 5 | -2.9762 | 0.2138 | 0.2374 | 0.2387 | -6.2292 | 0.7135 | 0.1776 | 0.8116 |
| 21 | -1.3147 | 0.1430 | 1.6428 | -0.0916 | -4.5677 | 0.6427 | -0.1527 | 2.2170 |
| 22 | 0.3516 | -0.0301 | -0.1817 | -0.0085 | -2.9014 | 0.4696 | -0.0696 | 0.3925 |
| 30 | 0.5435 | -0.0356 | 0.0648 | -0.0574 | -2.7095 | 0.4641 | -0.1185 | 0.6389 |
| 32 | 0.3003 | -0.0383 | -0.5532 | 0.0437 | -2.9526 | 0.4614 | -0.0174 | 0.0210 |
| 33 | -1.3558 | 0.1130 | 0.6018 | 0.0455 | -4.6088 | 0.6127 | -0.0156 | 1.1760 |
| 34 | -2.3806 | 0.2194 | 1.7234 | -0.0056 | -5.6335 | 0.7191 | -0.0666 | 2.2975 |
| 41 | 0.3893 | -0.0330 | -0.1914 | -0.0107 | -2.8636 | 0.4667 | -0.0718 | 0.3827 |
| 42 | 0.7952 | -0.0694 | -0.4514 | -0.0141 | -2.4578 | 0.4303 | -0.0751 | 0.1227 |
| 44 | -0.0580 | 0.0073 | 0.1038 | -0.0081 | -3.3110 | 0.5070 | -0.0691 | 0.6779 |
| 47 | 0.0862 | -0.0244 | -0.5847 | 0.0671 | -3.1668 | 0.4753 | 0.0061 | -0.0105 |
| 49 | 0.4641 | -0.0717 | -1.2531 | 0.1186 | -2.7888 | 0.4280 | 0.0575 | -0.6790 |
| 51 | 2.0034 | -0.1547 | -0.5020 | -0.1168 | -1.2496 | 0.3450 | -0.1779 | 0.0722 |
| 72 | 0.5627 | -0.0412 | -0.0706 | -0.0418 | -2.6902 | 0.4585 | -0.1029 | 0.5036 |
| 73 | 1.6616 | -0.1553 | -1.2722 | 0.0128 | -1.5914 | 0.3444 | -0.0483 | -0.6981 |
| 85 | 1.2298 | -0.0877 | -0.0785 | -0.1012 | -2.0232 | 0.4120 | -0.1622 | 0.4956 |
| 95 | 0.7575 | -0.0699 | -0.5502 | 0.0020 | -2.4954 | 0.4298 | -0.0591 | 0.0240 |
| 102 | 1.9650 | -0.1701 | -1.0730 | -0.0402 | -1.2879 | 0.3297 | -0.1013 | -0.4989 |
| 108 | -0.8444 | 0.0753 | 0.5333 | 0.0080 | -4.0973 | 0.5751 | -0.0531 | 1.1074 |

Table 8.3: Random Effects and Random Coefficients for WNBA Sample

$$\hat{\Sigma}_{\boldsymbol{\beta}} = \begin{bmatrix} 2.61510 & -0.21188 & -0.96962 & -0.11169 \\ -0.21188 & 0.01789 & 0.10090 & 0.00616 \\ -0.96962 & 0.10090 & 1.06872 & -0.04892 \\ -0.11169 & 0.00616 & -0.04892 & 0.01635 \end{bmatrix}$$

$\nabla$

### 8.2.2 Tests Regarding Elements of $\Sigma_\beta$

To test whether variance components (variances of regression coefficients) are 0, fit the model with and without the random effect(s) for the component(s) of interest. Obtain the log-likelihood with and without the random effect(s) of interest, along with the degrees of freedom. In R, these are obtained using the **logLik()** function. A conservative test is conducted as follows.

$$H_0 : \sigma^2_{\beta_j} = \cdots \sigma^2_{\beta_k} = 0 \quad TS : X^2_{LR} = -2\left[\text{lnL}_R - \text{lnL}_C\right] \quad RR : X^2_{LR} \geq \chi^2_{\alpha, df_C - df_R} \quad P : P\left(\chi^2_{df_C - df_R} \geq X^2_{LR}\right)$$

In R, the **lmerTest** package takes the output of the **lmer** function, and conducts the Likelihood-Ratio test, one-at-a-time for the regression coefficients (not including the intercept). To test multiple coefficients

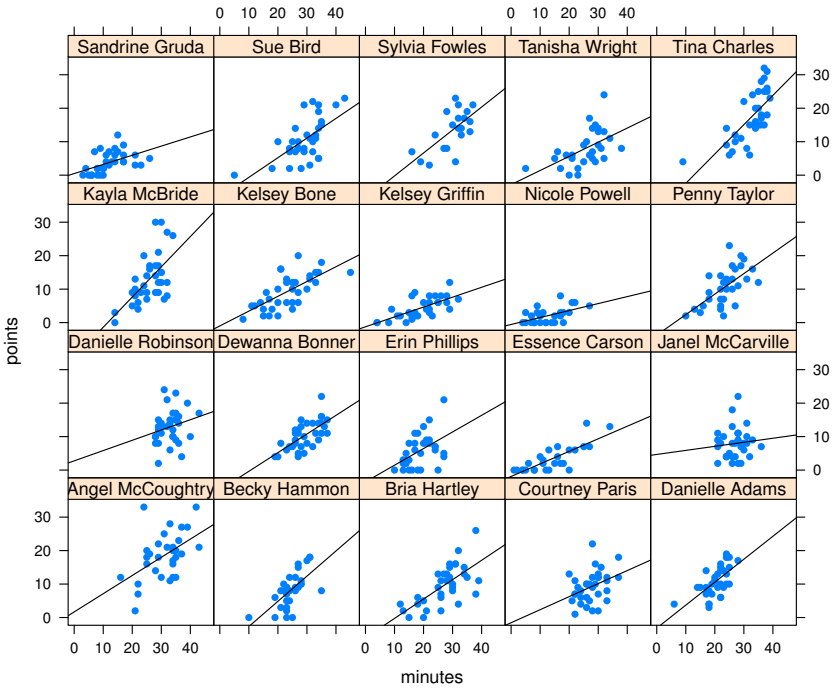


Figure 8.3: WNBA Data - Points versus Minutes with Simple Linear Regression

simultaneously (and/or the intercept), multiple models need to be fit, and their log-likelihoods can be compared using the test above.

## Example: Women's NBA Player's Points per Game

The following R program and output conducts Likelihood Ratio Tests regarding the variance components $\sigma^2_{\beta_j}$. Note that the variable the centered opponent average has been renamed **coppav**, as the **rand** command tries to treat c, opp, and av as 3 separate variables due to the underscore characters.

```
## Program
wnba1 <- read.csv("http://www.stat.ufl.edu/~winner/data/wnba2014a1.csv",
   header=T)
attach(wnba1); names(wnba1)
set.seed(54321)
max.id <- max(id2min)
id2min <- factor(id2min)
wnba1[1:300,]
wnba.samp <- wnba1[is.element(id2min,sample(levels(id2min),20)),]
detach(wnba1); attach(wnba.samp)

library(lmerTest)
library(lattice)
require(optimx)
coppav <- c_opp_av
```

```
wnba.rcr2 <- lmer(points~ minutes + coppav + home +
(1 + minutes + home + coppav|player_id),REML=TRUE,
control=lmerControl(opt="optimx",optCtrl  = list(method="nlminb")))
summary(wnba.rcr2)
(lnL2 <- logLik(wnba.rcr2))
rand(wnba.rcr2)

wnba.rcr3 <- lmer(points~ minutes + coppav + home +
(0 + minutes + home + coppav|player_id),REML=TRUE,
control=lmerControl(opt="optimx",optCtrl  = list(method="nlminb")))
summary(wnba.rcr3)
(lnL3 <- logLik(wnba.rcr3))
(X2.intercept <- -2*(lnL3 - lnL2))
(P.X2.intercept <- 1 - pchisq(X2.intercept,4))

wnba.rcr4 <- lmer(points~ minutes + coppav + home +
(0 + minutes|player_id),REML=TRUE,
control=lmerControl(opt="optimx",optCtrl  = list(method="nlminb")))
summary(wnba.rcr4)
(lnL4 <- logLik(wnba.rcr4))
(X2.inthmoppav <- -2*(lnL4 - lnL2))
(P.X2.inthmoppav <- 1 - pchisq(X2.inthmoppav,9))
rand(wnba.rcr4)

#### Output
> summary(wnba.rcr2)
Random effects:
 Groups    Name        Variance Std.Dev. Corr
 player_id (Intercept)  2.61510 1.6171
           minutes      0.01789 0.1337   -0.98
           home         1.06872 1.0338   -0.58  0.73
           coppav       0.01635 0.1279   -0.54  0.36 -0.37
 Residual              18.30500 4.2784
Number of obs: 631, groups:  player_id, 20

Fixed effects:
            Estimate Std. Error       df t value Pr(>|t|)
(Intercept) -3.25295    0.72662 23.55900  -4.477 0.000163 ***
minutes      0.49971    0.03900 18.44700  12.814 1.26e-10 ***
coppav      -0.06108    0.05285 21.05100  -1.156 0.260747
home         0.57415    0.41355 28.85100   1.388 0.175661
> (lnL2 <- logLik(wnba.rcr2))
'log Lik.' -1842.915 (df=15)
> rand(wnba.rcr2)
Analysis of Random effects Table:
                 Chi.sq Chi.DF p.value
minutes:player_id  17.11      4   0.002 **
home:player_id      6.93      4   0.140
coppav:player_id    2.30      4   0.682
>

> summary(wnba.rcr3)
Random effects:
 Groups    Name    Variance  Std.Dev. Corr
 player_id minutes  0.005715 0.0756
           home     0.981320 0.9906    0.69
           coppav   0.017320 0.1316    0.29 -0.49
 Residual          18.441315 4.2943
Number of obs: 631, groups:  player_id, 20

Fixed effects:
             Estimate Std. Error        df t value Pr(>|t|)
(Intercept)  -2.92394    0.66746 582.20000  -4.381  1.4e-05 ***
minutes       0.49310    0.03154  90.10000  15.633  < 2e-16 ***
```

```
coppav       -0.05986    0.05356  20.00000  -1.118    0.277
home          0.56485    0.40891  29.60000   1.381    0.177
> (lnL3 <- logLik(wnba.rcr3))
'log Lik.' -1844.804 (df=11)
>
> (X2.intercept <- -2*(lnL3 - lnL2))
'log Lik.' 3.777442 (df=11)
> (P.X2.intercept <- 1 - pchisq(X2.intercept,4))
'log Lik.' 0.4369629 (df=11)
>
> summary(wnba.rcr4)
Random effects:
 Groups    Name    Variance  Std.Dev.
 player_id minutes  0.008065 0.08981
 Residual          18.936503 4.35161
Number of obs: 631, groups:  player_id, 20

Fixed effects:
             Estimate Std. Error        df t value Pr(>|t|)
(Intercept)  -3.02526    0.66085 586.00000  -4.578 5.74e-06 ***
minutes       0.49626    0.03333  77.40000  14.889  < 2e-16 ***
coppav       -0.06675    0.04499 613.80000  -1.484    0.138
home          0.57061    0.34746 608.90000   1.642    0.101
> (lnL4 <- logLik(wnba.rcr4))
'log Lik.' -1848.065 (df=6)
>
> (X2.inthmoppav <- -2*(lnL4 - lnL2))
'log Lik.' 10.30022 (df=6)
> (P.X2.inthmoppav <- 1 - pchisq(X2.inthmoppav,9))
'log Lik.' 0.3267316 (df=6)
>
> rand(wnba.rcr4)
Analysis of Random effects Table:
                  Chi.sq Chi.DF p.value
minutes:player_id    108      1  <2e-16 ***
```

For the "one-at-a-time" tests, we have the following test statistics and $P$-values, controlling for all other variance components being in the model.

$$H_0^0 : \sigma^2_{\beta_0} = 0 \qquad TS : X^2_{obs} = 3.777 \qquad df = 4 \qquad P = P\left(\chi^2_4 \geq 3.777\right) = 0.4370$$

$$H_0^1 : \sigma^2_{\beta_1} = 0 \qquad TS : X^2_{obs} = 17.11 \qquad df = 4 \qquad P = P\left(\chi^2_4 \geq 17.11\right) = 0.002$$

$$H_0^2 : \sigma^2_{\beta_2} = 0 \qquad TS : X^2_{obs} = 6.93 \qquad df = 4 \qquad P = P\left(\chi^2_4 \geq 6.93\right) = 0.140$$

$$H_0^3 : \sigma^2_{\beta_3} = 0 \qquad TS : X^2_{obs} = 2.30 \qquad df = 4 \qquad P = P\left(\chi^2_4 \geq 2.30\right) = 0.682$$

It appears the only important random effect is the minutes played effect. We can test simultaneously that all other variance components are 0.

$$H_0^{023}: \sigma^2_{\beta_0} = \sigma^2_{\beta_2} = \sigma^2_{\beta_3} = 0 \qquad TS: X^2_{obs} = 10.30 \qquad df = 9 \qquad P = P\left(\chi^2_9 \geq 10.30\right) = 0.327$$

The variance component for minutes played is very significant ($X^2_{obs} = 108$, $df = 1$, $P \approx 0$). We now work with the following model.

$$Y_{ij} = \beta_0 + \beta_{i1}X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij3} + \epsilon_{ij} = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij3} + (\beta_{i1} - \beta_1) X_{ij1} + \epsilon_{ij}$$

$\nabla$

### 8.2.3 Tests Regarding $\boldsymbol{\beta}$

General Linear Tests regarding elements of $\boldsymbol{\beta}$ can be conducted as follow. There are two "classes" of tests. The first is conducted directly from the linear hypothesis to be tested, and can make use of either the ML or REML estimator and variance-covariance matrix. The second involves fitting a "Complete" and a "Reduced" model which imposes restrictions on the parameters; this method must be based on ML, due to the fact that the two models will have different sets of predictors. In each case, we are testing $H_0: \mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$, where $\mathbf{K}'$ has $q \leq p'$ linearly independent rows. The first version is a Wald test, the second is a likelihood-ratio test.

$$\text{Class 1: } X^2_{\text{ML}} = \left(\mathbf{K}'\hat{\boldsymbol{\beta}}_{\text{ML}} - m\right)' \left[\mathbf{K}'\hat{V}_{\text{ML}}\left\{\hat{\boldsymbol{\beta}}\right\}\mathbf{K}\right]^{-1} \left(\mathbf{K}'\hat{\boldsymbol{\beta}}_{\text{ML}} - m\right)$$

Under the null hypothesis, $X^2_{\text{ML}}$ and $X^2_{\text{RML}}$ are approximately chi-square with $q$ degrees of freedom. When fitting a Complete and Reduced model with respect to $\boldsymbol{\beta}$, we obtain $\hat{\boldsymbol{\theta}}^C_{\text{ML}}$ and $\hat{\boldsymbol{\theta}}^R_{\text{ML}}$ and evaluate -2 times the log-likelihood for each case.

$$\text{Class 2: } -2\left[\ln\lambda\left(\hat{\boldsymbol{\theta}}^R_{\text{ML}}\right) - \ln\lambda\left(\hat{\boldsymbol{\theta}}^C_{\text{ML}}\right)\right]$$

Under the null (Reduced) model, the test statistic is approximately chi-square with $q$ degrees of freedom.

**Example: Women's NBA Player's Points per Game**

Suppose we wish to test whether there is neither a home or opponent's average effect. That is, we are testing $H_0: \beta_2 = \beta_3 = 0$. Using the Wald test, with the REML estimate of the minutes played variance component, we have the following elements of the test.

$$\mathbf{K}' = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad \hat{\boldsymbol{\beta}} = \begin{bmatrix} -3.02526 \\ 0.49626 \\ -0.06675 \\ 0.57061 \end{bmatrix} \qquad \mathbf{K}'\hat{\boldsymbol{\beta}} = \begin{bmatrix} -0.06675 \\ 0.57061 \end{bmatrix}$$

$$\hat{V}_{\text{RML}}\left\{\hat{\boldsymbol{\beta}}\right\} = \begin{bmatrix} 0.43671784 & -0.01628390 & -0.00061881 & -0.06007691 \\ -0.01628390 & 0.00111084 & -0.00000494 & 0.00005909 \\ -0.00061881 & -0.00000494 & 0.00202397 & 0.00038255 \\ -0.06007691 & 0.00005909 & 0.00038255 & 0.12072720 \end{bmatrix}$$

$$\mathbf{K}'\hat{V}_{\text{RML}}\left\{\hat{\boldsymbol{\beta}}\right\}\mathbf{K} = \begin{bmatrix} 0.00202397 & 0.00038255 \\ 0.00038255 & 0.12072720 \end{bmatrix} \qquad \left[\mathbf{K}'\hat{V}_{\text{ML}}\left\{\hat{\boldsymbol{\beta}}\right\}\mathbf{K}\right]^{-1} = \begin{bmatrix} 494.3746 & -1.5665 \\ -1.5665 & 8.2881 \end{bmatrix}$$

$$\left(\mathbf{K}'\hat{\boldsymbol{\beta}}_{\text{ML}} - m\right)' \left[\mathbf{K}'\hat{V}_{\text{ML}}\left\{\hat{\boldsymbol{\beta}}\right\}\mathbf{K}\right]^{-1} \left(\mathbf{K}'\hat{\boldsymbol{\beta}}_{\text{ML}} - m\right) =$$

$$\begin{bmatrix} -0.06675 & 0.57061 \end{bmatrix} \begin{bmatrix} 494.3746 & -1.5665 \\ -1.5665 & 8.2881 \end{bmatrix} \begin{bmatrix} -0.06675 \\ 0.57061 \end{bmatrix} = 5.0206$$

$$\Rightarrow \qquad TS: X^2_{obs} = 5.0206 \qquad RR: X^2_{obs} =\geq \chi^2_{.05,2} = 5.991 \qquad P = P\left(\chi^2_2 \geq 5.0206\right) = .0812$$

The R program and output to compute both forms of the test are given below.

```
### Program
wnba.rcr4 <- lmer(points~ minutes + coppav + home +
(0 + minutes|player_id),REML=TRUE,
control=lmerControl(opt="optimx",optCtrl  = list(method="nlminb")))
beta <- fixef(wnba.rcr4)
v.beta <- vcov(wnba.rcr4)
Kp <- matrix(c(0,0,1,0,0,0,0,1),byrow=T,ncol=4)
(X2.obs.1 <-
(t(Kp %*% beta) %*% solve(Kp %*% v.beta %*% t(Kp)) %*% (Kp %*% beta)))
(P.X2.obs.1 <- 1 - pchisq(as.numeric(X2.obs.1),nrow(Kp)))

wnba.rcr5 <- lmer(points~ minutes + coppav + home +
(0 + minutes|player_id),REML=FALSE,
control=lmerControl(opt="optimx",optCtrl  = list(method="nlminb")))
logLik(wnba.rcr5)

wnba.rcr6 <- lmer(points~ minutes +
(0 + minutes|player_id),REML=FALSE,
control=lmerControl(opt="optimx",optCtrl  = list(method="nlminb")))
logLik(wnba.rcr6)

(X2.obs.2 <- -2*(logLik(wnba.rcr6)-logLik(wnba.rcr5)))
(P.X2.obs.2 <- 1 - pchisq(X2.obs.2,2))

### Output
> (X2.obs.1 <-
+ (t(Kp %*% beta) %*% solve(Kp %*% v.beta %*% t(Kp)) %*% (Kp %*% beta)))
1 x 1 Matrix of class "dgeMatrix"
         [,1]
[1,] 5.020417
> (P.X2.obs.1 <- 1 - pchisq(as.numeric(X2.obs.1),nrow(Kp)))
[1] 0.08125131

> logLik(wnba.rcr5)
'log Lik.' -1843.272 (df=6)
>
> logLik(wnba.rcr6)
'log Lik.' -1845.789 (df=4)
>
> (X2.obs.2 <- -2*(logLik(wnba.rcr6)-logLik(wnba.rcr5)))
```

```
'log Lik.' 5.033893 (df=4)
> (P.X2.obs.2 <- 1 - pchisq(X2.obs.2,2))
'log Lik.' 0.08070568 (df=4)
```

$\nabla$

### 8.2.4 Correlated Errors

Note that we have so far assumed that the $\epsilon^s$ within individuals are "conditionally" independent given $\mathbf{X}_i\boldsymbol{\beta}_i$. In practice, these measurements can be collected over time, and thus the errors may be autocorrelated. If the observations are made at equally spaced time points, we may use an AR(1) process on the errors within individuals.

$$\mathbf{R}_i = V\left\{\boldsymbol{\varepsilon}_i\right\} = \frac{\sigma^2}{1-\rho^2}\begin{bmatrix} 1 & \rho & \cdots & \rho^{t_i-1} \\ \rho & 1 & \cdots & \rho^{t_i-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{t_i-1} & \rho^{t_i-2} & \cdots & 1 \end{bmatrix}$$

$$V\left\{\mathbf{Y}\right\} = \begin{bmatrix} \mathbf{X}_1\boldsymbol{\Sigma}\mathbf{X}_1' + \mathbf{R}_1 & \mathbf{0}_{t_2} & \cdots & \mathbf{0}_{t_n} \\ \mathbf{0}_{t_1} & \mathbf{X}_2\boldsymbol{\Sigma}\mathbf{X}_2' + \mathbf{R}_2 & \cdots & \mathbf{0}_{t_n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{t_1} & \mathbf{0}_{t_2} & \cdots & \mathbf{X}_n\boldsymbol{\Sigma}\mathbf{X}_n' + \mathbf{R}_n \end{bmatrix}$$

This will change the variance-covariance matrix for $\hat{\boldsymbol{\beta}}$. Statistical software packages, such as SAS' Proc Mixed can handle these structures.

## 8.3 Nonlinear Models

The methods described for linear models can also be applied to nonlinear models. This topic is covered in detail in Davidian and Giltinan (1995). Without getting into the rather messy estimation methods, the form of the model is given here. As with the general linear case, statistical software procedures are needed to obtain the estimates. In general, there will be $n$ individuals, with the $i^{th}$ individual being observed on $t_i$ occasions.

$$Y_{ij} = g\left(\mathbf{x}_{ij}', \boldsymbol{\beta}_i\right) + \epsilon_{ij} \quad i = 1, \ldots, n; j = 1, \ldots, t_i \qquad \mathbf{x}_{ij}' = \begin{bmatrix} X_{ij1} & \cdots & X_{ijp} \end{bmatrix} \qquad \boldsymbol{\beta}_i = \begin{bmatrix} \beta_{i1} \\ \vdots \\ \beta_{ip} \end{bmatrix}$$

$$\mathbf{Y}_i = g\left(\mathbf{X}_i, \boldsymbol{\beta}_i\right) + \boldsymbol{\varepsilon}_i \quad i = 1, \ldots, n$$

$$\boldsymbol{\beta}_i \sim NID(\boldsymbol{\beta}, \boldsymbol{\Sigma}_\beta) \qquad \boldsymbol{\varepsilon}_i \sim NID\left(\mathbf{0}, \sigma^2\mathbf{I}\right) \qquad \{\boldsymbol{\beta}_i\} \perp \{\boldsymbol{\varepsilon}_i\}$$

**Example: Cumulative Revenues for 8 Harry Potter Films**

There were 8 Harry Potter films released between 2001 and 2011 in the United States. These films were wildly popular with huge initial box office success with exponential decay in terms of revenues per week. We consider an asymptotic model for cumulative domestic revenue by week. The model is fit through the origin.

$$Y_{it} = \beta_{i1}\left[1 - \exp\{-\exp\{(\beta_{i2})\, t\}\}\right] = \left[\beta_1 + (\beta_{i1} - \beta_1)\right]\left[1 - \exp\{-\exp\{(\beta_2 + (\beta_{i2} - \beta_2))\, t\}\}\right]$$

R has a built-in function **SSasympOrig** that uses a self starting algorithm for starting values of parameters. We make use of the **nlme** package and function to fit the model for the $n = 8$ films. Note that the data must be "grouped" within the nlme package. In this case, we group the data by their film number. The data are given in Table 8.4 and plotted in Figure 8.4. We consider two models: the first assumes that $\beta_{i1}$ and $\beta_{i2}$ are independent, the second does not. The R program and output are given below.

```
### Program
hp <- read.csv("E:\\coursenotes\\harrypotter.csv",header=T)
hp$film <- factor(hp$film)
hp$revperday <- hp$revperday/1000
hp$cumerev <- hp$cumerev/1000000
attach(hp); names(hp)

library(nlme)
library(lattice)
hp.grp <- groupedData(cumerev ~ weeknum | film, data=hp)
plot(hp.grp, aspect=2)

hp01.lis <- nlsList(cumerev ~ SSasympOrig(weeknum,beta1,beta2) | film,
     data = hp.grp)
hp01.lis
plot(intervals(hp01.lis))

hp01.nlme <- nlme(hp01.lis, random=pdDiag(beta1 + beta2 ~ 1))
hp01.nlme

hp03.nlme <- update(hp01.nlme,random = beta1 + beta2 ~ 1)
hp03.nlme

(ll01 <- logLik(hp01.nlme))
(ll03 <- logLik(hp03.nlme))
(LRT013 <- -2*(ll01 - ll03))
(P013 <- 1 - pchisq(LRT013,1))

plot(hp03.nlme, id=.05, adj = -1)
qqnorm(hp03.nlme)
plot(augPred(hp03.nlme, level=0:1))
summary(hp03.nlme)

#### Output
> hp01.lis <- nlsList(cumerev ~ SSasympOrig(weeknum,beta1,beta2) | film,
```

```
+      data = hp.grp)
> hp01.lis
Call:
  Model: cumerev ~ SSasympOrig(weeknum, beta1, beta2) | film
   Data: hp.grp

Coefficients:
     beta1      beta2
3 246.9838 -0.5222464
2 259.9362 -0.7314847
4 285.5305 -0.4913518
5 288.8013 -0.2321239
7 290.3397 -0.2863040
6 298.6841 -0.1390780
1 314.4891 -0.8694231
8 376.8952 -0.2204142

Degrees of freedom: 175 total; 159 residual
Residual standard error: 6.584572

> hp01.nlme
Nonlinear mixed-effects model fit by maximum likelihood
  Model: cumerev ~ SSasympOrig(weeknum, beta1, beta2)
  Data: hp.grp
  Log-likelihood: -618.3634
  Fixed: list(beta1 ~ 1, beta2 ~ 1)
      beta1       beta2
295.2046097  -0.4376524

Random effects:
 Formula: list(beta1 ~ 1, beta2 ~ 1)
 Level: film
 Structure: Diagonal
           beta1     beta2 Residual
StdDev: 36.71555 0.2429555   6.5842

Number of Observations: 175
Number of Groups: 8

> hp03.nlme
Nonlinear mixed-effects model fit by maximum likelihood
  Model: cumerev ~ SSasympOrig(weeknum, beta1, beta2)
  Data: hp.grp
  Log-likelihood: -617.917
  Fixed: list(beta1 ~ 1, beta2 ~ 1)
      beta1       beta2
295.2086443  -0.4381722

Random effects:
 Formula: list(beta1 ~ 1, beta2 ~ 1)
 Level: film
 Structure: General positive-definite, Log-Cholesky parametrization
         StdDev    Corr
beta1    36.676401 beta1
beta2     0.242754 0.331
Residual  6.584466

Number of Observations: 175
Number of Groups: 8
>
> (ll01 <- logLik(hp01.nlme))
'log Lik.' -618.3634 (df=5)
> (ll03 <- logLik(hp03.nlme))
'log Lik.' -617.917 (df=6)
```

```
> (LRT013 <- -2*(ll01 - ll03))
'log Lik.' 0.892802 (df=5)
> (P013 <- 1 - pchisq(LRT013,1))
'log Lik.' 0.3447191 (df=5)
>
> plot(hp03.nlme, id=.05, adj = -1)
> plot(augPred(hp03.nlme, level=0:1))
> summary(hp03.nlme)
Nonlinear mixed-effects model fit by maximum likelihood
  Model: cumerev ~ SSasympOrig(weeknum, beta1, beta2)
 Data: hp.grp
       AIC      BIC   logLik
  1247.834 1266.823 -617.917

Random effects:
 Formula: list(beta1 ~ 1, beta2 ~ 1)
 Level: film
 Structure: General positive-definite, Log-Cholesky parametrization
         StdDev    Corr
beta1    36.676401 beta1
beta2     0.242754 0.331
Residual  6.584466

Fixed effects: list(beta1 ~ 1, beta2 ~ 1)
          Value Std.Error  DF  t-value p-value
beta1 295.20864 13.054724 166 22.61317       0
beta2  -0.43817  0.087527 166 -5.00615       0
 Correlation:
      beta1
beta2 0.323

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-2.9453481 -0.2079161  0.2300135  0.4534575  3.1874196

Number of Observations: 175
Number of Groups: 8
```

The Likelihood Ratio Test does not reject the null hypothesis that $\text{COV}(\beta_{i1}, \beta_{i2}) = 0$, however we use the more complex model for plots and the final summary. The estimates for the fixed effects and the variance components for the random effects are as follow. The fixed and random fitted curve are given in Figure 8.5.

