1 Simple Linear Regression I – Least Squares Estimation

Textbook Sections: 18.1–18.3

Previously, we have worked with a random variable x that comes from a population that is normally distributed with mean μ and variance σ^2. We have seen that we can write x in terms of μ and a random error component ε, that is, $x = \mu + \varepsilon$. For the time being, we are going to change our notation for our random variable from x to y. So, we now write $y = \mu + \varepsilon$. We will now find it useful to call the random variable y a **dependent** or **response variable**. Many times, the response variable of interest may be related to the value(s) of one or more known or controllable **independent** or **predictor variables**. Consider the following situations:

LR1 A college recruiter would like to be able to **predict** a potential incoming student's first–year GPA (y) based on known information concerning high school GPA (x_1) and college entrance examination score (x_2). She feels that the student's first–year GPA will be related to the values of these two known variables.

LR2 A marketer is interested in the **effect** of changing shelf height (x_1) and shelf width (x_2) on the weekly sales (y) of her brand of laundry detergent in a grocery store.

LR3 A psychologist is interested in testing whether the amount of time to become proficient in a foreign language (y) is related to the child's age (x).

In each case we have at least one variable that is known (in some cases it is controllable), and a response variable that is a random variable. We would like to fit a model that relates the response to the known or controllable variable(s). The main reasons that scientists and social researchers use linear regression are the following:

1. **Prediction** – To predict a future response based on known values of the predictor variables and past data related to the process.
2. **Description** – To measure the effect of changing a controllable variable on the mean value of the response variable.
3. **Control** – To confirm that a process is providing responses (results) that we 'expect' under the present operating conditions (measured by the level(s) of the predictor variable(s)).

1.1 A Linear Deterministic Model

Suppose you are a vendor who sells a product that is in high demand (e.g. cold beer on the beach, cable television in Gainesville, or life jackets on the *Titanic*, to name a few). If you begin your day with 100 items, have a profit of \$10 per item, and an overhead of \$30 per day, you know exactly how much profit you will make that day, namely 100(10)-30=\$970. Similarly, if you begin the day with 50 items, you can also state your profits with certainty. In fact for any number of items you begin the day with (x), you can state what the day's profits (y) will be. That is,

$$y = 10 \cdot x - 30.$$

This is called a **deterministic** model. In general, we can write the equation for a straight line as

$$y = \beta_0 + \beta_1 x,$$

where β_0 is called the **y–intercept** and β_1 is called the **slope**. β_0 is the value of y when $x = 0$, and β_1 is the change in y when x increases by 1 unit. In many real–world situations, the response of interest (in this example it's profit) cannot be explained perfectly by a deterministic model. In this case, we make an adjustment for random variation in the process.

1.2 A Linear Probabilistic Model

The adjustment people make is to write the **mean response** as a linear function of the predictor variable. This way, we allow for variation in individual responses (y), while associating the mean linearly with the predictor x. The model we fit is as follows:

$$E(y|x) = \beta_0 + \beta_1 x,$$

and we write the individual responses as

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

We can think of y as being broken into a systematic and a random component:

$$y = \underbrace{\beta_0 + \beta_1 x}_{\text{systematic}} + \underbrace{\varepsilon}_{random}$$

where x is the level of the predictor variable corresponding to the response, β_0 and β_1 are unknown **parameters**, and ε is the random error component corresponding to the response whose distribution we assume is $N(0, \sigma)$, as before. Further, we assume the error terms are independent from one another, we discuss this in more detail in a later chapter. Note that β_0 can be interpreted as the mean response when x=0, and β_1 can be interpreted as the change in the mean response when x is increased by 1 unit. Under this model, we are saying that $y|x \sim N(\beta_0 + \beta_1 x, \sigma)$. Consider the following example.

Example 1.1 – Coffee Sales and Shelf Space

A marketer is interested in the relation between the width of the shelf space for her brand of coffee (x) and weekly sales (y) of the product in a suburban supermarket (assume the height is always at eye level). Marketers are well aware of the concept of 'compulsive purchases', and know that the more shelf space their product takes up, the higher the frequency of such purchases. She believes that in the range of 3 to 9 feet, the **mean weekly sales** will be linearly related to the width of the shelf space. Further, among weeks with the same shelf space, she believes that sales will be normally distributed with unknown standard deviation σ (that is, σ measures how variable weekly sales are at a given amount of shelf space). Thus, she would like to fit a model relating weekly sales y to the amount of shelf space x her product receives that week. That is, she is fitting the model:

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

so that $y|x \sim N(\beta_0 + \beta_1 x, \sigma)$.

One limitation of linear regression is that we must restrict our interpretation of the model to the range of values of the predictor variables that we observe in our data. We cannot assume this linear relation continues outside the range of our sample data.

We often refer to $\beta_0 + \beta_1 x$ as the *systematic component* of y and ε as the *random component.*

1.3 Least Squares Estimation of β_0 and β_1

We now have the problem of using sample data to compute estimates of the parameters β_0 and β_1. First, we take a sample of n subjects, observing values y of the response variable and x of the predictor variable. We would like to choose as estimates for β_0 and β_1, the values b_0 and b_1 that 'best fit' the sample data. Consider the coffee example mentioned earlier. Suppose the marketer conducted the experiment over a twelve week period (4 weeks with 3' of shelf space, 4 weeks with 6', and 4 weeks with 9'), and observed the sample data in Table 1.

Shelf Space x	Weekly Sales y	Shelf Space x	Weekly Sales y
6	526	6	434
3	421	3	443
6	581	9	590
9	630	6	570
3	412	3	346
9	560	9	672

Table 1: Coffee sales data for $n = 12$ weeks

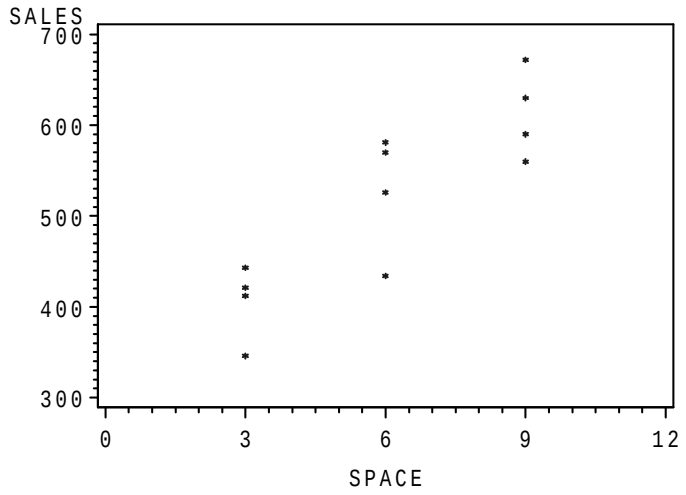

Figure 1: Plot of coffee sales vs amount of shelf space

Now, look at Figure 1. Note that while there is some variation among the weekly sales at 3', 6', and 9', respectively, there is a trend for the mean sales to increase as shelf space increases. If we define the **fitted equation** to be an equation:

$$\hat{y} = b_0 + b_1 x,$$

we can choose the estimates b_0 and b_1 to be the values that minimize the distances of the data points to the fitted line. Now, for each observed response y_i, with a corresponding predictor variable x_i, we obtain a **fitted value** $\hat{y}_i = b_0 + b_1 x_i$. So, we would like to minimize the sum of the squared distances of each observed response to its fitted value. That is, we want to minimize the **error**

sum of squares, SSE, where:

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2.$$

A little bit of calculus can be used to obtain the estimates:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{SS_{xy}}{SS_{xx}},$$

and

$$b_0 = \overline{y} - \hat{\beta}_1 \overline{x} = \frac{\sum_{i=1}^{n} y_i}{n} - b_1 \frac{\sum_{i=1}^{n} x_i}{n}.$$

An alternative formula, but exactly the same mathematically, is to compute the sample covariance of x and y, as well as the sample variance of x, then taking the ratio. This is the the approach your book uses, but is extra work from the formula above.

$$cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1} = \frac{SS_{xy}}{n-1} \qquad s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} = \frac{SS_{xx}}{n-1} \qquad b_1 = \frac{cov(x,y)}{s_x^2}$$

Some shortcut equations, known as the corrected sums of squares and crossproducts, that while not very intuitive are very useful in computing these and other estimates are:

- $SS_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}$
- $SS_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}$
- $SS_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}$

Example 1.1 Continued – Coffee Sales and Shelf Space
For the coffee data, we observe the following summary statistics in Table 2.

Week	Space (x)	Sales (y)	x^2	xy	y^2
1	6	526	36	3156	276676
2	3	421	9	1263	177241
3	6	581	36	3486	337561
4	9	630	81	5670	396900
5	3	412	9	1236	169744
6	9	560	81	5040	313600
7	6	434	36	2604	188356
8	3	443	9	1329	196249
9	9	590	81	5310	348100
10	6	570	36	3420	324900
11	3	346	9	1038	119716
12	9	672	81	6048	451584
	$\sum x = 72$	$\sum y = 6185$	$\sum x^2 = 504$	$\sum xy = 39600$	$\sum y^2 = 3300627$

Table 2: Summary Calculations — Coffee sales data

From this, we obtain the following sums of squares and crossproducts.

$$SS_{xx} = \sum (x - \overline{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n} = 504 - \frac{(72)^2}{12} = 72$$

$$SS_{xy} = \sum (x - \overline{x})(y - \overline{y}) = \sum xy - \frac{(\sum x)(\sum y)}{n} = 39600 - \frac{(72)(6185)}{12} = 2490$$

$$SS_{yy} = \sum (y - \overline{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n} = 3300627 - \frac{(6185)^2}{12} = 112772.9$$

From these, we obtain the least squares estimate of the true linear regression relation $(\beta_0 + \beta_1 x)$.

$$b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{2490}{72} = 34.5833$$

$$b_0 = \frac{\sum y}{n} - b_1 \frac{\sum x}{n} = \frac{6185}{12} - 34.5833(\frac{72}{12}) = 515.4167 - 207.5000 = 307.967.$$

$$\hat{y} \quad = \quad b_0 + b_1 x \quad = \quad 307.967 + 34.583x$$

So the fitted equation, estimating the mean weekly sales when the product has x feet of shelf space is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 307.967 + 34.5833x$. Our interpretation for b_1 is "the estimate for the increase in mean weekly sales due to increasing shelf space by 1 foot is 34.5833 bags of coffee". Note that this should only be interpreted within the range of x values that we have observed in the "experiment", namely $x = 3$ to 9 feet.

Example 1.2 – Computation of a Stock Beta

A widely used measure of a company's performance is their beta. This is a measure of the firm's stock price volatility relative to the overall market's volatility. One common use of beta is in the capital asset pricing model (CAPM) in finance, but you will hear them quoted on many business news shows as well. It is computed as (*Value Line*):

> The "beta factor" is derived from a least squares regression analysis between weekly percent changes in the price of a stock and weekly percent changes in the price of all stocks in the survey over a period of five years. In the case of shorter price histories, a smaller period is used, but never less than two years.

In this example, we will compute the stock beta over a 28-week period for Coca-Cola and Anheuser-Busch, using the S&P500 as 'the market' for comparison. Note that this period is only about 10% of the period used by *Value Line*. Note: While there are 28 weeks of data, there are only n=27 weekly changes.

Table 3 provides the dates, weekly closing prices, and weekly percent changes of: the S&P500, Coca-Cola, and Anheuser-Busch. The following summary calculations are also provided, with x representing the S&P500, y_C representing Coca-Cola, and y_A representing Anheuser-Busch. All calculations should be based on 4 decimal places. Figure 2 gives the plot and least squares regression line for Anheuser-Busch, and Figure 3 gives the plot and least squares regression line for Coca-Cola.

$$\sum x = 15.5200 \qquad \sum y_C = -2.4882 \qquad \sum y_A = 2.4281$$

$$\sum x^2 = 124.6354 \qquad \sum y_C^2 = 461.7296 \qquad \sum y_A^2 = 195.4900$$

$$\sum xy_C = 161.4408 \qquad \sum xy_A = 84.7527$$

a) Compute SS_{xx}, SS_{xy_C}, and SS_{xy_A}.

b) Compute the stock betas for Coca-Cola and Anheuser-Busch.

Closing Date	S&P Price	A-B Price	C-C Price	S&P % Chng	A-B % Chng	C-C % Chng
05/20/97	829.75	43.00	66.88	–	–	–
05/27/97	847.03	42.88	68.13	2.08	-0.28	1.87
06/02/97	848.28	42.88	68.50	0.15	0.00	0.54
06/09/97	858.01	41.50	67.75	1.15	-3.22	-1.09
06/16/97	893.27	43.00	71.88	4.11	3.61	6.10
06/23/97	898.70	43.38	71.38	0.61	0.88	-0.70
06/30/97	887.30	42.44	71.00	-1.27	-2.17	-0.53
07/07/97	916.92	43.69	70.75	3.34	2.95	-0.35
07/14/97	916.68	43.75	69.81	-0.03	0.14	-1.33
07/21/97	915.30	45.50	69.25	-0.15	4.00	-0.80
07/28/97	938.79	43.56	70.13	2.57	-4.26	1.27
08/04/97	947.14	43.19	68.63	0.89	-0.85	-2.14
08/11/97	933.54	43.50	62.69	-1.44	0.72	-8.66
08/18/97	900.81	42.06	58.75	-3.51	-3.31	-6.28
08/25/97	923.55	43.38	60.69	2.52	3.14	3.30
09/01/97	899.47	42.63	57.31	-2.61	-1.73	-5.57
09/08/97	929.05	44.31	59.88	3.29	3.94	4.48
09/15/97	923.91	44.00	57.06	-0.55	-0.70	-4.71
09/22/97	950.51	45.81	59.19	2.88	4.11	3.73
09/29/97	945.22	45.13	61.94	-0.56	-1.48	4.65
10/06/97	965.03	44.75	62.38	2.10	-0.84	0.71
10/13/97	966.98	43.63	61.69	0.20	-2.50	-1.11
10/20/97	944.16	42.25	58.50	-2.36	-3.16	-5.17
10/27/97	941.64	40.69	55.50	-0.27	-3.69	-5.13
11/03/97	914.62	39.94	56.63	-2.87	-1.84	2.04
11/10/97	927.51	40.81	57.00	1.41	2.18	0.65
11/17/97	928.35	42.56	57.56	0.09	4.29	0.98
11/24/97	963.09	43.63	63.75	3.74	2.51	10.75

Table 3: Weekly closing stock prices – S&P 500, Anheuser-Busch, Coca-Cola

Example 1.3 – Estimating Cost Functions of a Hosiery Mill

The following (approximate) data were published by Joel Dean, in the 1941 article: "Statistical Cost Functions of a Hosiery Mill," (*Studies in Business Administration*, vol. 14, no. 3).

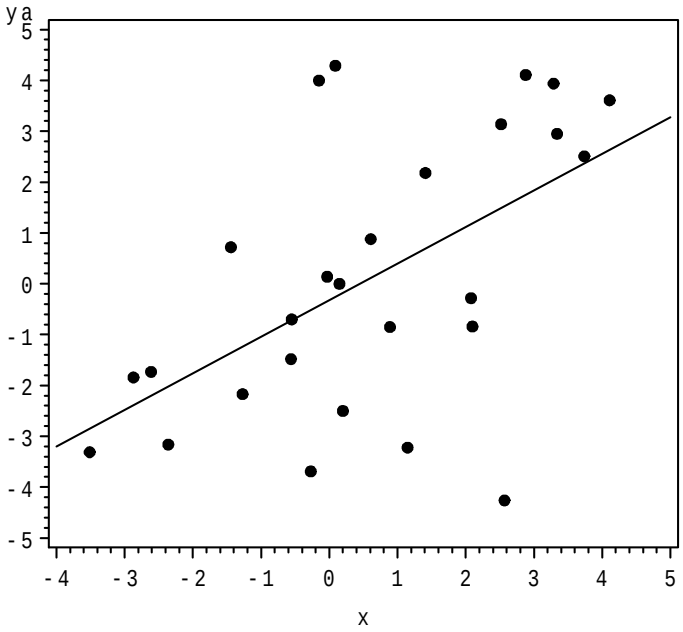

Figure 2: Plot of weekly percent stock price changes for Anheuser-Busch versus S&P 500 and least squares regression line

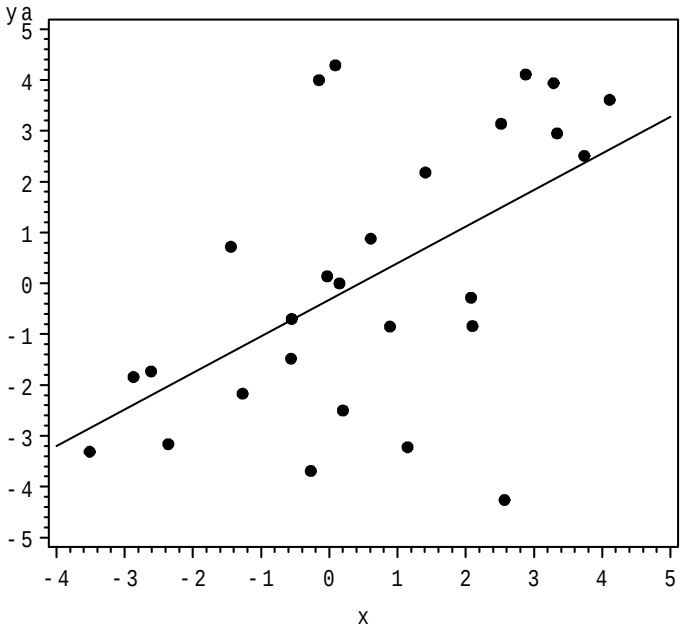

Figure 3: Plot of weekly percent stock price changes for Coca-Cola versus S&P 500 and least squares regression line

y — Monthly total production cost (in \$1000s).

x — Monthly output (in thousands of dozens produced).

A sample of $n = 48$ months of data were used, with x_i and y_i being measured for each month. The parameter β_1 represents the change in mean cost per unit increase in output (unit variable cost), and β_0 represents the true mean cost when the output is 0, without shutting plant (fixed cost). The data are given in Table 1.3 (the order is arbitrary as the data are printed in table form, and were obtained from visual inspection/approximation of plot).

i	x_i	y_i	i	x_i	y_i	i	x_i	y_i
1	46.75	92.64	17	36.54	91.56	33	32.26	66.71
2	42.18	88.81	18	37.03	84.12	34	30.97	64.37
3	41.86	86.44	19	36.60	81.22	35	28.20	56.09
4	43.29	88.80	20	37.58	83.35	36	24.58	50.25
5	42.12	86.38	21	36.48	82.29	37	20.25	43.65
6	41.78	89.87	22	38.25	80.92	38	17.09	38.01
7	41.47	88.53	23	37.26	76.92	39	14.35	31.40
8	42.21	91.11	24	38.59	78.35	40	13.11	29.45
9	41.03	81.22	25	40.89	74.57	41	9.50	29.02
10	39.84	83.72	26	37.66	71.60	42	9.74	19.05
11	39.15	84.54	27	38.79	65.64	43	9.34	20.36
12	39.20	85.66	28	38.78	62.09	44	7.51	17.68
13	39.52	85.87	29	36.70	61.66	45	8.35	19.23
14	38.05	85.23	30	35.10	77.14	46	6.25	14.92
15	39.16	87.75	31	33.75	75.47	47	5.45	11.44
16	38.59	92.62	32	34.29	70.37	48	3.79	12.69

Table 4: Production costs and Output – Dean (1941)

.

