Linear Regression

March 28, 2013

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.

Simple Linear Regression

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. The model is:

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

Here β_0 is the intercept (mean of Y when X=0) and β_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 (β_0), and thus Y and X are not associated.

Estimation of Model Parameters

We obtain a sample of pairs (X_i, Y_i) $i = 1, \ldots, n$. Our goal is to choose estimators of β_0 and β_1 that minimize the error sum of squares: $Q = \sum_{i=1}^{n} \epsilon_i^2$. The resulting estimators are (from calculus):

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

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

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

The (unbiased) estimate of the error variance σ^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: β_0 and β_1.

The estimators $\hat{\beta}_1$ and $\hat{\beta}_0$ are linear functions of $Y_1, \ldots, Y_n$ and thus using basic rules of mathematical statistics, their sampling distributions 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 σ^2 replaced by MSE.

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

Inference Regarding β_1

Primarily of interest is inferences regarding β_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 β_1 and its estimated standard

error. The t-test is conducted as follows. Note that the null value β_{10} is almost always 0, and that software packages that report these tests always are treating β_{10} as 0. Here, and in all other tests, TS represents Test Statistic, and RR represents Rejection Region.

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

One-sided tests use the same test statistic, but adjusts 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 β_1 is obtained as:

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

Note that the confidence interval represents the values of β_{10} that the two-sided test: $H_0 : \beta_1 = \beta_{10} \quad H_A : \beta_1 \neq \beta_{10}$ fails to reject the null hypothesis.

Inferences regarding β_0 are rarely of interest, but can be conducted in analogous manner, using the estimate $\hat{\beta}_0$ and its estimated standard error $SE\{\hat{\beta}_0\}$.

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

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

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

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

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 σ). The point prediction is the same as for the mean. The estimate, standard error of prediction, and $(1-\alpha)100\%$ Prediction Interval are given below:

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

Note that the Prediction Interval will tend to be much wider than the Confidence Interval for the mean.

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 Sum of Squares**.

When there is an association between Y and X ($\beta_1 \neq 0$), the best predictor of each observation is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ (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 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**.

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

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 $df_{\text{Error}} = n - 2$ (for simple regression). The **Regression Degrees of Freedom** is $df_{\text{Regression}} = 1$ (for simple regression).

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

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 1: Analysis of Variance Table for Simple Linear Regression

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

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

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.

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 ($100r^2$).

Correlation

The regression coefficient β_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 X, and does not distinguish which is the dependent and which is the independent variables. This makes it a widely reported measure when researchers are interested in how 2 random variables vary together in a population. The population correlation coefficient is labeled ρ, 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 testig thether $\beta_1 = 0$. The 2-sided test is given below:

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

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:

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

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

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

Model Diagnostics

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.

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 need to 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**:

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n_j}\sum_{j=1}^{c}\left(Y_{ij} - \overline{Y}_j\right)^2 + \sum_{j=1}^{c} n_j\left(\overline{Y}_j - \hat{Y}_j\right) \qquad SSE = SSPE + SSLF$$

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:

$$TS: F_{obs} = \frac{[SSLF/(c-2)]}{[SSPE/(n-c)]} = \frac{MSLF}{MSPE} \qquad RR: F_{obs} \geq F_{\alpha,c-2,n-c} \qquad P\text{-Value } : P\left(F_{c-2,n-c} \geq F_{obs}\right)$$

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.

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, reporting 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 $Y_1, \ldots, Y_n$ are all 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 λ between -2 and 2 by 0.1 are run, and the value of λ that has the smallest Error Sum of Squares (equivalently Maximum Likelihood) is identified. Software packages will present a confidence interval for λ.

Checking Equal Variance

Aplot 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 test.

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:

$$d_{ij} = |e_i j - \tilde{e}_i| \quad i = 1, 2; j = 1, \ldots, n_i \qquad \overline{d}_i = \frac{\sum_{i=1}^{n_i} d_{ij}}{n_i} \qquad s_i^2 = \frac{\sum_{i=1}^{n_i} \left(d_{ij} - \overline{d}_i\right)^2}{n_i - 1} \qquad s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

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 - Fits a regression of the squared residuals on X and tests whether the (natural) log of the variance is linearly related to X. 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 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 can use the Box-Cox transformation to obtain 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.

When the distribution of Y is a from a known family (e.g. Binomial, Poisson, Gamma), we can fit a **Generalized Linear Model**.

Checking Independence

When the data are a time (or spatial) series, the errors can be correlated over time (or space), referred to as **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. The test is obtained as follows:

$$TS: DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \qquad \text{Decision Rule: } DW < d_L \text{Reject } H_0 \quad DW > d_U \text{Accept } H_0 \quad \text{Otherwise Inconclusive}$$

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).

When errors are not independent, estimated standard errors of estimates tend to be too small, making 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.