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 295.20864 \\ -0.43817 \end{bmatrix} \quad \hat{\Sigma}_{\boldsymbol{\beta}} = \begin{bmatrix} 36.6764^2 = 1345.16 & 36.6764(0.2428)(0.331) = 2.95 \\ 2.95 & 0.2428^2 = 0.0589 \end{bmatrix} \quad \hat{V}\{\epsilon\} = 6.5845^2 = 43.36$$

$\nabla$

| Week | HP1 | HP2 | HP3 | HP4 | HP5 | HP6 | HP7 | HP8 |
|---|---|---|---|---|---|---|---|---|
| 1 | 129.5 | 106.1 | 123.1 | 146.3 | 175.4 | 191.8 | 170.0 | 226.1 |
| 2 | 196.0 | 168.0 | 172.9 | 209.4 | 224.7 | 237.8 | 227.5 | 296.5 |
| 3 | 224.9 | 203.9 | 200.3 | 233.8 | 251.5 | 264.9 | 249.2 | 330.6 |
| 4 | 243.3 | 216.3 | 217.2 | 246.6 | 266.6 | 278.7 | 260.7 | 350.1 |
| 5 | 257.0 | 224.5 | 228.7 | 256.7 | 275.1 | 286.8 | 269.8 | 361.5 |
| 6 | 274.7 | 233.8 | 235.6 | 269.4 | 280.8 | 291.8 | 278.9 | 368.2 |
| 7 | 294.5 | 247.4 | 240.1 | 278.7 | 284.3 | 295.1 | 285.3 | 372.2 |
| 8 | 301.6 | 253.0 | 242.9 | 282.2 | 287.2 | 297.9 | 288.4 | 376.1 |
| 9 | 306.2 | 256.0 | 244.6 | 284.6 | 288.5 | 299.1 | 290.4 | 377.5 |
| 10 | 310.2 | 258.1 | 245.8 | 285.8 | 289.5 | 299.7 | 291.6 | 378.4 |
| 11 | 311.9 | 259.1 | 246.5 | 286.4 | 289.9 | 300.1 | 292.4 | 379.0 |
| 12 | 313.2 | 259.7 | 247.1 | 286.9 | 290.2 | 300.3 | 292.9 | 379.3 |
| 13 | 313.9 | 260.1 | 247.8 | 287.3 | 290.8 | 300.9 | 293.5 | 380.0 |
| 14 | 314.4 | 260.6 | 248.4 | 288.2 | 291.1 | 301.2 | 294.1 | 380.4 |
| 15 | 314.7 | 260.9 | 248.7 | 288.9 | 291.4 | 301.5 | 294.4 | 380.6 |
| 16 | 315.1 | 261.2 | 248.9 | 289.3 | 291.6 | 301.6 | 294.6 | 380.8 |
| 17 | 315.4 | 261.5 | 249.1 | 289.5 | 291.7 | 301.7 | 294.8 | 380.9 |
| 18 | 315.7 | 261.6 | 249.2 | 289.8 | 291.8 | 301.8 | 294.9 | 381.0 |
| 19 | 316.2 | 261.8 | 249.2 | 289.9 | 291.9 | 301.9 | 294.9 | 381.0 |
| 20 | 316.6 | 261.9 | 249.3 | 290.0 | 292.0 | 301.9 | 295.0 | |
| 21 | 316.9 | 261.9 | 249.3 | | 292.0 | 302.0 | | |
| 22 | 317.1 | 262.0 | 249.4 | | 292.0 | 302.0 | | |
| 23 | 317.3 | 262.0 | | | | | | |
| 24 | 317.4 | 262.0 | | | | | | |
| 25 | 317.5 | | | | | | | |
| 26 | 317.5 | | | | | | | |

Table 8.4: Harry Potter Films' Cumulative U.S. Revenues

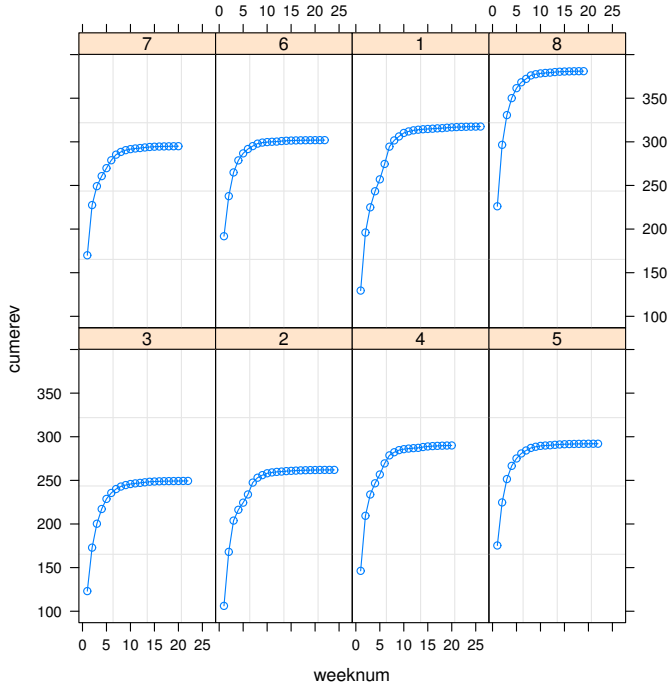


Figure 8.4: Cumulative U.S. Revenues by Week - Harry Potter Films

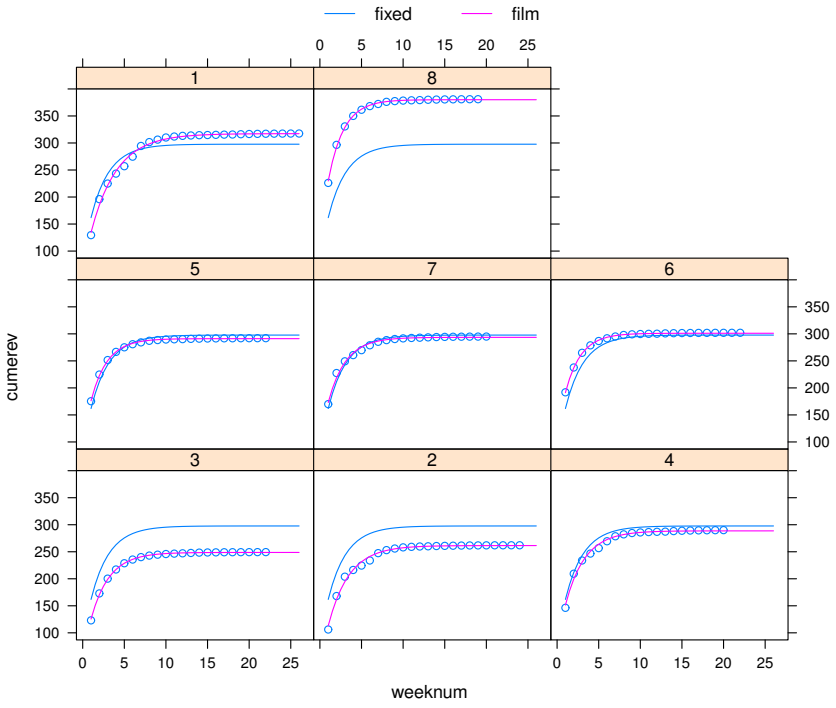


Figure 8.5: Fixed and Random Fitted Curves for Harry Potter Revenues

# Chapter 9

# Alternative Regression Models

## 9.1 Introduction

The normal theory based Regression model is based on the assumptions that the error terms are independent, Normally distributed with mean 0, and variance $\sigma^2$. We have considered those models in detail, and checked assumptions, as well as considered transformations to remedy any problems, and Estimated Generalized Least Squares when data are not independent.

Alternative types regression models are also used in specific situations, including when data arise from a non-normal family of distributions (e.g. Binomial, Poisson, Negative Binomial, Gamma, and Beta). While some of these distributions are part of exponential families and are analyzed as **Generalized Linear Models**, (see e.g. McCullaugh and Nelder (1989), and Agresti (2002)), we will consider each family separately, and work through the computations directly for ML estimation. When appropriate (Binomial, Poisson, and Gamma), we make use of software developed for Generalized Linear Models. Other regression models can be used to model quantiles (not just the mean) for data with skewed distributions.

In all cases, we will have a single response variable, and $p$ predictors, as in the case of linear regression. These predictors can include dummy variables for categorical predictors, quadratic terms, and cross-product terms to model interactions.

## 9.2 Binary Responses - Logistic Regression

Suppose we have $m$ distinct levels of the independent variable(s) with $n_i$ trials at the $i^{th}$ level and $y_i$ "successes" at that level. The goal is to model the probability of success, $\pi$, as a function of the predictors. A problem arises in that probabilities are required to lie between 0 and 1, while if we treat $\pi = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$, then there is no restriction. One possible (of several) **link functions** is the **logit link**.

$$g(\pi) = \ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p = \mathbf{x}'\boldsymbol{\beta} \quad \Rightarrow \quad \pi(\mathbf{x}) = \frac{\exp\{\mathbf{x}'\boldsymbol{\beta}\}}{1+\exp\{\mathbf{x}'\boldsymbol{\beta}\}}$$

Define the $\mathbf{X}$ matrix and $\boldsymbol{\beta}$ vector:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_m' \end{bmatrix} \qquad \mathbf{x}_i' = \begin{bmatrix} 1 & X_{i1} & \cdots & X_{ip} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}.$$

Set up the likelihood and log-likelihood functions, and take the derivatives of the log-likelihood.

$$L(\boldsymbol{\beta}|(n_1,y_1),(n_2,y_2),\ldots,(n_m,y_m)) = \prod_{i=1}^{m}\left(\frac{n_i!}{y_i!(n_i-y_i)!}\right)\left(\frac{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{1+\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}\right)^{y_i}\left(\frac{1}{1+\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}\right)^{n_i-y_i}$$

$$l = \ln(L) = \sum_{i=1}^{m}\left[\ln(n_i!) - \ln(y_i!) - \ln((n_i-y_i)!)\right] + \sum_{i=1}^{m} y_i\mathbf{x}_i'\boldsymbol{\beta} - \sum_{i=1}^{m} n_i \ln(1+\exp\{\mathbf{x}_i'\boldsymbol{\beta}\})$$

$$g_{\boldsymbol{\beta}} = \frac{\partial l}{\partial \beta} = \sum_{i=1}^{m} y_i\mathbf{x}_i - \sum_{i=1}^{m} n_i \frac{1}{1+\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}\mathbf{x}_i = \sum_{i=1}^{m} n_i(y_i - n_i\pi_i)\mathbf{x}_i = \mathbf{X}'(\mathbf{Y}-\mu)$$

$$\pi_i = \frac{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{1+\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}} \qquad \boldsymbol{\mu} = \begin{bmatrix} n_1\pi_1 \\ n_2\pi_2 \\ \vdots \\ n_m\pi_m \end{bmatrix}$$

$$G_{\boldsymbol{\beta}} = \frac{\partial^2 l}{\partial\beta\partial\beta'} = -\sum_{i=1}^{m} n_i\mathbf{x}_i \frac{[(1+\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}]\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}\mathbf{x}_i'] - [\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}\mathbf{x}_i']}{(1+\exp\{\mathbf{x}_i'\boldsymbol{\beta}\})^2} =$$

$$-\sum_{i=1}^{m} n_i\left[\frac{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{(1+\exp\{\mathbf{x}_i'\boldsymbol{\beta}\})^2}\right]\mathbf{x}_i\mathbf{x}_i' = -\sum_{i=1}^{m} n_i\pi_i(1-\pi_i)\mathbf{x}_i\mathbf{x}_i' = -\mathbf{X}'\mathbf{W}\mathbf{X} \qquad \mathbf{W} = \text{diag}[n_i\pi_i(1-\pi_i)]$$

The Newton-Raphson algorithm can be used to obtain Maximum Likelihood estimates of the elements of $\boldsymbol{\beta}$.

$$\tilde{\boldsymbol{\beta}}^{(k)} = \tilde{\boldsymbol{\beta}}^{(k-1)} - \left[G_{\tilde{\boldsymbol{\beta}}^{(k-1)})}\right]^{-1} g_{\tilde{\boldsymbol{\beta}}^{(k-1)}}$$

After setting staring values, iterate until convergence. A simple way to obtain starting values is as follows.

$$\tilde{\beta}_1^{(0)} = \cdots = \tilde{\beta}_p^{(0)} = 0 \qquad \tilde{\beta}_0^{(0)} = \ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) \qquad \hat{\pi} = \frac{\sum_{i=1}^m y_i}{\sum_{i=1}^m n_i}$$

The estimated variance of $\hat{\boldsymbol{\beta}}$ is $-E\{G_{\hat{\boldsymbol{\beta}}}\}$.

$$\hat{V}\left\{\hat{\boldsymbol{\beta}}\right\} = \left(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X}\right)^{-1}$$

In software packages, such as SAS, R, SPSS, and STATA, data can be entered at the individual (Bernoulli trial) level, or grouped by distinct combinations of $\mathbf{x}'$ levels (either numbers of successes and failures (e.g. R glm) or successes and trials (e.g. SAS Proc Genmod).

### 9.2.1 Interpreting the Slope Coefficients

In linear regression, the slope parameter(s) represent the change in the mean response when the independent variable(s) are increased by 1 unit, controlling for all other independent variables when there are multiple predictors. In the case of logistic regression, the slope parameters represent the change in the logit (aka log(odds)) when the independent variable(s) increase by 1 unit. This is not of primary interest to researchers. Consider the **odds** of a Success, which is defined as the number of Successes per 1 Failure: odds $= \pi/(1-\pi)$.

$$\pi(\mathbf{x}) = \frac{\exp\{\mathbf{x}'\boldsymbol{\beta}\}}{1+\exp\{\mathbf{x}'\boldsymbol{\beta}\}} \quad \Rightarrow \quad \text{odds}(\mathbf{x}) = \frac{\left[\frac{\exp\{\mathbf{x}'\boldsymbol{\beta}\}}{1+\exp\{\mathbf{x}'\boldsymbol{\beta}\}}\right]}{\left[\frac{1}{1+\exp\{\mathbf{x}'\boldsymbol{\beta}\}}\right]} = \exp\{\mathbf{x}'\boldsymbol{\beta}\} = \exp\{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\}$$

The **Odds Ratio** for an independent variable is the ratio of the odds when the independent variable is increased by 1 unit to the odds when it is held constant, while holding all other variables constant. There is an Odds Ratio for each predictor.

$$OR_j = \frac{\exp\{\beta_0 + \beta_1 X_1 + \cdots + \beta_j(X_j+1) + \cdots + \beta_p X_p\}}{\exp\{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\}} = \exp\{\beta_j\}$$

The odds of a Success change multiplicatively by $e^{\beta_j}$ when $X_j$ is increased by 1 unit. Note that if $\beta_j = 0$, the odds (and probability) of Success are the same for all levels of $X_j$, and we say that $X_j$ is not associated with the probability of Success, after controlling for all other predictors.

### 9.2.2 Inferences for the Regression Parameters

Tests for the regression coefficients are typically conducted as Likelihood Ratio and Wald tests. While being asymptotically equivalent, Likelihood Ratio tests tend to work better in small samples (Agresti (2002, p.12)).

Likelihood Ratio tests are based on the log-likelihood computed under the null hypothesis, $l_0 = \ln L_0$, and the log-likelihood under the alternative hypothesis (no restrictions) $l_A = \ln L_A$. The Likelihood Ratio test statistic, rejection region and $P$-value are given below, where $k$ is the number of parameter restrictions under the null hypothesis.

$$X^2_{LR} = -2\left[l_0 - l_A\right] \qquad RR : X^2_{LR} \geq \chi^2_{\alpha,k} \qquad P = P\left(\chi^2_k \geq X^2_{LR}\right)$$

In R, the **glm** function prints two deviance measures: **Null** and **Residual**. These represent the differences $D_0 = -2\left[l_0 - l_S\right]$ and $D_A = -2\left[l_A - l_S\right]$, where $l_S$ is the log-likelihood under the saturated model, where $\hat{\pi}_i = y_i/n_i$. It is computed as follows.

$$l_S = \sum_{i=1}^{m}\left[\ln\left(n_i!\right) - \ln\left(y_i!\right) - \ln\left((n_i - y_i)!\right)\right] + \sum_{i=1}^{m} y_i \ln\left(\frac{y_i}{n_i}\right) + \sum_{i=1}^{m}\left(n_i - y_i\right)\ln\left(\frac{n_i - y_i}{n_i}\right) \qquad \text{s.t. } 0\ln(0) = 0$$

The Likelihood Ratio test statistic for testing $H_0 : \beta_1 = \cdots = \beta_p = 0$ is the difference between the Null and Residual Deviances, with $p$ degrees of freedom.

Wald tests of the form $H_0 : \beta_j = \beta_{j0}$ are of the forms (chi-square and Z) given below.

$$X^2_W = \frac{\left(\hat{\beta}_j - \beta_{j0}\right)^2}{\hat{V}\left\{\hat{\beta}_j\right\}} \qquad RR : X^2_W \geq \chi^2_{\alpha,1} \qquad P = P\left(\chi^2_1 \geq X^2_W\right)$$

$$z_W = \frac{\hat{\beta}_j - \beta_{j0}}{\hat{SE}\left\{\hat{\beta}_j\right\}} \qquad RR : |z_W| \geq z_{\alpha/2} \qquad P = 2P\left(Z \geq |z_W|\right)$$

For general linear tests among the regression coefficients, $H_0 : \mathbf{K}'\boldsymbol{\beta} - \mathbf{m} = \mathbf{0}$, the Wald Test is conducted as a chi-square test, with $k$ restrictions (number of rows in $\mathbf{K}'$).

$$X^2_W = \left(\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m}\right)'\left[\mathbf{K}'\hat{V}\{\hat{\boldsymbol{\beta}}\}\mathbf{K}\right]^{-1}\left(\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m}\right) \qquad RR : X^2_W \geq \chi^2_{\alpha,k} \qquad P = P\left(\chi^2_k \geq X^2_W\right)$$

Confidence intervals for the regression coefficients and their corresponding odds ratios are obtained by using the approximate normality of the estimated regression coefficients, and then exponentiating the end points for the odds ratios.

$$(1-\alpha)100\%\text{ CI for }\beta_j: \quad \hat{\beta}_j \pm z_{\alpha/2}\hat{SE}\{\hat{\beta}_j\} = (\beta_{jL}, \beta_{jH}) \qquad (1-\alpha)100\%\text{ CI for } OR_j: \left(e^{\beta_{jL}}, e^{\beta_{jH}}\right)$$

### 9.2.3 Goodness of Fit Tests and Measures

There are two commonly reported goodness-of-fit tests for grouped (with respect to independent variable(s)) data. In each case, the null hypothesis is that the current model is appropriate.

The first test is based on the model's Deviance residuals and their corresponding chi-square statistic (Deviance), and compares it with the chi-square distribution with $m - p'$ degrees of freedom, where $p'$ is the number of parameters in the proposed model. Note that $X^2_D$ is the (Residual) Deviance for the current model.

$$\text{Deviance Residuals:} \quad e_{iD} = \text{sgn}\{y_i - n_i\hat{\pi}_i\}\left[2\left[y_i \ln\left(\frac{y_i}{n_i\hat{\pi}_i}\right) + (n_i - y_i)\ln\left(\frac{n_i - y_i}{n_i(1-\hat{\pi}_i)}\right)\right]\right]^{1/2} \quad \text{s.t.} \quad 0\ln(0) = 0$$

$$X^2_D = \sum_{i=1}^{m} e^2_{iD}$$

The second method is based on Pearson residuals and their corresponding chi-square statistic, also with $m - p'$ degrees of freedom. For both tests, the approximate $P$-value is the area greater than equal to the chi-square statistic for the $\chi^2_{m-p'}$ distribution.

$$\text{Pearson Residuals:} \quad e_{iP} = \frac{y_i - \hat{y}_i}{\hat{SE}\{y_i\}} = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}} \qquad X^2_P = \sum_{i=1}^{m} e^2_{iP}$$

When the independent variable(s) are continuous, and do not have exact groupings, the data can be collapsed into $g$ groups based on their predicted values (see e.g. Hosmer and Lemeshow (1989), Section 5.2.2). The test makes use of the numbers of observed successes in each derived group ($o_i$), the size of the group ($n_i$), and the average predicted probability for the group ($\overline{\hat{\pi}}_i$). They suggest using $g = 10$ derived groups, however this clearly will depend on the number of individual cases in the dataset. Under the null hypothesis, the statistic is approximately chi-square with $g - 2$ degrees of freedom.

$$X^2_{HL} = \sum_{i=1}^{g} \frac{\left(o_i - n_i\overline{\hat{\pi}}_i\right)^2}{n_i\overline{\hat{\pi}}_i\left(1-\overline{\hat{\pi}}_i\right)}$$

There are various measures that attempt to replicate $R^2$ for linear models, comparing the fit of a particular model to the null model. One possibility is to measure the deviance explained by the model. A second measure is due to Nagelkerke and is considered a better statistic (see e.g. Faraway (2006, p. 41)).

$$R^2_{\text{Dev}} = 1 - \frac{D_A}{D_0} \qquad R^2_{\text{Nag}} = \frac{1 - \exp\{(D_A - D_0)/n\}}{1 - \exp\{-D_0/n\}}$$

where $n$ is the number of individual Bernoulli trials.

We will analyze the data directly making use of the matrix form in R, then make use of the glm function, first grouped, then individually.

**Example: National Football League Field Goal Attempts - 2008 Regular Season**

In the National Football League (NFL), teams can attempt field goals for 3 points. The kicker must kick the ball between the uprights for a successful kick. In this example, we aggregate all kicks for the 2008 regular season (240 total games, with $N = 1039$ attempts. Table 9.1 contains the numbers of attempts ($n_i$) and successes ($y_i$) for the $m = 45$ distances attempted. Note that a Generalized Linear Mixed Model could be fit that includes random kicker effects. A plot of the the sample proportions and fitted model is given in Figure 9.1.