This dataset has $n = 48$ observations with a mean output (in 1000s of dozens) of $\overline{x} = 31.0673$, and a mean monthly cost (in \$1000s) of $\overline{y} = 65.4329$.

$$\sum_{i=1}^{n} x_i = 1491.23 \quad \sum_{i=1}^{n} x_i^2 = 54067.42 \quad \sum_{i=1}^{n} y_i = 3140.78 \quad \sum_{i=1}^{n} y_i^2 = 238424.46 \sum_{i=1}^{n} x_i y_i = 113095.80$$

From these quantites, we get:

- $SS_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} = 54067.42 - \frac{(1491.23)^2}{48} = 54067.42 - 46328.48 = 7738.94$
- $SS_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n} = 113095.80 - \frac{(1491.23)(3140.78)}{48} = 113095.80 - 97575.53 = 15520.27$
- $SS_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n} = 238424.46 - \frac{(3140.78)^2}{48} = 238424.46 - 205510.40 = 32914.06$

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} = \frac{SS_{xy}}{SS_{xx}} = \frac{15520.27}{7738.94} = 2.0055$$

$$b_0 = \overline{y} - b_1\overline{x} = 65.4329 - (2.0055)(31.0673) = 3.1274$$
$$\hat{y}_i = b_0 + b_1 x_i = 3.1274 + 2.0055x_i \qquad i = 1, \ldots, 48$$
$$e_i = y_i - \hat{y}_i = y_i - (3.1274 + 2.0055x_i) \qquad i = 1, \ldots, 48$$

Table 1.3 gives the raw data, their fitted values, and residuals.

A plot of the data and regression line are given in Figure 4.

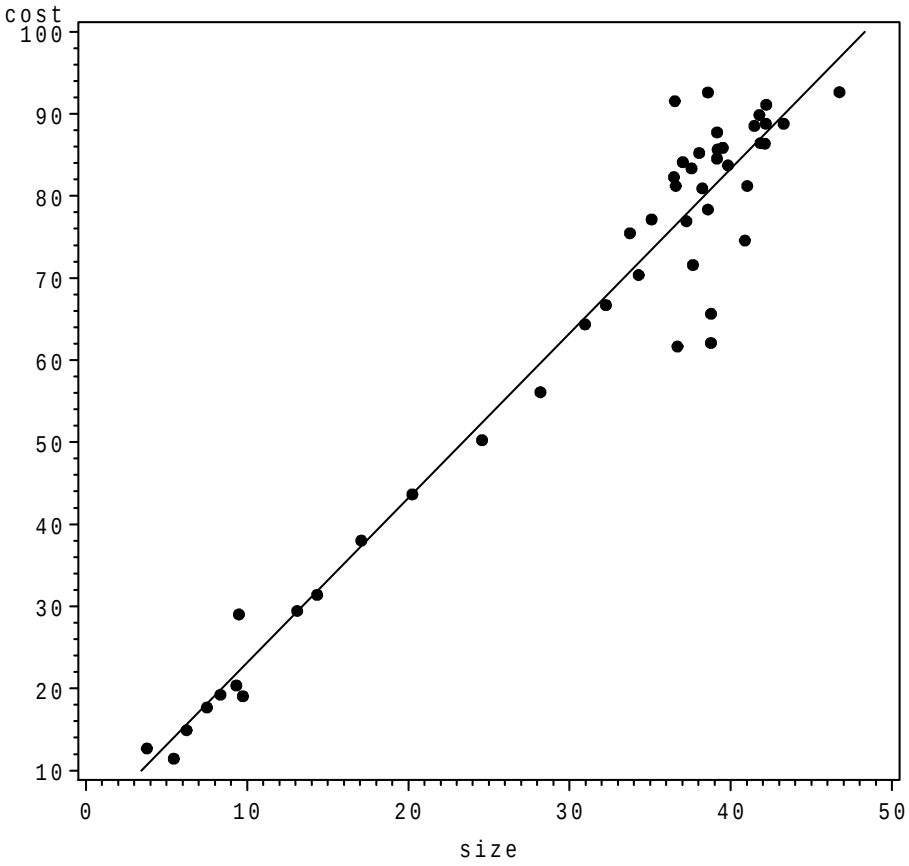

Figure 4: Estimated cost function for hosiery mill (Dean, 1941)

We have seen now, how to estimate β_0 and β_1. Now we can obtain an estimate of the variance of the responses at a given value of x. Recall from your previous statistics course, you estimated the variance by taking the 'average' squared deviation of each measurement from the sample (estimated) mean. That is, you calculated $s_y^2 = \frac{\sum_{i=1}^n (y_i - \overline{y})^2}{n-1}$. Now that we fit the regression model, we know longer use $\overline{y}$ to estimate the mean for each y_i, but rather $\hat{y}_i = b_0 + b_1 x_i$ to estimate the mean. The estimate we use now looks similar to the previous estimate except we replace $\overline{y}$ with $\hat{y}_i$ and we replace $n-1$ with $n-2$ since we have estimated 2 parameters, β_0 and β_1. The new estimate (which we will refer as to the estimated error variance) is:

$$s_e^2 = MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)}{n-2} = \frac{SS_{yy} - \frac{(SS_{xy})^2}{SS_{xx}}}{n-2} = \left(\frac{n-1}{n-2}\right)\left(s_y^2 - \frac{[cov(x,y)]^2}{s_x^2}\right)$$

i	x_i	y_i	$\hat{y}_i$	e_i
1	46.75	92.64	96.88	-4.24
2	42.18	88.81	87.72	1.09
3	41.86	86.44	87.08	-0.64
4	43.29	88.80	89.95	-1.15
5	42.12	86.38	87.60	-1.22
6	41.78	89.87	86.92	2.95
7	41.47	88.53	86.30	2.23
8	42.21	91.11	87.78	3.33
9	41.03	81.22	85.41	-4.19
10	39.84	83.72	83.03	0.69
11	39.15	84.54	81.64	2.90
12	39.20	85.66	81.74	3.92
13	39.52	85.87	82.38	3.49
14	38.05	85.23	79.44	5.79
15	39.16	87.75	81.66	6.09
16	38.59	92.62	80.52	12.10
17	36.54	91.56	76.41	15.15
18	37.03	84.12	77.39	6.73
19	36.60	81.22	76.53	4.69
20	37.58	83.35	78.49	4.86
21	36.48	82.29	76.29	6.00
22	38.25	80.92	79.84	1.08
23	37.26	76.92	77.85	-0.93
24	38.59	78.35	80.52	-2.17
25	40.89	74.57	85.13	-10.56
26	37.66	71.60	78.65	-7.05
27	38.79	65.64	80.92	-15.28
28	38.78	62.09	80.90	-18.81
29	36.70	61.66	76.73	-15.07
30	35.10	77.14	73.52	3.62
31	33.75	75.47	70.81	4.66
32	34.29	70.37	71.90	-1.53
33	32.26	66.71	67.82	-1.11
34	30.97	64.37	65.24	-0.87
35	28.20	56.09	59.68	-3.59
36	24.58	50.25	52.42	-2.17
37	20.25	43.65	43.74	-0.09
38	17.09	38.01	37.40	0.61
39	14.35	31.40	31.91	-0.51
40	13.11	29.45	29.42	0.03
41	9.50	29.02	22.18	6.84
42	9.74	19.05	22.66	-3.61
43	9.34	20.36	21.86	-1.50
44	7.51	17.68	18.19	-0.51
45	8.35	19.23	19.87	-0.64
46	6.25	14.92	15.66	-0.74
47	5.45	11.44	14.06	-2.62
48	3.79	12.69	10.73	1.96

Table 5: Approximated Monthly Outputs, total costs, fitted values and residuals – Dean (1941)

.

This estimated error variance s_e^2 can be thought of as the 'average' squared distance from each observed response to the fitted line. The word average is in quotes since we divide by $n-2$ and not n. The closer the observed responses fall to the line, the smaller s_e^2 is and the better our predicted values will be.

Example 1.1 (Continued) – Coffee Sales and Shelf Space
For the coffee data,

$$s_e^2 = \frac{112772.9 - \frac{(2490)^2}{72}}{12-2} = \frac{112772.9 - 86112.5}{10} = 2666.04,$$

and the estimated residual standard error (deviation) is $S_e = \sqrt{2666.04} = 51.63$. We now have estimates for all of the parameters of the regression equation relating the mean weekly sales to the amount of shelf space the coffee gets in the store. Figure 5 shows the 12 observed responses and the estimated (fitted) regression equation.

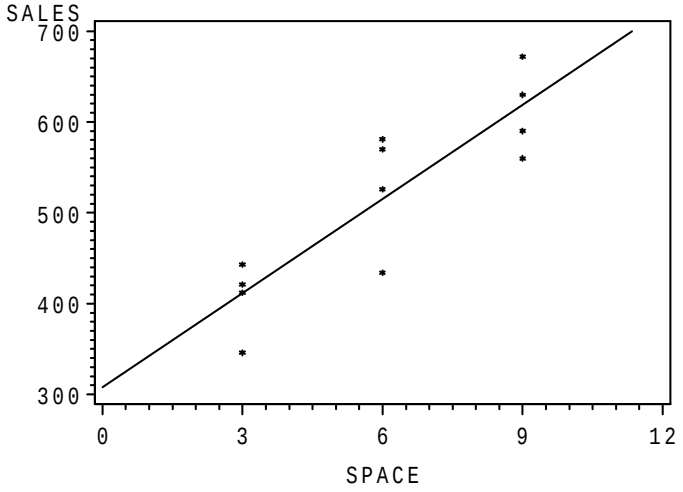

Figure 5: Plot of coffee data and fitted equation

Example 10.3 (Continued) – Estimating Cost Functions of a Hosiery Mill

For the cost function data:

- $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = SS_{yy} - \frac{SS_{xy}^2}{SS_{xx}} = 32914.06 - \frac{(15520.27)^2}{7738.94} = 32914.06 - 31125.55 = 1788.51$
- $s_e^2 = MSE = \frac{SSE}{n-2} = \frac{1788.51}{48-2} = 38.88$
- $s_e = \sqrt{38.88} = 6.24$

2 Simple Regression II — Inferences Concerning β_1

Textbook Section: 18.5 (and some supplementary material)

Recall that in our regression model, we are stating that $E(y|x) = \beta_0 + \beta_1 x$. In this model, β_1 represents the change in the mean of our response variable y, as the predictor variable x increases by 1 unit. Note that if $\beta_1 = 0$, we have that $E(y|x) = \beta_0 + \beta_1 x = \beta_0 + 0x = \beta_0$, which implies the mean of our response variable is the same at all values of x. In the context of the coffee sales example, this would imply that mean sales are the same, regardless of the amount of shelf space, so a marketer has no reason to purchase extra shelf space. This is like saying that knowing the level of the predictor variable does not help us predict the response variable.

Under the assumptions stated previously, namely that $y \sim N(\beta_0 + \beta_1 x, \sigma)$, our estimator b_1 has a sampling distribution that is normal with mean β_1 (the true value of the parameter), and standard error $\frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$. That is:

$$b_1 \sim N(\beta_1, \frac{\sigma}{\sqrt{SS_{xx}}})$$

We can now make inferences concerning β_1.

2.1 A Confidence Interval for β_1

First, we obtain the estimated standard error of b_1 (this is the standard deviation of its sampling distibution:

$$s_{b_1} = \frac{s_e}{\sqrt{SS_{xx}}} = \frac{s_e}{\sqrt{(n-1)s_x^2}}$$

The interval can be written:

$$b_1 \pm t_{\alpha/2,n-2} s_{b_1} \equiv b_1 \pm t_{\alpha/2,n-2} \frac{s_e}{\sqrt{SS_{xx}}}.$$

Note that $\frac{s_e}{\sqrt{SS_{xx}}}$ is the estimated standard error of b_1 since we use $s_e = \sqrt{MSE}$ to estimate σ. Also, we have $n-2$ degrees of freedom instead of $n-1$, since the estimate s_e^2 has 2 estimated paramters used in it (refer back to how we calculate it above).

Example 2.1 – Coffee Sales and Shelf Space

For the coffee sales example, we have the following results:

$$b_1 = 34.5833, \quad SS_{xx} = 72, \quad s_e = 51.63, \quad n = 12.$$

So a 95% confidence interval for the parameter β_1 is:

$$34.5833 \pm t_{.025,12-2} \frac{51.63}{\sqrt{72}} = 34.5833 \pm 2.228(6.085) = 34.583 \pm 13.557,$$

which gives us the range (21.026, 48.140). We are 95% confident that the true mean sales increase by between 21.026 and 48.140 bags of coffee per week for each extra foot of shelf space the brand

gets (within the range of 3 to 9 feet). Note that the entire interval is positive (above 0), so we are confident that in fact $\beta_1 > 0$, so the marketer is justified in pursuing extra shelf space.

Example 2.2 – Hosiery Mill Cost Function

$$b_1 = 2.0055, \quad SS_{xx} = 7738.94, \quad s_e = 6.24, \quad n = 48.$$

For the hosiery mill cost function analysis, we obtain a 95% confidence interval for average unit variable costs (β_1). Note that $t_{.025,48-2} = t_{.025,46} \approx 2.015$, since $t_{.025,40} = 2.021$ and $t_{.025,60} = 2.000$ (we could approximate this with $z_{.025} = 1.96$ as well).

$$2.0055 \pm t_{.025,46}\frac{6.24}{\sqrt{7738.94}} = 2.0055 \pm 2.015(.0709) = 2.0055 \pm 0.1429 = (1.8626, 2.1484)$$

We are 95% confident that the true average unit variable costs are between \$1.86 and \$2.15 (this is the incremental cost of increasing production by one unit, assuming that the production process is in place.

2.2 Hypothesis Tests Concerning β_1

Similar to the idea of the confidence interval, we can set up a test of hypothesis concerning β_1. Since the confidence interval gives us the range of 'believable' values for β_1, it is more useful than a test of hypothesis. However, here is the procedure to test if β_1 is equal to some value, say β_1^0.

- $H_0 : \beta_1 = \beta_1^0$ (β_1^0 specified, usually 0)
- (1) $H_a : \beta_1 \neq \beta_1^0$
  (2) $H_a : \beta_1 > \beta_1^0$
  (3) $H_a : \beta_1 < \beta_1^0$
- $TS : t_{obs} = \frac{b_1 - \beta_1^0}{\frac{s_e}{\sqrt{SS_{xx}}}} = \frac{b_1 - \beta_1^0}{s_{b_1}}$
- (1) $RR : |t_{obs}| \geq t_{\alpha/2,n-2}$
  (2) $RR : t_{obs} \geq t_{\alpha,n-2}$
  (3) $RR : t_{obs} \leq -t_{\alpha,n-2}$
- (1) P–value: $2 \cdot P(t \geq |t_{obs}|)$
  (2) P–value: $P(t \geq t_{obs})$
  (3) P–value: $P(t \leq t_{obs})$

  Using tables, we can only place bounds on these p–values.

Example 2.1 (Continued) – Coffee Sales and Shelf Space

Suppose in our coffee example, the marketer gets a set amount of space (say 6′) for free, and she must pay extra for any more space. For the extra space to be profitable (over the long run),

the mean weekly sales must increase by more than 20 bags, otherwise the expense outweighs the increase in sales. She wants to test to see if it is worth it to buy more space. She works under the assumption that it is not worth it, and will only purchase more if she can show that it is worth it. She sets $\alpha = .05$.

1. $H_0 : \beta_1 = 20 \quad H_A : \beta_1 > 20$
2. T.S.: $t_{obs} = \frac{34.5833-20}{\frac{51.63}{\sqrt{72}}} = \frac{14.5833}{6.085} = 2.397$
3. R.R.: $t_{obs} > t_{.05,10} = 1.812$
4. p-value: $P(T > 2.397) < P(T > 2.228) = .025$ and $P(T > 2.397) > P(T > 2.764) = .010$, so $.01 < p - value < .025$.

So, she has concluded that $\beta_1 > 20$, and she will purchase the shelf space. Note also that the entire confidence interval was over 20, so we already knew this.

Example 2.2 (Continued) – Hosiery Mill Cost Function

Suppose we want to test whether average monthly production costs increase with monthly production output. This is testing whether unit variable costs are positive ($\alpha = 0.05$).

- $H_0 : \beta_1 = 0$ (Mean Monthly production cost is not associated with output)
- $H_A : \beta_1 > 0$ (Mean monthly production cost increases with output)
- $TS : t_{obs} = \frac{2.0055-0}{\frac{6.24}{\sqrt{7738.94}}} = \frac{2.0055}{0.0709} = 28.29$
- $RR : t_{obs} > t_{0.05,46} \approx 1.680$ (or use $z_{0.05} = 1.645$)
- p-value: $P(T > 28.29) \approx 0$

We have overwhelming evidence of positive unit variable costs.

2.3 The Analysis of Variance Approach to Regression

Consider the deviations of the individual responses, y_i, from their overall mean $\overline{y}$. We would like to break these deviations into two parts, the deviation of the observed value from its fitted value, $\hat{y}_i = b_0 + b_1 x_i$, and the deviation of the fitted value from the overall mean. See Figure 6 corresponding to the coffee sales example. That is, we'd like to write:

$$y_i - \overline{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \overline{y}).$$

Note that all we are doing is adding and subtracting the fitted value. It so happens that algebraically we can show the same equality holds once we've squared each side of the equation and summed it over the n observed and fitted values. That is,

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2.$$

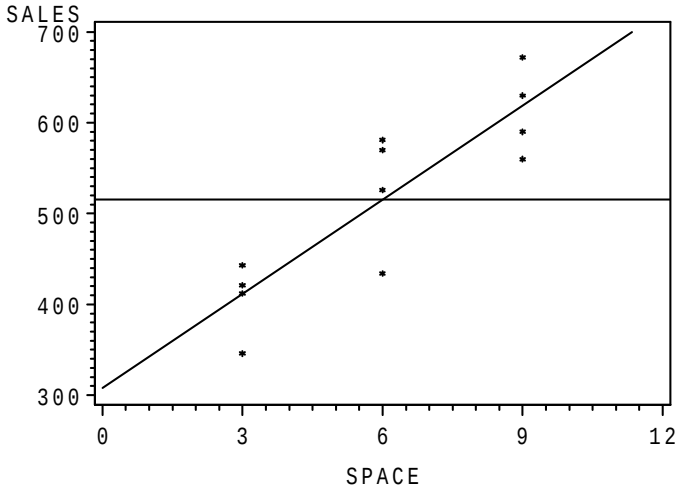

Figure 6: Plot of coffee data, fitted equation, and the line $\overline{y} = 515.4167$

These three pieces are called the **total**, **error**, and **model sums of squares**, respectively. We denote them as SS_{yy}, SSE, and SSR, respectively. We have already seen that SS_{yy} represents the total variation in the observed responses, and that SSE represents the variation in the observed responses around the fitted regression equation. That leaves SSR as the amount of the total variation that is 'accounted for' by taking into account the predictor variable X. We can use this decomposition to test the hypothesis $H_0 : \beta_1 = 0$ vs $H_A : \beta_1 \neq 0$. We will also find this decomposition useful in subsequent sections when we have more than one predictor variable. We first set up the **Analysis of Variance (ANOVA) Table** in Table 6. Note that we will have to make minimal calculations to set this up since we have already computed SS_{yy} and SSE in the regression analysis.

ANOVA				
Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
MODEL	$SSR = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$	1	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
ERROR	$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$n-2$	$MSE = \frac{SSE}{n-2}$	
TOTAL	$SS_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2$	$n-1$		

Table 6: The Analysis of Variance Table for simple regression

The procedure of testing for a linear association between the response and predictor variables using the analysis of variance involves using the F–distribution, which is given in Table 6 (pp B-11–B-16) of your text book. This is the same distribution we used in the chapteron the 1-Way ANOVA.

The testing procedure is as follows:

1. $H_0 : \beta_1 = 0 \quad H_A : \beta_1 \neq 0$ (This will always be a 2–sided test)
2. T.S.: $F_{obs} = \frac{MSR}{MSE}$
3. R.R.: $F_{obs} > F_{1,n-2,\alpha}$

4. p-value: $P(F > F_{obs})$ (You can only get bounds on this, but computer outputs report them exactly)

Note that we already have a procedure for testing this hypothesis (see the section on Inferences Concerning β_1), but this is an important lead–in to multiple regression.

Example 2.1 (Continued) – Coffee Sales and Shelf Space

Referring back to the coffee sales data, we have already made the following calculations:

$$SS_{yy} = 112772.9, \quad SSE = 26660.4, \quad n = 12.$$

We then also have that $SSR = SS_{yy} - SSE = 86112.5$. Then the Analysis of Variance is given in Table 7.

ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
MODEL	$SSR = 86112.5$	1	$MSR = \frac{86112.5}{1} = 86112.5$	$F = \frac{86112.5}{2666.04} = 32.30$
ERROR	$SSE = 26660.4$	$12 - 2 = 10$	$MSE = \frac{26660.4}{10} = 2666.04$	
TOTAL	$SS_{yy} = 112772.9$	$12 - 1 = 11$		

Table 7: The Analysis of Variance Table for the coffee data example

To test the hypothesis of no linear association between amount of shelf space and mean weekly coffee sales, we can use the F-test described above. Note that the null hypothesis is that there is no effect on mean sales from increasing the amount of shelf space. We will use $\alpha = .01$.

1. $H_0 : \beta_1 = 0 \quad H_A : \beta_1 \neq 0$
2. T.S.: $F_{obs} = \frac{MSR}{MSE} = \frac{86112.5}{2666.04} = 32.30$
3. R.R.: $F_{obs} > F_{1,n-2,\alpha} = F_{1,10,.01} = 10.04$
4. p-value: $P(F > F_{obs}) = P(F > 32.30) \approx 0$

We reject the null hypothesis, and conclude that $\beta_1 \neq 0$. There is an effect on mean weekly sales when we increase the shelf space.

Example 2.2 (Continued) – Hosiery Mill Cost Function

For the hosiery mill data, the sums of squares for each source of variation in monthly production costs and their corresponding degrees of freedom are (from previous calculations):

Total SS – $SS_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2 = 32914.06 \qquad df_{Total} = n - 1 = 47$

Error SS – $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 1788.51 \qquad df_E = n - 2 = 46$

Model SS – $SSR = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 = SS_{yy} - SSE = 32914.06 - 1788.51 = 31125.55 \qquad df_R = 1$

ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
MODEL	$SSR = 31125.55$	1	$MSR = \frac{31125.55}{1} = 31125.55$	$F = \frac{31125.55}{38.88} = 800.55$
ERROR	$SSE = 1788.51$	$48 - 2 = 46$	$MSE = \frac{1788.51}{46} = 38.88$	
TOTAL	$SS_{yy} = 32914.06$	$48 - 1 = 47$		

Table 8: The Analysis of Variance Table for the hosiery mill cost example

The Analysis of Variance is given in Table 8.

To test whether there is a linear association between mean monthly costs and monthly production output, we conduct the F-test ($\alpha = 0.05$).

1. $H_0 : \beta_1 = 0 \quad H_A : \beta_1 \neq 0$
2. T.S.: $F_{obs} = \frac{MSR}{MSE} = \frac{31125.55}{38.88} = 800.55$
3. R.R.: $F_{obs} > F_{1,n-2,\alpha} = F_{1,46,.05} \approx 4.06$
4. p-value: $P(F > F_{obs}) = P(F > 800.55) \approx 0$

We reject the null hypothesis, and conclude that $\beta_1 \neq 0$.

2.3.1 Coefficient of Determination

A measure of association that has a clear physical interpretation is R^2, the **coefficient of determination**. This measure is always between 0 and 1, so it does not reflect whether y and x are positively or negatively associated, and it represents the proportion of the total variation in the response variable that is 'accounted' for by fitting the regression on x. The formula for R^2 is:

$$R^2 = (R)^2 = 1 - \frac{SSE}{SS_{yy}} = \frac{SSR}{SS_{yy}} = \frac{[cov(x,y)]^2}{s_x^2 s_y^2}.$$

Note that $SS_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2$ represents the total variation in the response variable, while $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ represents the variation in the observed responses about the fitted equation (after taking into account x). This is why we sometimes say that R^2 is "proportion of the variation in y that is 'explained' by x."

Example 2.1 (Continued) – Coffee Sales and Shelf Space

For the coffee data, we can calculate R^2 using the values of $SS_{xy}, SS_{xx}, SS_{yy}$, and SSE we have previously obtained.

$$R^2 = 1 - \frac{26660.4}{112772.9} = \frac{86112.5}{112772.9} = .7636$$

Thus, over 3/4 of the variation in sales is "explained" by the model using shelf space to predict sales.

Example 2.2 (Continued) – Hosiery Mill Cost Function

For the hosiery mill data, the model (regression) sum of squares is $SSR = 31125.55$ and the total sum of squares is $SS_{yy} = 32914.06$. To get the coefficient of determination:

$$R^2 = \frac{31125.55}{32914.06} = 0.9457$$

Almost 95% of the variation in monthly production costs is "explained" by the monthly production output.

3 Simple Regression III – Estimating the Mean and Prediction at a Particular Level of x, Correlation

Textbook Sections: 18.7,18.8

We sometimes are interested in estimating the mean response at a particular level of the predictor variable, say $x = x_g$. That is, we'd like to estimate $E(y|x_g) = \beta_0 + \beta_1 x_g$. The actual estimate 9**point prediction**) is just $\hat{y} = b_0 + b_1 x_g$, which is simply where the fitted line crosses $x = x_g$. Under the previously stated normality assumptions, the estimator $\hat{y}_0$ is normally distributed with mean $\beta_0 + \beta_1 x_g$ and standard error of estimate $\sigma\sqrt{\frac{1}{n} + \frac{(x_g - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}$. That is:

$$\hat{y}_0 \sim N(\beta_0 + \beta_1 x_g, \quad \sigma\sqrt{\frac{1}{n} + \frac{(x_g - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}).$$

Note that the standard error of the estimate is smallest at $x_g = \overline{x}$, that is at the mean of the sampled levels of the predictor variable. The standard error increases as the value x_g goes away from this mean.

For instance, our marketer may wish to estimate the mean sales when she has $6'$ of shelf space, or $7'$, or $4'$. She may also wish to obtain a confidence interval for the mean at these levels of x.

3.1 A Confidence Interval for $E(y|x_g) = \beta_0 + \beta_1 x_g$

Using the ideas described in the previous section, we can write out the general form for a $(1-\alpha)100\%$ confidence interval for the mean response when x_g.

$$(b_0 + b_1 x_g) \pm t_{\alpha/2,n-2} S_e \sqrt{\frac{1}{n} + \frac{(x_g - \overline{x})^2}{SS_{xx}}}$$

Example 3.1 – Coffee Sales and Shelf Space

Suppose our marketer wants to compute 95% confidence intervals for the mean weekly sales at x=4,6, and 7 feet, respectively (these are not simultaneous confidence intervals as were computed based on Bonferroni's Method previously). Each of these intervals will depend on $t_{\alpha/2,n-2} = t_{.05,10} = 2.228$ and $\overline{x} = 6$. These intervals are:

$$(307.967 + 34.5833(4)) \pm 2.228(51.63)\sqrt{\frac{1}{12} + \frac{(4-6)^2}{72}} = 446.300 \pm 115.032\sqrt{.1389}$$

$$= 446.300 \pm 42.872 \equiv (403.428, 489.172)$$

$$(307.967 + 34.5833(6)) \pm 2.228(51.63)\sqrt{\frac{1}{12} + \frac{(6-6)^2}{72}} = 515.467 \pm 115.032\sqrt{.0833}$$

$$= 515.467 \pm 33.200 \equiv (482.267, 548.667)$$

$$(307.967 + 34.5833(7)) \pm 2.228(51.63)\sqrt{\frac{1}{12} + \frac{(7-6)^2}{72}} = 550.050 \pm 115.032\sqrt{.0972}$$

$$= 550.050 \pm 35.863 \equiv (514.187, 585.913)$$

Notice that the interval is the narrowest at $x_g = 6$. Figure 7 is a computer generated plot of the data, the fitted equation and the confidence limits for the mean weekly coffee sales at each value of x. Note how the limits get wider as x goes away from $\overline{x} = 6$.

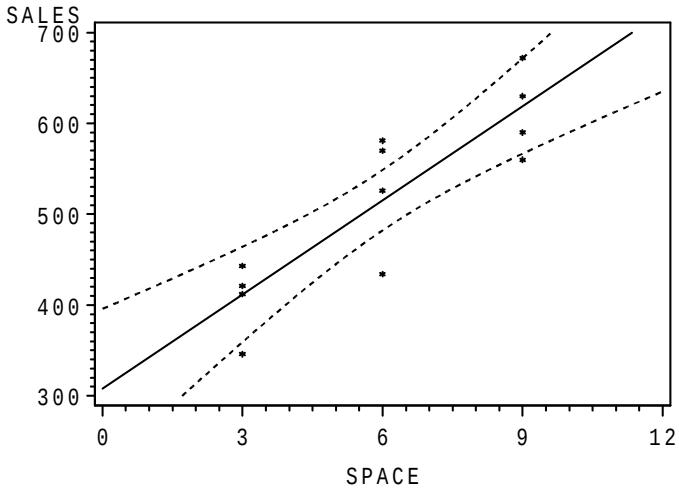

Figure 7: Plot of coffee data, fitted equation, and 95% confidence limits for the mean

Example 3.2 – Hosiery Mill Cost Function

Suppose the plant manager is interested in mean costs among months where output is 30,000 items produced ($x_g = 30$). She wants a 95% confidence interval for this true unknown mean. Recall:

$$b_0 = 3.1274 \quad b_1 = 2.0055 \quad s_e = 6.24 \quad n = 48 \quad \overline{x} = 31.0673 \quad SS_{xx} = 7738.94 t_{.025,46} \approx 2.015$$

Then the interval is obtained as:

$$3.1274 + 2.0055(30) \pm 2.015(6.24)\sqrt{\frac{1}{48} + \frac{(30 - 31.0673)^2}{7738.94}}$$

$$\equiv \quad 63.29 \pm 2.015(6.24)\sqrt{0.0210} \quad \equiv 63.29 \pm 1.82 \quad \equiv \quad (61.47, 65.11)$$

We can be 95% confident that the mean production costs among months where 30,000 items are produced is between \$61,470 and \$65,110 (recall units were thousands for x and thousands for y). A plot of the data, regression line, and 95% confidence bands for mean costs is given in Figure 8.

3.2 Predicting a Future Response at a Given Level of x

In many situations, a researcher would like to **predict** the outcome of the response variable at a specific level of the predictor variable. In the previous section we estimated the mean response, in this section we are interested in predicting a single outcome. In the context of the coffee sales example, this would be like trying to predict next week's sales given we know that we will have $6'$ of shelf space.

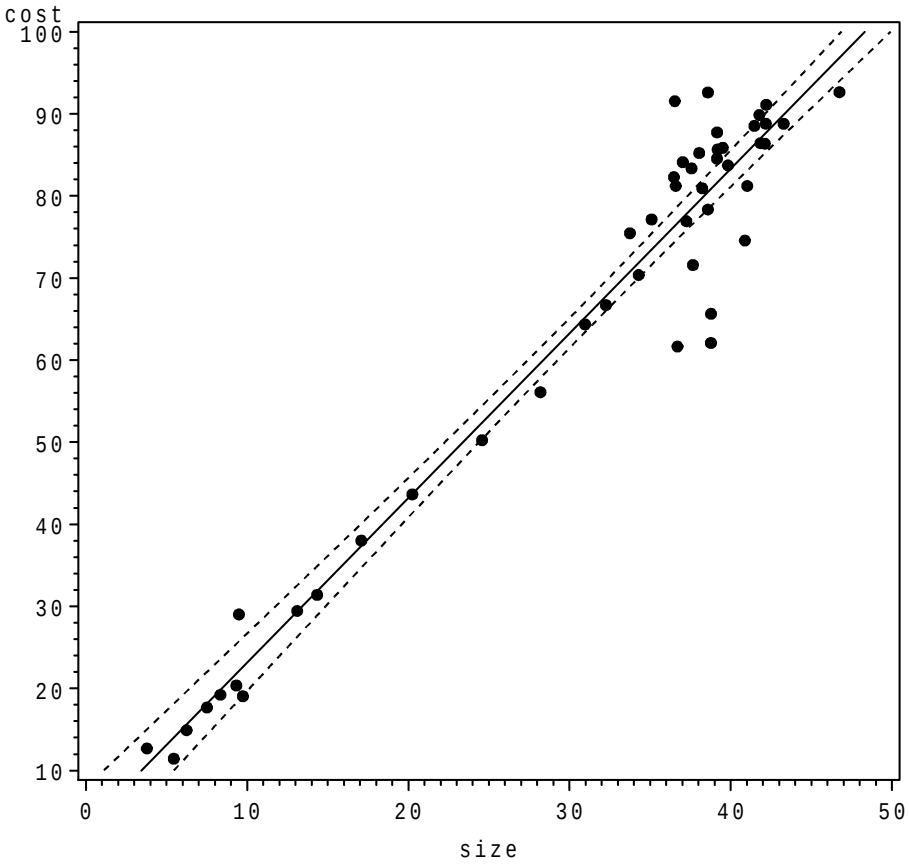

Figure 8: Plot of hosiery mill cost data, fitted equation, and 95% confidence limits for the mean

First, suppose you know the parameters β_0 and β_1 Then you know that the response variable, for a fixed level of the predictor variable ($x = x_g$), is normally distributed with mean $E(y|x_g) = \beta_0 + \beta_1 x_g$ and standard deviation σ. We know from previous work with the normal distribution that approximately 95% of the measurements lie within 2 standard deviations of the mean. So if we know β_0, β_1, and σ, we would be very confident that our response would lie between $(\beta_0+\beta_1 x_g)-2\sigma$ and $(\beta_0 + \beta_1 x_g) + 2\sigma$. Figure 9 represents this idea.

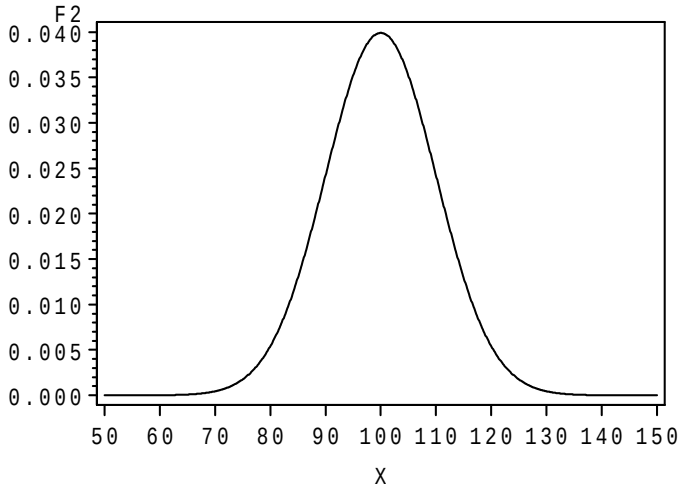

Figure 9: Distribution of response variable with known β_0, β_1, and σ

We rarely, if ever, know these parameters, and we must estimate them as we have in previous sections. There is uncertainty in what the mean response at the specified level, x_g, of the response variable. We do, however know how to obtain an interval that we are very confident contains the true mean $\beta_0 + \beta_1 x_g$. If we apply the method of the previous paragraph to all 'believable' values of this mean we can obtain a **prediction interval** that we are very confident will contain our future response. Since σ is being estimated as well, instead of 2 standard deviations, we must use $t_{\alpha/2,n-2}$ estimated standard deviations. Figure 10 portrays this idea.

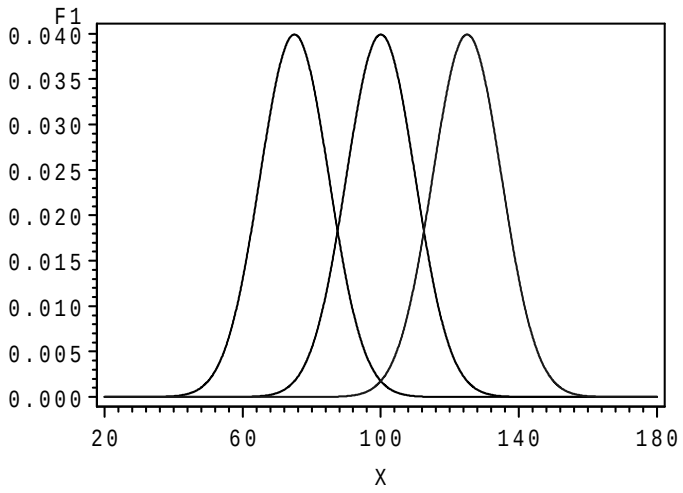

Figure 10: Distribution of response variable with estimated β_0, β_1, and σ

Note that all we really need are the two extreme distributions from the confidence interval for

the mean response. If we use the method from the last paragraph on each of these two distributions, we can obtain the prediction interval by choosing the left–hand point of the 'lower' distribution and the right–hand point of the 'upper' distribution. This is displayed in Figure 11.

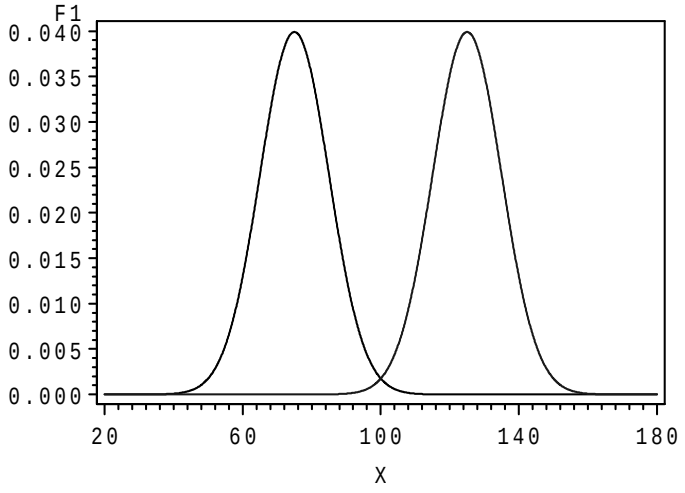

Figure 11: Upper and lower prediction limits when we have estimated the mean

The general formula for a $(1-\alpha)100\%$ prediction interval of a future response is similar to the confidence interval for the mean at x_g, except that it is wider to reflect the variation in individual responses. The formula is:

$$(b_0 + b_1 x_g) \pm t_{\alpha/2,n-2} s \sqrt{1 + \frac{1}{n} + \frac{(x_g - \overline{x})^2}{SS_{xx}}}.$$

Example 3.1 (Continued) – Coffee Sales and Shelf Space

For the coffee example, suppose the marketer wishes to predict next week's sales when the coffee will have $5'$ of shelf space. She would like to obtain a 95% prediction interval for the number of bags to be sold. First, we observe that $t_{.025,10} = 2.228$, all other relevant numbers can be found in the previous example. The prediction interval is then:

$$(307.967 + 34.5833(5)) \pm 2.228(51.63)\sqrt{1 + \frac{1}{12} + \frac{(5-6)^2}{72}} = 480.883 \pm 93.554\sqrt{1.0972}$$

$$= 480.883 \pm 97.996 \equiv (382.887, 578.879).$$

This interval is relatively wide, reflecting the large variation in weekly sales at each level of x. Note that just as the width of the confidence interval for the mean response depends on the distance between x_g and $\overline{x}$, so does the width of the prediction interval. This should be of no surprise, considering the way we set up the prediction interval (see Figure 10 and Figure 11). Figure 12 shows the fitted equation and 95% prediction limits for this example.

It must be noted that a prediction interval for a future response is only valid if conditions are similar when the response occurs as when the data was collected. For instance, if the store is being boycotted by a bunch of animal rights activists for selling meat next week, our prediction interval will not be valid.