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 intercept, β_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 are obtained from statistical software packages. Their matrix forms are not given here (see references).

Also, many 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):

$$MSE_{(i)} = \frac{SSE_{(i)}}{n - p' - 1} = \frac{SSE - e_i^2}{n - p' - 1} \qquad \text{for simple regression } p' = 2$$

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).

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.

DFFITS - These measure how much an individual case's fitted value shifts when it is included in the regression fit, and when it is excluded. 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.

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 and excluded 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.

Cook's D - Is a 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.

COVRATIO - This measures each case's influence on the estimated standard errors of the regression coefficients (inflating or deflating them). Cases with COVRATIO outside of $1 \pm 3p'/n$ are considered influential.

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:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \qquad \epsilon \sim N(0, \sigma^2) \text{ independent}$$

We then obtain least squares (and maximum likelihood) estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ that minimize the error sum of squares. The fitted values, residuals, and error sum of squares are obtained as:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots \hat{\beta}_p X_{ip} \qquad e_i = Y_i - \hat{Y}_i \qquad SSE = \sum_{i=1}^{n} e_i^2$$

The degrees of freedom for error are now $n - p' = n - (p + 1)$, as we have now estimated $p' = p + 1$ parameters.

In the multiple linear regression model, β_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**.

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 coefficients).

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):

$$H_0 : \beta_j = \beta_{j0} \qquad H_A : \beta_j \neq \beta_{j0} \quad TS : t_{obs} = \frac{\hat{\beta}_j - \beta_{j0}}{SE\{\hat{\beta}_j\}} \qquad RR : |t_{obs}| \geq t_{\alpha/2,n-p'} \qquad P\text{-value} : P(t_{n-p'} \geq |t_{obs}|)$$

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 β_j is obtained as:

$$\hat{\beta}_j \pm t_{\alpha/2,n-p'} SE\{\hat{\beta}_j\}$$

Note that the confidence interval represents the values of β_{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.

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 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 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**.

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

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.

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

Error and Regression sums of squares have a **Mean Square**, which is the sum of squares divided by its corresponding degrees of freedom: $MSE = SSE/(n - p')$ and $MSR = SSR/p$. It can be shown that these mean squares have the following **Expected Values**, average values in repeated sampling at the same observed X levels:

$$E\{MSE\} = \sigma^2 \qquad E\{MSR\} \geq \sigma^2$$

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:

$$H_0 : \beta_1 = \cdots \beta_p = 0 \qquad H_A : \text{ Not all } \beta_j = 0 \quad TS : F_{obs} = \frac{MSR}{MSE} \qquad RR : F_{obs} \geq F_{\alpha,p,n-p'} \qquad P\text{-value} : P(F_{p,n-p'} \geq F_{obs})$$

The Analysis of Variance is typically set up in a table as in Table 2.

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 2: 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 ($100R^2$).

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 test statistic and rejection region:

$$TS : F_obs = \frac{\left[\frac{SSE(R)-SSE(F)}{(n-g')-(n-p')}\right]}{\left[\frac{SSE(F)}{n-p'}\right]} = \frac{\left[\frac{SSR(F)-SSE(R)}{p-g}\right]}{\left[\frac{SSE(F)}{n-p'}\right]} = \frac{\left[\frac{R_F^2-R_R^2}{p-g}\right]}{\left[\frac{1-R_F^2}{n-p'}\right]}$$

$$RR : F_{obs} \geq F_{\alpha,p-g,n-p'} \qquad P\text{-value} : P(F_{p-g,n-p'} \geq F_{obs})$$

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 is in 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. The β^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 $m-1$ β^s are 0, controlling for all other predictors.

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:

$$E\{Y\} = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_1X_2 = \beta_0 + \beta_2X_2 + + (\beta_1 + \beta_3X_2)\, X_1$$

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 t-test ($m-1=1$) or F-test ($m-1>2$) can be conducted to test for interactions among predictors.

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:

$$E\{Y\} = \beta_0 + \beta_1X + \beta_2X^2 + \beta_3X^3$$

If we wish to test whether the fit is linear, as opposed to "not linear," we could test $H_0 : \beta_2 = \beta_3 = 0$.

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:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{33} X_3^2 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{23} X_2 X_3$$

We typically wish to simplify the model, to make it more parsimonious, when possible.

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 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.

$$AIC(\text{Model}) = nln(SSE(\text{Model})) + 2p' - nln(n) \qquad BIC(\text{Model}) = nln(SSE(\text{Model})) + [ln(n)]p' - nln(n)$$

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.

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."

Forward Selection

This is a "bottom-up, 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."

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.

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 $C_p \leq p'$

AIC - $AIC(\text{Model}) = nln(SSE(\text{Model})) + 2p' - nln(n)$ - Goal is to minimize

BIC - $BIC(\text{Model}) = nln(SSE(\text{Model})) + [ln(n)]p' - nln(n)$ - Goal is to minimize