The matrix form and the Newton-Raphson algorithm are given below in R.

```
### Program
fga2008 <- read.csv("http://www.stat.ufl.edu/~winner/data/nfl2008_fga.csv",
                    header=T)
attach(fga2008); names(fga2008)

mindist <- min(distance); maxdist <- max(distance)
n.dist <- numeric(maxdist-mindist+1)
y.dist <- numeric(maxdist-mindist+1)
pi_hat.dist <- numeric(maxdist-mindist+1)
fga.dist <- numeric(maxdist-mindist+1)
cnt.dist <- 0

for (i in mindist:maxdist) {
cnt.dist <- cnt.dist+1
n.dist[cnt.dist] <- length(distance[distance==i])
y.dist[cnt.dist] <- sum(GOOD[distance==i])
fga.dist[cnt.dist] <- i
if (n.dist[cnt.dist] == 0) pi_hat.dist[cnt.dist] <- NA
else pi_hat.dist[cnt.dist] <- y.dist[cnt.dist] / n.dist[cnt.dist]
}

################ Direct Computations

m.dist <- length(n.dist[n.dist > 0])
X0 <- rep(1,m.dist)
X <- cbind(X0,fga.dist[n.dist > 0])
Y <- y.dist[n.dist > 0]
n <- n.dist[n.dist > 0]

(pi.hat.all <- sum(y.dist[n.dist > 0]) / sum(n.dist[n.dist > 0]))

beta.old <- matrix(c(log(pi.hat.all/(1-pi.hat.all)),0),ncol=1)
beta.diff <- 1000
num.iter <- 0

while (beta.diff > 0.00001) {
num.iter <- num.iter + 1
pi.hat <- exp(X %*% beta.old) / (1 + exp(X %*% beta.old))
w.v <-  n.dist[n.dist > 0] * pi.hat * (1-pi.hat)
W.M <- matrix(rep(0,m.dist^2),ncol=m.dist)
```

```
for (i in 1:m.dist) W.M[i,i] <- w.v[i,1]
mu.Y <- n.dist[n.dist > 0] * pi.hat
g.beta <- t(X) %*% (Y - mu.Y)
G.beta <- -t(X) %*% W.M %*% X
beta.new <- beta.old - solve(G.beta) %*% g.beta
beta.diff <- sum((beta.new - beta.old)^2)
beta.old <- beta.new
print(beta.old)
}

num.iter
V.beta <- -solve(G.beta)
SE.beta <- sqrt(diag(V.beta))
z.beta <- beta.old / SE.beta
pv.beta <- 2*(1-pnorm(abs(z.beta),0,1))

beta.est <- cbind(beta.old,SE.beta,z.beta,pv.beta,
beta.old - qnorm(.975)*SE.beta,
beta.old + qnorm(.975)*SE.beta)
colnames(beta.est) <- c("Estimate", "Std. Error", "z", "Pr(>|z|)",
  "LL","UL")
rownames(beta.est) <- c("Intercept","Distance")
beta.est

pi.hat <- exp(X %*% beta.old) / (1 + exp(X %*% beta.old))
l.0 <- rep(0,m.dist); l.A <- rep(0,m.dist); l.S <- rep(0,m.dist)
pearson.r <- rep(0,m.dist)

for (i in 1:m.dist) {
l.0[i] <- Y[i] * log(pi.hat.all) + (n[i] - Y[i]) * log(1-pi.hat.all)
l.A[i] <- Y[i] * log(pi.hat[i]) + (n[i] - Y[i]) * log(1-pi.hat[i])
if (Y[i] == 0) l.S[i] <- 0
else if (Y[i] == n[i]) l.S[i] <- 0
else l.S[i] <- Y[i] * log(Y[i]/n[i]) + (n[i] - Y[i]) * log(1-Y[i]/n[i])
pearson.r[i] <- (Y[i] - n[i]*pi.hat[i]) / sqrt(n[i]*pi.hat[i]*(1-pi.hat[i]))
}

D.0 <- -2*(sum(l.0) - sum(l.S))
D.A <- -2*(sum(l.A) - sum(l.S))

X2.LR <- D.0 - D.A
p.X2.LR <- 1 - pchisq(X2.LR,1)

LR.test <- cbind(D.0,D.A,X2.LR,p.X2.LR)
colnames(LR.test) <- c("Null Dev","Resid Dev","LR X2 Stat","P(>X2)")
LR.test

X2.pearson <- sum(pearson.r^2)
p.X2.pearson <- 1-pchisq(X2.pearson,m.dist-2)
p.X2.deviance <- 1-pchisq(D.A,m.dist-2)

GOF.test <- cbind(X2.pearson,p.X2.pearson,D.A,p.X2.deviance)
colnames(GOF.test) <- c("Pearson X2","P(>X2)","Deviance X2","P(>X2)")
GOF.test

(R2.Dev <- 1 - D.A/D.0)
(R2.Nag <- (1-exp((D.A-D.0)/sum(n))) / (1-exp(-D.0/sum(n))))

### Output
> (pi.hat.all <- sum(y.dist[n.dist > 0]) / sum(n.dist[n.dist > 0]))
[1] 0.8662175

> beta.est
            Estimate Std. Error         z Pr(>|z|)         LL          UL
```

```
Intercept  6.7627078 0.54442506 12.421742        0  5.6956543  7.82976136
Distance  -0.1208357 0.01228512 -9.835941        0 -0.1449141 -0.09675729

> LR.test
     Null Dev Resid Dev LR X2 Stat P(>X2)
[1,] 170.9969  40.20124   130.7957      0

> GOF.test
     Pearson X2    P(>X2) Deviance X2   P(>X2)
[1,]   36.07857 0.7635178    40.20124 0.593378

> (R2.Dev <- 1 - D.A/D.0)
[1] 0.7649008
> (R2.Nag <- (1-exp((D.A-D.0)/sum(n))) / (1-exp(-D.0/sum(n))))
[1] 0.7794778
```

The fitted equation is given below, followed by the R program based on aggregated data (fga.10) and individual kick data (fga.1), based on the **glm** function.

$$\hat{\pi} = \frac{\exp(7.7627 - 0.1208X)}{1 + \exp(7.7627 - 0.1208X)} \qquad 95\%\text{CI for } \beta_1 : -0.1208 \pm 1.96(0.0123) \quad \equiv \quad (-0.1449, -0.0967)$$

$$\hat{OR}_i = e^{-0.1208} = 0.8862 \qquad 95\% \text{ CI for } OR_1 : \quad \left(e^{-0.1449}, e^{-0.0967}\right) \quad \equiv \quad (0.8561, 0.9078)$$

The odds of a successful field goal decreases multiplicatively by 0.8862 for each added yard of distance (95% CI: ((0.8561,0.9078)). This corresponds to a 100(0.8862-1) = -11.38% decrease per yard.

```
### Program
fga2008 <- read.csv("http://www.stat.ufl.edu/~winner/data/nfl2008_fga.csv",
                    header=T)
attach(fga2008); names(fga2008)

mindist <- min(distance); maxdist <- max(distance)
n.dist <- numeric(maxdist-mindist+1)
y.dist <- numeric(maxdist-mindist+1)
pi_hat.dist <- numeric(maxdist-mindist+1)
fga.dist <- numeric(maxdist-mindist+1)
cnt.dist <- 0

for (i in mindist:maxdist) {
cnt.dist <- cnt.dist+1
n.dist[cnt.dist] <- length(distance[distance==i])
y.dist[cnt.dist] <- sum(GOOD[distance==i])
fga.dist[cnt.dist] <- i
if (n.dist[cnt.dist] == 0) pi_hat.dist[cnt.dist] <- NA
else pi_hat.dist[cnt.dist] <- y.dist[cnt.dist] / n.dist[cnt.dist]
}

y1.dist <- y.dist
y0.dist <- n.dist - y.dist
y10 <- cbind(y1.dist,y0.dist)

fga.10 <- glm(y10 ~ fga.dist, binomial("logit"))
summary(fga.10)
```

```
confint(fga.10)
logLik(fga.10)
(X2.P1 <- sum(resid(fga.10,type="pearson")^2))
(P.X2.P1 <- 1-pchisq(X2.P1,45-2))
(X2.D1 <- deviance(fga.10))
(P.X2.D1 <- 1-pchisq(X2.D1,45-2))

fga.100 <- glm(y10 ~ 1, binomial("logit"))
summary(fga.100)
logLik(fga.100)
(X2.P0 <- sum(resid(fga.100,type="pearson")^2))
(P.X2.P0 <- 1-pchisq(X2.P0,45-1))
(X2.D0 <- deviance(fga.100))
(P.X2.D0 <- 1-pchisq(X2.D0,45-1))

anova(fga.100, fga.10, test="Chisq")

fga.dist1 <- seq(16,80,0.01)
plot(fga.dist,pi_hat.dist,pch=16)
lines(fga.dist1,predict(fga.10,list(fga.dist=fga.dist1),type="response"))

fga.1 <- glm(GOOD ~ distance, family=binomial("logit"))
summary(fga.1)
confint(fga.1)

### Output
> summary(fga.10)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.76271    0.54441  12.422   <2e-16 ***
fga.dist    -0.12084    0.01228  -9.836   <2e-16 ***

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 170.997  on 44  degrees of freedom
Residual deviance:  40.201  on 43  degrees of freedom
AIC: 132.65

> confint(fga.10)
                2.5 %      97.5 %
(Intercept)  5.7399741  7.87764341
fga.dist    -0.1457751 -0.09754001
> logLik(fga.10)
'log Lik.' -64.32263 (df=2)
>
> (X2.P1 <- sum(resid(fga.10,type="pearson")^2))
[1] 36.07857
> (P.X2.P1 <- 1-pchisq(X2.P1,45-2))
[1] 0.7635178
> (X2.D1 <- deviance(fga.10))
[1] 40.20124
> (P.X2.D1 <- 1-pchisq(X2.D1,45-2))
[1] 0.593378

> fga.100 <- glm(y10 ~ 1, binomial("logit"))
> summary(fga.100)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.86792    0.09113    20.5   <2e-16 ***

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 171  on 44  degrees of freedom
```

```
Residual deviance: 171  on 44  degrees of freedom
AIC: 261.44

> logLik(fga.100)
'log Lik.' -129.7205 (df=1)
> (X2.P0 <- sum(resid(fga.100,type="pearson")^2))
[1] 181.3516
> (P.X2.P0 <- 1-pchisq(X2.P0,45-1))
[1] 0
> (X2.D0 <- deviance(fga.100))
[1] 170.9969
> (P.X2.D0 <- 1-pchisq(X2.D0,45-1))
[1] 1.110223e-16

> anova(fga.100, fga.10, test="Chisq")
Analysis of Deviance Table

Model 1: y10 ~ 1
Model 2: y10 ~ fga.dist
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        44    170.997
2        43     40.201  1    130.8 < 2.2e-16 ***

> summary(fga.1)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.76271    0.54443  12.422   <2e-16 ***
distance    -0.12084    0.01229  -9.836   <2e-16 ***

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 817.72  on 1038  degrees of freedom
Residual deviance: 686.93  on 1037  degrees of freedom
AIC: 690.93

> confint(fga.1)
                 2.5 %      97.5 %
(Intercept)  5.7399740  7.87764350
distance    -0.1457751 -0.09754001
```

The Likelihood Ratio test for $H_0 : \beta_1 = 0$ (probability of success is not related to distance) is given below.

$$X^2_{LR} = -2\left[(-129.7205) - (-64.32263)\right] = 130.80 \qquad RR : X^2_{LR} \geq \chi^2_{.05,1} = 3.841 \qquad P = P\left(\chi^2_1 \geq 130.80\right) \approx 0$$

The computations to obtain the Null and Residual Deviances are given in Table 9.2.

$$\begin{aligned} \text{Null Deviance:} \quad & -2\left[l_0 - l_S\right] = -2\left[(-129.72) - (-44.22)\right] = 171.00 \\ \text{Residual Deviance:} \quad & -2\left[l_A - l_S\right] = -2\left[(-64.32) - (-44.22)\right] = 40.20 \end{aligned}$$

Based on the Goodness-of-Fit tests, we fail to reject the null hypothesis that the model with Distance provides a good fit (Pearson $p = .7635$, Deviance $p = .5934$). The intercept only model clearly does not provide a good fit (both P-values $\approx 0$).

$\nabla$

| Distance | Attempts | Successes | Distance | Attempts | Successes | Distance | Attempts | Successes |
|---|---|---|---|---|---|---|---|---|
| 18 | 2 | 2 | 33 | 37 | 35 | 48 | 37 | 23 |
| 19 | 7 | 7 | 34 | 33 | 31 | 49 | 27 | 19 |
| 20 | 23 | 22 | 35 | 34 | 33 | 50 | 22 | 15 |
| 21 | 25 | 25 | 36 | 22 | 19 | 51 | 26 | 16 |
| 22 | 27 | 27 | 37 | 34 | 30 | 52 | 12 | 9 |
| 23 | 33 | 33 | 38 | 50 | 47 | 53 | 18 | 11 |
| 24 | 19 | 19 | 39 | 24 | 22 | 54 | 15 | 10 |
| 25 | 29 | 29 | 40 | 29 | 23 | 55 | 2 | 2 |
| 26 | 36 | 34 | 41 | 30 | 27 | 56 | 5 | 4 |
| 27 | 32 | 31 | 42 | 34 | 29 | 57 | 3 | 2 |
| 28 | 32 | 31 | 43 | 40 | 33 | 58 | 1 | 0 |
| 29 | 19 | 19 | 44 | 31 | 28 | 59 | 1 | 0 |
| 30 | 34 | 31 | 45 | 27 | 25 | 68 | 1 | 0 |
| 31 | 29 | 27 | 46 | 23 | 13 | 69 | 1 | 0 |
| 32 | 36 | 34 | 47 | 36 | 23 | 76 | 1 | 0 |

Table 9.1: NFL Field Goal Attempts and Successes by Distance - 2008 Regular Season

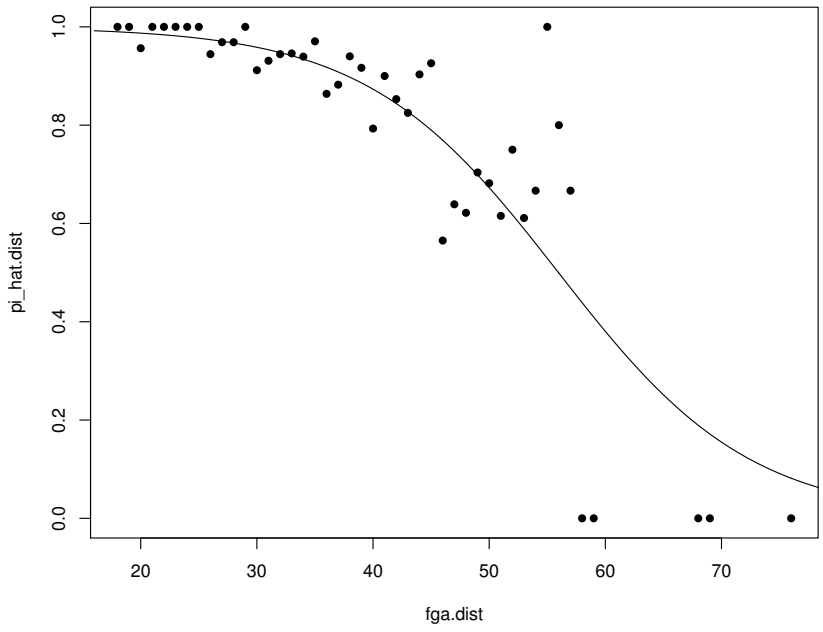


Figure 9.1: NFL Field Goal Attempts and P(Success) - 2008 Regular Season

| distance | $n$ | $y$ | $\hat{\pi}_A$ | $\hat{\pi}_0$ | $l_A$ | $l_0$ | $y\ln(y/n)$ | $(n-y)\ln((n-y)/n)$ | $l_S$ |
|---|---|---|---|---|---|---|---|---|---|
| 18 | 2 | 2 | 0.9899 | 0.8662 | -0.0203 | -0.2872 | 0 | 0 | 0 |
| 19 | 7 | 7 | 0.9886 | 0.8662 | -0.0799 | -1.0053 | 0 | 0 | 0 |
| 20 | 23 | 22 | 0.9872 | 0.8662 | -1.5066 | -2.0357 | -0.9779 | -3.1355 | -0.9779 |
| 21 | 25 | 25 | 0.9856 | 0.8662 | -0.3629 | -3.5905 | 0 | 0 | 0 |
| 22 | 27 | 27 | 0.9838 | 0.8662 | -0.4419 | -3.8777 | 0 | 0 | 0 |
| 23 | 33 | 33 | 0.9817 | 0.8662 | -0.6088 | -4.7394 | 0 | 0 | 0 |
| 24 | 19 | 19 | 0.9794 | 0.8662 | -0.3951 | -2.7288 | 0 | 0 | 0 |
| 25 | 29 | 29 | 0.9768 | 0.8662 | -0.6796 | -4.165 | 0 | 0 | 0 |
| 26 | 36 | 34 | 0.9739 | 0.8662 | -1.7468 | -2.4604 | -1.9434 | -5.7807 | -1.2784 |
| 27 | 32 | 31 | 0.9707 | 0.8662 | -0.9863 | -2.998 | -0.9842 | -3.4657 | -0.9842 |
| 28 | 32 | 31 | 0.9671 | 0.8662 | -0.9857 | -2.998 | -0.9842 | -3.4657 | -0.9842 |
| 29 | 19 | 19 | 0.963 | 0.8662 | -0.7168 | -2.7288 | 0 | 0 | 0 |
| 30 | 34 | 31 | 0.9584 | 0.8662 | -2.1601 | -1.79 | -2.8636 | -7.2832 | -1.45 |
| 31 | 29 | 27 | 0.9533 | 0.8662 | -1.4134 | -1.8944 | -1.9294 | -5.3483 | -1.2713 |
| 32 | 36 | 34 | 0.9476 | 0.8662 | -1.2821 | -2.4604 | -1.9434 | -5.7807 | -1.2784 |
| 33 | 37 | 35 | 0.9413 | 0.8662 | -1.2866 | -2.5485 | -1.9449 | -5.8355 | -1.2792 |
| 34 | 33 | 31 | 0.9343 | 0.8662 | -1.283 | -2.2062 | -1.9381 | -5.6067 | -1.2758 |
| 35 | 34 | 33 | 0.9265 | 0.8662 | -1.6044 | -3.2246 | -0.9851 | -3.5264 | -0.9851 |
| 36 | 22 | 19 | 0.9178 | 0.8662 | -1.7858 | -1.4238 | -2.7855 | -5.9773 | -1.4232 |
| 37 | 34 | 30 | 0.9082 | 0.8662 | -1.6967 | -1.6102 | -3.7549 | -8.5603 | -1.5706 |
| 38 | 50 | 47 | 0.8976 | 0.8662 | -2.0305 | -2.9014 | -2.9081 | -8.4402 | -1.4651 |
| 39 | 24 | 22 | 0.886 | 0.8662 | -1.3859 | -1.5623 | -1.9143 | -4.9698 | -1.2637 |
| 40 | 29 | 23 | 0.8732 | 0.8662 | -2.4377 | -2.3014 | -5.3314 | -9.4532 | -1.7135 |
| 41 | 30 | 27 | 0.8592 | 0.8662 | -1.6699 | -1.6034 | -2.8447 | -6.9078 | -1.4436 |
| 42 | 34 | 29 | 0.8439 | 0.8662 | -1.6719 | -1.6864 | -4.6129 | -9.5846 | -1.6612 |
| 43 | 40 | 33 | 0.8273 | 0.8662 | -1.8088 | -2.0792 | -6.3483 | -12.2008 | -1.808 |
| 44 | 31 | 28 | 0.8094 | 0.8662 | -2.4835 | -1.6452 | -2.8499 | -7.0061 | -1.4453 |
| 45 | 27 | 25 | 0.79 | 0.8662 | -3.1529 | -1.7528 | -1.924 | -5.2054 | -1.2686 |
| 46 | 23 | 13 | 0.7693 | 0.8662 | -4.1254 | -8.0324 | -7.4171 | -8.3291 | -1.7961 |
| 47 | 36 | 23 | 0.7471 | 0.8662 | -3.0177 | -7.8924 | -10.3046 | -13.2414 | -1.9851 |
| 48 | 37 | 23 | 0.7236 | 0.8662 | -2.9116 | -8.9321 | -10.9347 | -13.606 | -2.0081 |
| 49 | 27 | 19 | 0.6988 | 0.8662 | -1.7962 | -4.208 | -6.6766 | -9.7312 | -1.7947 |
| 50 | 22 | 15 | 0.6728 | 0.8662 | -1.7181 | -4.1883 | -5.7449 | -8.0159 | -1.7141 |
| 51 | 26 | 16 | 0.6457 | 0.8662 | -1.8893 | -6.9279 | -7.7681 | -9.5551 | -1.8378 |
| 52 | 12 | 9 | 0.6176 | 0.8662 | -1.8277 | -1.9336 | -2.5891 | -4.1589 | -1.3544 |
| 53 | 18 | 11 | 0.5887 | 0.8662 | -1.6793 | -5.2926 | -5.4172 | -6.6112 | -1.6605 |
| 54 | 15 | 10 | 0.5591 | 0.8662 | -1.9015 | -3.4865 | -4.0547 | -5.4931 | -1.5403 |
| 55 | 2 | 2 | 0.5292 | 0.8662 | -1.273 | -0.2872 | 0 | 0 | 0 |
| 56 | 5 | 4 | 0.499 | 0.8662 | -1.8624 | -0.9766 | -0.8926 | -1.6094 | -0.8926 |
| 57 | 3 | 2 | 0.4688 | 0.8662 | -1.0491 | -1.2002 | -0.8109 | -1.0986 | -0.8109 |
| 58 | 1 | 0 | 0.4389 | 0.8662 | -0.5778 | -2.0115 | 0 | 0 | 0 |
| 59 | 1 | 0 | 0.4094 | 0.8662 | -0.5266 | -2.0115 | 0 | 0 | 0 |
| 68 | 1 | 0 | 0.1894 | 0.8662 | -0.2099 | -2.0115 | 0 | 0 | 0 |
| 69 | 1 | 0 | 0.1715 | 0.8662 | -0.1882 | -2.0115 | 0 | 0 | 0 |
| 76 | 1 | 0 | 0.0816 | 0.8662 | -0.0851 | -2.0115 | 0 | 0 | 0 |
| Sum | 1039 | 900 | | | -64.3228 | -129.72 | | | -44.2219 |

Table 9.2: NFL Field Goal Attempts and Successes by Distance - Computations for Null and Residual Deviance

## 9.3 Count Data - Poisson and Negative Binomial Regression

When the response variable is the count of some item or event occurring in some time and/or space, regression models typically are based on the Poisson or Negative Binomial distributions. The Poisson has a particular restriction, that the mean and variance are equal. Typically (but not always) when that does not hold, the variance exceeds the mean, and the data are said to be "overdispersed." There are methods to adjust tests of regression coefficients when overdispersion is present. An alternative possibility is to fit a Negative Binomial model, which has two parameters and allows for the variance to exceed the mean.

### 9.3.1 Poisson Regression

The Poisson mean $\lambda$ is required to be positive, so it is natural (but not necessary) to model the log of the mean as being a linear function of $X_1, \ldots, X_p$. This is referred to as a **log link**.

$$g(\lambda) = \ln(\lambda) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p = \mathbf{x}'\boldsymbol{\beta} \quad \Rightarrow \quad \lambda = \exp\{\mathbf{x}'\boldsymbol{\beta}\}$$

The likelihood and log likelihood for the Poisson, are given below, along with the relevant derivatives.

$$L = \prod_{i=1}^{n} \frac{\exp\{-\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}\}\left[\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}\right]^{y_i}}{y_i!}$$

$$l = \ln(L) = \sum_{i=1}^{n} \left[-\exp\{\mathbf{x}_i'\boldsymbol{\beta}\} + y_i\mathbf{x}_i'\boldsymbol{\beta} - \ln(y_i!)\right]$$

$$g_{\boldsymbol{\beta}} = \frac{\partial l}{\partial \beta} = \sum_{i=1}^{n} \left[-\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}\mathbf{x}_i + y_i\mathbf{x}_i\right] = \sum_{i=1}^{n} (y_i - \lambda_i)\,\mathbf{x}_i = \mathbf{X}'(\mathbf{Y} - \lambda) \qquad \lambda = \mathrm{diag}\left[\lambda_i\right]$$

$$G_{\boldsymbol{\beta}} = \frac{\partial^2 l}{\partial\beta\partial\beta'} = -\sum_{i=1}^{n} \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}\mathbf{x}_i\mathbf{x}_i' = -\mathbf{X}'\mathbf{W}\mathbf{X} \qquad \mathbf{W} = \mathrm{diag}\left[\lambda_i\right] \qquad \lambda_i = \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}$$

The Newton-Raphson algorithm can be used to obtain Maximum Likelihood estimates of the elements of $\boldsymbol{\beta}$.

$$\tilde{\boldsymbol{\beta}}^{(k)} \quad = \quad \tilde{\boldsymbol{\beta}}^{(k-1)} - \left[G_{\tilde{\boldsymbol{\beta}}^{(k-1)}}\right]^{-1} g_{\tilde{\boldsymbol{\beta}}^{(k-1)}}$$

After setting staring values, iterate until convergence. A simple way to obtain starting values is as follows.

$$\tilde{\beta}_1^{(0)} = \cdots = \tilde{\beta}_p^{(0)} = 0 \qquad \tilde{\beta}_0^{(0)} = \ln\left(\overline{Y}\right)$$

The estimated variance of $\hat{\boldsymbol{\beta}}$ is $-E\{G_{\hat{\boldsymbol{\beta}}}\}$.

$$\hat{V}\left\{\hat{\boldsymbol{\beta}}\right\} = \left(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X}\right)^{-1}$$

**Interpreting the Regression Coefficients**

The estimated mean of $Y$ is $\exp\{\mathbf{x}'\hat{\boldsymbol{\beta}}\}$. When $X_j$ is increased by 1 unit, holding all other independent variables constant, we obtain the following ratio for the mean.

$$\frac{\exp\{\beta_0 + \beta_1 X_1 + \cdots + \beta_{j-1}X_{j-1} + \beta_j(X_j+1) + \beta_{j+1}X_{j+1} + \cdots + \beta_p X_p\}}{\exp\{\beta_0 + \beta_1 X_1 + \cdots + \beta_{j-1}X_{j-1} + \beta_j X_j + \beta_{j+1}X_{j+1} + \cdots + \beta_p X_p\}} = e^{\beta_j}$$

When $\beta_j = 0$, the multiplicative effect of increasing $X_j$ by 1 unit, holding all other predictors constant, is to change the mean by $e^{\beta_j}$. When $\beta_j = 0$, the multiplicative change is $e^0 = 1$, and we say $Y$ is not associated with $X_j$, controlling for all other variables.

In many studies, researchers are interested in **Risk Ratios**, the ratio of the mean response at $\mathbf{x}_i'$ and $\mathbf{x}_j'$. This can be written in the following form.

$$\frac{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{\exp\{\mathbf{x}_j'\boldsymbol{\beta}\}} = \exp\{\left(\mathbf{x}_i' - \mathbf{x}_j'\right)\boldsymbol{\beta}\}$$

**Inferences Regarding Regression Parameters**

The Likelihood-Ratio and Wald tests are conducted in a similar manner as in Logistic Regression, making use of the Poisson likelihood, ML estimates, and estimated variances.

Likelihood Ratio tests are based on the log-likelihood computed under the null hypothesis, $l_0 = \ln L_0$, and the log-likelihood under the alternative hypothesis (no restrictions) $l_A = \ln L_A$. The Likelihood Ratio test statistic, rejection region and $P$-value are given below, where $k$ is the number of parameter restrictions under the null hypothesis.

$$X^2_{LR} = -2\left[l_0 - l_A\right] \qquad RR: X^2_{LR} \geq \chi^2_{\alpha,k} \qquad P = P\left(\chi^2_k \geq X^2_{LR}\right)$$

In R, the **glm** function prints two deviance measures: **Null** and **Residual**. These represent the differences $-2\left[l_0 - l_S\right]$ and $-2\left[l_A - l_S\right]$, where $l_S$ is the log-likelihood under the saturated model, where $\hat{\lambda}_i = y_i$. It is computed as follows.

$$l = \sum_{i=1}^{n}\left[-\lambda_i + y_i \ln\left(\lambda_i\right) - \ln\left(y_i!\right)\right] \qquad \Rightarrow \qquad l_S = \sum_{i=1}^{n}\left[-y_i + y_i \ln\left(y_i\right) - \ln\left(y_i!\right)\right] \qquad \text{s.t. } 0\ln(0) = 0$$

The Likelihood Ratio test statistic for testing $H_0 : \beta_1 = \cdots = \beta_p = 0$ is the difference between the Null and Residual Deviances, with $p$ degrees of freedom.