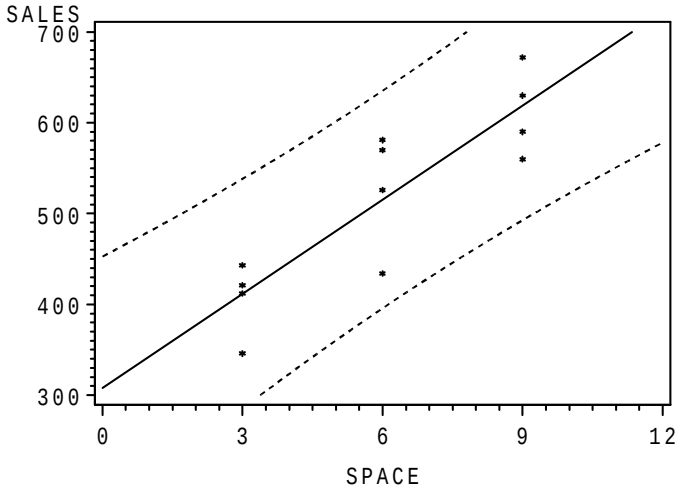

Figure 12: Plot of coffee data, fitted equation, and 95% prediction limits for a single response

Example 3.2 (Continued) – Hosiery Mill Cost Function

Suppose the plant manager knows based on purchase orders that this month, her plant will produce 30,000 items ($x_g = 30.0$). She would like to predict what the plant's production costs will be. She obtains a 95% prediction interval for this month's costs.

$$3.1274 + 2.0055(30) \pm 2.015(6.24)\sqrt{1 + \frac{1}{48} + \frac{(30 - 31.0673)^2}{7738.94}} \quad \equiv \quad 63.29 \pm 2.015(6.24)\sqrt{1.0210}$$

$$\equiv 63.29 \pm 12.70 \quad \equiv \quad (50.59, 75.99)$$

She predicts that the costs for this month will be between \$50,590 and \$75,990. This interval is much wider than the interval for the mean, since it includes random variation in monthly costs around the mean. A plot of the 95% prediction bands is given in Figure 13.

3.3 Coefficient of Correlation

In many situations, we would like to obtain a measure of the strength of the linear association between the variables y and x. One measure of this association that is reported in research journals from many fields is the Pearson product moment coefficient of correlation. This measure, denoted by r, is a number that can range from -1 to +1. A value of r close to 0 implies that there is very little association between the two variables (y tends to neither increase or decrease as x increases). A positive value of r means there is a positive association between y and x (y tends to increase as x increases). Similarly, a negative value means there is a negative association (y tends to decrease as x increases). If r is either +1 or -1, it means the data fall on a straight line ($SSE = 0$) that has either a positive or negative slope, depending on the sign of r. The formula for calculating r is:

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} = \frac{cov(x, y)}{s_x s_y}.$$

Note that the sign of r is always the same as the sign of b_1. We can test whether a population coefficient of correlation is 0, but since the test is mathematically equivalent to testing whether $\beta_1 = 0$, we won't cover this test.

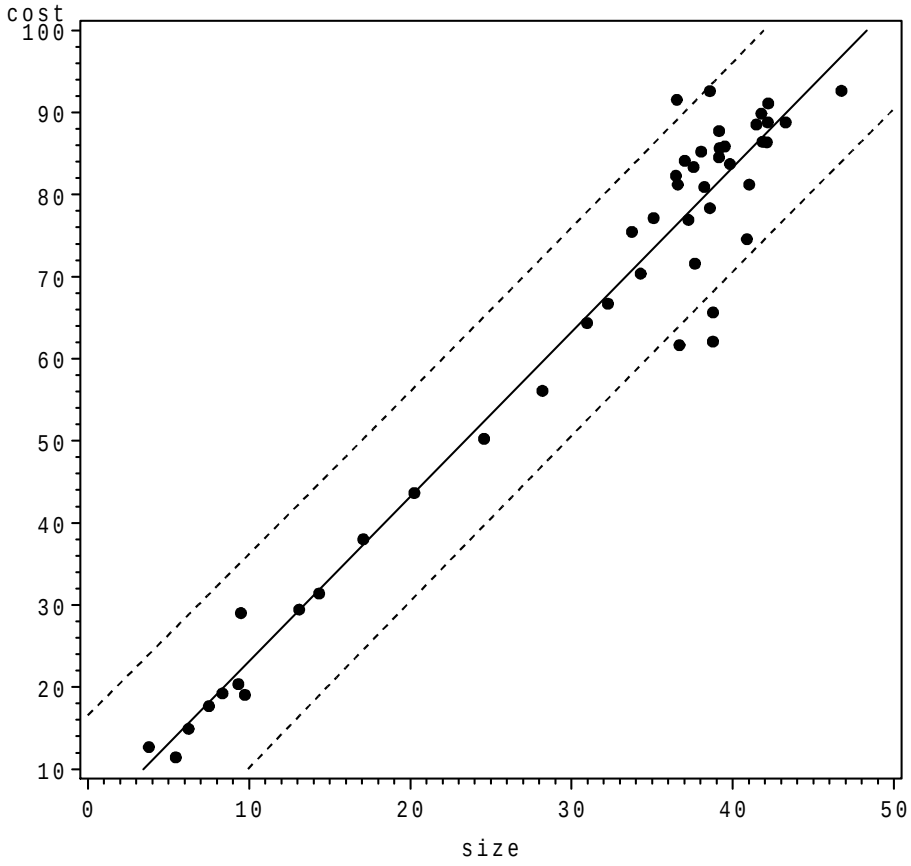

Figure 13: Plot of hosiery mill cost data, fitted equation, and 95% prediction limits for an individual outcome

Example 3.1 (Continued) – Coffee Sales and Shelf Space

For the coffee data, we can calculate r using the values of $SS_{xy}, SS_{xx}, SS_{yy}$ we have previously obtained.

$$r = \frac{2490}{\sqrt{(72)(112772.9)}} = \frac{2490}{2849.5} = .8738$$

Example 3.2 (Continued) – Hosiery Mill Cost Function

For the hosiery mill cost function data, we have:

$$r = \frac{15520.27}{\sqrt{(7738.94)(32914.06)}} = \frac{15520.27}{15959.95} = .9725$$

Computer Output for Coffee Sales Example (SAS System)

Dependent Variable: SALES

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	86112.50000	86112.50000	32.297	0.0002
Error	10	26662.41667	2666.24167		
C Total	11	112774.91667			

Root MSE	51.63566	R-square	0.7636
Dep Mean	515.41667	Adj R-sq	0.7399

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	307.916667	39.43738884	7.808	0.0001
SPACE	1	34.583333	6.08532121	5.683	0.0002

Obs	Dep Var SALES	Predict Value	Std Err Predict	Lower95% Mean	Upper95% Mean	Lower95% Predict
1	421.0	411.7	23.568	359.2	464.2	285.2
2	412.0	411.7	23.568	359.2	464.2	285.2
3	443.0	411.7	23.568	359.2	464.2	285.2
4	346.0	411.7	23.568	359.2	464.2	285.2
5	526.0	515.4	14.906	482.2	548.6	395.7
6	581.0	515.4	14.906	482.2	548.6	395.7
7	434.0	515.4	14.906	482.2	548.6	395.7
8	570.0	515.4	14.906	482.2	548.6	395.7
9	630.0	619.2	23.568	566.7	671.7	492.7
10	560.0	619.2	23.568	566.7	671.7	492.7
11	590.0	619.2	23.568	566.7	671.7	492.7
12	672.0	619.2	23.568	566.7	671.7	492.7

Obs	Upper95% Predict	Residual	Obs	Upper95% Predict	Residual
1	538.1	9.3333	7	635.2	-81.4167
2	538.1	0.3333	8	635.2	54.5833
3	538.1	31.3333	9	745.6	10.8333
4	538.1	-65.6667	10	745.6	-59.1667
5	635.2	10.5833	11	745.6	-29.1667
6	635.2	65.5833	12	745.6	52.8333

4 Logistic Regression

Often, the outcome is nominal (or binary), and we wish to relate the **probability** that an outcome has the characteristic of interest to an interval scale predictor variable. For instance, a local service provider may be interested in the probability that a customer will redeem a coupon that is mailed to him/her as a function of the amount of the coupon. We would expect that as the value of the coupon increases, so does the proportion of coupons redeemed. An experiment could be conducted as follows.

- Choose a range of reasonable coupon values (say x=\$0 (flyer only), \$1, \$2, \$5, \$10)
- Identify a sample of customers (say 200 households)
- Randomly assign customers to coupon values (say 40 per coupon value level)
- Send out coupons, and determine whether each coupon was redeemed by the expiration date ($y = 1$ if yes, 0 if no)
- Tabulate results and fit estimated regression equation.

Note that probabilities are bounded by 0 and 1, so we cannot fit a linear regression, since it will provide fitted values outside this range (unless b_0 is between 0 and 1 and b_1 is 0). We consider the following model, that does force fitted probabilities to lie between 0 and 1:

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} \qquad e = 2.71828\ldots$$

Unfortunately, unlike the case of simple linear regression, where there are close form equations for least squares estimates of β_0 and β_1, computer software must be used to obtain *maximum likelihood* estimates of β_0 and β_1, as well as their standard errors. Fortunately, many software packages (e.g. SAS, SPSS, Statview) offer procedures to obtain the estimates, standard errors and tests. We will give estimates and standard errors in this section, obtained from one of these packages. Once the estimates of β_0 and β_1 are obtained, which we will label as b_0 and b_1 respectively, we obtain the fitted equation:

$$\hat{P}(x) = \frac{e^{b_0+b_1 x}}{1 + e^{b_0+b_1 x}} \qquad e = 2.71828\ldots$$

Example 4.1 – Viagra Clinical Trial

In a clinical trial for Viagra, patients suffering from erectile dysfunction were randomly assigned to one of four daily doses ($0mg$, $25mg$, $50mg$, and $100mg$). One measure obtained from the patients was whether the patient had improved erections after 24 weeks of treatment ($y = 1$ if yes, $y = 0$ if no). Table 9 gives the number of subjects with $y = 1$ an $y = 0$ for each dose level.

Source: I. Goldstein, *et al*, (1998), "Oral Sildenafil in the Treatment of Erectile Dysfunction", *NEJM*, 338:1397-1404.

Based on an analysis using SAS software, we obtain the following estimates and standard errors for the logistic regression model:

$$b_0 = -0.8311 \quad s_{b_0} = .1354 \qquad\qquad b_1 = 0.0313 \quad s_{b_1} = .0034$$

		# Responding	
Dose (x)	n	$y = 1$	$y = 0$
0	199	50	149
25	96	54	42
50	105	81	24
100	101	85	16

Table 9: Patients showing improvement ($y = 1$) and not showing improvement ($y = 0$) by Viagra dose (x)

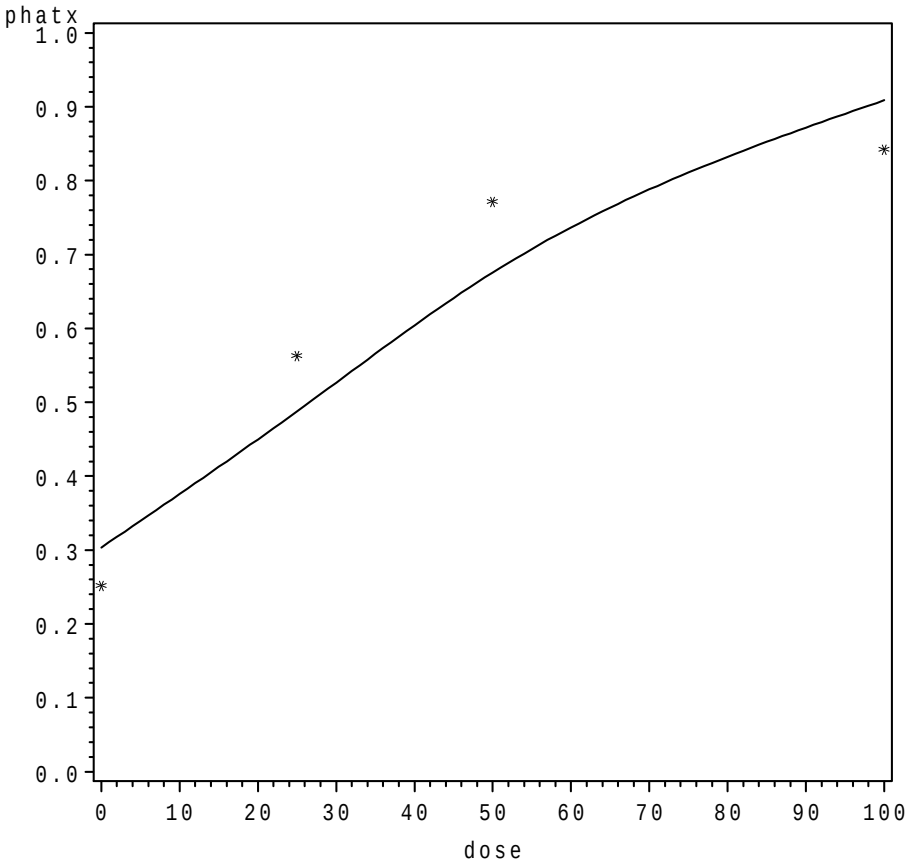

Figure 14: Plot of estimated logistic regression equation - Viagra data

A plot of the fitted equation (line) and the sample proportions at each dose (dots) are given in Figure 14.

4.1 Testing for Association between Outcome Probabilities and x

Consider the logistic regression model:

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \qquad e = 2.71828$$

Note that if $\beta_1 = 0$, then the equation becomes $p(x) = e^{\beta_0}/(1+e^{\beta_0})$. That is, the probability that the outcome is the characteristic of interest is not related to x, the predictor variable. In terms of the Viagra example, this would mean that the probability a patient shows improvement is independent of dose. This is what we would expect if the drug were not effective (still allowing for a placebo effect).

Futher, note that if $\beta_1 > 0$, the probability of the characteristic of interest occurring increases in x, and if $\beta_1 < 0$, the probability decreases in x. We can test whether $\beta_1 = 0$ as follows:

- $H_0 : \beta_1 = 0$ (Probability of outcome is independent of x)
- $H_A : \beta_1 \neq 0$ (Probability of outcome is associated with x)
- Test Statistic: $X^2_{obs} = [b_1/s_{b_1}]^2$
- Rejection Region: $X^2_{obs} \geq \chi^2_{\alpha,1}$ (=3.841, for $\alpha = 0.05$).
- P-value: Area in χ^2_1 distribution above X^2_{obs}

Note that if we reject H_0, we determine direction of association (positive/negative) by the sign of b_1.

Example 4.1 (Continued) – Viagra Clinical Trial

For this data, we can test whether the probability of showing improvement is associated with dose as follows:

- $H_0 : \beta_1 = 0$ (Probability of improvement is independent of dose)
- $H_A : \beta_1 \neq 0$ (Probability of improvement is associated with dose)
- Test Statistic: $X^2_{obs} = [b_1/s_{b_1}]^2 = [.0313/.0034]^2 = (9.2059)^2 = 84.75$
- Rejection Region: $X^2_{obs} \geq \chi^2_{\alpha,1}$ (=3.841, for $\alpha = 0.05$).
- P-value: Area in χ^2_1 distribution above X^2_{obs} (virtually 0)

Thus, we have strong evidence of a positive association (since $b_1 > 0$ and we reject H_0) between probability of improvement and dose.

5 Multiple Linear Regression I

Textbook Sections: 19.1-19.3 and Supplement

In most situations, we have more than one independent variable. While the amount of math can become overwhelming and involves matrix algebra, many computer packages exist that will provide the analysis for you. In this chapter, we will analyze the data by interpreting the results of a computer program. It should be noted that simple regression is a special case of multiple regression, so most concepts we have already seen apply here.

5.1 The Multiple Regression Model and Least Squares Estimates

In general, if we have k predictor variables, we can write our response variable as:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon.$$

Again, x is broken into a systematic and a random component:

$$y = \underbrace{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}_{\text{systematic}} + \underbrace{\varepsilon}_{random}$$

We make the same assumptions as before in terms of ε, specifically that they are independent and normally distributed with mean 0 and standard deviation σ. That is, we are assuming that y, at a given set of levels of the k independent variables $(x_1, \ldots, x_k)$ is normal with mean $E[y|x_1, \ldots, x_k] = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ and standard deviation σ. Just as before, $\beta_0, \beta_1, \ldots, \beta_k$, and σ are unknown parameters that must be estimated from the sample data. The parameters β_i represent the change in the mean response when the i^{th} predictor variable changes by 1 unit and all other predictor variables are held constant.

In this model:

- y — Random outcome of the dependent variable
- β_0 — Regression constant ($E(y|x_1 = \cdots = x_k = 0)$ if appropriate)
- β_i — Partial regression coefficient for variable x_i (Change in $E(y)$ when x_i increases by 1 unit and all other $x^{'s}$ are held constant)
- ε — Random error term, assumed (as before) that $\varepsilon \sim N(0, \sigma)$
- k — The number of independent variables

By the method of least squares (choosing the b_i values that minimize $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$), we obtain the fitted equation:

$$\hat{Y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

and our estimate of σ:

$$s_e = \sqrt{\frac{\sum(y - \hat{y})^2}{n - k - 1}} = \sqrt{\frac{SSE}{n - k - 1}}$$

The Analysis of Variance table will be very similar to what we used previously, with the only adjustments being in the degrees' of freedom. Table 10 shows the values for the general case when there are k predictor variables. We will rely on computer outputs to obtain the Analysis of Variance and the estimates b_0, b_1, and b_k.

ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
MODEL	$SSR = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$	k	$MSR = \frac{SSR}{k}$	$F = \frac{MSR}{MSE}$
ERROR	$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$n-k-1$	$MSE = \frac{SSE}{n-k-1}$	
TOTAL	$SS_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2$	$n-1$		

Table 10: The Analysis of Variance Table for multiple regression

5.2 Testing for Association Between the Response and the Full Set of Predictor Variables

To see if the set of predictor variables is useful in predicting the response variable, we will test $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$. Note that if H_0 is true, then the mean response does not depend on the levels of the predictor variables. We interpret this to mean that there is no association between the response variable and the set of predictor variables. To test this hypothesis, we use the following method:

1. $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$
2. H_A : Not every $\beta_i = 0$
3. T.S.: $F_{obs} = \frac{MSR}{MSE}$
4. R.R.: $F_{obs} > F_{\alpha,k,n-k-1}$
5. p-value: $P(F > F_{obs})$ (You can only get bounds on this, but computer outputs report them exactly)

The computer automatically performs this test and provides you with the p-value of the test, so in practice you really don't need to obtain the rejection region explicitly to make the appropriate conclusion. However, we will do so in this course to help reinforce the relationship between the test's decision rule and the p-value. Recall that we reject the null hypothesis if the p-value is less than α.

5.3 Testing Whether Individual Predictor Variables Help Predict the Response

If we reject the previous null hypothesis and conclude that not all of the β_i are zero, we may wish to test whether individual β_i are zero. Note that if we fail to reject the null hypothesis that β_i is zero, we can drop the predictor x_i from our model, thus simplifying the model. Note that this test is testing whether x_i is useful **given that we are already fitting a model containing the remaining $k-1$ predictor variables**. That is, does this variable contribute anything once we've taken into account the other predictor variables. These tests are t-tests, where we compute $t = \frac{b_i}{s_{b_i}}$ just as we did in the section on making inferences concerning β_1 in simple regression. The procedure for testing whether $\beta_i = 0$ (the i^{th} predictor variable does not contribute to predicting the response given the other $k-1$ predictor variables are in the model) is as follows:

- $H_0 : \beta_i = 0$ (y is not associated with x_i after controlling for all other independent variables)

- (1) $H_A : \beta_i \neq 0$
  (2) $H_A : \beta_i > 0$
  (3) $H_A : \beta_i < 0$
- T.S.: $t_{obs} = \frac{b_i}{S_{b_i}}$
- R.R.: (1) $|t_{obs}| > t_{\alpha/2,n-k-1}$
  (2) $t_{obs} > t_{\alpha,n-k-1}$
  (3) $t_{obs} < -t_{\alpha,n-k-1}$
- (1) p–value: $2P(T > |t_{obs}|)$
  (2) p–value: $P(T > t_{obs})$
  (3) p–value: $P(T < t_{obs})$

Computer packages print the test statistic and the p-value based on the two-sided test, so to conduct this test is simply a matter of interpreting the results of the computer output.

5.4 Testing for an Association Between a Subset of Predictor Variables and the Response

We have seen the two extreme cases of testing whether all regression coefficients are simultaneously 0 (the F-test), and the case of testing whether a single regression coefficient is 0, controlling for all other predictors (the t-test). We can also test whether a subset of the k regression coefficients are 0, controlling for all other predictors. Note that the two extreme cases can be tested using this very general procedure.