Wald tests of the form $H_0 : \beta_j = \beta_{j0}$ are of the forms (chi-square and Z) given below.

$$X^2_W = \frac{\left(\hat{\beta}_j - \beta_{j0}\right)^2}{\hat{V}\left\{\hat{\beta}_j\right\}} \qquad RR : X^2_W \geq \chi^2_{\alpha,1} \qquad P = P\left(\chi^2_1 \geq X^2_W\right)$$

$$z_W = \frac{\hat{\beta}_j - \beta_{j0}}{\hat{SE}\left\{\hat{\beta}_j\right\}} \qquad RR : |z_W| \geq z_{\alpha/2} \qquad P = 2P\left(Z \geq |z_W|\right)$$

For general linear tests among the regression coefficients, $H_0 : \mathbf{K}'\boldsymbol{\beta} - \mathbf{m} = \mathbf{0}$, the Wald Test is conducted as a chi-square test, with $k$ restrictions (number of rows in $\mathbf{K}'$).

$$X^2_W = \left(\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m}\right)'\left[\mathbf{K}'\hat{V}\{\hat{\boldsymbol{\beta}}\}\mathbf{K}\right]^{-1}\left(\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m}\right) \qquad RR : X^2_W \geq \chi^2_{\alpha,k} \qquad P = P\left(\chi^2_k \geq X^2_W\right)$$

Confidence intervals for the regression coefficients and their corresponding risk ratios (RR) are obtained by using the approximate normality of the estimated regression coefficients, and then exponentiating the end points for the odds ratios.

$$(1-\alpha)100\%\text{ CI for }\beta_j : \quad \hat{\beta}_j \pm z_{\alpha/2}\hat{SE}\{\hat{\beta}_j\} = (\beta_{jL}, \beta_{jH}) \qquad (1-\alpha)100\%\text{ CI for } RR_j : \left(e^{\beta_{jL}}, e^{\beta_{jH}}\right)$$

In general, for more complex Risk Ratios of the following form, we can set up Tests and Confidence Intervals based on the following result.

$$\hat{RR} = \exp\{(\mathbf{x}_i - \mathbf{x}_j)'\hat{\boldsymbol{\beta}}\} \qquad \Rightarrow \qquad \ln\{\hat{RR}\} = (\mathbf{x}_i - \mathbf{x}_j)'\hat{\boldsymbol{\beta}}$$

$$\hat{V}\left\{\ln\{\hat{RR}\}\right\} = (\mathbf{x}_i - \mathbf{x}_j)'\hat{V}\{\hat{\boldsymbol{\beta}}\}(\mathbf{x}_i - \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)'\left(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X}\right)^{-1}(\mathbf{x}_i - \mathbf{x}_j)$$

Tests and Confidence Intervals regarding $\ln\{\hat{RR}\}$ can be made, and Confidence Intervals for the Risk Ratio can be obtained by exponentiating the endpoints.

### 9.3.2 Goodness of Fit Tests

When there is a fixed number of $n$ distinct $\mathbf{x}$ levels, the Pearson and Likelihood-Ratio (Deviance) Goodness-of-Fit statistics given below have approximate chi-square distributions with $n - p'$ degrees of freedom (see e.g. Agresti (1996, pp.89-90)).

$$\text{Pearson: } X_P^2 = \sum_{i=1}^{n} \frac{\left(y_i - \hat{\lambda}_i\right)^2}{\hat{\lambda}_i} = \sum_{i=1}^{n} e_{iP}^2 \qquad e_{iP} = \frac{y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}}$$

$$\text{Deviance: } X_D^2 = G^2 = 2\sum_{i=1}^{n}\left[y_i \ln \frac{y_i}{\hat{\lambda}_i} - \left(y_i - \hat{\lambda}_i\right)\right] = \sum_{i=1}^{n} e_{iD}^2 \qquad e_{iD} = \text{sgn}\{y_i - \hat{\lambda}_i\}\left[2y_i \ln \frac{y_i}{\hat{\lambda}_i} - \left(y_i - \hat{\lambda}_i\right)\right]^{1/2}$$

When the independent variable(s) have many distinct level(s) or combinations, observations can be grouped based on their $X$ levels in cases where there is $p = 1$ independent variable, or grouped based on their predicted means in general. The sums of events ($y$) and their corresponding predicted values $\left(\hat{\lambda}_i\right)$ are obtained for each group. Pearson residuals are obtained for each group based on the sums of their observed and predicted values. If we have $g$ groups and $p$ predictors, the approximate Pearson chi-square statistic will have $g - p'$ degrees of freedom (see e.g. Agresti (1996, p. 90)).

**Example: NASCAR Crashes - 1972-1979 Seasons**

NASCAR is a professional stock car racing organization in the United States. The top division is currently called the Sprint Cup. We consider all races during the 1972-1979 season (Winner (2006)). The response is the number of Caution Flags (a proxy for crashes) for each race ($Y$), and the predictors considered are: Track Length ($X_1$), Number of Drivers ($X_2$), and Number of Laps ($X_3$). During this period, there were $n = 151$ races. Table 9.3 contains summaries (quantiles) and correlations for $X_1$, $X_2$,$X_3$, and $Y$. The pairs of independent variables with high correlations are Track Length is highly correlated with Laps ($-.901$) and Drivers (.731). The following model is fit, first in matrix form, then using R's **glm** function.

$$Y_i \sim \text{Poisson}P\left(\lambda_i\right) \qquad \ln\left(\lambda_i\right) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3}$$

**R Program and Output - Matrix Form**

```
### Program
race1 <- read.fwf("http://www.stat.ufl.edu/~winner/data/race7579.dat", width=c(8,8,8,8,8,8,8,8,8,12,40),
col.names=c('srace', 'yr', 'yrace', 'drivers', 'trklen', 'laps', 'roadtrk',
 'cautions', 'leadchng', 'trkid', 'track'))

race <- data.frame(drivers=race1$drivers, trklen=race1$trklen, laps=race1$laps,
 cautions=race1$cautions)
attach(race)
```

```
############## Matrix Form
n.race <- length(cautions)
X0 <- rep(1,n.race)
X <- cbind(X0,drivers,trklen,laps)
Y <- cautions

beta.old <- matrix(c(log(mean(cautions)),0,0,0),ncol=1)
beta.diff <- 1000
num.iter <- 0

while (beta.diff > 0.00001) {
num.iter <- num.iter + 1
lambda.hat <- exp(X %*% beta.old)
w.v <-  lambda.hat
W.M <- matrix(rep(0,n.race^2),ncol=n.race)
for (i in 1:n.race) W.M[i,i] <- w.v[i,1]
mu.Y <- lambda.hat
g.beta <- t(X) %*% (Y - mu.Y)
G.beta <- -t(X) %*% W.M %*% X
beta.new <- beta.old - solve(G.beta) %*% g.beta
beta.diff <- sum((beta.new - beta.old)^2)
beta.old <- beta.new
print(beta.old)
}

num.iter
lambda.hat <- exp(X %*% beta.old)
w.v <-  lambda.hat
W.M <- matrix(rep(0,n.race^2),ncol=n.race)
for (i in 1:n.race) W.M[i,i] <- w.v[i,1]
g.beta <- t(X) %*% (Y - lambda.hat)
G.beta <- -t(X) %*% W.M %*% X

V.beta <- -solve(G.beta)
SE.beta <- sqrt(diag(V.beta))
z.beta <- beta.old / SE.beta
pv.beta <- 2*(1-pnorm(abs(z.beta),0,1))

beta.est <- cbind(beta.old,SE.beta,z.beta,pv.beta,
beta.old - 1.96*SE.beta, beta.old + 1.96*SE.beta)
colnames(beta.est) <- c("Estimate","Robust SE","z","Pr(>|z|)","LL","UL")
beta.est

### Output

> num.iter
[1] 3

> beta.est
            Estimate    Robust SE          z    Pr(>|z|)           LL          UL
X0      -0.796270269 0.4117015586 -1.9340958 0.053101346 -1.603205324 0.010664786
drivers  0.036525310 0.0124933565  2.9235786 0.003460328  0.012038331 0.061012288
trklen   0.114498691 0.1684265631  0.6798137 0.496622410 -0.215617373 0.444614754
laps     0.002596319 0.0007893063  3.2893681 0.001004126  0.001049279 0.004143359
```

### R Program and Output - glm Function

```
### Program
race.mod <- glm(formula = cautions ~ drivers + trklen + laps, family=poisson("log"))
```

```
summary(race.mod)
anova(race.mod, test="Chisq")

muhat <- predict(race.mod, type="response")

#print(cbind(cautions, muhat))
(pearson.x2 <- sum((cautions - muhat)^2/muhat))
(pearson.x2a <- sum(resid(race.mod,type="pearson")^2))
(deviance.x2 <- sum(resid(race.mod)^2))

### Output
> summary(race.mod)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.7962699  0.4116942  -1.934  0.05310 .
drivers      0.0365253  0.0124932   2.924  0.00346 **
trklen       0.1144986  0.1684236   0.680  0.49662
laps         0.0025963  0.0007893   3.289  0.00100 **

(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 215.49  on 150  degrees of freedom
Residual deviance: 171.22  on 147  degrees of freedom
AIC: 671.11
Number of Fisher Scoring iterations: 4

> anova(race.mod, test="Chisq")
Analysis of Deviance Table
Model: poisson, link: log
Response: cautions
Terms added sequentially (first to last)
        Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                      150     215.49
drivers  1   5.1673       149     210.32  0.023016 *
trklen   1  28.1912       148     182.13 1.099e-07 ***
laps     1  10.9167       147     171.22  0.000953 ***

> muhat <- predict(race.mod, type="response")
> #print(cbind(cautions, muhat))
> (pearson.x2 <- sum((cautions - muhat)^2/muhat))
[1] 158.8281
> (pearson.x2a <- sum(resid(race.mod,type="pearson")^2))
[1] 158.8281
> (deviance.x2 <- sum(resid(race.mod)^2))
[1] 171.2162
```

Note that the **anova** given in the second portion of the output is testing terms sequentially. Given that drivers is in the model, trklen is highly significant ($p$ = 1.099e-07). Then when laps is added to a model with drivers and trklen, it is highly significant ($p$ = .000953). However, when looking at the $z$-tests in the **summary** portion, which test for each predictor given all other predictors, trklen is no longer significant ($p$ = .49662). This is an example of collinearity among the predictors.

A "grouped" (12 groups) goodness of fit test is conducted based on the model with drivers and laps. We see that the null hypothesis of the Poisson model being appropriate is rejected ($p$ = .0188).

```
### Program
race.mod1 <- glm(formula = cautions ~ drivers + laps, family=poisson("log"))
summary(race.mod1)
anova(race.mod1, test="Chisq")
muhat <- predict(race.mod1, type="response")
```

```
mean.grp <- rep(0,length(cautions))
for (i in 1:length(cautions)) {
if (muhat[i] < 3.50) mean.grp[i] <- 1
else  if (muhat[i] < 3.70) mean.grp[i] <- 2
else  if (muhat[i] < 4.00) mean.grp[i] <- 3
else  if (muhat[i] < 4.15) mean.grp[i] <- 4
else  if (muhat[i] < 4.30) mean.grp[i] <- 5
else  if (muhat[i] < 4.40) mean.grp[i] <- 6
else  if (muhat[i] < 4.70) mean.grp[i] <- 7
else  if (muhat[i] < 5.25) mean.grp[i] <- 8
else  if (muhat[i] < 5.50) mean.grp[i] <- 9
else  if (muhat[i] < 6.00) mean.grp[i] <- 10
else  if (muhat[i] < 6.80) mean.grp[i] <- 11
else mean.grp[i] <- 12
}

count.mg <- rep(0,max(mean.grp))
sum.mg <- rep(0,max(mean.grp))
sum.muhat.mg <- rep(0,max(mean.grp))

for (i in 1:max(mean.grp)) {
count.mg[i] <- length(cautions[mean.grp == i])
sum.mg[i] <- sum(cautions[mean.grp == i])
sum.muhat.mg[i] <- sum(muhat[mean.grp == i])
}

gof.grp <- cbind(count.mg,sum.mg,sum.muhat.mg,pearson.r)
colnames(gof.grp) <- c("# Races","Total Obs", "Total Exp", "Pearson r")
gof.grp

### Output
> summary(race.mod1)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.6876335  0.3776880  -1.821   0.0687 .
drivers      0.0428077  0.0084250   5.081 3.75e-07 ***
laps         0.0021136  0.0003435   6.153 7.59e-10 ***

(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 215.49  on 150  degrees of freedom
Residual deviance: 171.68  on 148  degrees of freedom
AIC: 669.57
Number of Fisher Scoring iterations: 4

> pearson.r <- (sum.mg - sum.muhat.mg) / sqrt(sum.muhat.mg)
> (pearson.X2.mg <- sum(pearson.r^2))
[1] 19.85854
> qchisq(.95,12-2-1)
[1] 16.91898
> (pval.mg <- 1-pchisq(pearson.X2.mg,12-2-1))
[1] 0.01880574
> gof.grp
      # Races Total Obs Total Exp  Pearson r
 [1,]      15        37  46.15526 -1.3475973
 [2,]      11        50  39.37362  1.6934908
 [3,]      11        39  42.13560 -0.4830542
 [4,]      14        58  57.20983  0.1044684
 [5,]      16        53  67.80602 -1.7980605
 [6,]       4        20  17.30750  0.6471999
 [7,]      17        80  76.85249  0.3590364
 [8,]      21       129 108.13139  2.0068626
 [9,]       2        10  10.81521 -0.2478863
[10,]       8        40  45.35976 -0.7958108
```

| | Quantiles | | | | | | Correlations | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Variable | Min | 25% | Med | Mean | 75% | Max | TrkLen | Drivers | Laps | Cautions |
| TrkLen | 0.526 | 0.625 | 1.366 | 1.446 | 2.500 | 2.660 | 1 | 0.731 | -0.901 | -0.317 |
| Drivers | 22 | 30 | 36 | 35.2 | 40 | 50 | 0.731 | 1 | -0.508 | 0.160 |
| Laps | 95 | 200 | 367 | 339.7 | 420 | 500 | -0.901 | -0.508 | 1 | 0.297 |
| Cautions | 0 | 3 | 5 | 4.795 | 6 | 12 | -0.317 | 0.160 | 0.297 | 1 |

Table 9.3: NASCAR Caution Flag, Track Length, Drivers, Laps: Summaries and Correlations

```
[11,]      24       167 154.12482  1.0370914
[12,]       8        41  58.72850 -2.3133835
```

$\nabla$

### 9.3.3 Overdispersion

The Poisson distribution has a restriction that the mean and variance are equal. In practice, actual data may demonstrate that the variance exceeds the mean (overdispersion), or less frequently in practice, that the variance is smaller than the mean (underdispersion). There are several ways of checking for overdispersion (or underdispersion). If the model is appropriate, then the Pearson chi-square statistic should be approximately equal to its degrees of freedom $(n - p')$. If the chi-square statistic is much larger, then there is evidence of overdispersion. If the chi-square statistic is much smaller, this is evidence of underdispersion. Three possibilities that can be used in the presence of overdispersion are given here. The first involves fitting a quasipoisson model by adjusting the standard errors of the regression coefficients. The second involves using robust standard errors for the regression coefficients, computed in the manner of White's robust standard errors for the linear regression model. The third involves fitting a two parameter Negative Binomial model which allows the variance to exceed the mean.

The quasipoisson model corrects the Poisson assumption of variance being equal to the mean. The estimated regression coefficients remain the same, however the variances and standard errors are inflated (when overdispesion is present) or deflated (underdispersion). The correction factor is given below. The R function **glm** has a **quasipoisson** option for the "distribution family," and uses this correction.

$$\phi = \frac{\chi^2_P}{n - p'} \qquad \hat{V}\{\hat{\boldsymbol{\beta}}\} = \phi \left(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X}\right)^{-1}$$

Robust standard errors for the regression coefficients are another option (see Cameron and Trivedi (2010, pp. 574-575)). This is an extension of White's robust variance for normal based regression models. Let $\hat{\mathbf{E}}^2$ be a diagonal matrix of the squared residuals from the Poisson regression model: $\left(Y_i - \hat{\lambda}_i\right)^2$. Then the robust variance-covariance matrix for the estimated regression vector is given below. This method can be run in R, making use of **sandwich** and **msm** packages.

$$\hat{V}_R\{\hat{\boldsymbol{\beta}}\} = \left(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X}\right)^{-1} \left(\mathbf{X}'\hat{\mathbf{E}}^2\mathbf{X}\right) \left(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X}\right)^{-1}$$

The Negative Binomial model will be described in the following section.

**Example: NASCAR Crashes - 1972-1979 Seasons**

The model with Drivers and Laps is fit below, with and without correction for overdispersion. First, the model is fit directly making use of the matrix form. Then it is fit making use of R functions and packages.

```
### Program
############## Matrix Form
n.race <- length(cautions)
X0 <- rep(1,n.race)
X <- cbind(X0,drivers,laps)
Y <- cautions

beta.old <- matrix(c(log(mean(cautions)),0,0),ncol=1)
beta.diff <- 1000
num.iter <- 0

while (beta.diff > 0.0000001) {
num.iter <- num.iter + 1
lambda.hat <- exp(X %*% beta.old)
w.v <-  lambda.hat
W.M <- matrix(rep(0,n.race^2),ncol=n.race)
for (i in 1:n.race) W.M[i,i] <- w.v[i,1]
g.beta <- t(X) %*% (Y - lambda.hat)
G.beta <- -t(X) %*% W.M %*% X
beta.new <- beta.old - solve(G.beta) %*% g.beta
beta.diff <- sum((beta.new - beta.old)^2)
beta.old <- beta.new
print(beta.old)
}

num.iter
lambda.hat <- exp(X %*% beta.old)
w.v <-  lambda.hat
W.M <- matrix(rep(0,n.race^2),ncol=n.race)
for (i in 1:n.race) W.M[i,i] <- w.v[i,1]
g.beta <- t(X) %*% (Y - lambda.hat)
G.beta <- -t(X) %*% W.M %*% X

V.beta <- -solve(G.beta)
SE.beta <- sqrt(diag(V.beta))
z.beta <- beta.old / SE.beta
pv.beta <- 2*(1-pnorm(abs(z.beta),0,1))

beta.est <- cbind(beta.old,SE.beta,z.beta,pv.beta,
beta.old - 1.96*SE.beta, beta.old + 1.96*SE.beta)
colnames(beta.est) <- c("Estimate","Std Error","z","Pr(>|z|)","LL","UL")
rownames(beta.est) <- c("Intercept","Drivers","Laps")
beta.est

### quasipoisson model
pearson.res.m <- (Y - lambda.hat) / sqrt(lambda.hat)
phi <- sum(pearson.res.m^2)/(n.race-ncol(X))
SE.beta.qp <- SE.beta * sqrt(phi)
z.beta.qp <- beta.old / SE.beta.qp
pv.beta.qp <- 2*(1-pnorm(abs(z.beta.qp)))

qp.est <- cbind(beta.old, SE.beta.qp, z.beta.qp, pv.beta.qp,
```

```
beta.old - 1.96 * SE.beta.qp,
beta.old + 1.96 * SE.beta.qp)
colnames(qp.est) <- c("Estimate","QP SE","QP z","Pr(>|z|)","LL","UL")
rownames(qp.est) <- c("Intercept","Drivers","Laps")
qp.est

### Robust Variance-Covariance matrix and SEs
e2.v <- (Y-lambda.hat)^2
e2.M <- matrix(rep(0,n.race^2),ncol=n.race)
for (i in 1:n.race) e2.M[i,i] <- e2.v[i]
X.e.X.M <- t(X) %*% e2.M %*% X
V.beta.R <- V.beta %*% X.e.X.M %*% V.beta
SE.beta.R <- sqrt(diag(V.beta.R))
z.beta.R <- beta.old / SE.beta.R
pv.beta.R <- 2*(1-pnorm(abs(z.beta.R),0,1))

robust.est <- cbind(beta.old, SE.beta.R, z.beta.R, pv.beta.R,
beta.old - 1.96 * SE.beta.R,
beta.old + 1.96 * SE.beta.R)
colnames(robust.est) <- c("Estimate","Robust SE","z","Pr(>|z|)","LL","UL")
rownames(robust.est) <- c("Intercept","Drivers","Laps")
robust.est

#### Output

> beta.est
             Estimate    Std Error         z     Pr(>|z|)          LL          UL
Intercept -0.687633518 0.3776937513 -1.820611 6.866596e-02 -1.42791327 0.052646235
Drivers    0.042807669 0.0084250966  5.080971 3.755110e-07  0.02629448 0.059320858
Laps       0.002113642 0.0003435008  6.153237 7.591725e-10  0.00144038 0.002786904

> qp.est
             Estimate        QP SE      QP z     Pr(>|z|)           LL          UL
Intercept -0.687633518 0.3916890187 -1.755560 7.916359e-02 -1.455343994 0.080076959
Drivers    0.042807669 0.0087372847  4.899425 9.611763e-07  0.025682591 0.059932747
Laps       0.002113642 0.0003562291  5.933378 2.967639e-09  0.001415433 0.002811851

> robust.est
             Estimate    Robust SE         z     Pr(>|z|)           LL          UL
Intercept -0.687633518 0.4030183158 -1.706209 8.796916e-02 -1.477549417 0.102282381
Drivers    0.042807669 0.0087162900  4.911226 9.050875e-07  0.025723741 0.059891598
Laps       0.002113642 0.0003694014  5.721803 1.053995e-08  0.001389615 0.002837669
```

Note that there is little evidence of overdispersion, as $\chi^2_P = 159.101$, $n - p' = 151 - 3 = 148$, and $\phi = 1.076$. Thus, the effects of the quasipoisson model and the robust standard errors are very small.

```
### Program
require(sandwich)
require(msm)

race.1 <- glm(formula = cautions ~ drivers + laps, family=poisson("log"))
summary(race.1)

race.2 <- glm(formula = cautions ~ drivers + laps,
family=quasipoisson("log"))
summary(race.2)

cov.race1 <- vcovHC(race.1, type="HC0")
std.err.R <- sqrt(diag(cov.race1))
z.R <- coef(race.1) / std.err.R
```

```
r.est <- cbind(coef(race.1),std.err.R,z.R,
2 * pnorm(abs(z.R), lower.tail=FALSE),
coef(race.1) - 1.96 * std.err.R,
coef(race.1) + 1.96 * std.err.R)
colnames(r.est) <- c("Estimate","Robust SE","z","Pr(>|z|)","LL","UL")
r.est

### Output
> summary(race.1)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.6876335  0.3776880  -1.821   0.0687 .
drivers      0.0428077  0.0084250   5.081 3.75e-07 ***
laps         0.0021136  0.0003435   6.153 7.59e-10 ***

(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 215.49  on 150  degrees of freedom
Residual deviance: 171.68  on 148  degrees of freedom
AIC: 669.57
Number of Fisher Scoring iterations: 4

> summary(race.2)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.6876335  0.3916893  -1.756   0.0812 .
drivers      0.0428077  0.0087373   4.899 2.49e-06 ***
laps         0.0021136  0.0003562   5.933 2.02e-08 ***

(Dispersion parameter for quasipoisson family taken to be 1.075517)
    Null deviance: 215.49  on 150  degrees of freedom
Residual deviance: 171.68  on 148  degrees of freedom
AIC: NA
Number of Fisher Scoring iterations: 4

> r.est
               Estimate    Robust SE         z     Pr(>|z|)           LL          UL
(Intercept) -0.687633511 0.4030185239 -1.706208 8.796932e-02 -1.477549818 0.102282795
drivers      0.042807669 0.0087163133  4.911213 9.051480e-07  0.025723695 0.059891643
laps         0.002113642 0.0003694006  5.721815 1.053923e-08  0.001389617 0.002837667
```

$\nabla$

### 9.3.4 Models with Varying Exposures

In many studies, interest is in comparing rates of events in groups or observations with different amounts of exposure to the outcome of interest. In these cases, the response is the number of of observations per unit of exposure. A log linear model is assumed for the expectation of the ratio. The fixed exposure (in the log model) is referred to as an **offset**. The model is as follows.

$$\text{Sample Rate: } \frac{Y_i}{t_i} \qquad E\left\{\frac{Y_i}{t_i}\right\} = \frac{\lambda_i}{t_i} \qquad \log\left(\frac{\lambda_i}{t_i}\right) = \log(\lambda_i) - \log(t_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$

Note that we will place $\log(t_i)$ on the right-hand side of the equal sign, but do not want to put a

regression coefficient on it. In statistical software packages, an offset option is typically available. The predicted values for the observations are given below.

$$\hat{Y}_i = t_i\hat{\lambda}_i = \exp\{\log(t_i) + \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}\} = t_i \exp\{\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}\}$$

Inferences are conducted as in the previously described Poisson model.

**Example: Friday the 13th, Gender, and Traffic Deaths**

A study reported incidences of traffic deaths by gender on Friday the 13th and other Fridays in Finland over the years 1971-1997 (Nayha (2002)). The response was the number of traffic deaths and the exposure was the number of person-days (100000s). The groups were the 4 combinations of Friday type ($X_{i1} = 1$ if Friday the 13th, 0 otherwise) and Gender ($X_{i2} = 1$ if Female, 0 if Male). The model contains an interaction term, $X_{i3} = X_{i1}X_{i2}$, which allows the Friday the 13th effect to differ by Gender (and vice versa). Table 9.4 gives the data, exposure, the independent variables, the predicted mean, and Total Death Rate per 100000 exposures for the four classifications/groups.

For Males and Females, the Friday the 13th effects are given below.

$$\text{Males: } \frac{\exp\{\beta_0 + \beta_1\}}{\exp\{\beta_0\}} = \exp\{\beta_1\} \qquad \text{Females: } \frac{\exp\{\beta_0 + \beta_1 + \beta_2 + \beta_3\}}{\exp\{\beta_0 + \beta_2\}} = \exp\{\beta_1 + \beta_3\}$$

Thus, a significant interaction ($\beta_3 \neq 0$) implies the Friday the 13th effect is not the same among Males and Females. To obtain 95% Confidence Intervals for the Male and Female Friday the 13th effects, first obtain 95% CIs for $\beta_1$ and $\beta_1 + \beta_3$, then exponentiate the endpoints.

$$\beta_1: \quad \hat{\beta}_1 \pm 1.96\hat{SE}\left\{\hat{\beta}_1\right\} \qquad \beta_1 + \beta_3: \quad \hat{\beta}_1 + \hat{\beta}_3 \pm 1.96\sqrt{\hat{V}\left\{\hat{\beta}_1\right\} + \hat{V}\left\{\hat{\beta}_3\right\} + 2\hat{\text{COV}}\left\{\hat{\beta}_1, \hat{\beta}_3\right\}}$$

The R Program and Output are given below. The Friday the 13th effect is not significant for Males, as the 95% CI for their Risk Ratio (0.8442,1.3110) contains 1. For Females, there is evidence of a Friday the 13th effect, as the 95% CI for their Risk Ratio (1.1793,2.2096) is entirely above 1.

```
### Program
Y.13 <- c(41,82,789,2423)
t.13 <- c(86.5,79.9,2687.1,2483.7)
X0.13 <- c(1,1,1,1)
X1.13 <- c(1,1,0,0)
X2.13 <- c(1,0,1,0)
X3.13 <- c(1,0,0,0)

f13.mod1 <- glm(Y.13 ~ X1.13 + X2.13 + X3.13, offset=log(t.13),
    family=poisson("log"))
summary(f13.mod1)
vcov(f13.mod1)
```

| Group | $i$ | $Y_i$ | $t_i$ | $X_{i1}$ | $X_{i2}$ | $X_{i3}$ | $\log\left(\hat{\lambda}_i\right)$ | $TDR_i = \hat{\lambda}_i$ |
|---|---|---|---|---|---|---|---|---|
| Friday 13th/Female | 1 | 41 | 86.5 | 1 | 1 | 1 | -0.7466 | 0.4740 |
| Friday 13th/Male | 2 | 82 | 79.9 | 1 | 0 | 0 | 0.0259 | 1.0263 |
| Other Friday/Female | 3 | 789 | 2687.1 | 0 | 1 | 0 | -1.2255 | 0.2936 |
| Other Friday/Male | 4 | 2423 | 2483.7 | 0 | 0 | 0 | -0.0247 | 0.9756 |

Table 9.4: Friday the 13th and Gender for Finland Traffic Deaths

```
beta1.hat <- coef(f13.mod1)[2]
se.beta1.hat <- sqrt(vcov(f13.mod1)[2,2])
beta13.hat <- beta1.hat + coef(f13.mod1)[4]
se.beta13.hat <- sqrt(vcov(f13.mod1)[2,2]+vcov(f13.mod1)[4,4]+
   2*vcov(f13.mod1)[2,4])
ll.beta1 <- beta1.hat - 1.96*se.beta1.hat
ul.beta1 <- beta1.hat + 1.96*se.beta1.hat
ll.beta13 <- beta13.hat - 1.96*se.beta13.hat
ul.beta13 <- beta13.hat + 1.96*se.beta13.hat

f13.eff.m <- cbind(beta1.hat,se.beta1.hat,ll.beta1,ul.beta1,
      exp(beta1.hat),exp(ll.beta1),exp(ul.beta1))
f13.eff.f <- cbind(beta13.hat,se.beta13.hat,ll.beta13,ul.beta13,
      exp(beta13.hat),exp(ll.beta13),exp(ul.beta13))
f13.eff <- rbind(f13.eff.m,f13.eff.f)
rownames(f13.eff) <- c("Males","Females")
colnames(f13.eff) <- c("Estimate","Std Err","LL","UL",
  "Risk Ratio", "LL RR", "UL RR")
round(f13.eff,4)

### Output
> summary(f13.mod1)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.02474    0.02032  -1.218   0.2232
X1.13        0.05069    0.11228   0.451   0.6517
X2.13       -1.20071    0.04099 -29.293   <2e-16 ***
X3.13        0.42819    0.19562   2.189   0.0286 *

(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 1.0258e+03  on 3  degrees of freedom
Residual deviance: 1.8230e-13  on 0  degrees of freedom
AIC: 37.942
Number of Fisher Scoring iterations: 2

> vcov(f13.mod1)
              (Intercept)         X1.13         X2.13         X3.13
(Intercept)  0.0004127115 -0.0004127115 -0.0004127115  0.0004127115
X1.13       -0.0004127115  0.0126078244  0.0004127115 -0.0126078244
X2.13       -0.0004127115  0.0004127115  0.0016801386 -0.0016801386
X3.13        0.0004127115 -0.0126078244 -0.0016801386  0.0382654231

> round(f13.eff,4)
        Estimate Std Err      LL     UL Risk Ratio  LL RR  UL RR
Males     0.0507  0.1123 -0.1694 0.2708     1.0520 0.8442 1.3110
Females   0.4789  0.1602  0.1649 0.7928     1.6143 1.1793 2.2096
```

$\nabla$

## 9.4 Negative Binomial Regression

An alternative to the Poisson Regression model for count data is the Negative Binomial Regression model. This distribution has 2 parameters, and allows for the variance to exceed the mean. The model arises as a Poisson distribution with parameters $\lambda$ that follow a Gamma distribution (see e.g. Agresti (2002) and Cameron and Trivedi (2010)). Suppose that $Y$ is Poisson with mean $\lambda\nu$, where $\nu$ is distributed Gamma$(\alpha^{-1}, \alpha)$, with mean 1, and variance $\alpha$. Then we have the following results.

$$p(y|\lambda,\nu) = \frac{e^{-\lambda\nu}(\lambda\nu)^y}{y!} \qquad y = 0, 1, \ldots \qquad\qquad p(\nu|\alpha) = \frac{1}{\Gamma(\alpha^{-1})\alpha^{\alpha^{-1}}}\nu^{\alpha^{-1}-1}\exp\left\{-\frac{\nu}{\alpha}\right\} \qquad \nu > 0$$

$$\Rightarrow \qquad p(y|\lambda,\nu,\alpha) = \frac{\exp\left\{-\nu\left(\lambda+\alpha^{-1}\right)\right\}\lambda^y\nu^{y+\alpha^{-1}-1}}{\Gamma(y+1)\Gamma(\alpha^{-1})\alpha^{\alpha^{-1}}}$$

$$\Rightarrow \qquad p(y|\lambda,\alpha) = \frac{\lambda^y}{\Gamma(y+1)\Gamma(\alpha^{-1})\alpha^{\alpha^{-1}}}\int_0^\infty \nu^{y+\alpha^{-1}-1}\exp\left\{-\frac{\nu}{(\lambda+\alpha^{-1})^{-1}}\right\}d\nu =$$

$$\frac{\lambda^y}{\Gamma(y+1)\Gamma(\alpha^{-1})\alpha^{\alpha^{-1}}}\Gamma\left(y+\alpha^{-1}\right)\left[\left(\lambda+\alpha^{-1}\right)^{-1}\right]^{y+\alpha^{-1}} = \frac{\Gamma\left(y+\alpha^{-1}\right)}{\Gamma(y+1)\Gamma(\alpha^{-1})}\left(\frac{\lambda}{\lambda+\alpha^{-1}}\right)^y\left(\frac{\alpha^{-1}}{\lambda+\alpha^{-1}}\right)^{\alpha^{-1}}$$

That is, given $\lambda$ and $\alpha$, $Y$ is distributed Negative Binomial$(r, p)$, with the following parameters and mean and variance (see e.g. Casella and Berger (1990)).

$$r = \alpha^{-1} \qquad p = \frac{\alpha^{-1}}{\lambda+\alpha^{-1}} \qquad E\{Y\} = \frac{r(1-p)}{p} = \lambda \qquad V\{Y\} = \frac{r(1-p)}{p^2} = \lambda(1+\lambda\alpha) = \lambda + \frac{\lambda^2}{\alpha^{-1}}$$

The model can be parameterized with a log link function, with $\log(\lambda)$ being linearly related to a set of predictors, and with $a^* = \ln\left(\alpha^{-1}\right)$. This way both estimates $\lambda$ and $\alpha^{-1}$ will be positive. The likelihood and log-likelihood functions are given below for data $y_1, \ldots, y_n$ with means $\lambda_1, \ldots, \lambda_n$.

$$L_i = \frac{\Gamma\left(y_i+\alpha^{-1}\right)}{\Gamma(y_i+1)\Gamma(\alpha^{-1})}\left(\frac{\lambda_i}{\lambda_i+\alpha^{-1}}\right)^{y_i}\left(\frac{\alpha^{-1}}{\lambda_i+\alpha^{-1}}\right)^{\alpha^{-1}} =$$

$$\frac{\left(y_i+\alpha^{-1}-1\right)\cdots\alpha^{-1}\Gamma\left(\alpha^{-1}\right)}{\Gamma(\alpha^{-1})\Gamma(y_i+1)}\left(\frac{\lambda_i}{\lambda_i+\alpha^{-1}}\right)^{y_i}\left(\frac{\alpha^{-1}}{\lambda_i+\alpha^{-1}}\right)^{\alpha^{-1}} =$$

$$\frac{\left(y_i+\alpha^{-1}-1\right)\cdots\alpha^{-1}}{y_i!}\left(\frac{\lambda_i}{\lambda_i+\alpha^{-1}}\right)^{y_i}\left(\frac{\alpha^{-1}}{\lambda_i+\alpha^{-1}}\right)^{\alpha^{-1}} =$$

$$\frac{\left(y_i+e^{a^*}-1\right)\cdots e^{a^*}}{y_i!}\left(\frac{\lambda_i}{\lambda_i+e^{a^*}}\right)^{y_i}\left(\frac{e^{a^*}}{\lambda_i+e^{a^*}}\right)^{e^{a^*}} = \frac{\prod_{j=0}^{y_i-1}\left(e^{a^*}+j\right)}{y_i!}\left(\frac{\exp\{\mathbf{x_i}'\boldsymbol{\beta}\}}{\exp\{\mathbf{x_i}'\boldsymbol{\beta}\}+e^{a^*}}\right)^{y_i}\left(\frac{e^{a^*}}{\exp\{\mathbf{x_i}'\boldsymbol{\beta}\}+e^{a^*}}\right)^{e^{a^*}}$$

$$l_i = \ln L_i = \sum_{j=0}^{y_i-1} \ln\left(e^{a^*} + j\right) - \ln\left(y_i!\right) + y_i\left[\mathbf{x_i}'\boldsymbol{\beta} - \ln\left(\exp\{\mathbf{x_i}'\boldsymbol{\beta}\} + e^{a^*}\right)\right] + e^{a^*}\left[a^* - \ln\left(\exp\{\mathbf{x_i}'\boldsymbol{\beta}\} + e^{a^*}\right)\right]$$

Taking the necessary derivatives of $l_i$ with respect to $a^*$ and $\boldsymbol{\beta}$, we obtain the following results.

$$\frac{\partial l_i}{\partial \boldsymbol{\beta}} = e^{a^*}\left(\frac{y_i - \exp\{\mathbf{x_i}'\boldsymbol{\beta}\}}{\exp\{\mathbf{x_i}'\boldsymbol{\beta}\} + e^{a^*}}\right)\mathbf{x_i}$$

$$\frac{\partial l_i}{\partial a^*} = e^{a^*}\left[\sum_{j=0}^{y_i-1}\frac{1}{e^{a^*}+j} + a^* + 1 - \frac{e^{a^*}+y_i}{e^{a^*}+\exp\{\mathbf{x_i}'\boldsymbol{\beta}\}} - \ln\left(e^{a^*} + \exp\{\mathbf{x_i}'\boldsymbol{\beta}\}\right)\right]$$

$$\frac{\partial^2 l_i}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} = -\left(e^{a^*} + y_i\right)\frac{\exp\{a^* + \mathbf{x_i}'\boldsymbol{\beta}\}\,\mathbf{x_i}\mathbf{x_i}'}{\left(e^{a^*} + \exp\{\mathbf{x_i}'\boldsymbol{\beta}\}\right)^2}$$

$$\frac{\partial^2 l_i}{\partial\left(a^*\right)^2} = e^{a^*}\left[\sum_{j=0}^{y_i-1}\frac{j}{\left(e^{a^*}+j\right)^2} + a^* + 2 - \frac{2e^{a^*}+y_i}{e^{a^*}+\exp\{\mathbf{x_i}'\boldsymbol{\beta}\}} + \frac{\left(y_i - \exp\{\mathbf{x_i}'\boldsymbol{\beta}\}\right)e^{a^*}}{\left(e^{a^*}+\exp\{\mathbf{x_i}'\boldsymbol{\beta}\}\right)^2} - \ln\left(e^{a^*} + \exp\{\mathbf{x_i}'\boldsymbol{\beta}\}\right)\right]$$

$$\frac{\partial^2 l_i}{\partial \boldsymbol{\beta}\partial a^*} = \exp\{a^* + \mathbf{x_i}'\boldsymbol{\beta}\}\left(\frac{y_i - \exp\{\mathbf{x_i}'\boldsymbol{\beta}\}}{\left(e^{a^*} + \exp\{\mathbf{x_i}'\boldsymbol{\beta}\}\right)^2}\right)\mathbf{x_i}$$

To obtain Newton-Raphson estimates of $a^*$ and $\boldsymbol{\beta}$, we construct the following matrix $(G_{\beta a^*})$ and vector $(g_{\beta a^*})$.

$$g_{a^*} = \sum_{i=1}^{n}\frac{\partial l_i}{\partial a^*} \qquad g_{\beta} = \sum_{i=1}^{n}\frac{\partial l_i}{\partial \beta} \qquad g_{\beta a^*} = \begin{bmatrix} g_{\beta} \\ g_{a^*} \end{bmatrix}$$

$$G_{aa} = \sum_{i=1}^{n}\frac{\partial^2 l_i}{\partial\left(a^*\right)^2} \qquad G_{bb} = \sum_{i=1}^{n}\frac{\partial^2 l_i}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} \qquad G_{ba} = \sum_{i=1}^{n}\frac{\partial^2 l_i}{\partial \boldsymbol{\beta}\partial a^*} \qquad G_{\beta a^*} = \begin{bmatrix} G_{bb} & G_{ba} \\ G_{ba}' & G_{aa} \end{bmatrix}$$

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\beta} \\ a^* \end{bmatrix}$$

To obtain starting values for the Newton-Raphson algorithm, first set $a^* = 0$, which corresponds to $\alpha^{-1} = 1$, and obtain an estimate of $\boldsymbol{\beta}$. Then set $\boldsymbol{\beta}$ to this estimate and obtain an estimate of $a^*$. Then use these for the process of estimating $\boldsymbol{\theta}$.

$$\tilde{\boldsymbol{\beta}}^{(i)} = \tilde{\boldsymbol{\beta}}^{(i-1)} - [G_{bb}]^{-1} g_{\beta} \qquad \tilde{a}^{*(i)} = \tilde{a}^{*(i-1)} - [G_{aa}]^{-1} g_{a^*} \qquad \tilde{\boldsymbol{\theta}}^{(i)} = \tilde{\boldsymbol{\theta}}^{(i-1)} - [G_{\beta a^*}]^{-1} g_{\beta a^*}$$

Once the maximum likelihood estimate of $\boldsymbol{\theta}$ is obtained, its estimated variance-covariance matrix is obtained as follows.

$$\hat{V}\left\{\hat{\boldsymbol{\theta}}_{ML}\right\} = -\left[E\left\{\frac{\partial^2 l}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right\}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{ML}}\right]^{-1} \qquad E\left\{\frac{\partial^2 l}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right\} = \begin{bmatrix} E_{bb} & E_{ba} \\ E_{ba}' & E_{aa} \end{bmatrix}$$

$$\text{where: } E_{bb} = \sum_{i=1}^{n}\left[-\left(\frac{\exp\{a^* + \mathbf{x_i}'\boldsymbol{\beta}\}}{\exp\{\mathbf{x_i}'\boldsymbol{\beta}\} + e^{a^*}}\right)\mathbf{x_i}\mathbf{x_i}'\right] = -\mathbf{X}'\mathbf{A}\mathbf{X} \qquad \mathbf{A} = \text{diag}\left[\frac{\exp\{a^* + \mathbf{x_i}'\boldsymbol{\beta}\}}{\exp\{\mathbf{x_i}'\boldsymbol{\beta}\} + e^{a^*}}\right]$$

$$E_{ba} = \mathbf{0}_{p'}$$

$$E_{aa} = \sum_{i=1}^{n} e^{a^*}\left[\sum_{j=0}^{y_i-1}\frac{j}{(e^{a^*}+j)^2} + a^* + 1 - \frac{e^{a^*}}{e^{a^*} + \exp\{\mathbf{x_i}'\boldsymbol{\beta}\}} - \ln\left(e^{a^*} + \exp\{\mathbf{x_i}'\boldsymbol{\beta}\}\right)\right]$$

$$\Rightarrow \qquad \hat{V}\left\{\hat{\boldsymbol{\theta}}_{ML}\right\} = \begin{bmatrix} (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} & \mathbf{0}_{p'} \\ \mathbf{0}_{p'}' & -\frac{1}{E_{aa}} \end{bmatrix}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{ML}}$$

A conservative test of whether the Poisson Regression is appropriate ($H_0$) versus the Negative Binomial Regression is appropriate ($H_A$) compares -2(log-likelihood) for the 2 models, and compares the difference with the chi-square distribution with 1 degree of freedom.

$$TS: X^2_{obs} = -2\left[lnL_P - lnL_{NB}\right] \qquad RR: X^2_{obs} \geq \chi^2_{\alpha,1}$$

**Example: NASCAR Lead Changes - 1972-1979 Seasons**

In the same NASCAR races described in Crash data example, Lead Changes are also measured. In racing, lead changes typically occur much more frequently than crashes, due to teams having to make pit stops for fuel and tire changes. We first fit the Negative Binomial Regression model relating the number of lead changes ($Y$) to the 3 predictors: numbers of drivers, track length, and laps based on direct computations of the matrix form. Then we fit and compare the Poisson and the Negative Binomial models based on the **glm** and **glm.nb** functions in R, where **glm.nb** is contained in the **MASS** package.

**R Program**

```
race1 <- read.fwf("http://www.stat.ufl.edu/~winner/data/race7579.dat", width=c(8,8,8,8,8,8,8,8,8,12,40),
col.names=c('srace', 'yr', 'yrace', 'n.drivers', 'trklen', 'laps', 'roadtrk',
 'cautions', 'leadchng', 'trkid', 'track'))
race <- data.frame(n.drivers=race1$n.drivers, trklen=race1$trklen, laps=race1$laps,
 cautions=race1$cautions, leadchng = race1$leadchng)
attach(race)

### Direct Computations
n.lc <- length(leadchng)
X <- cbind(rep(1,n.lc), n.drivers, trklen, laps)
Y <- leadchng

### Preliminary estimate of Beta
a.st <- 0
beta.old0 <- matrix(c(1, 0, 0, 0),ncol=1)
beta.diff0 <- 1000
num.iter.b0 <- 0
while (beta.diff0 >= 0.000001) {
num.iter.b0 <- num.iter.b0 + 1
mu0 <- X %*% beta.old0
g.b01 <- (exp(a.st) * (leadchng - exp(mu0)))  / (exp(a.st) + exp(mu0))
g.b0 <- t(X) %*% g.b01
G.bb01 <- -(exp(a.st) + leadchng)*exp(a.st + mu0)/
    ((exp(a.st) + exp(mu0))^2)
G.bb02 <- matrix(rep(0,n.lc^2),ncol=n.lc)
for (i in 1:n.lc) G.bb02[i,i] <- G.bb01[i]
G.bb0 <- t(X) %*% G.bb02 %*% X
beta.new0 <- beta.old0 - solve(G.bb0) %*% g.b0
# print(beta.new0)
beta.diff0 <- t(beta.new0 - beta.old0) %*% (beta.new0 - beta.old0)
beta.old0 <- beta.new0
}
beta.new0
num.iter.b0

### Preliminary estimate of a*
ast.old0<- 0
# beta.a <- matrix(c(1, 0, 0, 0),ncol=1)
beta.a <- beta.new0
ast.diff0 <- 1000
num.iter.ast0 <- 0
while (ast.diff0 >= 0.000001) {
num.iter.ast0 <- num.iter.ast0 + 1
mu0.a <- X %*% beta.a
g.a01 <- rep(0,n.lc)
for (i in 1:n.lc) {
if (leadchng[i] == 0) g.a01[i] <- 0
else for (i1 in 0:(leadchng[i]-1)) g.a01[i] <- g.a01[i] + 1/(exp(ast.old0)+i1)
}
g.a02 <- exp(ast.old0)*(g.a01 + ast.old0 + 1 - ((exp(ast.old0)+leadchng) /
        (exp(ast.old0) + exp(mu0.a))) - log(exp(ast.old0) + exp(mu0.a)))
g.a0 <- sum(g.a02)
G.aa01 <- rep(0,n.lc)
for (i in 1:n.lc) {
if (leadchng[i] == 0) G.aa01[i] <- 0
else for (i1 in 0:(leadchng[i]-1)) G.aa01[i] <- G.aa01[i] +
   i1/((exp(ast.old0)+i1)^2)
}
G.aa02 <- exp(ast.old0)*(G.aa01 + ast.old0 + 2 - ((2*exp(ast.old0)+leadchng) /
        (exp(ast.old0) + exp(mu0.a))) - log(exp(ast.old0) + exp(mu0.a)) +
     ((leadchng-exp(mu0.a))*exp(ast.old0))/((exp(ast.old0) + exp(mu0.a))^2) )
G.aa0 <- sum(G.aa02)
ast.new0 <- ast.old0 - g.a0/G.aa0
ast.diff0 <- t(ast.new0 - ast.old0) %*% (ast.new0 - ast.old0)
```

```
ast.old0 <- ast.new0
print(ast.new0)
}
ast.new0
num.iter.ast0

################ Combined estimation of theta
beta.old <- beta.new0
ast.old <- ast.new0
theta.old <- rbind(beta.old, ast.old)
theta.diff <- 1000
num.iter.theta <- 0
while (theta.diff >= .000000001) {
num.iter.theta <- num.iter.theta + 1
mu <- X %*% beta.old
#  print(mu)
g.b1 <- (exp(ast.old) * (leadchng - exp(mu))) / (exp(ast.old) + exp(mu))
g.b <- t(X) %*% g.b1
G.bb1 <- -(exp(ast.old) + leadchng)*exp(ast.old + mu)/
    ((exp(ast.old) + exp(mu))^2)
G.bb2 <- matrix(rep(0,n.lc^2),ncol=n.lc)
for (i in 1:n.lc) G.bb2[i,i] <- G.bb1[i]
G.bb <- t(X) %*% G.bb2 %*% X
g.a1 <- rep(0,n.lc)
for (i in 1:n.lc) {
if (leadchng[i] == 0) g.a1[i] <- 0
else for (i1 in 0:(leadchng[i]-1)) g.a1[i] <- g.a1[i] + 1/(exp(ast.old)+i1)
}
g.a2 <- exp(ast.old)*(g.a1 + ast.old + 1 - ((exp(ast.old)+leadchng) /
        (exp(ast.old) + exp(mu))) - log(exp(ast.old) + exp(mu)))
g.a <- sum(g.a2)
G.aa1 <- rep(0,n.lc)
for (i in 1:n.lc) {
if (leadchng[i] == 0) G.aa1[i] <- 0
else for (i1 in 0:(leadchng[i]-1)) G.aa1[i] <- G.aa1[i] + i1/((exp(ast.old)+i1)^2)
}
G.aa2 <- exp(ast.old)*(G.aa1 + ast.old + 2 - ((2*exp(ast.old)+leadchng) /
        (exp(ast.old) + exp(mu))) - log(exp(ast.old) + exp(mu)) +
     ((leadchng-exp(mu))*exp(ast.old))/((exp(ast.old) + exp(mu))^2) )
G.aa <- sum(G.aa2)
G.ba1 <- exp(ast.old + mu)*(leadchng - exp(mu)) / ((exp(ast.old) + exp(mu))^2)
G.ba <- t(X) %*% G.ba1
G.theta <- rbind( cbind(G.bb,G.ba) , cbind(t(G.ba),G.aa) )
g.theta <- rbind(g.b , g.a)
theta.new <- theta.old - solve(G.theta) %*% g.theta
theta.diff <- t(theta.new - theta.old) %*% (theta.new - theta.old)
theta.old <- theta.new
#  print(theta.old)
beta.old <- theta.old[1:ncol(X)]
ast.old <- theta.old[ncol(X)+1]
#  print(beta.old)
# print(ast.old)
}
theta.new
num.iter.theta

### Plug in ML estimates for Variance and Standard Error Estimates
theta.old <- theta.new
#  print(theta.old)
beta.old <- theta.old[1:ncol(X)]
ast.old <- theta.old[ncol(X)+1]
mu <- X %*% beta.old
#  print(mu)
g.b1 <- (exp(ast.old) * (leadchng - exp(mu))) / (exp(ast.old) + exp(mu))
```

```
g.b <- t(X) %*% g.b1
G.bb1 <- -(exp(ast.old) + leadchng)*exp(ast.old + mu)/
    ((exp(ast.old) + exp(mu))^2)
G.bb2 <- matrix(rep(0,n.lc^2),ncol=n.lc)
for (i in 1:n.lc) G.bb2[i,i] <- G.bb1[i]
G.bb <- t(X) %*% G.bb2 %*% X
g.a1 <- rep(0,n.lc)
for (i in 1:n.lc) {
if (leadchng[i] == 0) g.a1[i] <- 0
else for (i1 in 0:(leadchng[i]-1)) g.a1[i] <- g.a1[i] + 1/(exp(ast.old)+i1)
}
g.a2 <- exp(ast.old)*(g.a1 + ast.old + 1 - ((exp(ast.old)+leadchng) /
        (exp(ast.old) + exp(mu))) - log(exp(ast.old) + exp(mu)))
g.a <- sum(g.a2)
G.aa1 <- rep(0,n.lc)
for (i in 1:n.lc) {
if (leadchng[i] == 0) G.aa1[i] <- 0
else for (i1 in 0:(leadchng[i]-1)) G.aa1[i] <- G.aa1[i] + i1/((exp(ast.old)+i1)^2)
}
G.aa2 <- exp(ast.old)*(G.aa1 + ast.old + 2 - ((2*exp(ast.old)+leadchng) /
        (exp(ast.old) + exp(mu))) - log(exp(ast.old) + exp(mu)) +
     ((leadchng-exp(mu))*exp(ast.old))/((exp(ast.old) + exp(mu))^2) )
G.aa <- sum(G.aa2)
G.ba1 <- exp(ast.old + mu)*(leadchng - exp(mu)) / ((exp(ast.old) + exp(mu))^2)
G.ba <- t(X) %*% G.ba1
G.theta <- rbind( cbind(G.bb,G.ba) , cbind(t(G.ba),G.aa) )
g.theta <- rbind(g.b , g.a)

d2l.bb.1 <- -(exp(ast.old + mu))/(exp(ast.old)+exp(mu))
d2l.bb.2 <- matrix(rep(0,n.lc^2),ncol=n.lc)
for (i in 1:n.lc) d2l.bb.2[i,i] <- d2l.bb.1[i]
d2l.bb <- t(X) %*% d2l.bb.2 %*% X
sqrt(diag(solve(-d2l.bb)))
d2l.ba1 <- rep(0,n.lc)
d2l.ba <- t(X) %*% d2l.ba1
d2l.aa1 <- exp(ast.old)*(G.aa1 + ast.old + 1 - (exp(ast.old) /
     (exp(ast.old) + exp(mu))) - log(exp(ast.old) + exp(mu)))
d2l.aa <- sum(d2l.aa1)
d2l.tt <- rbind( cbind(d2l.bb , d2l.ba) , cbind(t(d2l.ba), d2l.aa) )

SE.nb <- sqrt(diag(solve(-d2l.tt)))
SE.beta.nb <- SE.nb[1:ncol(X)]
z.beta.nb <- beta.old / SE.beta.nb
pv.beta.nb <- 2*(1-pnorm(abs(z.beta.nb)))

nb.est <- cbind(beta.old, SE.beta.nb, z.beta.nb, pv.beta.nb,
beta.old - 1.96 * SE.beta.nb,
beta.old + 1.96 * SE.beta.nb)
colnames(nb.est) <- c("Estimate","NB SE","NB z","Pr(>|z|)","LL","UL")
print(round(nb.est,6))

alpha.est <- cbind(ast.old, exp(ast.old))
colnames(alpha.est) <- c("log(alpha^(-1))", "alpha^(-1)")
alpha.est
```