To make the notation as simple as possible, suppose our model consists of k predictor variables, of which we'd like to test whether q ($q \leq k$) are simultaneously not associated with y, after controlling for the remaining $k-q$ predictor variables. Further assume that the $k-q$ remaining predictors are labelled $x_1, x_2, \ldots, x_{k-q}$ and that the q predictors of interest are labelled $x_{k-q+1}, x_{k-q+2}, \ldots, x_k$.

This test is of the form:

$$H_0 : \beta_{k-q+1} = \beta_{k-q+2} = \cdots = \beta_k = 0 \qquad H_A : \beta_{k-q+1} \neq 0 \text{ and/or } \beta_{k-q+2} \neq 0 \text{ and/or } \ldots \text{ and/or } \beta_k \neq 0$$

The procedure for obtaining the numeric elements of the test is as follows:

1. Fit the model under the null hypothesis ($\beta_{k-q+1} = \beta_{k-q+2} = \cdots = \beta_k = 0$). It will include only the first $k-q$ predictor variables. This is referred to as the **Reduced** model. Obtain the error sum of squares ($SSE(R)$) and the error degrees of freedom $df_E(R) = n-(k-q)-1$.
2. Fit the model with all k predictors. This is referred to as the **Complete** or **Full** model (and was used for the F-test for all regression coefficients). Obtain the error sum of squares ($SSE(F)$) and the error degrees of freedom ($df_E(F) = n-k-1$).

By definition of the least squares citerion, we know that $SSE(R) \geq SSE(F)$. We now obtain the test statistic:

$$TS: \qquad F_{obs} = \frac{\frac{SSE(R)-SSE(F)}{(n-(k-q)-1)-(n-k-1)}}{\frac{SSE(F)}{n-k-1}} = \frac{(SSE(R)-SSE(F))/q}{MSE(F)}$$

and our rejection region is values of $F_{obs} \geq F_{\alpha,q,n-k-1}$.

Example 5.1 – Texas Weather Data

In this example, we will use regression in the context of predicting an outcome. A construction company is making a bid on a project in a remote area of Texas. A certain component of the project will take place in December, and is very sensitive to the daily high temperatures. They would like to estimate what the average high temperature will be at the location in December. They believe that temperature at a location will depend on its latitude (measure of distance from the equator) and its elevation. That is, they believe that the response variable (mean daily high temperature in December at a particular location) can be written as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon,$$

where x_1 is the latitude of the location, x_2 is the longitude, and x_3 is its elevation (in feet). As before, we assume that $\varepsilon \sim N(0, \sigma)$. Note that higher latitudes mean farther north and higher longitudes mean farther west.

To estimate the parameters $\beta_0, \beta_1, \beta_2, \beta_3$, and σ, they gather data for a sample of $n = 16$ counties and fit the model described above. The data, including one other variable are given in Table 11.

COUNTY	LATITUDE	LONGITUDE	ELEV	TEMP	INCOME
HARRIS	29.767	95.367	41	56	24322
DALLAS	32.850	96.850	440	48	21870
KENNEDY	26.933	97.800	25	60	11384
MIDLAND	31.950	102.183	2851	46	24322
DEAF SMITH	34.800	102.467	3840	38	16375
KNOX	33.450	99.633	1461	46	14595
MAVERICK	28.700	100.483	815	53	10623
NOLAN	32.450	100.533	2380	46	16486
ELPASO	31.800	106.40	3918	44	15366
COLLINGTON	34.850	100.217	2040	41	13765
PECOS	30.867	102.900	3000	47	17717
SHERMAN	36.350	102.083	3693	36	19036
TRAVIS	30.300	97.700	597	52	20514
ZAPATA	26.900	99.283	315	60	11523
LASALLE	28.450	99.217	459	56	10563
CAMERON	25.900	97.433	19	62	12931

Table 11: Data corresponding to 16 counties in Texas

The results of the Analysis of Variance are given in Table 12 and the parameter estimates, estimated standard errors, t-statistics and p-values are given in Table 13. Full computer programs and printouts are given as well.

We see from the Analysis of Variance that at least one of the variables, latitude and elevation, are related to the response variable temperature. This can be seen by setting up the test $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ as described previously. The elements of this test, provided by the computer output, are detailed below, assuming $\alpha = .05$.

1. $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	p-value
MODEL	$SSR = 934.328$	$k = 3$	$MSR = \frac{934.328}{3}$ =311.443	$F = \frac{311.443}{0.634}$ =491.235	.0001
ERROR	$SSE = 7.609$	$n - k - 1 =$ $16 - 3 - 1 = 12$	$MSE = \frac{7.609}{12}$ =0.634		
TOTAL	$SS_{yy} = 941.938$	$n - 1 = 15$			

Table 12: The Analysis of Variance Table for Texas data

PARAMETER	ESTIMATE	t FOR H_0: β_i=0	P-VALUE	STANDARD ERROR OF ESTIMATE
INTERCEPT (β_0)	b_0=109.25887	36.68	.0001	2.97857
LATITUDE (β_1)	$b_1 = -1.99323$	-14.61	.0001	0.13639
LONGITUDE (β_2)	$b_2 = -0.38471$	-1.68	.1182	0.22858
ELEVATION (β_3)	$b_3 = -0.00096$	-1.68	.1181	0.00057

Table 13: Parameter estimates and tests of hypotheses for individual parameters

2. H_A : Not all $\beta_i = 0$
3. T.S.: $F_{obs} = \frac{MSR}{MSE} = \frac{311.443}{0.634} = 491.235$
4. R.R.: $F_{obs} > F_{2,13,.05} = 3.81$ (This is not provided on the output, the p-value takes the place of it).
5. p-value: $P(F > 644.45) = .0001$ (Actually it is less than .0001, but this is the smallest p-value the computer will print).

We conclude that we are sure that at least one of these three variables is related to the response variable temperature.

We also see from the individual t-tests that latitude is useful in predicting temperature, even after taking into account the other predictor variables.

The formal test (based on $\alpha = 0.05$ significance level) for determining whetheher temperature is associated with latitude after controlling for longitude and elevation is given here:

- $H_0 : \beta_1 = 0$ (TEMP (y) is not associated with LAT (x_1) after controlling for LONG (x_2) and ELEV (x_3))
- $H_A : \beta_1 \neq 0$ (TEMP is associated with LAT after controlling for LONG and ELEV)
- T.S.: $t_{obs} = \frac{b_1}{S_{b_1}} = \frac{-1.99323}{0.136399} = -14.614$
- R.R.: $|t_{obs}| > t_{\alpha/2,n-k-1} = t_{.025,12} = 2.179$
- p–value: $2P(T > |t_{obs}|) = 2P(T > 14.614) < .0001$

Thus, we can conclude that there is an association between temperature and latitude, controlling for longitude and elevation. Note that the coeficient is negative, so we conclude that temperature decreases as latitude increases (given a level of longitude and elevation).

Note from Table 13 that neither the coefficient for LONGITUDE (X_2) or ELEVATION (X_3) are significant at the $\alpha = 0.05$ significance level (p-values are .1182 and .1181, respectively). Recall these are testing whether each term is 0 **controlling for LATITUDE and the other term**.

Before concluding that neither LONGITUDE (x_2) or ELEVATION (x_3) are useful predictors, controlling for LATITUDE, we will test whether they are both simultaneously 0, that is:

$$H_0 : \beta_2 = \beta_3 = 0 \qquad vs \qquad H_A : \beta_2 \neq 0 \text{ and/or } \beta_3 \neq 0$$

First, note that we have:

$$n = 16 \quad k = 3 \quad q = 2 \quad SSE(F) = 7.609 \quad df_E(F) = 16 - 3 - 1 = 12 \quad MSE(F) = 0.634$$

$$df_E(R) = 16 - (3 - 2) - 1 = 14 \quad F_{.05,2,12} = 3.89$$

Next, we fit the model with only LATITUDE (x_1) and obtain the error sum of squares: $SSE(R) = 60.935$ and get the following test statistic:

$$TS: \qquad F_{obs} \quad = \quad \frac{(SSE(R) - SSE(F))/q}{MSE(F)} = \frac{(60.935 - 7.609)/2}{0.634} = \frac{26.663}{0.634} = 42.055$$

Since $42.055 >> 3.89$, we reject H_0, and conclude that LONGITUDE (x_2) and/or ELEVATION (x_3) are associated with TEMPERATURE (y), after controlling for LATITUDE (x_1).

The reason we failed to reject $H_0 : \beta_2 = 0$ and $H_0 : \beta_3 = 0$ individually based on the t-tests is that ELEVATION and LONGITUDE are highly correlated (Elevations rise as you go further west in the state. So, once you control for LONGITUDE, we observe little ELEVATION effect, and vice versa. We will discuss why this is the case later. In theory, we have little reason to believe that temperatures naturally increase or decrease with LONGITUDE, but we may reasonably expect that as ELEVATION increases, TEMPERATURE decreases.

We re–fit the more parsimonious (simplistic) model that uses ELEVATION (x_1) and LATITUDE (x_2) to predict TEMPERATURE (y). Note the new symbols for ELEVATION and LATITUDE. That is to show you that they are merely symbols. The results are given in Table 14 and Table 15.

ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	p-value
MODEL	$SSR = 932.532$	$k = 2$	$MSR = \frac{932.532}{2}$ =466.266	$F = \frac{466.266}{0.634}$ =644.014	.0001
ERROR	$SSE = 9.406$	$n - k - 1 =$ $16 - 2 - 1 = 13$	$MSE = \frac{9.406}{13}$ =0.724		
TOTAL	$SS_{yy} = 941.938$	$n - 1 = 15$			

Table 14: The Analysis of Variance Table for Texas data – without LONGITUDE

We see this by observing that the t-statistic for testing $H_0 : \beta_1 = 0$ (no latitude effect on temperature) is -17.65, corresponding to a p-value of .0001, and the t-statistic for testing $H_0 : \beta_2 = 0$ (no elevation effect) is -8.41, also corresponding to a p-value of .0001. Further note that both estimates are negative, reflecting that as elevation and latitude increase, temperature decreases. That should not come as any big surprise.

PARAMETER	ESTIMATE	t FOR H_0: β_i=0	P-VALUE	STANDARD ERROR OF ESTIMATE
INTERCEPT (β_0)	b_0=63.45485	36.68	.0001	0.48750
ELEVATION (β_1)	$b_1 = -0.00185$	-8.41	.0001	0.00022
LATITUDE (β_2)	$b_2 = -1.83216$	-17.65	.0001	0.10380

Table 15: Parameter estimates and tests of hypotheses for individual parameters – without LONGITUDE

The magnitudes of the estimated coefficients are quite different, which may make you believe that one predictor variable is more important than the other. This is not necessarily true, because the ranges of their levels are quite different (1 unit change in latitude represents a change of approximately 19 miles, while a unit change in elevation is 1 foot) and recall that β_i represents the change in the mean response when variable X_i is increased by 1 unit.

The data corresponding to the 16 locations in the sample are plotted in Figure 15 and the fitted equation for the model that does not include LONGITUDE is plotted in Figure 16. The fitted equation is a plane in three dimensions.

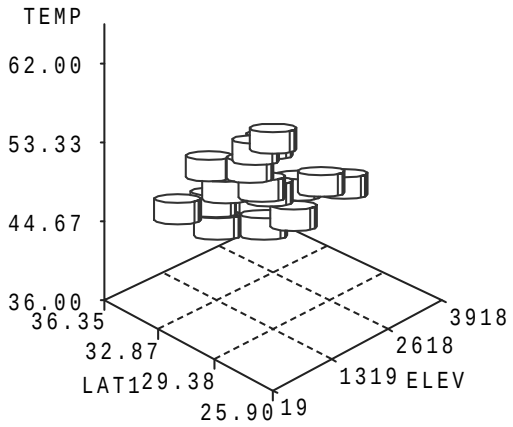

Figure 15: Plot of temperature data in 3 dimensions

Example 5.2 – Mortgage Financing Cost Variation (By City)

A study in the mid 1960's reported regional differences in mortgage costs for new homes. The sampling units were $n = 18$ metro areas (SMSA's) in the U.S. The dependent variable (y) is the average yield (in percent) on a new home mortgage for the SMSA. The independent variables (x_i) are given below.

Source: Schaaf, A.H. (1966), "Regional Differences in Mortgage Financing Costs," *Journal of Finance*, **21**:85-94.

x_1 – Average Loan Value / Mortgage Value Ratio (Higher x_1 means lower down payment and higher risk to lender).

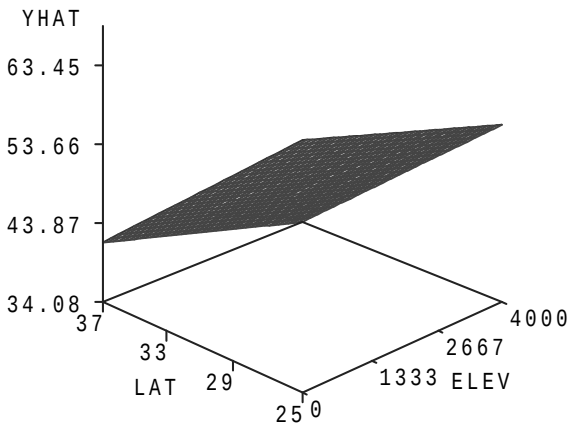

Figure 16: Plot of the fitted equation for temperature data

x_2 – Road Distance from Boston (Higher x_2 means further from Northeast, where most capital was at the time, and higher costs of capital).

x_3 – Savings per Annual Dwelling Unit Constructed (Higher x_3 means higher relative credit surplus, and lower costs of capital).

x_4 – Savings per Capita (does not adjust for new housing demand).

x_5 – Percent Increase in Population 1950–1960

x_6 – Percent of First Mortgage Debt Controlled by Inter-regional Banks.

The data, fitted values, and residuals are given in Table 16. The Analysis of Variance is given in Table 17. The regression coefficients, test statistics, and p-values are given in Table 18.

Show that the fitted value for Los Angeles is 6.19, based on the fitted equation, and that the residual is -0.02.

Based on the large F-statistic, and its small corresponding P-value, we conclude that this set of predictor variables is associated with the mortgage rate. That is, at least one of these independent variables is associated with y.

Based on the t-tests, while none are strictly significant at the $\alpha = 0.05$ level, there is some evidence that x_1 (Loan Value/Mortgage Value, $P = .0515$), x_3 (Savings per Unit Constructed, $P = .0593$), and to a lesser extent, x_4 (Savings per Capita, $P = .1002$) are helpful in predicting mortgage rates. We can fit a reduced model, with just these three predictors, and test whether we can simultaneously drop x_2, x_5, and x_6 from the model. That is:

$$H_0 : \beta_2 = \beta_5 = \beta_6 = 0 \qquad vs \qquad H_A : \beta_2 \neq 0 \text{ and/or } \beta_5 \neq 0 \text{ and/or } \beta_6 \neq 0$$

First, we have the following values:

$$n = 18 \qquad k = 6 \qquad q = 3$$

$$SSE(F) = 0.10980 \qquad df_E(F) = 18 - 6 - 1 = 11 \qquad MSE(F) = 0.00998$$

$$df_E(R) = 18 - (6 - 3) - 1 = 14 \qquad F_{.05,3,11} = 3.59$$

SMSA	y	x_1	x_2	x_3	x_4	x_5	x_6	$\hat{y}$	$e = y - \hat{y}$
Los Angeles-Long Beach	6.17	78.1	3042	91.3	1738.1	45.5	33.1	6.19	-0.02
Denver	6.06	77.0	1997	84.1	1110.4	51.8	21.9	6.04	0.02
San Francisco-Oakland	6.04	75.7	3162	129.3	1738.1	24.0	46.0	6.05	-0.01
Dallas-Fort Worth	6.04	77.4	1821	41.2	778.4	45.7	51.3	6.05	-0.01
Miami	6.02	77.4	1542	119.1	1136.7	88.9	18.7	6.04	-0.02
Atlanta	6.02	73.6	1074	32.3	582.9	39.9	26.6	5.92	0.10
Houston	5.99	76.3	1856	45.2	778.4	54.1	35.7	6.02	-0.03
Seattle	5.91	72.5	3024	109.7	1186.0	31.1	17.0	5.91	0.00
New York	5.89	77.3	216	364.3	2582.4	11.9	7.3	5.82	0.07
Memphis	5.87	77.4	1350	111.0	613.6	27.4	11.3	5.86	0.01
New Orleans	5.85	72.4	1544	81.0	636.1	27.3	8.1	5.81	0.04
Cleveland	5.75	67.0	631	202.7	1346.0	24.6	10.0	5.64	0.11
Chicago	5.73	68.9	972	290.1	1626.8	20.1	9.4	5.60	0.13
Detroit	5.66	70.7	699	223.4	1049.6	24.7	31.7	5.63	0.03
Minneapolis-St Paul	5.66	69.8	1377	138.4	1289.3	28.8	19.7	5.81	-0.15
Baltimore	5.63	72.9	399	125.4	836.3	22.9	8.6	5.77	-0.14
Philadelphia	5.57	68.7	304	259.5	1315.3	18.3	18.7	5.57	0.00
Boston	5.28	67.8	0	428.2	2081.0	7.5	2.0	5.41	-0.13

Table 16: Data and fitted values for mortgage rate multiple regression example.

ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	p-value
MODEL	$SSR = 0.73877$	$k = 6$	$MSR = \frac{0.73877}{6}$ =0.12313	$F = \frac{0.12313}{0.00998}$ =12.33	.0003
ERROR	$SSE = 0.10980$	$n - k - 1 =$ $18 - 6 - 1 = 11$	$MSE = \frac{0.10980}{11}$ =0.00998		
TOTAL	$SS_{yy} = 0.84858$	$n - 1 = 17$			

Table 17: The Analysis of Variance Table for Mortgage rate regression analysis

PARAMETER	ESTIMATE	STANDARD ERROR	t-statistic	P-value
INTERCEPT (β_0)	b_0=4.28524	0.66825	6.41	.0001
x_1 (β_1)	$b_1 = 0.02033$	0.00931	2.18	.0515
x_2 (β_2)	$b_2 = 0.000014$	0.000047	0.29	.7775
x_3 (β_3)	$b_3 = -0.00158$	0.000753	-2.10	.0593
x_4 (β_4)	$b_4 = 0.000202$	0.000112	1.79	.1002
x_5 (β_5)	$b_5 = 0.00128$	0.00177	0.73	.4826
x_6 (β_6)	$b_6 = 0.000236$	0.00230	0.10	.9203

Table 18: Parameter estimates and tests of hypotheses for individual parameters – Mortgage rate regression analysis

ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	p-value
MODEL	$SSR = 0.73265$	$k - q = 3$	$MSR = \frac{0.73265}{3}$ =0.24422	$F = \frac{0.24422}{0.00828}$ =29.49	.0001
ERROR	$SSE = 0.11593$	$n - (k - q) - 1 = 18 - 3 - 1 = 14$	$MSE = \frac{0.11593}{14}$ =0.00828		
TOTAL	$SS_{yy} = 0.84858$	$n - 1 = 17$			

Table 19: The Analysis of Variance Table for Mortgage rate regression analysis (Reduced Model)

PARAMETER	ESTIMATE	STANDARD ERROR	t-statistic	P-value
INTERCEPT (β_0)	b_0=4.22260	0.58139	7.26	.0001
x_1 (β_1)	$b_1 = 0.02229$	0.00792	2.81	.0138
x_3 (β_3)	$b_3 = -0.00186$	0.00041778	-4.46	.0005
x_4 (β_4)	$b_4 = 0.000225$	0.000074	3.03	.0091

Table 20: Parameter estimates and tests of hypotheses for individual parameters – Mortgage rate regression analysis (Reduced Model)

Next, we fit the reduced model, with $\beta_2 = \beta_5 = \beta_6 = 0$. We get the Analysis of Variance in Table 19 and parameter estimates in Table 20.

Note first, that all three regression coefficients are significant now at the $\alpha = 0.05$ significance level. Also, our residual standard error, $S_e = \sqrt{MSE}$ has also decreased (0.09991 to 0.09100). This implies we have lost very little predictive ability by dropping x_2, x_5, and x_6 from the model. Now to formally test whether these three predictor variables' regression coefficients are simultaneously 0 (with $\alpha = 0.05$):

- $H_0 : \beta_2 = \beta_5 = \beta_6 = 0$
- $H_A : \beta_2 \neq 0$ and/or $\beta_5 \neq 0$ and/or $\beta_6 \neq 0$
- $TS : F_{obs} = \frac{(0.11593 - 0.10980)/2}{0.00998} = \frac{.00307}{.00998} = 0.307$
- $RR : F_{obs} \geq F_{0.05,3,11} = 3.59$

We fail to reject H_0, and conclude that none of x_2, x_5, or x_6 are associated with mortgage rate, after controlling for x_1, x_3, and x_4.

Example 5.3 – Store Location Characteristics and Sales

A study proposed using linear regression to describe sales at retail stores based on location characteristics. As a case study, the authors modelled sales at $n = 16$ liquor stores in Charlotte, N.C. Note that in North Carolina, all stores are state run, and do not practice promotion as liquor stores in Florida do. The response was **SALES** volume (for the individual stores) in the fiscal year 7/1/1979-6/30/1980. The independent variables were: **POP** (number of people living within 1.5 miles of store), **MHI** (mean household income among households within 1.5 miles of store), **DIS**, (distance to the nearest store), **TFL** (daily traffic volume on the street the store was located), and

EMP (the amount of employment within 1.5 miles of the store. The regression coefficients and standard errors are given in Table 5.4.

Source: Lord, J.D. and C.D. Lynds (1981), "The Use of Regression Models in Store Location Research: A Review and Case Study," *Akron Business and Economic Review*, Summer, 13-19.

Variable	Estimate	Std Error
POP	0.09460	0.01819
MHI	0.06129	0.02057
DIS	4.88524	1.72623
TFL	-2.59040	1.22768
EMP	-0.00245	0.00454

Table 21: Regression coefficients and standard errors for liquor store sales study

a) Do any of these variables fail to be associated with store sales after controlling for the others?

b) Consider the signs of the significant regression coefficients. What do they imply?

5.5 R^2 and Adjusted–R^2

As was discussed in the previous chapter, the coefficient of multiple determination represents the proportion of the variation in the dependent variable (y) that is "explained" by the regression on the collection of independent variables: (x_1,...,x_k). R^2 is computed exactly as before:

$$R^2 \quad = \quad \frac{SSR}{SS_{yy}} \quad = \quad 1 - \frac{SSE}{SS_{yy}}$$

One problem with R^2 is that when we continually add independent variables to a regression model, it continually increases (or at least, never decreases), even when the new variable(s) add little or no predictive power. Since we are trying to fit the simplest (most parsimonious) model that explains the relationship between the set of independent variables and the dependent variable, we need a measure that penalizes models that contain useless or redundant independent variables. This penalization takes into account that by including useless or redundant predictors, we are decreasing error degrees of freedom ($df_E = n - k - 1$). A second measure, that does not carry the proportion of variation explained criteria, but is useful for comparing models of varying degrees of complexity, is Adjusted-R^2:

$$\text{Adjusted} - R^2 \quad = \quad 1 - \frac{SSE/(n-k-1)}{SS_{yy}/(n-1)} \quad = \quad 1 - \frac{n-1}{n-k-1}\left(\frac{SSE}{SS_{yy}}\right)$$

Example 5.1 (Continued) – Texas Weather Data
Consider the two models we have fit:

Full Model — I.V.'s: LATITUDE, LONGITUDE, ELEVATION

Reduced Model — I.V.'s: LATITUDE, ELEVATION

For the Full Model, we have:

$$n = 16 \quad k = 3 \quad SSE = 7.609 \quad SS_{yy} = 941.938$$

and, we obtain R_F^2 and Adj-R_F^2:

$$R_F^2 = 1 - \frac{7.609}{941.938} = 1 - .008 = 0.992 \qquad \text{Adj} - R_F^2 = 1 - \frac{15}{12}\left(\frac{7.609}{941.938}\right) = 1 - 1.25(.008) = 0.9900$$

For the Reduced Model, we have:

$$n = 16 \quad k = 2 \quad SSE = 9.406 \quad SS_{yy} = 941.938$$

and, we obtain R_R^2 and Adj-R_R^2:

$$R_R^2 = 1 - \frac{9.406}{941.938} = 1 - .010 = 0.990 \qquad \text{Adj} - R_R^2 = 1 - \frac{15}{13}\left(\frac{9.406}{941.938}\right) = 1 - 1.15(.010) = 0.9885$$

Thus, by both measures the Full Model "wins", but it should be added that both appear to fit the data very well!

Example 5.2 (Continued) – Mortgage Financing Costs

For the mortgage data (with Total Sum of Squares $SS_{yy} = 0.84858$ and $n = 18$), when we include all 6 independent variables in the full model, we obtain the following results:

$$SSR = 0.73877 \qquad SSE = 0.10980 \qquad k = 6$$

From this full model, we compute R^2 and Adj-R^2:

$$R_F^2 = \frac{SSR_F}{SS_{yy}} = \frac{0.73877}{0.84858} = 0.8706 \qquad \text{Adj}{-}R_F^2 = 1{-}\frac{n-1}{n-k-1}\left(\frac{SSE_F}{SS_{yy}}\right) = 1{-}\frac{17}{11}\left(\frac{0.10980}{0.84858}\right) = 0.8000$$

Example 5.3 (Continued) – Store Location Characteristics and Sales

In this study, the authors reported that $R^2 = 0.69$. Note that although we are not given the Analysis of Variance, we can still conduct the F test for the overall model:

$$F = \frac{MSR}{MSE} = \frac{SSR/k}{SSE/(n-k-1)} = \frac{\frac{SSR}{SS_{yy}}/k}{\frac{SSE}{SS_{yy}}/(n-k-1)} = \frac{R^2/k}{(1-R^2)/(n-k-1)}$$

For the liquor store example, there were $n = 16$ stores and $k = 5$ variables in the full model. To test:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \qquad vs \qquad H_A : \text{ Not all } \beta_i = 0$$

we get the following test statistic and rejection region ($\alpha = 0.05$):

$$TS : F_{obs} = \frac{0.69/5}{(1-0.69)/(16-5-1)} = \frac{0.138}{0.031} = 4.45 \qquad RR : F_{obs} \geq F_{\alpha,k,n-k-1} = F_{0.05,5,10} = 3.33$$

Thus, at least one of these variables is associated with store sales.

What is Adjusted-R^2 for this analysis?

5.6 Multicollinearity

Textbook: Section 19.4, Supplement

Multicollinearity refers to the situation where independent variables are highly correlated among themselves. This can cause problems mathematically and creates problems in interpreting regression coefficients.

Some of the problems that arise include:

- Difficult to interpret regression coefficient estimates
- Inflated std errors of estimates (and thus small t–statistics)
- Signs of coefficients may not be what is expected.
- However, predicted values are not adversely affected

It can be thought that the independent variables are explaining "the same" variation in y, and it is difficult for the model to attribute the variation explained (recall partial regression coefficients).

Variance Inflation Factors provide a means of detecting whether a given independent variable is causing multicollinearity. They are calculated (for each independent variable) as:

$$VIF_i = \frac{1}{1 - R_i^2}$$

where R_i^2 is the coefficient of multiple determination when x_i is regressed on the $k-1$ other independent variables. One rule of thumb suggests that severe multicollinearity is present if $VIF_i > 10$ ($R_i^2 > .90$).

Example 5.1 Continued

First, we run a regression with ELEVATION as the dependent variable and LATITUDE and LONGITUDE as the independent variables. We then repeat the process with LATITUDE as the dependent variable, and finally with LONGITUDE as the dependent variable. Table **??** gives R^2 and VIF for each model.

Variable	R^2	VIF
ELEVATION	.9393	16.47
LATITUDE	.7635	4.23
LONGITUDE	.8940	9.43

Table 22: Variance Inflation Factors for Texas weather data

Note how large the factor is for ELEVATION. Texas elevation increases as you go West and as you go North. The Western rise is the more pronounced of the two (the simple correlation between ELEVATION and LONGITUDE is .89).

Consider the effects on the coefficients in Table 23 and Table 24 (these are subsets of previously shown tables).

Compare the estimate and estimated standard error for the coefficient for ELEVATION and LATITUDE for the two models. In particular, the ELEVATION coefficient doubles in absolute value and its standard error decreases by a factor of almost 3. The LATITUDE coefficient and standard error do not change very much. We choose to keep ELEVATION, as opposed to LONGITUDE, in the model due to theoretical considerations with respect to weather and climate.

PARAMETER	ESTIMATE	STANDARD ERROR OF ESTIMATE
INTERCEPT (β_0)	b_0=109.25887	2.97857
LATITUDE (β_1)	$b_1 = -1.99323$	0.13639
LONGITUDE (β_2)	$b_2 = -0.38471$	0.22858
ELEVATION (β_3)	$b_3 = -0.00096$	0.00057

Table 23: Parameter estimates and standard errors for the full model

PARAMETER	ESTIMATE	STANDARD ERROR OF ESTIMATE
INTERCEPT (β_0)	b_0=63.45485	0.48750
ELEVATION (β_1)	$b_1 = -0.00185$	0.00022
LATITUDE (β_2)	$b_2 = -1.83216$	0.10380

Table 24: Parameter estimates and standard errors for the reduced model

5.7 Autocorrelation

Textbook Section: 19.5

Recall a key assumption in regression: Error terms are independent. When data are collected over time, the errors are often serially correlated (Autocorrelated). Under first–Order Autocorrelation, consecutive error terms are linealy related:

$$\varepsilon_t \quad = \quad \rho\varepsilon_{t-1} + \nu_t$$

where ρ is the correlation between consecutive error terms, and ν_t is a normally distributed independent error term. When errors display a positive correlation, $\rho > 0$ (Consecutive error terms are associated). We can test this relation as follows, note that when $\rho = 0$, error terms are independent (which is the assumption in the derivation of the tests in the chapters on linear regression).

Durbin–Watson Test for Autocorrelation

$H_0 : \rho = 0$ No autocorrelation $\quad H_a : \rho > 0$ Postive Autocorrelation

$$D = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

$D \geq d_U \Longrightarrow$ Don't Reject H_0

$D \leq d_L \Longrightarrow$ Reject H_0

$d_L \leq D \leq d_U \Longrightarrow$ Withhold judgement

Values of d_L and d_U (indexed by n and k (the number of predictor variables)) are given in Table 11(a), p. B-22.

"Cures" for Autocorrelation:

- Additional independent variable(s) — A variable may be missing from the model that will eliminate the autocorrelation.
- Transform the variables — Take "first differences" $(y_{t+1}-y_t)$ and $(y_{t+1}-y_t)$ and run regression with transformed y and x.

Example 5.4 Spirits Sales and Income and Prices in Britain

A study was conducted relating annual spirits (liquor) sales (y) in Britain to per capita income (x_1) and prices (x_2), where all monetary values were in constant (adjusted for inflation) dollars for the years 1870-1938. The following output gives the results from the regression analysis and the Durbin-Watson statistic. Note that there are $n = 69$ observations and $k = 2$ predictors, and the approximate lower and upper bounds for the rejection region are $d_L = 1.55$ and $d_U = 1.67$ for an $\alpha = 0.05$ level test. Since the test statistic is $d = 0.247$ (see output below), we reject the null hypothesis of no autocorrelation among the residuals, and conclude that they are positively correlated. See Figure 17 for a plot of the residuals versus year.

Source: Durbin J., and Watson, G.S. (1950), "Testing for Serial Correlation in Least Squares Regression, I", *Biometrika*, 37:409-428.

The REG Procedure
Model: MODEL1
Dependent Variable: consume

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	4.80557	2.40278	712.27	<.0001
Error	66	0.22264	0.00337		
Corrected Total	68	5.02821			

Root MSE	0.05808	R-Square	0.9557
Dependent Mean	1.76999	Adj R-Sq	0.9544
Coeff Var	3.28143		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.61171	0.15262	30.22	<.0001
income	1	-0.11846	0.10885	-1.09	0.2804
price	1	-1.23174	0.05024	-24.52	<.0001

Durbin-Watson D	0.247
Number of Observations	69
1st Order Autocorrelation	0.852

Obs	consume	income	price	yhat	e
1	1.9565	1.7669	1.9176	2.04042	-0.08392
2	1.9794	1.7766	1.9059	2.05368	-0.07428
3	2.0120	1.7764	1.8798	2.08586	-0.07386
4	2.0449	1.7942	1.8727	2.09249	-0.04759
5	2.0561	1.8156	1.8984	2.05830	-0.00220
6	2.0678	1.8083	1.9137	2.04032	0.02748
7	2.0561	1.8083	1.9176	2.03552	0.02058
8	2.0428	1.8067	1.9176	2.03571	0.00709
9	2.0290	1.8166	1.9420	2.00448	0.02452
10	1.9980	1.8041	1.9547	1.99032	0.00768
11	1.9884	1.8053	1.9379	2.01087	-0.02247
12	1.9835	1.8242	1.9462	1.99841	-0.01491
13	1.9773	1.8395	1.9504	1.99142	-0.01412
14	1.9748	1.8464	1.9504	1.99060	-0.01580
15	1.9629	1.8492	1.9723	1.96330	-0.00040
16	1.9396	1.8668	2.0000	1.92709	0.01251
17	1.9309	1.8783	2.0097	1.91378	0.01712
18	1.9271	1.8914	2.0146	1.90619	0.02091
19	1.9239	1.9166	2.0146	1.90321	0.02069
20	1.9414	1.9363	2.0097	1.90691	0.03449
21	1.9685	1.9548	2.0097	1.90472	0.06378
22	1.9727	1.9453	2.0097	1.90585	0.06685
23	1.9736	1.9292	2.0048	1.91379	0.05981
24	1.9499	1.9209	2.0097	1.90874	0.04116
25	1.9432	1.9510	2.0296	1.88066	0.06254
26	1.9569	1.9776	2.0399	1.86482	0.09208
27	1.9647	1.9814	2.0399	1.86437	0.10033
28	1.9710	1.9819	2.0296	1.87700	0.09400
29	1.9719	1.9828	2.0146	1.89537	0.07653
30	1.9956	2.0076	2.0245	1.88024	0.11536
31	2.0000	2.0000	2.0000	1.91131	0.08869
32	1.9904	1.9939	2.0048	1.90612	0.08428
33	1.9752	1.9933	2.0048	1.90619	0.06901
34	1.9494	1.9797	2.0000	1.91372	0.03568
35	1.9332	1.9772	1.9952	1.91993	0.01327
36	1.9139	1.9924	1.9952	1.91813	-0.00423
37	1.9091	2.0117	1.9905	1.92163	-0.01253
38	1.9139	2.0204	1.9813	1.93193	-0.01803
39	1.8886	2.0018	1.9905	1.92280	-0.03420
40	1.7945	2.0038	1.9859	1.92823	-0.13373
41	1.7644	2.0099	2.0518	1.84634	-0.08194
42	1.7817	2.0174	2.0474	1.85087	-0.06917
43	1.7784	2.0279	2.0341	1.86601	-0.08761
44	1.7945	2.0359	2.0255	1.87565	-0.08115
45	1.7888	2.0216	2.0341	1.86675	-0.07795
46	1.8751	1.9896	1.9445	1.98091	-0.10581
47	1.7853	1.9843	1.9939	1.92069	-0.13539
48	1.6075	1.9764	2.2082	1.65766	-0.05016
49	1.5185	1.9965	2.2700	1.57916	-0.06066
50	1.6513	2.0652	2.2430	1.60428	0.04702
51	1.6247	2.0369	2.2567	1.59075	0.033946
52	1.5391	1.9723	2.2988	1.54655	-0.007450
53	1.4922	1.9797	2.3723	1.45514	0.037059
54	1.4606	2.0136	2.4105	1.40407	0.056527
55	1.4551	2.0165	2.4081	1.40669	0.048415
56	1.4425	2.0213	2.4081	1.40612	0.036383
57	1.4023	2.0206	2.4367	1.37097	0.031328
58	1.3991	2.0563	2.4284	1.37697	0.022134

59	1.3798	2.0579	2.4310	1.37357	0.006226
60	1.3782	2.0649	2.4363	1.36622	0.011983
61	1.3366	2.0582	2.4552	1.34373	-0.007131
62	1.3026	2.0517	2.4838	1.30927	-0.006673
63	1.2592	2.0491	2.4958	1.29480	-0.035600
64	1.2365	2.0766	2.5048	1.28046	-0.043957
65	1.2549	2.0890	2.5017	1.28281	-0.027906
66	1.2527	2.1059	2.4958	1.28807	-0.035372
67	1.2763	2.1205	2.4838	1.30112	-0.024823
68	1.2906	2.1205	2.4636	1.32600	-0.035404
69	1.2721	2.1182	2.4580	1.33317	-0.061074

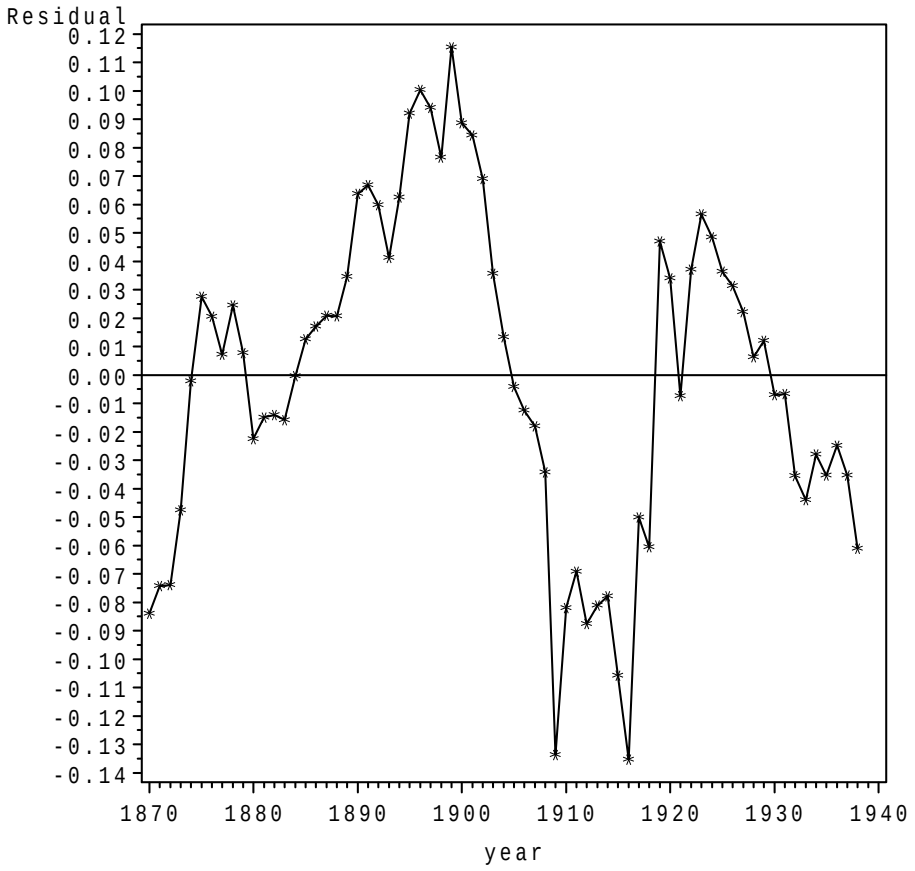

Figure 17: Plot of the residuals versus year for British spirits data

6 Special Cases of Multiple Regression

Textbook Sections: 20.2,20.3

In this section, we will look at three special cases that are frequently used methods of multiple regression. The ideas such as the Analysis of Variance, tests of hypotheses, and parameter estimates are exactly the same as before and we will concentrate on their interpretation through specific examples. The four special cases are:

1. Polynomial Regression
2. Regression Models with Nominal (Dummy) Variables
3. Regression Models Containing Interaction Terms

6.1 Polynomial Regression

While certainly not restricted to this case, it is best to describe polynomial regression in the case of a model with only one predictor variable. In many real–world settings relationships will not be linear, but will demonstrate nonlinear associations. In economics, a widely described phenomenon is "diminishing marginal returns". In this case, y may increase with x, but the rate of increase decreases over the range of x. By adding quadratic terms, we can test if this is the case. Other situations may show that the rate of increase in y is increasing in x.