### R Output

```
> beta.new0
                 [,1]
          -0.525373491
n.drivers  0.062055478
```

```
trklen     0.492484225
laps       0.001656048
> num.iter.b0
[1] 6

> ast.new0
[1] 1.657059
> num.iter.ast0
[1] 4

> theta.new
                  [,1]
          -0.503775021
n.drivers  0.059689981
trklen     0.515298326
laps       0.001742154
ast.old    1.657923341
> num.iter.theta
[1] 3

> print(round(nb.est,6))
           Estimate    NB SE      NB z Pr(>|z|)       LL       UL
          -0.503775 0.448626 -1.122928 0.261468 -1.383083 0.375533
n.drivers  0.059690 0.013602  4.388283 0.000011  0.033030 0.086350
trklen     0.515298 0.178934  2.879831 0.003979  0.164589 0.866008
laps       0.001742 0.000867  2.009160 0.044520  0.000043 0.003442

> alpha.est
     log(alpha^(-1)) alpha^(-1)
[1,]        1.657923     5.2484
```

The model fit based on the **glm** (Poisson and quasipoisson) and **glm.nb** (Negative Binomial) functions is given below, as well as a test between the the Poisson and Negative Binomial models.

```
### Program

race1 <- read.fwf("http://www.stat.ufl.edu/~winner/data/race7579.dat", width=c(8,8,8,8,8,8,8,8,8,12,40),
col.names=c('srace', 'yr', 'yrace', 'n.drivers', 'trklen', 'laps', 'roadtrk',
 'cautions', 'leadchng', 'trkid', 'track'))
race <- data.frame(n.drivers=race1$n.drivers, trklen=race1$trklen, laps=race1$laps,
 cautions=race1$cautions, leadchng = race1$leadchng)
attach(race)

race.mod <- glm(leadchng ~ n.drivers + trklen + laps, family=poisson("log"))
summary(race.mod)
anova(race.mod, test="Chisq")
muhat <- predict(race.mod, type="response")
(pearson.x2 <- sum((leadchng - muhat)^2/muhat))
(pearson.x2a <- sum(resid(race.mod,type="pearson")^2))
phi <- pearson.x2a / df.residual(race.mod)
round(c(phi,sqrt(phi)),4)
qchisq(.95,df.residual(race.mod))
deviance(race.mod)

race.mod1 <- glm(leadchng ~ n.drivers + trklen + laps, family=quasipoisson)
summary(race.mod1)

library(MASS)
race.mod2 <- glm.nb(leadchng ~ n.drivers + trklen + laps)
summary(race.mod2)
```

```
X2.p.nb <- -2*(logLik(race.mod) - logLik(race.mod2))
X2.p.nb

### Output

> summary(race.mod)
Call:
glm(formula = leadchng ~ n.drivers + trklen + laps, family = poisson("log"))
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.490253   0.217796  -2.251   0.0244 *
n.drivers    0.051612   0.005679   9.089  < 2e-16 ***
trklen       0.610414   0.082949   7.359 1.85e-13 ***
laps         0.002138   0.000415   5.152 2.58e-07 ***

(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 1388.59  on 150  degrees of freedom
Residual deviance:  687.61  on 147  degrees of freedom
AIC: 1395.2

> anova(race.mod, test="Chisq")
Analysis of Deviance Table
Model: poisson, link: log
Response: leadchng
Terms added sequentially (first to last)
          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                        150    1388.59
n.drivers  1   627.31       149     761.28 < 2.2e-16 ***
trklen     1    46.94       148     714.34 7.308e-12 ***
laps       1    26.73       147     687.61 2.343e-07 ***

> (pearson.x2 <- sum((leadchng - muhat)^2/muhat))
[1] 655.6059
> (pearson.x2a <- sum(resid(race.mod,type="pearson")^2))
[1] 655.6059
> phi <- pearson.x2a / df.residual(race.mod)
> round(c(phi,sqrt(phi)),4)
[1] 4.4599 2.1118

> summary(race.mod1)
Call:
glm(formula = leadchng ~ n.drivers + trklen + laps, family = quasipoisson)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.4902529  0.4599522  -1.066  0.28823
n.drivers    0.0516117  0.0119924   4.304 3.05e-05 ***
trklen       0.6104140  0.1751757   3.485  0.00065 ***
laps         0.0021381  0.0008764   2.440  0.01590 *

(Dispersion parameter for quasipoisson family taken to be 4.459904)
    Null deviance: 1388.59  on 150  degrees of freedom
Residual deviance:  687.61  on 147  degrees of freedom
AIC: NA

> summary(race.mod2)
Call:
glm.nb(formula = leadchng ~ n.drivers + trklen + laps, init.theta = 5.248400381,
    link = log)
 Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.5037750  0.4486264  -1.123  0.26147
n.drivers    0.0596900  0.0136021   4.388 1.14e-05 ***
trklen       0.5152983  0.1789336   2.880  0.00398 **
laps         0.0017422  0.0008671   2.009  0.04452 *
```

```
(Dispersion parameter for Negative Binomial(5.2484) family taken to be 1)
    Null deviance: 308.06  on 150  degrees of freedom
Residual deviance: 162.80  on 147  degrees of freedom
AIC: 1098.1

Number of Fisher Scoring iterations: 1
              Theta:  5.248
          Std. Err.:  0.809
 2 x log-likelihood:  -1088.054
>
> X2.p.nb <- -2*(logLik(race.mod) - logLik(race.mod2))
> X2.p.nb
'log Lik.' 299.1397

### Plot of mean-variance for qp and nb models

pred.nb <- predict(race.mod2,type="response")
grp.lc <- cut(pred.nb, breaks=quantile(pred.nb,seq(0,100,100/10)/100))
mean.lc <- tapply(leadchng,grp.lc,mean)
var.lc <- tapply(leadchng,grp.lc,var)

plot(mean.lc, var.lc, xlab="Mean", ylab="Variance",
 main="Mean-Variance Relationship - Lead Change Data")

x.lc <- seq(2,50.1)
lines(x.lc, phi*x.lc, lty=2)
lines(x.lc, x.lc+x.lc^2/race.mod2$theta, lty=1)
legend("topleft",lty=c(2,1),legend=c("quasipoisson", "Negative Binomial"))
```

Clearly the Negative Binomial fits better than the Poisson ($X^2_{obs}$=299.1, df=1). The relationship between the variance and the mean for the quasipoisson and the Negative Binomial models are obtained as follow.

quasipoisson: $\hat{V}\left\{\hat{\mu}_i\right\} = \phi\hat{\mu}_i = 4.46\hat{\mu}_i$ Negative Binomial: $\hat{V}\left\{\hat{\mu}_i\right\} = \hat{\mu}_i + \frac{\hat{\mu}_i^2}{\theta} = \hat{\mu}_i + \frac{\hat{\mu}_i^2}{5.25} = \hat{\mu}_i + 0.19\hat{\mu}_i^2$

Figure 9.2 is based on 10 "groups" of races and plots the predicted and observed variances versus means for the two models. It is not clear whether one model is better than the other based on the plot.

$\nabla$

## 9.5 Gamma Regression

When data are purely positive, and are right-skewed, they are frequently modeled as following a gamma distribution. Regression models can be fit utilizing this distribution, with potential link functions being the identity, inverse, or log. That is, we could model the mean as a function of the predictors as follow.

$$g_1(\mu) = \mu = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \qquad g_2(\mu) = \frac{1}{\mu} \qquad g_3(\mu) = \log(\mu)$$

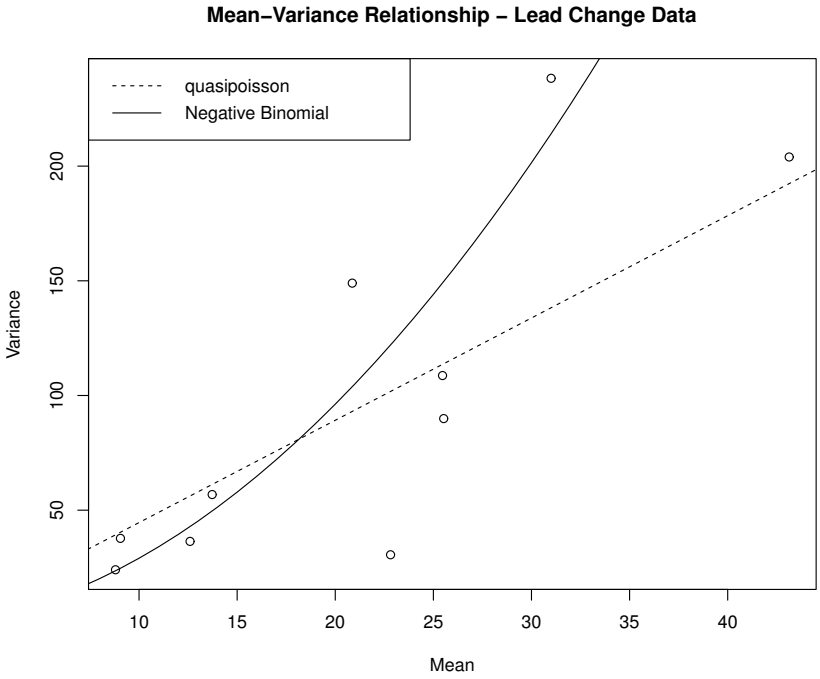


Figure 9.2: Variance versus Mean - Grouped Lead Change Race Data

The density function for the Gamma family can be parameterized in various ways. The parameterization with $E\{Y\} = \alpha\beta$ is used here.

$$f(y_i|\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} y_i^{\alpha-1} \exp\left\{-\frac{y_i}{\beta}\right\} \qquad y_i > 0; \quad \alpha,\beta > 0$$

Letting $\mu = \alpha\beta$, $\phi = 1/\alpha$, and $\beta = \mu\phi$, the following parameterization is obtained for the likelihood function for observation $i$.

$$g(\mu_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} = \mathbf{x_i}'\boldsymbol{\beta}$$

$$L_i = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} y_i^{\alpha-1} \exp\left\{-\frac{y_i}{\beta}\right\} = \frac{1}{y_i\Gamma\left(\frac{1}{\phi}\right)} \left(\frac{y_i}{\mu_i\phi}\right)^{1/\phi} \exp\left\{-\frac{y_i}{\mu_i\phi}\right\}$$

### 9.5.1 Estimating Model Parameters

The likelihood and log-likelihood under the inverse link function, where $\mu_i = 1/\mathbf{x_i}'\boldsymbol{\beta}$ are given below.

$$L_i = \frac{1}{y_i \Gamma\left(\frac{1}{\phi}\right)} \left(\frac{y_i \mathbf{x_i}'\boldsymbol{\beta}}{\phi}\right)^{1/\phi} \exp\left\{-\frac{y_i \mathbf{x_i}'\boldsymbol{\beta}}{\phi}\right\}$$

$$l_i = \ln L_i = -\ln(y_i) - \ln\left(\Gamma\left(\frac{1}{\phi}\right)\right) + \frac{1}{\phi}\left[\ln(y_i) + \ln(\mathbf{x_i}'\boldsymbol{\beta}) - \ln(\phi)\right] - \frac{y_i \mathbf{x_i}'\boldsymbol{\beta}}{\phi}$$

Taking the necessary derivatives of $l_i$ with respect to $\boldsymbol{\beta}$, we obtain the following results.

$$\frac{\partial l_i}{\partial \boldsymbol{\beta}} = \frac{1}{\phi}\left[\frac{1}{\mathbf{x_i}'\boldsymbol{\beta}}\mathbf{x_i} - y_i\mathbf{x_i}\right] = \frac{1}{\phi}\left[\frac{1}{\mathbf{x_i}'\boldsymbol{\beta}} - y_i\right]\mathbf{x_i} \qquad \Rightarrow \qquad g(\boldsymbol{\beta}) = \sum_{i=1}^{n}\frac{\partial l_i}{\partial \boldsymbol{\beta}} = \frac{1}{\phi}\sum_{i=1}^{n}\left[\frac{1}{\mathbf{x_i}'\boldsymbol{\beta}} - y_i\right]\mathbf{x_i}$$

$$\frac{\partial^2 l_i}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} = -\frac{1}{\phi}\frac{1}{(\mathbf{x_i}'\boldsymbol{\beta})^2}\mathbf{x_i}\mathbf{x_i}' \quad \Rightarrow \quad G(\boldsymbol{\beta}) = \sum_{i=1}^{n}\frac{\partial^2 l_i}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} = -\frac{1}{\phi}\sum_{i=1}^{n}\frac{1}{(\mathbf{x_i}'\boldsymbol{\beta})^2}\mathbf{x_i}\mathbf{x_i}' = -\frac{1}{\phi}\mathbf{X}'\mathbf{W}\mathbf{X} \qquad \mathbf{W} = \operatorname{diag}\left\{\frac{1}{(\mathbf{x_i}'\boldsymbol{\beta})^2}\right\}$$

The iterative procedure for the Newton-Raphson algorithm goes as follows.

$$\tilde{\boldsymbol{\beta}}^{\text{New}} = \tilde{\boldsymbol{\beta}}^{\text{Old}} - \left[G\left(\tilde{\boldsymbol{\beta}}^{\text{Old}}\right)\right]^{-1} g\left(\tilde{\boldsymbol{\beta}}^{\text{Old}}\right)$$

For starting values, set $\beta_1 = \cdots = \beta_p = 0$ and $\beta_0 = 1/\overline{Y}$. Then the procedure is iterated to convergence with the ML estimate and its estimated variance-covariance matrix.

$$\tilde{\boldsymbol{\beta}}^0 = \begin{bmatrix} 1/\overline{Y} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \hat{V}\left\{\hat{\boldsymbol{\beta}}\right\} = \left[E\left\{-G(\boldsymbol{\beta})\right\}\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}\right]^{-1} = \phi\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}$$

The likelihood and log-likelihood under the log link, where $\mu_i = \exp\{\mathbf{x_i}'\boldsymbol{\beta}\}$ is given below.

$$L_i = \frac{1}{y_i \Gamma\left(\frac{1}{\phi}\right)} \left(\frac{y_i}{\exp\{\mathbf{x_i}'\boldsymbol{\beta}\}\phi}\right)^{1/\phi} \exp\left\{-\frac{y_i}{\exp\{\mathbf{x_i}'\boldsymbol{\beta}\}\phi}\right\} = \frac{1}{y_i \Gamma\left(\frac{1}{\phi}\right)} \left(\frac{y_i \exp\{-\mathbf{x_i}'\boldsymbol{\beta}\}}{\phi}\right)^{1/\phi} \exp\left\{-\frac{y_i \exp\{-\mathbf{x_i}'\boldsymbol{\beta}\}}{\phi}\right\}$$

$$l_i = \ln L_i = -\ln(y_i) - \ln\left(\Gamma\left(\frac{1}{\phi}\right)\right) + \frac{1}{\phi}\left[\ln(y_i) - \mathbf{x_i}'\boldsymbol{\beta} - \ln(\phi)\right] - \frac{y_i \exp\{-\mathbf{x_i}'\boldsymbol{\beta}\}}{\phi}$$

Taking the necessary derivatives of $l_i$ with respect to $\boldsymbol{\beta}$, we obtain the following results.

$$\frac{\partial l_i}{\partial \boldsymbol{\beta}} = \frac{1}{\phi}\left[y_i \exp\left\{-\mathbf{x_i}'\boldsymbol{\beta}\right\} - 1\right]\mathbf{x_i} \qquad \Rightarrow \qquad g\left(\boldsymbol{\beta}\right) = \frac{1}{\phi}\sum_{i=1}^{n}\left[y_i \exp\left\{-\mathbf{x_i}'\boldsymbol{\beta}\right\} - 1\right]\mathbf{x_i}$$

$$\frac{\partial^2 l_i}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} = -\frac{1}{\phi}y_i \exp\left\{-\mathbf{x_i}'\boldsymbol{\beta}\right\}\mathbf{x_i}\mathbf{x_i}' \qquad \Rightarrow \qquad G\left(\boldsymbol{\beta}\right) = -\frac{1}{\phi}\sum_{i=1}^{n} y_i \exp\left\{-\mathbf{x_i}'\boldsymbol{\beta}\right\}\mathbf{x_i}\mathbf{x_i}' = -\frac{1}{\phi}\mathbf{X}'\mathbf{W}\mathbf{X}$$

$$\mathbf{W} = \operatorname{diag}\left\{y_i \exp\left\{-\mathbf{x_i}'\boldsymbol{\beta}\right\}\right\} \qquad E\left\{-G\left(\boldsymbol{\beta}\right)\right\} = \frac{1}{\phi}\sum_{i=1}^{n}\mu_i \exp\left\{-\mathbf{x_i}'\boldsymbol{\beta}\right\}\mathbf{x_i}\mathbf{x_i}' = \frac{1}{\phi}\sum_{i=1}^{n}\mathbf{x_i}\mathbf{x_i}' = \frac{1}{\phi}\mathbf{X}'\mathbf{X}$$

The iterative procedure for the Newton-Raphson algorithm goes as follows.

$$\tilde{\boldsymbol{\beta}}^{\text{New}} = \tilde{\boldsymbol{\beta}}^{\text{Old}} - \left[G\left(\tilde{\boldsymbol{\beta}}^{\text{Old}}\right)\right]^{-1} g\left(\tilde{\boldsymbol{\beta}}^{\text{Old}}\right)$$

For starting values, set $\beta_1 = \cdots = \beta_p = 0$ and $\beta_0 = \ln\left(\overline{Y}\right)$. Then the procedure is iterated to convergence with the ML estimate and its estimated variance-covariance matrix.

$$\tilde{\boldsymbol{\beta}}^0 = \begin{bmatrix} \ln\left(\overline{Y}\right) \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \hat{V}\left\{\hat{\boldsymbol{\beta}}\right\} = \left[E\left\{-G\left(\boldsymbol{\beta}\right)\right\}\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}\right]^{-1} = \phi\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

The parameter $\phi$ can be estimated by the method of moments as follows.

$$E\left\{Y_i\right\} = \mu_i \qquad \Rightarrow \qquad E\left\{\frac{Y_i}{\mu_i}\right\} = 1 \qquad \Rightarrow \qquad E\left\{\frac{Y_i - \mu_i}{\mu_i}\right\} = 0$$

$$V\left\{Y_i\right\} = \phi\mu_i^2 \qquad \Rightarrow \qquad V\left\{\frac{Y_i}{\mu_i}\right\} = \phi \qquad \Rightarrow \qquad V\left\{\frac{Y_i - \mu_i}{\mu_i}\right\} = \phi$$

$$\Rightarrow \qquad E\left\{\left(\frac{Y_i - \mu_1}{\mu_i}\right)^2\right\} = \phi \qquad \Rightarrow \qquad \hat{\phi} = \frac{1}{n-p'}\sum_{i=1}^{n}\left(\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}\right)^2$$

For the inverse and log links we have the following estimated means.

$$\text{Inverse Link:} \quad \hat{\mu}_i = \frac{1}{\mathbf{x_i}'\hat{\boldsymbol{\beta}}} \qquad\qquad \text{Log Link:} \quad \hat{\mu}_i = \exp\left\{\mathbf{x_i}'\hat{\boldsymbol{\beta}}\right\}$$

### 9.5.2 Inferences for Model Parameters and Goodness of Fit Test

Wald tests for the individual regression coefficients are obtained by obtaining the ratio between the point estimates and their estimated standard errors ($z$-tests), or equivalently the square of the ratio ($\chi^2$-tests).

$$H_0 : \beta_k = 0 \qquad H_A : \beta_k \neq 0 \qquad TS : z_{obs} = \frac{\hat{\beta}_k}{\hat{SE}\left\{\hat{\beta}_k\right\}} \qquad RR : |z_{obs}| \geq z_{\alpha/2}$$

$$H_0 : \beta_k = 0 \qquad H_A : \beta_k \neq 0 \qquad TS : X^2_{obs} = \left(\frac{\hat{\beta}_k}{\hat{SE}\left\{\hat{\beta}_k\right\}}\right)^2 \qquad RR : X^2_{obs} \geq \chi^2_{\alpha,1}$$

Confidence Intervals for the individual regression coefficients can be obtained as follow.

$$(1-\alpha)100\% \text{ Confidence Interval for } \beta_k : \qquad \hat{\beta}_k \pm z_{\alpha/2}\hat{SE}\left\{\hat{\beta}_k\right\}$$

To obtain the Likelihood Ratio test for comparison of two nested models, first define the Deviance and Scaled Deviance for a given model. The following results are obtained.

$$\text{Deviance} = DEV = -2\sum_{i=1}^{n}\left[\ln\left(\frac{y_i}{\hat{\mu}_i}\right) - \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}\right]$$

$$\text{Scaled Deviance} = -2\left[l\left(\hat{\mu}, \phi, y\right) - l\left(y, \phi, y\right)\right] \qquad \text{where:}$$

$$l\left(\hat{\mu}, \phi, y\right) = -\sum_{i=1}^{n} y_i - n\ln\left(\Gamma\left(\frac{1}{\phi}\right)\right) + \frac{1}{\phi}\left[\sum_{i=1}^{n}\ln\left(y_i\right) - \sum_{i=1}^{n}\ln\left(\hat{\mu}_i\right) - n\ln(\phi)\right] - \frac{1}{\phi}\sum_{i=1}^{n}\frac{y_i}{\hat{\mu}_i}$$

$$l\left(y, \phi, y\right) = -\sum_{i=1}^{n} y_i - n\ln\left(\Gamma\left(\frac{1}{\phi}\right)\right) + \frac{1}{\phi}\left[\sum_{i=1}^{n}\ln\left(y_i\right) - \sum_{i=1}^{n}\ln\left(y_i\right) - n\ln(\phi)\right] - \frac{1}{\phi}\sum_{i=1}^{n}\frac{y_i}{y_i}$$

$$\Rightarrow \qquad l\left(\hat{\mu}, \phi, y\right) - l\left(y, \phi, y\right) = \frac{1}{\phi}\left[-\sum_{i=1}^{n}\ln\left(\hat{\mu}_i\right) + \sum_{i=1}^{n}\ln\left(y_i\right) - \sum_{i=1}^{n}\frac{y_i}{\hat{\mu}_i} + \sum_{i=1}^{n}\frac{y_i}{y_i}\right] = \frac{1}{\phi}\left[\sum_{i=1}^{n}\ln\left(\frac{y_i}{\hat{\mu}_i}\right) - \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}\right]$$

To test $H_0 : \beta_{k+1} \ldots \beta_p = 0$, construct the full model with all $p$ predictors, the reduced model with $k < p$ predictors, and compute the deviance for each model: $DEV_F$ and $DEV_R$. Then the test is conducted as follows.

$$H_0 : \beta_{k+1} \ldots \beta_p = 0 \qquad TS : X^2_{obs} = \frac{DEV_R - DEV_F}{\hat{\phi}_F} \qquad RR : X^2_{obs} \geq \chi^2_{\alpha,p-k} \qquad P = P\left\{\chi^2_{p-k} \geq X^2_{obs}\right\}$$

To test whether the current model is correct, the test statistic is obtained by dividing the Deviance of the model by $\hat{\phi}$ and comparing that with the chi-square distribution with $n - p'$ degrees of freedom.

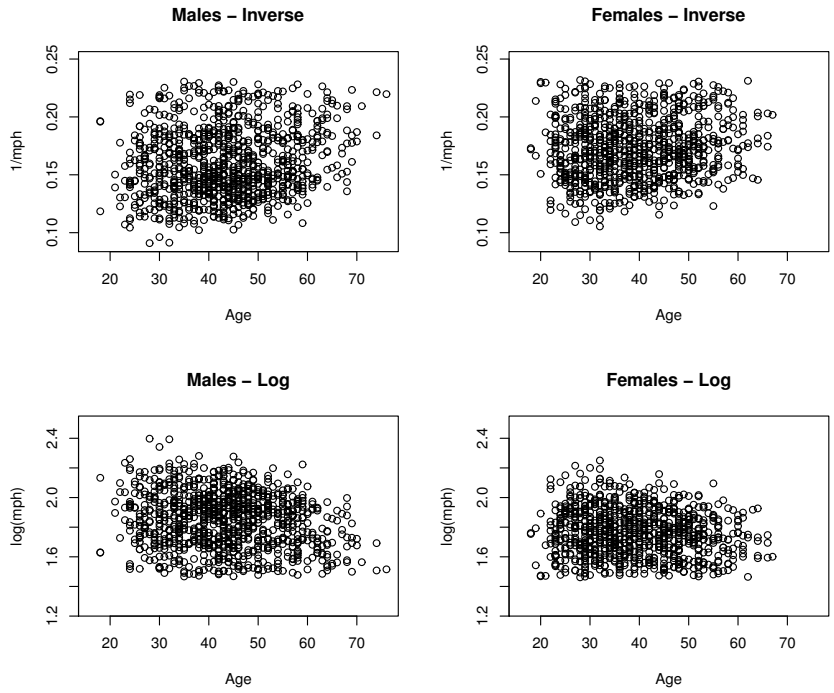


Figure 9.3: Plot of Inverse and Log MPH by Age by Gender - 2015 NAPA Marathon

$$TS: X^2_{GOF} = \frac{DEV}{\hat{\phi}} \qquad RR: X^2_{GOF} \geq \chi^2_{\alpha, n-p'} \qquad P = P\left\{\chi^2_{n-p'} \geq X^2_{GOF}\right\}$$

**Example: Napa Marathon Velocities - 2015**

The Napa Valley marathon in 2015 had 977 Males and 905 Females complete the 26.2 mile race. Consider a model relating runners' speeds in miles per hour ($Y$=mph) to Gender ($X_1 = 1$ if Male, 0 if Female), Age ($X_2$, in Years), and an interaction term ($X_1X_2$, allowing for different slopes with respect to age for Males and Females. Figure 9.3 plots the reciprocal of mph and the log of mph separately for Males and Females. Both the inverse link and log link models are fit in matrix form, then using the **glm** function in R. Note that the "default" link for the gamma regression model in the glm function is the inverse link.