Example 6.1 – Health Club Demand

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon.$$

Again, we assume that $\varepsilon \sim N(0, \sigma)$. In this model, the number of people attending in a day when there are x machines is nomally distributed with mean $\beta_0 + \beta_1 x + \beta_2 x^2$ and standard deviation σ. Note that we are no longer saying that the mean is linearly related to x, but rather that it is approximately quadratically related to x (curved). Suppose she leases varying numbers of machines over a period of $n = 12$ Wednesdays (always advertising how many machines will be there on the following Wednesday), and observes the number of people attending the club each day, and obtaining the data in Table 25.

In this case, we would like to fit the multiple regression model:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon,$$

which is just like our previous model except instead of a second predictor variable x_2, we are using the variable x^2, the effect is that the fitted equation $\hat{y}$ will be a curve in 2 dimensions, not a plane in 3 dimensions as we saw in the weather example. First we will run the regression on the computer, obtaining the Analysis of Variance and the parameter estimates, then plot the data and fitted equation. Table 26 gives the Analysis of Variance for this example and Table 27 gives the parameter estimates and their standard errors. Note that even though we have only one predictor variable, it is being used twice and could in effect be treated as two different predictor variables, so $k = 2$.

The first test of hypothesis is whether the attendance is associated with the number of machines. This is a test of $H_0 : \beta_1 = \beta_2 = 0$. If the null hypothesis is true, that implies mean daily attendance

Week	# Machines (x)	Attendance (y)
1	3	555
2	6	776
3	1	267
4	2	431
5	5	722
6	4	635
7	1	218
8	5	692
9	3	534
10	2	459
11	6	810
12	4	671

Table 25: Data for health club example

ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	p-value
MODEL	$SSR = 393933.12$	$k = 2$	$MSR = \frac{393933.12}{2}$ =196966.56	$F = \frac{196966.56}{776.06}$ =253.80	.0001
ERROR	$SSE = 6984.55$	$n - k - 1 =$ =12-2-1=9	$MSE = \frac{6984.55}{9}$ =776.06		
TOTAL	$SS_{yy} = 400917.67$	$n - 1 = 11$			

Table 26: The Analysis of Variance Table for health club data

PARAMETER	ESTIMATE	t FOR H_0: β_i=0	P-VALUE	STANDARD ERROR OF ESTIMATE
INTERCEPT (β_0)	b_0=72.0500	2.04	.0712	35.2377
MACHINES (β_1)	$b_1 = 199.7625$	8.67	.0001	23.0535
MACHINES SQ (β_2)	$b_2 = -13.6518$	-4.23	.0022	3.2239

Table 27: Parameter estimates and tests of hypotheses for individual parameters

is unrelated to the number of machines, thus the club owner would purchase very few (if any) of the machines. As before this test is the F-test from the Analysis of Variance table, which we conduct here at $\alpha = .05$.

1. $H_0 : \beta_1 = \beta_2 = 0$
2. H_A : Not both $\beta_i = 0$
3. T.S.: $F_{obs} = \frac{MSR}{MSE} = \frac{196966.56}{776.06} = 253.80$
4. R.R.: $F_{obs} > F_{2,9,.05} = 4.26$ (This is not provided on the output, the p-value takes the place of it).
5. p-value: $P(F > 253.80) = .0001$ (Actually it is less than .0001, but this is the smallest p-value the computer will print).

Another test with an interesting interpretation is $H_0 : \beta_2 = 0$. This is testing the hypothesis that the mean increases linearly with x (since if $\beta_2 = 0$ this becomes the simple regression model (refer back to the coffee data example)). The t-test in Table 27 for this hypothesis has a test statistic $t_{obs} = -4.23$ which corresponds to a p-value of .0022, which since it is below .05, implies we reject H_0 and conclude $\beta_2 \neq 0$. Since b_2 is is negative, we will conclude that β_2 is negative, which is in agreement with her theory that once you get to a certain number of machines, it does not help to keep adding new machines. This is the idea of 'diminishing returns'. Figure 18 shows the actual data and the fitted equation $\hat{y} = 72.0500 + 199.7625x - 13.6518x^2$.

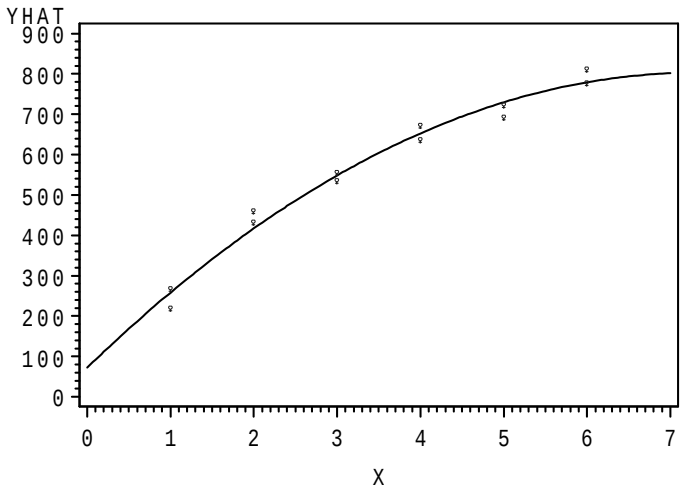

Figure 18: Plot of the data and fitted equation for health club example

6.2 Regression Models With Nominal (Dummy) Variables

All of the predictor variables we have used so far were numeric or what are often called quantitative variables. Other variables also can be used that are called qualitative variables. Qualitative variables measure characteristics that cannot be described numerically, such as a person's sex, race, religion, or blood type; a city's region or mayor's political affiliation; the list of possibilities is endless. In this case, we frequently have some numeric predictor variable(s) that we believe is (are)

related to the response variable, but we believe this relationship may be different for different levels of some qualitative variable of interest.

If a qualitative variable has m levels, we create $m-1$ **indicator** or **dummy variables**. Consider an example where we are interested in health care expenditures as related to age for men and women, separately. In this case, the response variable is health care expenditures, one predictor variable is age, and we need to create a variable representing sex. This can be done by creating a variable x_2 that takes on a value 1 if a person is female and 0 if the person is male. In this case we can write the mean response as before:

$$E[y|x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon.$$

Note that for women of age x_1, the mean expenditure is $E[y|x_1, 1] = \beta_0 + \beta_1 x_1 + \beta_2(1) = (\beta_0 + \beta_2) + \beta_1 x_1$, while for men of age x_1, the mean expenditure is $E[y|x_1, 0] = \beta_0 + \beta_1 x_1 + \beta_0(0) = \beta_0 + \beta_1 x_1$. This model allows for different means for men and women, but requires they have the same slope (we will see a more general case in the next section). In this case the interpretation of $\beta_2 = 0$ is that the means are the same for both sexes, this is a hypothesis a health care professional may wish to test in a study. In this example the variable sex had two variables, so we had to create $2 - 1 = 1$ dummy variable, now consider a second example.

Example 6.2

We would like to see if annual per capita clothing expenditures is related to annual per capita income in cities across the U.S. Further, we would like to see if there is any differences in the means across the 4 regions (Northeast, South, Midwest, and West). Since the variable region has 4 levels, we will create 3 dummy variables x_2, x_3, and x_4 as follows (we leave x_1 to represent the predictor variable per capita income):

$$x_2 = \begin{cases} 1 & \text{if region=South} \\ 0 & \text{otherwise} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{if region=Midwest} \\ 0 & \text{otherwise} \end{cases}$$

$$x_4 = \begin{cases} 1 & \text{if region=West} \\ 0 & \text{otherwise} \end{cases}$$

Note that cities in the Northeast have $x_2 = x_3 = x_4 = 0$, while cities in other regions will have either x_2, x_3, or x_4 being equal to 1. Northeast cities act like males did in the previous example. The data are given in Table 28.

The Analysis of Variance is given in Table 29, and the parameter estimates and standard errors are given in Table 30.

Note that we would fail to reject $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ at $\alpha = .05$ significance level if we looked only at the F-statistic and it's p-value ($F_{obs} = 2.93$, p-value=.0562). This would lead us to conclude that there is no association between the predictor variables income and region and the response variable clothing expenditures. This is where you need to be careful when using multiple regression with many predictor variables. Look at the test of $H_0 : \beta_1 = 0$, based on the t-test in Table 30. Here we observe t_{obs}=3.11, with a p-value of .0071. We thus conclude $\beta_1 \neq 0$, and that clothing expenditures is related to income, as we would expect. However, we do fail to reject $H_0 : \beta_2 = 0$, $H_0 : \beta_3 = 0$,and $H_0 : \beta_4 = 0$, so we fail to observe any differences among the regions in terms of clothing expenditures after 'adjusting' for the variable income. Figure 19 and Figure 20 show the original data using region as the plotting symbol and the 4 fitted equations corresponding to the 4

PER CAPITA INCOME & CLOTHING EXPENDITURES (1990)						
Metro Area	Region	Income x_1	Expenditure y	x_2	x_3	x_4
New York City	Northeast	25405	2290	0	0	0
Philadelphia	Northeast	21499	2037	0	0	0
Pittsburgh	Northeast	18827	1646	0	0	0
Boston	Northeast	24315	1659	0	0	0
Buffalo	Northeast	17997	1315	0	0	0
Atlanta	South	20263	2108	1	0	0
Miami/Ft Laud	South	19606	1587	1	0	0
Baltimore	South	21461	1978	1	0	0
Houston	South	19028	1589	1	0	0
Dallas/Ft Worth	South	19821	1982	1	0	0
Chicago	Midwest	21982	2108	0	1	0
Detroit	Midwest	20595	1262	0	1	0
Cleveland	Midwest	19640	2043	0	1	0
Minneapolis/St Paul	Midwest	21330	1816	0	1	0
St Louis	Midwest	20200	1340	0	1	0
Seattle	West	21087	1667	0	0	1
Los Angeles	West	20691	2404	0	0	1
Portland	West	18938	1440	0	0	1
San Diego	West	19588	1849	0	0	1
San Fran/Oakland	West	25037	2556	0	0	1

Table 28: Clothes Expenditures and income example

ANOVA					
Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	p-value
MODEL	1116419.0	4	279104.7	2.93	.0562
ERROR	1426640.2	15	95109.3		
TOTAL	2543059.2	19			

Table 29: The Analysis of Variance Table for clothes expenditure data

PARAMETER	ESTIMATE	t FOR H_0: β_i=0	P-VALUE	STANDARD ERROR OF ESTIMATE
INTERCEPT (β_0)	−657.428	−0.82	.4229	797.948
x_1 (β_1)	0.113	3.11	.0071	0.036
x_2 (β_2)	237.494	1.17	.2609	203.264
x_3 (β_3)	21.691	0.11	.9140	197.536
x_4 (β_4)	254.992	1.30	.2130	196.036

Table 30: Parameter estimates and tests of hypotheses for individual parameters

regions. Recall that the fitted equation is $\hat{y} = -657.428 + 0.113x_1 + 237.494x_2 + 21.691x_3 + 254.992x_4$, and each of the regions has a different set of levels of variables x_2, x_3, and x_4.

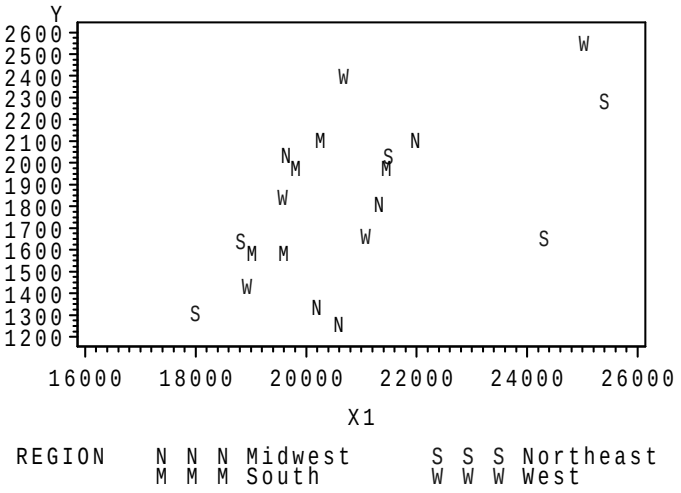

Figure 19: Plot of clothing data, with plotting symbol region

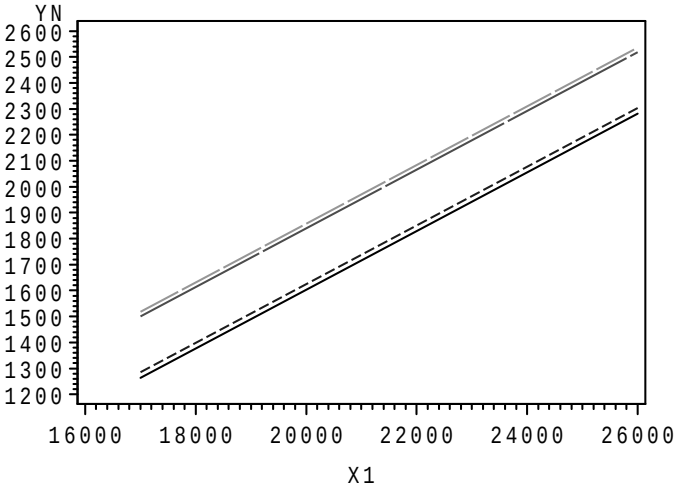

Figure 20: Plot of fitted equations for each region

6.3 Regression Models With Interactions

In some situations, two or more predictor variables may **interact** in terms of their effects on the mean response. That is, the effect on the mean response of changing the level of one predictor variable depends on the level of another predictor variable. This idea is easiest understood in the case where one of the variables is qualitative.

Example 6.3 – Truck and SUV Safety Ratings

Several years ago, *The Wall Street Journal* reported safety scores on 33 models of SUV's and

trucks. Safety scores (y) were reported, as well as the vehicle's weight (x_1) and and indicator of whether the vehicle has side air bags ($x_2 = 1$ if it does, 0 if not). We fit a model, relating safety scores to weight, presence of side airbags, and an interaction term that allows the effect of weight to depend on whether side airbags are present:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon.$$

We can write the equations for the two side airbag types as follows:

$$\text{Side airbags: } y = \beta_0 + \beta_1 x_1 + \beta_2(1) + \beta_3 x_1(1) + \varepsilon = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x_1 + \varepsilon,$$

and

$$\text{No side airbags: } y = \beta_0 + \beta_1 x_1 + \beta_2(0) + \beta_3 x_1(0) + \varepsilon = \beta_0 + \beta_1 x_1 + \varepsilon.$$

The data for years the 33 models are given in Table 31.

The Analysis of Variance table for this example is given in Table 32. Note that $R^2 = .5518$. Table 33 provides the parameter estimates, standard errors, and individual t-tests. Note that the F-test for testing $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ rejects the null hypothesis (F=11.90, P-value=.0001), but none of the individual t-tests are significant (all P-values exceed 0.05). This can happen due to the nature of the partial regression coefficients. It can be seen that weight is a very good predictor, and that the presence of side airbags and the interaction term do not contribute much to the model (SSE for a model containing only Weight (x_1) is 3493.7, use this to test $H_0 : \beta_2 = \beta_3 = 0$).

For vehicles with side airbags the fitted equation is:

$$\hat{y}_{airbags} = (b_0 + b_2) + (b_1 + b_3)x_1 = 44.18 + 0.02162x_1,$$

while for vehicles without airbags, the fitted equation is:

$$\hat{y}_{noairbags} = b_0 + b_1 x_1 = 76.09 + 0.01262x_1.$$

Figure 21 shows the two fitted equations for the safety data.

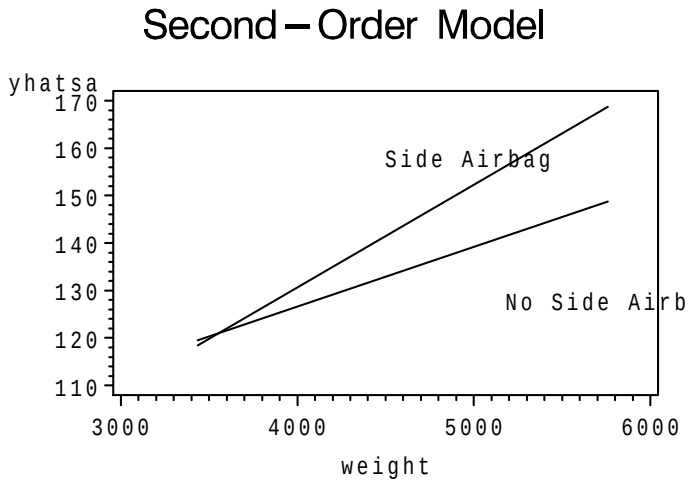

Figure 21: Plot of fitted equations for each vehicle type

SUV/Truck Safety Ratings				
Make	Model	Safety (y)	Weight (x_1)	Airbag (x_2)
TOYOTA	AVALON	111.34	3437	1
CHEVROLET	IMPALA	119.22	3454	1
FORD	RANGER	113.39	3543	0
BUICK	LESABRE	124.6	3610	1
MAZDA	MPV	117.13	3660	1
PLYMOUTH	VOYAGER	117.29	3665	0
VOLVO	S80	136.66	3698	1
AUDI	A8	138.62	3751	1
DODGE	DAKOTA	120.49	3765	0
ACURA	RL	113.05	3824	1
PONTIAC	TRANSPORT	118.83	3857	1
CHRYSLER	TOWN&COUNTRY	122.62	3918	0
FORD	F-150	118.7	3926	0
TOYOTA	4RUNNER	130.96	3945	0
MERCURY	GRAND MARQUIS	136.37	3951	0
ISUZU	RODEO	126.92	3966	0
TOYOTA	SIENNA	138.54	3973	0
MERCURY	VILLAGER	123.07	4041	0
LINCOLN	TOWN CAR	120.83	4087	1
FORD	F-150X	132.01	4125	0
FORD	WINDSTAR	152.48	4126	1
NISSAN	PATHFINDER	137.67	4147	1
OLDSMOBILE	BRAVADO	117.61	4164	0
HONDA	ODYSSEY	156.84	4244	0
MERCURY	MOUNTAINEER	136.27	4258	1
TOYOTA	TUNDRA	118.27	4356	0
MERCEDES-BENZ	ML320	140.57	4396	1
FORD	ECONOLINE	140.72	4760	0
DODGE	RAM	120.08	4884	0
LINCOLN	NAVIGATOR	144.57	4890	1
DODGE	RAM	144.75	4896	0
CADILLAC	ESCALANTE	158.82	5372	1
CHEVROLET	SUBURBAN	170.26	5759	1

Table 31: Safety ratings for trucks and SUV's

ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	p-value
MODEL	3838.03	3	1279.34	11.90	.0001
ERROR	3116.88	29	107.48		
TOTAL	6954.91	32			

Table 32: The Analysis of Variance Table for truck/SUV safety data

PARAMETER	ESTIMATE	STANDARD ERROR OF ESTIMATE	t FOR H_0: β_i=0	P-VALUE
INTERCEPT (β_0)	76.09	27.04	2.81	.0087
x_1 (β_1)	0.01262	0.0065	1.93	.0629
x_2 (β_2)	-31.91	31.78	-1.00	.3236
x_3 (β_3)	0.0090	.0076	1.18	.2487

Table 33: Parameter estimates and tests of hypotheses for individual parameters – Safety data

7 Introduction to Time Series and Forecasting

Textbook Sections: 21.1-21.6

In the remainder of the course, we consider data that are collected over time. Many economic and financial models are based on **time series**. First, we will describe means of smoothing series, then some simple ways to decompose a series, then we will describe some simple methods used to predict future outcomes based on past values.

7.1 Time Series Components

Textbook Section: 21.2

Time series can be broken into five components: **level**, **long-term trend**, **Cyclical variation**, **seasonal variation**, and **random variation**. A brief description of each is given below:

Level – Horizontal sales history in absence of other sources of variation (long run average).

Trend – Continuing pattern of increasing/decreasing values in the form of a line or curve.

Cyclical – Wavelike patterns that represent business cycles over multiple periods such as economic expansions and recessions.

Seasonal – Patterns that occur over repetitive calendar periods such as quarters, months, weeks, or times of the day.

Random – Short term irregularities that cause variation in individual outcomes above and beyond the other sources of variation.