The model with the Inverse link is fit first the R Program and Output are given below.

```
napaf2015 <- read.csv("http://www.stat.ufl.edu/~winner/data/napa_marathon_fm2015.csv",
header=T)
attach(napaf2015); names(napaf2015)
#### Matrix form - Inverse Link

n.napa <- length(mph)
Y <- mph
```

```
X0 <- rep(1,n.napa)
male <- rep(0,n.napa)
for (i in 1:n.napa) {
if (Gender[i] == "M") male[i] <- 1
}
Age.male <- Age*male
X <- cbind(X0,Age,male,Age.male)
beta.old <- matrix(rep(0,ncol(X)),ncol=1)
beta.old[1]  <- 1/mean(Y)
beta.diff <- 10000
num.iter <- 0

while (beta.diff > 0.000001) {
num.iter <- num.iter + 1
mu <- 1/(X%*%beta.old)
g_beta <- t(X) %*% (mu - Y)
W <- matrix(rep(0,n.napa^2),ncol=n.napa)
for (i in 1:n.napa) W[i,i] <- mu[i]^2
G_beta <- -t(X) %*% W %*% X
beta.new <- beta.old - solve(G_beta) %*% g_beta
beta.diff <- t(beta.new-beta.old) %*% (beta.new-beta.old)
beta.old <- beta.new
}
num.iter
beta.new
beta.old <- beta.new
mu <- 1/(X%*%beta.old)
phi.hat <- (1/(n.napa))*(sum(((Y-mu)/mu)^2))
W <- matrix(rep(0,n.napa^2),ncol=n.napa)
for (i in 1:n.napa) W[i,i] <- mu[i]^2
V_beta <- phi.hat * solve(t(X) %*% W %*% X)
SE_beta <- sqrt(diag(V_beta))
z_beta <- beta.new/SE_beta
p_z_beta <- 2*(1-pnorm(abs(z_beta)))
beta.out1 <- cbind(beta.new, SE_beta, z_beta, p_z_beta)
colnames(beta.out1) <- c("Estimate", "Std. Error", "z", "Pr(>|z|)")
round(beta.out1,6)

phi.hat
deviance <- -2*sum(log(Y/mu) - (Y-mu)/mu)
deviance

mu0 <- rep(mean(Y),n.napa)
deviance.0 <- -2*sum(log(Y/mu0) - (Y-mu0)/mu0)
deviance.0
X2.obs <- (deviance.0-deviance)/phi.hat
p.X2.obs <- 1 - pchisq(X2.obs,3)
LR.out1 <- cbind(X2.obs, p.X2.obs)
colnames(LR.out1) <- c("LR Stat", "P(>LR)")
LR.out1

X2.GOF <- deviance/phi.hat
p.X2.GOF <- 1-pchisq(X2.GOF,n.napa-ncol(X))
GOF.out1 <- cbind(X2.GOF, p.X2.GOF)
colnames(GOF.out1) <- c("GOF Stat", "P(>GOF)")
GOF.out1

### Output
> round(beta.out1,6)
          Estimate Std. Error         z Pr(>|z|)
X0        0.157079   0.003707 42.376417 0.000000
Age       0.000302   0.000093  3.264476 0.001097
male     -0.022808   0.005062 -4.505580 0.000007
Age.male  0.000175   0.000121  1.444146 0.148698
```

```
> phi.hat
[1] 0.02872085
> deviance <- -2*sum(log(Y/mu) - (Y-mu)/mu)
> deviance
[1] 53.6987
>
> deviance.0
[1] 58.58576
> LR.out1
      LR Stat P(>LR)
[1,] 170.1571      0
>
> GOF.out1
     GOF Stat   P(>GOF)
[1,] 1869.677 0.5497886
```

In terms of the individual regression coefficients, Age and male are highly significant, while the interaction term is not significant. An additive model with Age and male appears to be appropriate, and will be fit with the **glm** function. The Likelihood ratio test of whether any of the 3 predictors are related to inverse mph, yields a a chi-square statistic of 170.1571 with 3 degrees of freedom, with a p-value of 0. The Goodness of Fit test ($X^2 = 1869.677$, df=1882-4=1878, p=.5498) does not reject the hypothesis that the gamma model is correct.

The model fit on the log link function is given below. It gives very similar conclusions to the inverse link model.

```
#### Matrix form - Log Link

n.napa <- length(mph)
Y <- mph
X0 <- rep(1,n.napa)
male <- rep(0,n.napa)
for (i in 1:n.napa) {
if (Gender[i] == "M") male[i] <- 1
}
Age.male <- Age*male
X <- cbind(X0,Age,male,Age.male)
beta.old <- matrix(rep(0,ncol(X)),ncol=1)
beta.old[1]  <- log(mean(Y))
beta.diff <- 10000
num.iter <- 0

while (beta.diff > 0.000001) {
num.iter <- num.iter + 1
mu <- exp(X%*%beta.old)
g_beta <- t(X) %*% (Y/mu - X0)
W <- matrix(rep(0,n.napa^2),ncol=n.napa)
for (i in 1:n.napa) W[i,i] <- Y[i]/mu[i]
G_beta <- -t(X) %*% W %*% X
beta.new <- beta.old - solve(G_beta) %*% g_beta
beta.diff <- t(beta.new-beta.old) %*% (beta.new-beta.old)
beta.old <- beta.new
}
num.iter
beta.new
beta.old <- beta.new
mu <- exp(X%*%beta.old)
phi.hat <- (1/(n.napa-ncol(X)))*(sum(((Y-mu)/mu)^2))
```

```
W <- matrix(rep(0,n.napa^2),ncol=n.napa)
for (i in 1:n.napa) W[i,i] <- mu[i]^2
V_beta <- phi.hat * solve(t(X) %*%  X)
SE_beta <- sqrt(diag(V_beta))
z_beta <- beta.new/SE_beta
p_z_beta <- 2*(1-pnorm(abs(z_beta)))
beta.out2 <- cbind(beta.new, SE_beta, z_beta, p_z_beta)
colnames(beta.out2) <- c("Estimate", "Std. Error", "z", "Pr(>|z|)")
round(beta.out2,6)

phi.hat
deviance <- -2*sum(log(Y/mu) - (Y-mu)/mu)
deviance

mu0 <- rep(mean(Y),n.napa)
deviance.0 <- -2*sum(log(Y/mu0) - (Y-mu0)/mu0)
deviance.0
X2.obs <- (deviance.0-deviance)/phi.hat
p.X2.obs <- 1 - pchisq(X2.obs,3)
LR.out2 <- cbind(X2.obs, p.X2.obs)
colnames(LR.out2) <- c("LR Stat", "P(>LR)")
LR.out2

X2.GOF <- deviance/phi.hat
p.X2.GOF <- 1-pchisq(X2.GOF,n.napa-ncol(X))
GOF.out2 <- cbind(X2.GOF, p.X2.GOF)
colnames(GOF.out2) <- c("GOF Stat", "P(>GOF)")
GOF.out2

### Output
> round(beta.out2,6)
          Estimate Std. Error         z Pr(>|z|)
X0        1.849418   0.022058 83.843194 0.000000
Age      -0.001812   0.000546 -3.317099 0.000910
male      0.152194   0.031508  4.830278 0.000001
Age.male -0.001339   0.000742 -1.803920 0.071244
>
> phi.hat
[1] 0.02876127
> deviance <- -2*sum(log(Y/mu) - (Y-mu)/mu)
> deviance
[1] 53.66784
>
> mu0 <- rep(mean(Y),n.napa)
> deviance.0 <- -2*sum(log(Y/mu0) - (Y-mu0)/mu0)
> deviance.0
[1] 58.58576
> X2.obs <- (deviance.0-deviance)/phi.hat
> p.X2.obs <- 1 - pchisq(X2.obs,3)
> LR.out2 <- cbind(X2.obs, p.X2.obs)
> colnames(LR.out2) <- c("LR Stat", "P(>LR)")
> LR.out2
      LR Stat P(>LR)
[1,] 170.9909      0
>
> X2.GOF <- deviance/phi.hat
> p.X2.GOF <- 1-pchisq(X2.GOF,n.napa-ncol(X))
> GOF.out2 <- cbind(X2.GOF, p.X2.GOF)
> colnames(GOF.out2) <- c("GOF Stat", "P(>GOF)")
> GOF.out2
     GOF Stat   P(>GOF)
[1,] 1865.976 0.5736719
```

The R Program and Output for the Additive and Interaction models are given below.

```
gender <- factor(Gender)

par(mfrow=c(2,2))
plot(Age[gender=="M"],1/mph[gender=="M"],xlab="Age",ylab="1/mph",
main="Males - Inverse",xlim=c(16,76),ylim=c(0.09,0.25))
plot(Age[gender=="F"],1/mph[gender=="F"],xlab="Age",ylab="1/mph",
main="Females - Inverse",xlim=c(16,76),ylim=c(0.09,0.25))
plot(Age[gender=="M"],log(mph[gender=="M"]),xlab="Age",ylab="log(mph)",
main="Males - Log",xlim=c(16,76),ylim=c(1.25,2.50))
plot(Age[gender=="F"],log(mph[gender=="F"]),xlab="Age",ylab="log(mph)",
main="Females - Log",xlim=c(16,76),ylim=c(1.25,2.50))
par(mfrow=c(1,1))

napa.mod5 <- glm(mph~Age + gender,family=Gamma)
summary(napa.mod5)
napa.mod6 <- glm(mph ~ Age + gender, family=Gamma(link="log"))
summary(napa.mod6)
napa.mod7 <- glm(mph~Age*gender,family=Gamma)
summary(napa.mod7)
napa.mod8 <- glm(mph ~ Age*gender, family=Gamma(link="log"))
summary(napa.mod8)

age1 <- min(Age):max(Age)
yhat.F <- exp(1.8494178 - 0.0018116*age1)
yhat.M <- exp((1.8494178+0.1521938) - (0.0018116+0.0013388)*age1)

plot(Age,mph,col=gender)
lines(age1,yhat.F,col=1)
lines(age1,yhat.M,col=2)

anova(napa.mod5,napa.mod7,test="Chisq")
anova(napa.mod6,napa.mod8,test="Chisq")

### R Output
> summary(napa.mod5)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.531e-01  2.496e-03   61.34  < 2e-16 ***
Age          4.048e-04  5.988e-05    6.76 1.83e-11 ***
genderM     -1.574e-02  1.296e-03  -12.15  < 2e-16 ***

(Dispersion parameter for Gamma family taken to be 0.02879748)

    Null deviance: 58.586  on 1881  degrees of freedom
Residual deviance: 53.759  on 1879  degrees of freedom

> summary(napa.mod6)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.877592   0.015512 121.044  < 2e-16 ***
Age         -0.002532   0.000370  -6.844 1.04e-11 ***
genderM      0.097190   0.007997  12.154  < 2e-16 ***

(Dispersion parameter for Gamma family taken to be 0.02879762)

    Null deviance: 58.586  on 1881  degrees of freedom
Residual deviance: 53.760  on 1879  degrees of freedom
AIC: 5475.7

> summary(napa.mod7)
Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.571e-01  3.711e-03  42.331  < 2e-16 ***
Age          3.025e-04  9.275e-05   3.261  0.00113 **
genderM     -2.281e-02  5.068e-03  -4.501 7.19e-06 ***
Age:genderM  1.751e-04  1.214e-04   1.443  0.14930

(Dispersion parameter for Gamma family taken to be 0.02878207)

    Null deviance: 58.586  on 1881  degrees of freedom
Residual deviance: 53.699  on 1878  degrees of freedom

> summary(napa.mod8)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.8494178  0.0220581  83.843  < 2e-16 ***
Age         -0.0018116  0.0005461  -3.317 0.000927 ***
genderM      0.1521938  0.0315083   4.830 1.47e-06 ***
Age:genderM -0.0013388  0.0007422  -1.804 0.071404 .

(Dispersion parameter for Gamma family taken to be 0.02876127)

    Null deviance: 58.586  on 1881  degrees of freedom
Residual deviance: 53.668  on 1878  degrees of freedom

> anova(napa.mod5,napa.mod7,test="Chisq")
Analysis of Deviance Table
Model 1: mph ~ Age + gender
Model 2: mph ~ Age * gender
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      1879     53.759
2      1878     53.699  1  0.05987   0.1492
>
> anova(napa.mod6,napa.mod8,test="Chisq")
Analysis of Deviance Table
Model 1: mph ~ Age + gender
Model 2: mph ~ Age * gender
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      1879     53.760
2      1878     53.668  1 0.091878  0.07389 .
```

Based on the Additive models, we have the following fitted models and estimated variances and standard deviations. Plots are shown in Figure 9.4.

$$\text{Inverse Link:} \quad \hat{\mu}_i = \frac{1}{0.1531 + 0.0004048A_i - 0.01574M_i} \qquad \hat{\sigma}_i^2 = 0.02879748\hat{\mu}_i \qquad \hat{\sigma}_i = 0.1697\sqrt{\hat{\mu}_i}$$

$$\text{Log Link:} \quad \hat{\mu}_i = \exp\{1.877592 - 0.002532A_i + 0.097190M_i\} \qquad \hat{\sigma}_i^2 = 0.02876127\hat{\mu}_i \qquad \hat{\sigma}_i = 0.1696\sqrt{\hat{\mu}_i}$$

$\nabla$

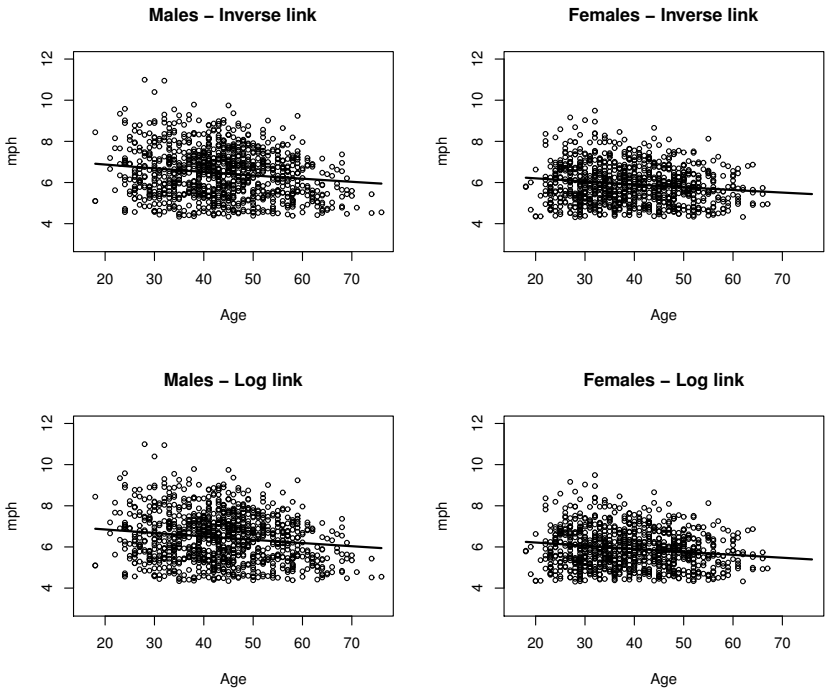


Figure 9.4: Data and Fitted Equations by Gender and Link Finction - 2015 Napa Marathon

## 9.6 Beta Regression

When data are rates or proportions, a regression model based on the Beta Distribution can be fit (Ferrari and Cribari-Neto (2004)). The density function and the mean and variance for a beta random variable are given below.

$$f\left(y_i|\alpha,\beta\right)=\frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}y_i^{\alpha-1}\left(1-y_i\right)^{\beta-1}\qquad 0\leq y_i\leq 1;\quad \alpha,\beta>0$$

$$E\left\{Y\right\}=\mu=\frac{\alpha}{\alpha+\beta}\qquad V\left\{Y\right\}=\frac{\alpha\beta}{\left(\alpha+\beta+1\right)\left(\alpha+\beta\right)^2}$$

A re-parameterization of the model is used in estimating the regression model. Below is the re-parameterized likelihood function.

$$\phi=\alpha+\beta\qquad\Rightarrow\qquad\alpha=\mu\phi\qquad\beta=\left(1-\mu\right)\phi\qquad\Rightarrow\qquad E\left\{Y\right\}=\mu\qquad V\left\{Y\right\}=\frac{\mu\left(1-\mu\right)}{\phi+1}$$

$$L_i=\frac{\Gamma\left(\phi\right)}{\Gamma\left(\mu_i\phi\right)\Gamma\left(\left(1-\mu_i\right)\phi\right)}y_i^{\mu_i\phi-1}\left(1-y_i\right)^{(1-\mu_i)\phi}\qquad g\left(\mu_i\right)=\beta_0+\beta_1X_{i1}+\cdots+\beta_pX_{ip}=\mathbf{x_i}'\boldsymbol{\beta}$$

### 9.6.1 Estimating Model Parameters

Using the logit as the link function, we obtain the following likelihood and log-likelihood functions.

$$g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) \qquad \Rightarrow \qquad \mu_i = \frac{\exp\{\mathbf{x_i}'\boldsymbol{\beta}\}}{1+\exp\{\mathbf{x_i}'\boldsymbol{\beta}\}} \qquad 1-\mu_i = \frac{1}{1+\exp\{\mathbf{x_i}'\boldsymbol{\beta}\}}$$

$$g(\mu_i) = \frac{1}{\left(\frac{\mu_i}{1-\mu_i}\right)}\left[\frac{(1-\mu_i)(1)-\mu_i(-1)}{(1-\mu_i)^2}\right] = \frac{1}{\mu_i(1-\mu_i)}$$

$$L_i = \frac{\Gamma(\phi)}{\Gamma\left(\frac{\exp\{\mathbf{x_i}'\boldsymbol{\beta}\}}{1+\exp\{\mathbf{x_i}'\boldsymbol{\beta}\}}\phi\right)\Gamma\left(\left(\frac{1}{1+\exp\{\mathbf{x_i}'\boldsymbol{\beta}\}}\right)\phi\right)} y_i^{\frac{\exp\{\mathbf{x_i}'\boldsymbol{\beta}\}}{1+\exp\{\mathbf{x_i}'\boldsymbol{\beta}\}}\phi-1}(1-y_i)^{\frac{1}{1+\exp\{\mathbf{x_i}'\boldsymbol{\beta}\}}\phi-1}$$

$$l_i = \ln(L_i) = \ln(\Gamma(\phi)) - \ln(\Gamma(\mu_i\phi)) - \ln(\Gamma((1-\mu_i)\phi)) + (\mu_i\phi-1)\ln(y_i) + ((1-\mu_1)\phi-1)\ln(1-y_i)$$

$$= \ln(\Gamma(\phi)) - \ln\left(\Gamma\left(\frac{\exp\{\mathbf{x_i}'\boldsymbol{\beta}\}}{1+\exp\{\mathbf{x_i}'\boldsymbol{\beta}\}}\phi\right)\right) - \ln\left(\Gamma\left(\left(\frac{1}{1+\exp\{\mathbf{x_i}'\boldsymbol{\beta}\}}\right)\phi\right)\right) +$$
$$\left(\frac{\exp\{\mathbf{x_i}'\boldsymbol{\beta}\}}{1+\exp\{\mathbf{x_i}'\boldsymbol{\beta}\}}\phi-1\right)\ln(y_i) + \left(\frac{1}{1+\exp\{\mathbf{x_i}'\boldsymbol{\beta}\}}\phi-1\right)\ln(1-y_i)$$

Making use of the following results based on the digamma ($\psi$) and trigamma ($\psi'$) functions (which are available in R), the relevant derivatives with respect to $l_i$ are obtained.

$$\frac{d\ln(\Gamma(z))}{dz} = \psi(z) \qquad\qquad \frac{d\psi(z)}{dz} = \psi'(z)$$

$$\frac{\partial l_i}{\partial \mu_i} = -\psi(\mu_i\phi)\phi + \psi((1-\mu_i)\phi)\phi + \phi\ln(y_i) - \phi\ln(1-y_i) = \phi\left[\ln\left(\frac{y_i}{1-y_i}\right) - (\psi(\mu_i\phi) - \psi((1-\mu_i)\phi))\right]$$

$$E\left\{\frac{\partial l_i}{\partial \mu_i}\right\} = 0 \qquad \Rightarrow \qquad E\left\{\left(\frac{y_i}{1-y_i}\right) - (\psi(\mu_i\phi) - \psi((1-\mu_i)\phi))\right\} = 0$$

$$\frac{\partial l_i}{\partial \boldsymbol{\beta}} = \frac{\exp\{\mathbf{x_i}'\boldsymbol{\beta}\}}{(1+\exp\{\mathbf{x_i}'\boldsymbol{\beta}\})^2}\phi\left[\psi\left(\frac{\exp\{\mathbf{x_i}'\boldsymbol{\beta}\}}{1+\exp\{\mathbf{x_i}'\boldsymbol{\beta}\}}\phi\right) + \psi\left(\frac{1}{1+\exp\{\mathbf{x_i}'\boldsymbol{\beta}\}}\phi\right) + \ln\left(\frac{y_i}{1-y_i}\right)\right]\mathbf{x_i}$$

$$\frac{\partial l_i}{\partial \phi} = \psi(\phi) - \psi\left(\frac{\exp\{\mathbf{x_i}'\boldsymbol{\beta}\}}{1+\exp\{\mathbf{x_i}'\boldsymbol{\beta}\}}\phi\right)\left[\frac{\exp\{\mathbf{x_i}'\boldsymbol{\beta}\}}{1+\exp\{\mathbf{x_i}'\boldsymbol{\beta}\}}\right] - \psi\left(\frac{1}{1+\exp\{\mathbf{x_i}'\boldsymbol{\beta}\}}\phi\right)\left[\frac{1}{1+\exp\{\mathbf{x_i}'\boldsymbol{\beta}\}}\right] +$$

$$\left(\frac{\exp\left\{\mathbf{x_i}'\boldsymbol{\beta}\right\}}{1+\exp\left\{\mathbf{x_i}'\boldsymbol{\beta}\right\}}\right)\ln\left(y_i\right)+\left(\frac{1}{1+\exp\left\{\mathbf{x_i}'\boldsymbol{\beta}\right\}}\right)\ln\left(1-y_i\right)$$

$$\boldsymbol{\theta}=\left[\begin{array}{c}\boldsymbol{\beta}\\ \phi\end{array}\right]\qquad g\left(\boldsymbol{\theta}\right)=\left[\begin{array}{c}g\left(\boldsymbol{\beta}\right)\\ g\left(\phi\right)\end{array}\right]\qquad\qquad g\left(\boldsymbol{\beta}\right)=\sum_{i=1}^{n}\frac{\partial l_i}{\partial\boldsymbol{\beta}}\qquad g\left(\phi\right)=\sum_{i=1}^{n}\frac{\partial l_i}{\partial\phi}$$

$$\frac{\partial^2 l_i}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}=\left[\mathbf{A_1}-\mathbf{A_2}\right]\mathbf{x_i}\mathbf{x_i}'\quad\text{where:}$$

$$\mathbf{A_1}=\left(\frac{\phi\exp\left\{\mathbf{x_i}'\boldsymbol{\beta}\right\}\left(1-\exp\left\{\mathbf{x_i}'\boldsymbol{\beta}\right\}\right)}{\left(1+\exp\left\{\mathbf{x_i}'\boldsymbol{\beta}\right\}\right)^3}\right)\left[-\psi\left(\frac{\exp\left\{\mathbf{x_i}'\boldsymbol{\beta}\right\}}{\left(1-\exp\left\{\mathbf{x_i}'\boldsymbol{\beta}\right\}\right)}\phi\right)+\psi\left(\frac{1}{\left(1-\exp\left\{\mathbf{x_i}'\boldsymbol{\beta}\right\}\right)}\phi\right)+\ln\left(\frac{y_i}{1-y_i}\right)\right]$$

$$\mathbf{A_2}=\left(\frac{\phi^2\exp\left\{2\mathbf{x_i}'\boldsymbol{\beta}\right\}\left(1-\exp\left\{\mathbf{x_i}'\boldsymbol{\beta}\right\}\right)}{\left(1+\exp\left\{\mathbf{x_i}'\boldsymbol{\beta}\right\}\right)^4}\right)\left[\psi'\left(\frac{\exp\left\{\mathbf{x_i}'\boldsymbol{\beta}\right\}}{\left(1-\exp\left\{\mathbf{x_i}'\boldsymbol{\beta}\right\}\right)}\phi\right)+\psi'\left(\frac{1}{\left(1-\exp\left\{\mathbf{x_i}'\boldsymbol{\beta}\right\}\right)}\phi\right)\right]$$

$$E\left\{\frac{\partial^2 l_i}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right\}=\left(\frac{\phi^2\exp\left\{2\mathbf{x_i}'\boldsymbol{\beta}\right\}\left(1-\exp\left\{\mathbf{x_i}'\boldsymbol{\beta}\right\}\right)}{\left(1+\exp\left\{\mathbf{x_i}'\boldsymbol{\beta}\right\}\right)^4}\right)\left[\psi'\left(\frac{\exp\left\{\mathbf{x_i}'\boldsymbol{\beta}\right\}}{\left(1-\exp\left\{\mathbf{x_i}'\boldsymbol{\beta}\right\}\right)}\phi\right)+\psi'\left(\frac{1}{\left(1-\exp\left\{\mathbf{x_i}'\boldsymbol{\beta}\right\}\right)}\phi\right)\right]\mathbf{x_i}\mathbf{x_i}'=$$

$$-\phi\left(\phi\mu_i^2\left(1-\mu_i\right)^2\right)\left[\psi'\left(\mu_i\phi\right)+\psi'\left(\left(1-\mu_i\right)\phi\right)\right]\mathbf{x_i}\mathbf{x_i}'=-\phi\left[\phi\frac{\left[\psi'\left(\mu_i\phi\right)+\psi'\left(\left(1-\mu_i\right)\phi\right)\right]}{\left[g'\left(\mu_i\right)\right]^2}\right]\mathbf{x_i}\mathbf{x_i}'$$

$$E\left\{\frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right\}=\sum_{i=1}^{n}E\left\{\frac{\partial^2 l_i}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right\}=-\phi\sum_{i=1}^{n}\left[\phi\frac{\left[\psi'\left(\mu_i\phi\right)+\psi'\left(\left(1-\mu_i\right)\phi\right)\right]}{\left[g'\left(\mu_i\right)\right]^2}\right]\mathbf{x_i}\mathbf{x_i}'=-\phi\mathbf{X}'\mathbf{W}\mathbf{X}$$

$$\mathbf{W}=\operatorname{diag}\left\{\left(\frac{\phi\exp\left\{2\mathbf{x_i}'\boldsymbol{\beta}\right\}\left(1-\exp\left\{\mathbf{x_i}'\boldsymbol{\beta}\right\}\right)}{\left(1+\exp\left\{\mathbf{x_i}'\boldsymbol{\beta}\right\}\right)^4}\right)\left[\psi'\left(\frac{\exp\left\{\mathbf{x_i}'\boldsymbol{\beta}\right\}}{\left(1-\exp\left\{\mathbf{x_i}'\boldsymbol{\beta}\right\}\right)}\phi\right)+\psi'\left(\frac{1}{\left(1-\exp\left\{\mathbf{x_i}'\boldsymbol{\beta}\right\}\right)}\phi\right)\right]\right\}=$$

$$\operatorname{diag}\left\{\left(\phi\mu_i^2\left(1-\mu_i\right)^2\right)\left[\psi'\left(\mu_i\phi\right)+\psi'\left(\left(1-\mu_i\right)\phi\right)\right]\right\}$$

$$\frac{\partial^2 l_i}{\partial\boldsymbol{\beta}\partial\phi}=\left[\mathbf{B_1}+\mathbf{B_2}\right]\mathbf{x_i}'\quad\text{where:}$$

$$\mathbf{B_1}=-\psi'\left(\mu_i\phi\right)\left[\mu_i\left(1-\mu_i\right)\phi\mu_i\right]+\psi'\left(\left(1-\mu_i\right)\phi\right)\left[\mu_i\left(1-\mu_i\right)\phi\left(1-\mu_i\right)\right]$$

$$\mathbf{B_2}=\mu_i\left(1-\mu_i\right)\left[-\psi\left(\mu_i\phi\right)+\psi\left(\left(1-\mu_i\right)\phi\right)+\ln\left(\frac{y_i}{1-y_i}\right)\right]$$

$$E\left\{\frac{\partial^2 l_i}{\partial\boldsymbol{\beta}\partial\phi}\right\}=\left[-\psi'\left(\mu_i\phi\right)\left[\mu_i\left(1-\mu_i\right)\phi\mu_i\right]+\psi'\left(\left(1-\mu_i\right)\phi\right)\left[\mu_i\left(1-\mu_i\right)\phi\left(1-\mu_i\right)\right]\right]\mathbf{x_i}=-\mathbf{X}'\mathbf{T}\mathbf{c}$$

$$\text{where:}\qquad\mathbf{T}=\operatorname{diag}\left\{\frac{1}{g'\left(\mu_i\right)}=\mu_i\left(1-\mu_i\right)\right\}\qquad\mathbf{c_i}=\phi\left[\psi'\left(\mu_i\phi\right)\mu_i+\psi'\left(\left(1-\mu_i\right)\phi\right)\left(1-\mu_i\right)\right]$$

$$\frac{\partial^2 l_i}{\partial \phi^2} = \psi'(\phi) - \psi'(\mu_i\phi)\mu_i^2 - \psi'((1-\mu_i)\phi)(1-\mu_i)^2 = E\left\{\frac{\partial^2 l_i}{\partial \phi^2}\right\}$$

$$E\left\{\frac{\partial^2 l_i}{\partial \phi^2}\right\} = -\text{trace}(\mathbf{D}) \qquad \mathbf{D} = \text{diag}\left\{\psi'(\mu_i\phi)\mu_i^2 + \psi'((1-\mu_i)\phi)(1-\mu_i)^2 - \psi'(\phi)\right\}$$

$$G(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial^2 l}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} & \frac{\partial^2 l}{\partial \boldsymbol{\beta}\partial \phi} \\ \frac{\partial^2 l}{\partial \phi \partial \boldsymbol{\beta}'} & \frac{\partial^2 l}{\partial \phi^2} \end{bmatrix} \qquad E\{G(\boldsymbol{\theta})\} = \begin{bmatrix} -\phi\mathbf{X}'\mathbf{W}\mathbf{X} & -\mathbf{X}'\mathbf{T}\mathbf{c} \\ -\mathbf{c}'\mathbf{T}\mathbf{X} & -\text{trace}(\mathbf{D}) \end{bmatrix}$$

The Fisher Scoring algorithm iterates the following equation to convergence for the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$.

$$\tilde{\boldsymbol{\theta}}^{\text{New}} = \tilde{\boldsymbol{\theta}}^{\text{Old}} - \left(E\left\{G\left(\tilde{\boldsymbol{\theta}}^{\text{Old}}\right)\right\}\right)^{-1} g\left(\tilde{\boldsymbol{\theta}}^{\text{Old}}\right)$$

Once the ML estimator has been obtained, its estimated Variance-Covariance matrix is obtained.

$$\hat{V}\left\{\hat{\boldsymbol{\beta}}\right\} = -\left(E\{G(\boldsymbol{\theta})\}\,|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right)^{-1}$$

To obtain starting values for the Fisher Scoring algorithm, consider the following model components.

$$\ln\left(\frac{\mu_i}{1-\mu_i}\right) = \mathbf{x_i}'\boldsymbol{\beta} \qquad V\{Y_i\} = \frac{\mu_i(1-\mu_i)}{\phi+1} \quad \Rightarrow \quad \phi = \frac{\mu_i(1-\mu_i)}{V\{Y_i\}} - 1$$

Let $h(y) = \ln(y/(1-y))$, then we have the following.

$$h(Y_i) \approx h(\mu_i) + (Y_i - \mu_i)h'(\mu_i) \quad \Rightarrow \quad V\{h(Y_i)\} \approx V\{Y_i\}[h'(\mu_i)]^2 \quad \Rightarrow \quad V\{Y_i\} \approx V\{h(Y_i)\}[h'(\mu_i)]^{-2}$$

First, fit a linear regression of $h(y)$ on $X_1, \ldots, X_p$, and obtain the residuals.

$$h(\hat{Y}_i) = \mathbf{x_i}'\tilde{\boldsymbol{\beta}}^{\text{Old}} \qquad \hat{e}_i = h(Y_i) - h(\hat{Y}_i)$$

Next, obtain the estimated means, and obtain an estimate of $V\{h(Y_i)\}$.

$$\hat{\mu}_i = \frac{\exp\left\{\mathbf{x_i}'\tilde{\boldsymbol{\beta}}^{\text{Old}}\right\}}{1+\exp\left\{\mathbf{x_i}'\tilde{\boldsymbol{\beta}}^{\text{Old}}\right\}} \qquad [h'(\hat{\mu}_i)]^{-2} = [\hat{\mu}_i(1-\hat{\mu}_i)]^2 \qquad V\{h(Y_i)\} \approx \frac{\hat{e}'\hat{e}}{n-p'}$$

Finally, obtain estimates of $V\{Y_i\}$ and $\phi$.

$$V\{Y_i\} \approx V\{h(Y_i)\}[\hat{\mu}_i(1-\hat{\mu}_i)]^2 \qquad \hat{\phi} = \frac{\hat{\mu}_i(1-\hat{\mu}_i)}{V\{Y_i\}} - 1 \approx \left[\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\mu}_i(1-\hat{\mu}_i)}{V\{Y_i\}}\right] - 1$$

### 9.6.2 Diagnostics and Influence Measures

A measure of the goodness-of-fit of the Beta Regression model is pseudo-$R^2$, which represents the squared correlation between $g(y_i)$ and $g(\hat{\mu}_i)$, where:

$$g(y_i) = \ln\left(\frac{y_i}{1-y_i}\right) \qquad g(\hat{\mu}_i) = \mathbf{x_i}'\hat{\boldsymbol{\beta}}.$$

The deviance and Pearson residuals for the individual observations are described below.

$$\text{Deviance Residuals:} \qquad r_{Di} = \text{sgn}\{y_i - \hat{\mu}_i\}\sqrt{2\left[l_i\left(y_i, y_i, \hat{\phi}\right) - l_i\left(y_i, \hat{\mu}_i, \hat{\phi}\right)\right]}$$

$$\text{where: } l_i\left(y_i, \mu_i, \hat{\phi}\right) = \ln\left[\frac{\Gamma(\phi)}{\Gamma(\mu_i\phi)\,\Gamma((1-\mu_i)\,\phi)}\right] + (\mu_i\phi - 1)\ln(y_i) + ((1-\mu_i)\,\phi - 1)\ln(1-y_i)$$

$$\Rightarrow \qquad l_i\left(y_i, y_i, \hat{\phi}\right) - l_i\left(y_i, \hat{\mu}_i, \hat{\phi}\right) = \ln\left[\frac{\Gamma\left(\hat{\mu}_i\hat{\phi}\right)}{\Gamma\left(y_i\hat{\phi}\right)}\right] + \ln\left[\frac{\Gamma\left((1-\hat{\mu}_i)\,\hat{\phi}\right)}{\Gamma\left((1-y_i)\,\hat{\phi}\right)}\right] + (y_i - \hat{\mu}_i)\,\hat{\phi}\ln\left(\frac{y_i}{1-y_i}\right)$$

$$\text{Pearson Residuals:} \qquad r_{Pi} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{V}\{Y_i\}}} = \frac{y_i - \hat{\mu}_i}{\sqrt{\frac{\hat{\mu}_i(1-\hat{\mu}_i)}{\hat{\phi}+1}}}$$

Computations for the Hat matrix, Cook's D, and the Generalized Leverage are given below.

$$\text{Hat Matrix:} \qquad \mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}^{1/2} = \{H_{ij}\}$$

$$\text{Cook's D:} \qquad D_i = \frac{H_{ii}r_{Pi}^2}{p'\left(1-H_{ii}\right)^2}$$

$$\text{Generalized Leverage:} \qquad GL(\boldsymbol{\beta}, \phi) = \mathbf{AM} + \frac{1}{\gamma\phi}\mathbf{AF}\left(\mathbf{f}'\mathbf{AM} - \mathbf{b}\right)$$

where:

$$\mathbf{W} = \text{diag}\left\{\left(\hat{\phi}\hat{\mu}_i^2\left(1-\hat{\mu}_i\right)^2\right)\left[\psi'\left(\hat{\mu}_i\hat{\phi}\right)+\psi'\left(\left(1-\hat{\mu}_i\right)\hat{\phi}\right)\right]\right\}$$

$$\mathbf{A} = \mathbf{TX}\left(\mathbf{X}'\mathbf{QX}\right)^{-1}\mathbf{X}'\mathbf{T} \qquad \mathbf{M} = \text{diag}\left\{\frac{1}{y_i\left(1-y_i\right)}\right\} \qquad \mathbf{b} = \begin{bmatrix} -\frac{y_1-\hat{\mu}_1}{y_1(1-y_1)} \\ \vdots \\ -\frac{y_n-\hat{\mu}_n}{y_n(1-y_n)} \end{bmatrix}$$

$$\mathbf{f} = \begin{bmatrix} c_1\left[\ln\left(\frac{y_1}{1-y_1}\right)-\left(\psi\left(\hat{\mu}_1\hat{\phi}\right)-\psi\left(\left(1-\hat{\mu}_1\right)\hat{\phi}\right)\right)\right] \\ \vdots \\ c_n\left[\ln\left(\frac{y_n}{1-y_n}\right)-\left(\psi\left(\hat{\mu}_n\hat{\phi}\right)-\psi\left(\left(1-\hat{\mu}_n\right)\hat{\phi}\right)\right)\right] \end{bmatrix}$$

$$\mathbf{T} = \text{diag}\left\{\frac{1}{g'\left(\hat{\mu}_i\right)}\right\} = \text{diag}\left\{\hat{\mu}_i\left(1-\hat{\mu}_i\right)\right\} \qquad \mathbf{c} = \begin{bmatrix} \hat{\phi}\left[\psi'\left(\hat{\mu}_1\hat{\phi}\right)\hat{\mu}_1-\psi'\left(\left(1-\hat{\mu}_1\right)\hat{\phi}\right)\left(1-\hat{\mu}_1\right)\right] \\ \vdots \\ \hat{\phi}\left[\psi'\left(\hat{\mu}_n\hat{\phi}\right)\hat{\mu}_n-\psi'\left(\left(1-\hat{\mu}_n\right)\hat{\phi}\right)\left(1-\hat{\mu}_n\right)\right] \end{bmatrix}$$

$$\mathbf{Q} = \text{diag}\left\{\left[\hat{\phi}\left(\psi'\left(\hat{\mu}_i\hat{\phi}\right)+\psi'\left(\left(1-\hat{\mu}_i\right)\hat{\phi}\right)\right)+\left(\ln\left(\frac{y_i}{1-y_i}\right)-\left(\psi\left(\hat{\mu}_i\hat{\phi}\right)-\psi\left(\left(1-\hat{\mu}_i\right)\hat{\phi}\right)\right)\right)\frac{g''\left(\hat{\mu}_i\right)}{g'\left(\hat{\mu}_i\right)}\right]\left(\frac{1}{g'\left(\hat{\mu}_i\right)}\right)^2\right\}$$

**Example: Ford Prize Winnings in NASCAR Races: 1992-2000**

The NASCAR Winston Cup series had $n = 267$ races during the years 1992-2000. For each race, we obtain $Y$, the proportion of the prize money won by Ford cars. Variables used as predictors include: $X_1$, the proportion of cars in the race that are Fords, $X_2$, the track length (miles), $X_3$, the bank of the turns of the track (degrees), $X_4$, the number of laps, and dummy variables for the years 1993-2000.

The matrix form of the program is very long and available on the course notes webpage. The R program, making use of the **betareg** package and function and partial output are given below. As the proportion of Ford cars increases, so does the proportion of Ford prize money (not surprisingly). As track length and laps increase, proportion of Ford prize money decreases. All years, with the exception of 1994, have significantly lower proportion of Ford prize money than the reference year of 1992. Summary plots are given in Figure 9.5.

```
### R Program
ford <- read.csv("http://www.stat.ufl.edu/~winner/data/nas_ford_1992_2000a.csv",header=T)
attach(ford); names(ford)

library(betareg)
```

```
Year <- factor(Year)
Track_id <- factor(Track_id)
g.Y <- log(FPrzp / (1-FPrzp))

beta.mod1 <- betareg(FPrzp ~ FDrvp + TrkLng + Bank + Laps + Year)
summary(beta.mod1)
resid(beta.mod1,type="pearson")
resid(beta.mod1,type="deviance")
cooks.distance(beta.mod1)
gleverage(beta.mod1)
hatvalues(beta.mod1)
vcov(beta.mod1)
(pseudo.R2 <- cor(g.Y, predict(beta.mod1))^2)

par(mfrow=c(2,2))
plot(beta.mod1,which=1:4,type="pearson")

### Output

Call:
betareg(formula = FPrzp ~ FDrvp + TrkLng + Bank + Laps + Year)

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-3.7996 -0.6602  0.0031  0.6532  5.2351

Coefficients (mean model with logit link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.7312457  0.2074226  -3.525 0.000423 ***
FDrvp        2.5440987  0.4128540   6.162 7.17e-10 ***
TrkLng      -0.1106140  0.0405332  -2.729 0.006353 **
Bank        -0.0019614  0.0015161  -1.294 0.195767
Laps        -0.0007601  0.0002544  -2.988 0.002808 **
Year1993    -0.2225534  0.0817810  -2.721 0.006502 **
Year1994    -0.0441460  0.0852535  -0.518 0.604584
Year1995    -0.1924006  0.0844577  -2.278 0.022722 *
Year1996    -0.2031800  0.0785715  -2.586 0.009712 **
Year1997    -0.1441680  0.0571996  -2.520 0.011721 *
Year1998    -0.1585144  0.0550342  -2.880 0.003973 **
Year1999    -0.1892330  0.0602789  -3.139 0.001694 **
Year2000    -0.1904757  0.0571511  -3.333 0.000860 ***

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)  102.877      8.861   11.61   <2e-16 ***

Type of estimator: ML (maximum likelihood)
Log-likelihood: 427.4 on 14 Df
Pseudo R-squared: 0.3906
Number of iterations: 23 (BFGS) + 2 (Fisher scoring)
```

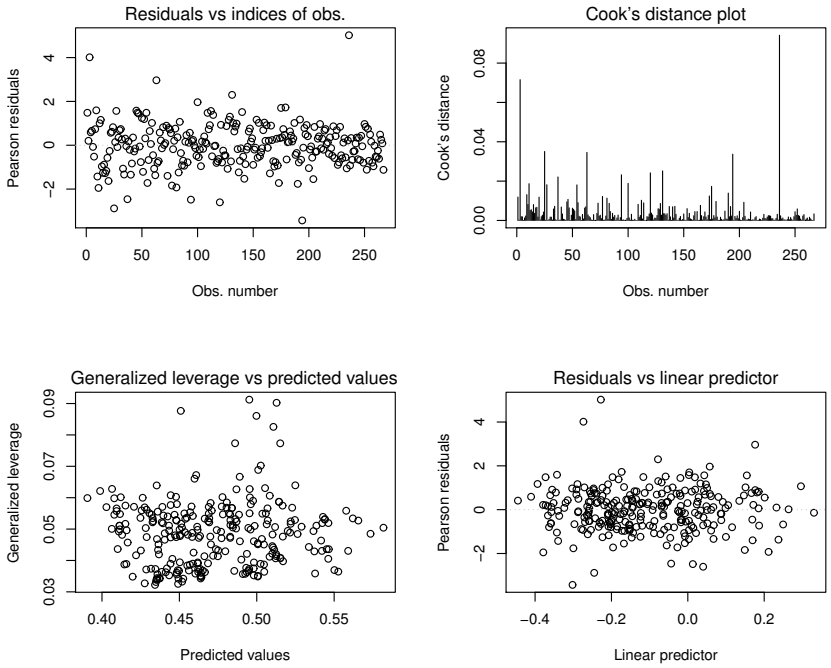


Figure 9.5: Caption for nas19922000

# Bibliography


[1] Agresti, A. (2002). *Categorical Data Analysis. 2nd Ed.* Wiley, New York.

[2] Agresti, A. (1996). *An Introduction to Categorical Data Analysis.* Wiley, New York.

[3] Alvord, W.G., J.H. Driver, L. Claxton, J.P. Creason (1990). "Methods for Comparing Salmonella Mutagenicity Data Sets Using Nonlinear Models," *Mutation Research,* Vol. 240, pp. 177-194.

[4] Cameron, A.C. and P.K. Trivedi (2005). *Microeconometrics: Methods and Applications.* Cambridge, Cambridge.

[5] Cameron, A.C. and P.K. Trivedi (2010). *Microeconometrics Using Stata, Revised Edition.* Stata Press, College Station, TX.

[6] Chanying, J. and P. Junzheng (1998). Relationship Between Bearing Capacity and Depth of Soft Layer of Paddy Fields in South China, *Journal of Terramechanics,* Vol. 35, pp. 225-228.

[7] Cobb, C.W. and P.H. Douglas (1928). "A Theory of Production," *American Economic Review,* Vol 18 (Supplement), pp:139-165.

[8] Colotti, V. (2016). "Mechanical Shear Strength Model for Reinforced Concrete Beams Strengthened with FRP Materials," *Construction and Building Materials,* Vol. 124, pp. 855-865.

[9] Crawley, M.J. (2013). *The R Book. 2nd Ed.* Wiley, Chichester, West Sussex, UK.

[10] Csapo, J., Z. Csapo-Kiss, T.G. Martin, S. Folestad, O. Orwar, A. Tivesten, and S. Nemethy (1995). "Age Estimation of Old Carpets Based on Cystine and Cysteic Acid Content," *Analytica Chimica Acta,* Vol. 300, pp. 313-320.

[11] Davidian, M. and D.M. Giltinan (1995). *Nonlinear Models for Repeated Measurement Data,* Chapman & Hall, London, UK.

[12] Denny, M.W. (2008). Limits to Running Speeds in Dogs, Horses, and Humans, *The Journal of Experimental Biology,* Vol. 211, pp. 3836-3849.

[13] Edwards, D.J., G.D. Holt, F.C. Harris (2000). "A Comparative Analysis Between the Multilayer Perceptron "Neural Network" and Multiple Regression Analysis for Predicting Construction Plant Maintenance Costs," *Journal of Quality in Maintenance Engineering,* Vol. 6, #1, pp. 45-61.

[14] Faraway, J.J. (2006). *Extending the Linear Model with R.* Chapman&Hall/CRC, Boca Raton, FL.

[15] Ferrari, S.L.P. and F. Cribari-Neto (2004). Beta Regression for Modeling Rates and Proportions, *Journal of Applied Statistics,* Vol. 31, #7, pp. 799-815.

[16] Gallant, A.R. (1987). *Nonlinear Statistical Models.* Wiley, New York.

[17] Gallant, A.R. and J.J. Goebel (1976). Nonlinear Regression with Autocorrelated Errors, *Journal of the American Statistical Association,* Vol. 71, pp.961-967.

[18] Greene, W.H. (2003). *Econometric Analysis. 5th Ed.* Prentice-Hall, Upper Saddle River, NJ.

[19] Gumpertz, M.L. and S.G. Pantula (1989). A Simple Approach to Inferences in Random Coefficient Models, *The American Statistician,* Vol. 43, pp. 203-210.

[20] Jones, G.V. and K-H. Storchmann (2001). "Wine Market Prices and Investment Under Uncertainty: An Econometric Model for Bordeaux Crus Classes," *Agricultural Economics,* Vol. 26, pp. 115-133.

[21] Kumar, Y.S., R.S. Praksam, O.V.S. Reddy (2009). "Optimisation of Fermentation Conditions for Mango (Mangifera indica L.) Wine Production by Employing Response Surface Methodology," *International Journal of Food Science & Technology,* Vol. 44, pp. 2320-2327.

[22] Kutner, M.H., C.J. Nachtsheim, J. Neter, and W. Li (2005). *Applied Linear Statistical Models. 5th Ed.* McGraw-Hill, New York.

[23] Lafi, S.Q. and J.B. Kaneene (1992). "An Explanation of the Use of Principal-Components Analysis to Detect and Correct for Multicollinearity." *Preventive Veterinary Medicine,* Vol. 13, Issue 4, pp. 261-275.

[24] Leike, A. (2002). Demonstration of the Exponential Decay Law Using Beer Froth, *European Journal of Physics,* Vol. 23, Number 1, pp. 21-26.

[25] Marichal, C. and M.S. Mantecon (1994). Silver and Situados: New Spain and the Financing of the Spanish Empire in the Caribbean in the Eighteenth Century, *Hispanic American Historical Review,* Vol. 74, #4, pp. 587-613.

[26] Miksa, S., D. Lutz, and C. Guy (2013). "In vitro/vivo Correlation and Repeatabilty According to Substrate," *Cosmetics and Toiletries,* Vol. 128, #9, pp. 648-657.

[27] Monahan, J.F. (2008). *A Primer on Linear Models.* Chapman&Hall/CRC. Boca Raton, FL.

[28] Myers, R.H. (1990). *Classical and Modern Regression with Applications. 2nd Ed.* PWS-Kent, Boston.

[29] Nyh, S. (2002). Traffic Deaths and Superstition on Friday the 13th, *American Journal of Psychiatry,* Vol. 159, #12, pp. 2110-2111.

[30] Ndaro, M.S., X-Y. Jin, T. Chen, C-W. Yu (2007). "Splitting of Islands-in-the-Sea Fibers (PA6/COPET) During Hydroentangling of Nonwovens," *Journal of Engineered Fibers and Fabrics,* Vol. 2, #4, pp. 1-9.

[31] Pinheiro, J.C. and D.M. Bates (2000). *Mixed Effects Models in S and S-Plus.* Springer Verlag, New York.

[32] Rawlings, J.O., S.G. Pantula, and D.A. Dickey (2001). *Applied Regression Analysis: A Research Tool. 2nd Ed.* Springer, New York.

[33] Rowe, W.F. and S.R. Hanson (1985). "Range-of-Fire Estimates from Regression Analysis Applied to the Spreads of Shotgun Pellets Patterns: Results of a Blind Study," *Forensic Science International,* Vol. 28, pp. 239-250.

[34] Saito, Y., Y. Goto, A. Dane, K. Strutt, and A. Raza (2003). Randomized Dose-Response Study of Rovustatin in Japanese Patients with Hypercholesterolemia, *Journal of Atherosclerosis and Thrombosis,* Vol. 10, #6, pp. 329-336.

[35] Spurgeon, J.C. (1978). "The Correlation of animal response Data with the Yields of Selected Thermal Decomposition Products for Typical Aircraft Interior Materials," U.S. D.O.T. Report No. FAA-RD-78-131.

[36] Wang, S. and Q. Meng (2012). "Sailing Speed Optimization for Container Ships in a Liner Shipping Network," *Transportation Research Part E*, Vol. 48, pp. 701-714

[37] Winner, L. (2006). NASCAR Winston Cup Race Results for 1975-2003, *Journal of Statistics Education,* Vol.14,#3, www.amstat.org/publications/jse/v14n3/datasets.winner.html

[38] Zhao, H., H. Li, G. Sun, B. Yang, M. Zhao (2013). "Assessment of Endogenous Antioxidative Compounds and Antioxidant Activities of Lager Beers," *Journal of the Science of Food and Agriculture,* Vol. 93, pp. 910-917.

[39] Zhu, Q. and X. Peng (2012). "The Impacts of Population Change on Carbon Emissions in China During 1978-2008," *Environmental Impact Assessment Review,* Vol. 36, pp. 1-8.