Example 7.1 - U.S. Cotton Production - 1978-2001

Figure 22 represents a plot of U.S. cotton production from 1978 to 2001 (Source: Cotton association web site). We can see that there has been a trend to higher production over time, with cyclical patterns arising as well along the way. Since the data are annual production, we cannot observe seasonal patterns.

Example 7.2 - Texas in-state Finance/Insurance/Real Estate Sales - 1989-2002

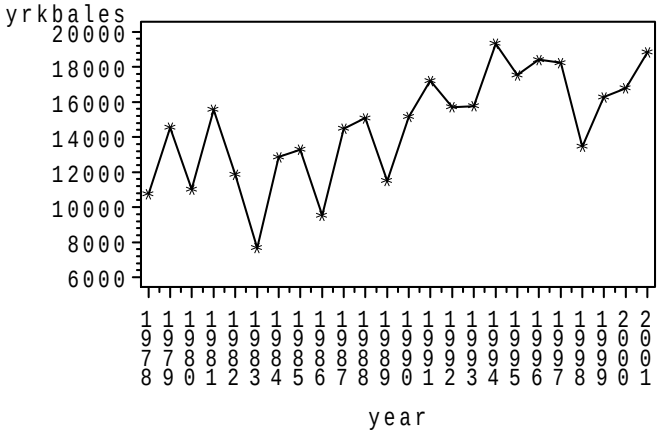

Figure 22: Plot of U.S. cotton production 1978-2001

Table 34 gives in-state gross sales for the Finance, Insurance, and Real Estate (FIRE) for the state of Texas for the 4 quarters of years 1989-2002 in hundreds of millions of dollars (Source: State of Texas web site). A plot of the data (with vertical lines delineating years) is shown in Figure 23. There is a clear positive trend in the series, and the fourth quarter tends to have much larger sales than the other three quarters. We will use the variables in the last two columns in a subsequent section.

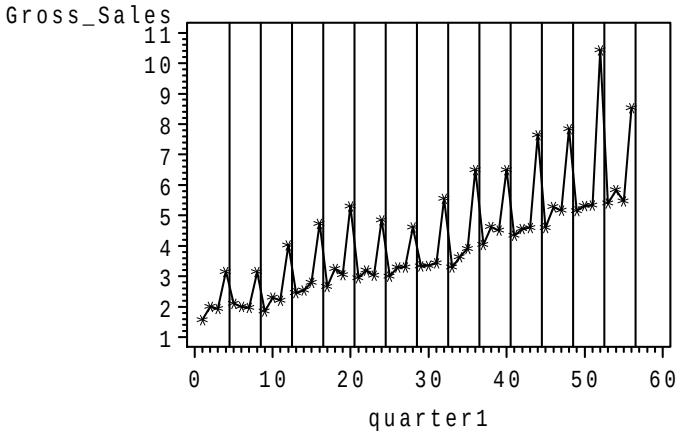

Figure 23: Plot of quarterly Texas in-state FIRE gross sales 1989-2002

7.2 Smoothing Techniques

Textbook Section: 21.3

Moving Averages are averages of values at a particular time period, and values that are near it in time. We will focus on odd numbered moving averages, as they are simpler to describe and

t	year	quarter	gross sales (y_t)	fitted sales ($\hat{y}_t$)	ratio ($y_t/\hat{y}_t$)
1	1989	1	1.567	1.725	0.908
2	1989	2	1.998	1.813	1.102
3	1989	3	1.929	1.900	1.015
4	1989	4	3.152	1.988	1.586
5	1990	1	2.108	2.075	1.016
6	1990	2	2.004	2.163	0.926
7	1990	3	1.965	2.250	0.873
8	1990	4	3.145	2.338	1.345
9	1991	1	1.850	2.425	0.763
10	1991	2	2.303	2.513	0.916
11	1991	3	2.209	2.600	0.850
12	1991	4	4.030	2.688	1.499
13	1992	1	2.455	2.776	0.884
14	1992	2	2.536	2.863	0.886
15	1992	3	2.800	2.951	0.949
16	1992	4	4.733	3.038	1.558
17	1993	1	2.666	3.126	0.853
18	1993	2	3.256	3.213	1.013
19	1993	3	3.050	3.301	0.924
20	1993	4	5.307	3.388	1.566
21	1994	1	2.950	3.476	0.849
22	1994	2	3.190	3.563	0.895
23	1994	3	3.025	3.651	0.829
24	1994	4	4.847	3.738	1.297
25	1995	1	3.005	3.826	0.785
26	1995	2	3.297	3.913	0.843
27	1995	3	3.301	4.001	0.825
28	1995	4	4.607	4.089	1.127
29	1996	1	3.333	4.176	0.798
30	1996	2	3.352	4.264	0.786
31	1996	3	3.430	4.351	0.788
32	1996	4	5.552	4.439	1.251
33	1997	1	3.297	4.526	0.728
34	1997	2	3.637	4.614	0.788
35	1997	3	3.909	4.701	0.832
36	1997	4	6.499	4.789	1.357
37	1998	1	4.047	4.876	0.830
38	1998	2	4.621	4.964	0.931
39	1998	3	4.509	5.051	0.893
40	1998	4	6.495	5.139	1.264
41	1999	1	4.334	5.226	0.829
42	1999	2	4.557	5.314	0.858
43	1999	3	4.596	5.401	0.851
44	1999	4	7.646	5.489	1.393
45	2000	1	4.596	5.577	0.824
46	2000	2	5.282	5.664	0.933
47	2000	3	5.158	5.752	0.897
48	2000	4	7.834	5.839	1.342
49	2001	1	5.155	5.927	0.870
50	2001	2	5.312	6.014	0.883
51	2001	3	5.331	6.102	0.874
52	2001	4	10.42	6.189	1.684
53	2002	1	5.397	6.277	0.860
54	2002	2	5.832	6.364	0.916
55	2002	3	5.467	6.452	0.847
56	2002	4	8.522	6.539	1.303

Table 34: Quarterly in-state gross sales for Texas FIRE firms

implement (the textbook also covers even numbered MA's as well). A 3-period moving averge involves averaging the value directly prior to the current time point, the current value, and the value directly after the current time point. There will not be values for either the first or last periods of the series. Similarly, a 5-period moving average will include the current time point, and the two prior time points and the two subsequent time points.

Example 7.3 - U.S. Internet Retail Sales - 1999q4-2003q1

The data in Table 35 gives the U.S. e-commerce sales for $n = 14$ quarters (quarter 1 is the 4th quarter of 1999 and quarter 14 is preliminary reported sales for the 1st quarter of 2003) in millions of dollars (Source: U.S. Census Bureau).

Quarter	Sales (y_t)	MA(3)	ES(0.1)	ES(0.5)
1	5393	.	5393	5393
2	5722	5788	5426	5558
3	6250	6350	5508	5904
4	7079	7526	5665	6491
5	9248	8112	6024	7870
6	8009	8387	6222	7939
7	7904	7936	6390	7922
8	7894	8862	6541	7908
9	10788	9384	6965	9348
10	9470	10006	7216	9409
11	9761	9899	7470	9585
12	10465	11332	7770	10025
13	13770	12052	8370	11897
14	11921	.	8725	11909

Table 35: Quarterly e-commerce sales and smoothed values for U.S. 1999q4-2003q1

To obatain the three period moving average (MA(3)) for the second quarter, we average the first, second, and third period sales:

$$\frac{5393 + 5722 + 6250}{3} = \frac{17365}{3} = 5788.3 \approx 5788$$

We can similarly obtain the three period moving average for quarters 3-13. The data and three period moving averages are given in Figure 24. The moving average is the dashed line, while the original series is the solid line.

Exponential Smoothing is an alternative means of smoothing the series. It makes use of all prior time points, with higher weights on more recent time points, and exponentially decaying weights on more distance time points. One advantage is that we have smoothed values for all time points. One drawback is that we must select a tuning parameter (although we would also have to choose the length of a moving average as well, for that method). One widely used convention is to set the first period's smoothed value to the first observation, then make subsequent smoothed values as a weighted average of the current observation and the previous value of the smoothed series. We use the notation S_t for the smoothed value at time t.

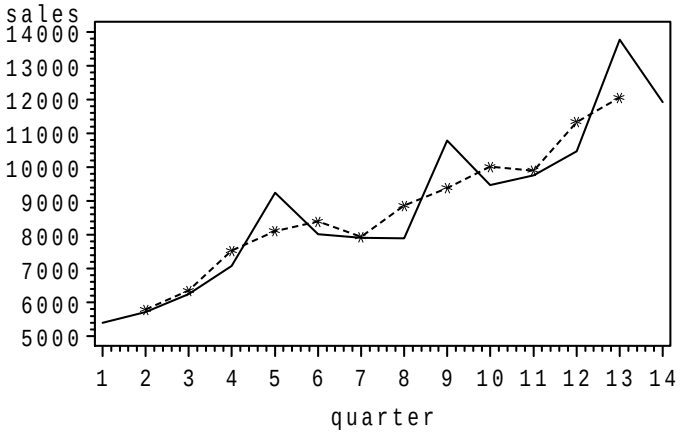

Figure 24: Plot of quarterly U.S. internet retail sales and 3-Period moving average

$$S_1 = y_1 \qquad S_t = wy_t + (1-w)S_{t-1} \quad t \geq 2$$

Example 7.3 (Continued)

Thus, for quarter 4 of 1999, we set $S_1 = y_1 = 5393$. In Table 35, we include smoothed values based on $w = 0.1$ and $w = 0.5$, respectively:

$$w = 0.1: \quad S_2 = 0.1 * y_2 + 0.9 * S_1 = 0.1(5722) + 0.9(5393) = 572.2 + 4853.7 = 5425.9 \approx 5426$$

$$w = 0.5: \quad S_2 = 0.5 * y_2 + 0.5 * S_1 = 0.5(5722) + 0.5(5393) = 2861.0 + 2696.5 = 5557.5 \approx 5558$$

The smoothed values are given in Table 35, as well as in Figure 25. The solid line is the original series, the smoothest line is $w = 0.1$, and the intermediate line is $w = 0.5$.

7.3 Estimating Trend and Seasonal Effects

Textbook Section: 21.4

While the **cyclical** patterns are difficult to predict and estimate, we can estimate **linear trend** and **seasonal indexes** fairly simply. Further, there is no added difficulty if the trend is nonlinear (quadratic), but we will consider only the linear case here.

First, we must identify seasons, these can be weeks, months, or quarters (or even times of the day or days of the week). Then we fit a linear trend for the entire series. This is followed by taking the ratio of the actual to the fitted value (from the regression equation) for each period. Next, we average these ratios for each season, and adjust so that the averages sum to 1.0.

Example 7.2 (Continued)

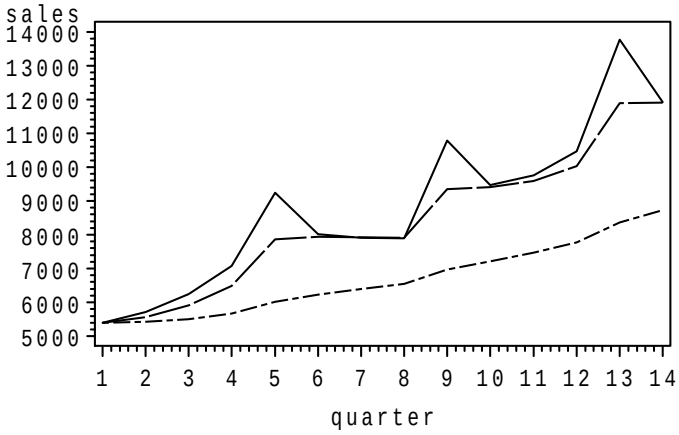

Figure 25: Plot of quarterly U.S. internet retail sales and 32 Exponentially smoothed series

Consider the Texas gross (in-state) sales for the FIRE industry. The seasons are the four quarters. Fitting a simple linear regression, relating sales to time period, we get:

$$\hat{y}_t = b_0 + b_1 t = 1.6376 + 0.08753t$$

The fitted values (as well as the observed values) have been shown previously in Table 34. Also for each outcome, we obtain the ratio of the observed to fitted value, also given in the table. Consider the first and last cases:

$$t = 1: \quad y_1 = 1.567 \qquad \hat{y}_1 = 1.6376 + 0.08753(1) = 1.725 \qquad \frac{y_1}{\hat{y}_1} = \frac{1.567}{1.725} = 0.908$$

$$t = 56: \quad y_{56} = 8.522 \qquad \hat{y}_1 = 1.6376 + 0.08753(56) = 6.539 \qquad \frac{y_1}{\hat{y}_1} = \frac{8.522}{6.539} = 1.303$$

Next, we take the mean of the observed-to-fitted ratio for each quarter. There are 14 years of data:

$$Q1: \quad \frac{0.908 + 1.016 + 0.763 + 0.884 + 0.853 + 0.849 + 0.785 + 0.798 + 0.728 + 0.830 + 0.829 + 0.824 + 0.870 + 0.860}{14} = 0.84$$

The means for the remaining three quarters are:

$$Q2: \quad 0.906 \qquad Q3: \quad 0.875 \qquad Q4: \quad 1.398$$

The means sum to 4.022, and have a mean of 4.022/4=1.0055. If we divide each mean by 1.0055, the indexes will sum to 1:

$$Q1: \quad 0.838 \qquad Q2: \quad 0.901 \qquad Q3: \quad 0.870 \qquad Q4: \quad 1.390$$

The seasonally adjusted time series is given by dividing each observed value by its seasonal index. This way, we can determine when there are real changes in the series, beyond seasonal fluctuations. Table 36 contains all components as well as the seasonally adjusted values.

t	year	quarter	gross sales (y_t)	fitted sales ($\hat{y}_t$)	ratio ($y_t/\hat{y}_t$)	season adjusted
1	1989	1	1.567	1.725	0.908	1.870
2	1989	2	1.998	1.813	1.102	2.218
3	1989	3	1.929	1.900	1.015	2.218
4	1989	4	3.152	1.988	1.586	2.268
5	1990	1	2.108	2.075	1.016	2.515
6	1990	2	2.004	2.163	0.926	2.224
7	1990	3	1.965	2.250	0.873	2.259
8	1990	4	3.145	2.338	1.345	2.263
9	1991	1	1.850	2.425	0.763	2.207
10	1991	2	2.303	2.513	0.916	2.556
11	1991	3	2.209	2.600	0.850	2.540
12	1991	4	4.030	2.688	1.499	2.899
13	1992	1	2.455	2.776	0.884	2.929
14	1992	2	2.536	2.863	0.886	2.815
15	1992	3	2.800	2.951	0.949	3.218
16	1992	4	4.733	3.038	1.558	3.405
17	1993	1	2.666	3.126	0.853	3.182
18	1993	2	3.256	3.213	1.013	3.614
19	1993	3	3.050	3.301	0.924	3.506
20	1993	4	5.307	3.388	1.566	3.818
21	1994	1	2.950	3.476	0.849	3.521
22	1994	2	3.190	3.563	0.895	3.540
23	1994	3	3.025	3.651	0.829	3.477
24	1994	4	4.847	3.738	1.297	3.487
25	1995	1	3.005	3.826	0.785	3.585
26	1995	2	3.297	3.913	0.843	3.660
27	1995	3	3.301	4.001	0.825	3.794
28	1995	4	4.607	4.089	1.127	3.314
29	1996	1	3.333	4.176	0.798	3.977
30	1996	2	3.352	4.264	0.786	3.720
31	1996	3	3.430	4.351	0.788	3.942
32	1996	4	5.552	4.439	1.251	3.994
33	1997	1	3.297	4.526	0.728	3.934
34	1997	2	3.637	4.614	0.788	4.037
35	1997	3	3.909	4.701	0.832	4.493
36	1997	4	6.499	4.789	1.357	4.675
37	1998	1	4.047	4.876	0.830	4.829
38	1998	2	4.621	4.964	0.931	5.129
39	1998	3	4.509	5.051	0.893	5.183
40	1998	4	6.495	5.139	1.264	4.672
41	1999	1	4.334	5.226	0.829	5.171
42	1999	2	4.557	5.314	0.858	5.058
43	1999	3	4.596	5.401	0.851	5.282
44	1999	4	7.646	5.489	1.393	5.501
45	2000	1	4.596	5.577	0.824	5.485
46	2000	2	5.282	5.664	0.933	5.862
47	2000	3	5.158	5.752	0.897	5.928
48	2000	4	7.834	5.839	1.342	5.636
49	2001	1	5.155	5.927	0.870	6.152
50	2001	2	5.312	6.014	0.883	5.896
51	2001	3	5.331	6.102	0.874	6.128
52	2001	4	10.42	6.189	1.684	7.498
53	2002	1	5.397	6.277	0.860	6.440
54	2002	2	5.832	6.364	0.916	6.473
55	2002	3	5.467	6.452	0.847	6.283
56	2002	4	8.522	6.539	1.303	6.131

Table 36: Quarterly in-state gross sales for Texas FIRE firms and seasonally adjusted series

7.4 Introduction to Forecasting

Textbook Section: 21.5

There are unlimited number of possibilities of ways of forecasting future outcomes, so we need means of comparing the various methods. First, we introduce some notation:

- y_t — Actual (random) outcome at time t, unknown prior to t
- F_t — Forecast of y_t, made prior to t
- e_t — Forecast error $e_t = y_t - F_t$ (Book does not use this notation).

Two commonly used measures of comparing forecasting methods are given below:

Mean Absolute Deviation (MAD) — MAD=$\frac{\sum |e_t|}{\text{number of forecasts}} = \frac{\sum_{t=1}^{n} |y_t - F_t|}{n}$

Sum of Square Errors (SSE) — SSE=$\sum e_t^2 = \sum_{t=1}^{n} (y_t - F_t)^2$

When comparing forecasting methods, we wish to minimize one or both of these quantities.

7.5 Simple Forecasting Techniques

Textbook Section: 21.6 and Supplement

In this section, we describe some simple methods of using past data to predict future outcomes. Most forecasts you hear reported are generally complex hybrids of these techniques.

7.5.1 Moving Averages

This method, which is not included in the tesxt, is a slight adjustment to the centered moving averages in the smoothing section. At time point t, we use the previous k observations to forecast y_t. We use the mean of the last k observations to forecast outcome at t:

$$F_t = \frac{X_{t-1} + X_{t-2} + \cdots + X_{t-k}}{k}$$

Problem: How to choose k?

Example 7.4 - Anhueser-Busch Annual Dividend Yields 1952-1995

Table 37 gives average dividend yields for Anheuser–Busch for the years 1952–1995 (Source: *Value Line*), forecasts and errors based on moving averages based on lags of 1, 2, and 3. Note that we don't have early year forecasts, and the longer the lag, the longer we must wait until we get our first forecast.

Here we compute moving averages for year=1963:

1–Year: $F_{1963} = y_{1962} = 3.2$

2–Year: $F_{1963} = \frac{y_{1962}+y_{1961}}{2} = \frac{3.2+2.8}{2} = 3.0$

3–Year: $F_{1963} = \frac{y_{1962}+y_{1961}+y_{1960}}{3} = \frac{3.2+2.8+4.4}{3} = 3.47$

Figure 26 displays raw data and moving average forecasts.

t	Year	y_t	$F_{1,t}$	$e_{1,t}$	$F_{2,t}$	$e_{2,t}$	$F_{3,t}$	$e_{3,t}$
1	1952	5.30	.	.	.	.	.	.
2	1953	4.20	5.30	-1.10	.	.	.	.
3	1954	3.90	4.20	-0.30	4.75	-0.85	.	.
4	1955	5.20	3.90	1.30	4.05	1.15	4.47	0.73
5	1956	5.80	5.20	0.60	4.55	1.25	4.43	1.37
6	1957	6.30	5.80	0.50	5.50	0.80	4.97	1.33
7	1958	5.60	6.30	-0.70	6.05	-0.45	5.77	-0.17
8	1959	4.80	5.60	-0.80	5.95	-1.15	5.90	-1.10
9	1960	4.40	4.80	-0.40	5.20	-0.80	5.57	-1.17
10	1961	2.80	4.40	-1.60	4.60	-1.80	4.93	-2.13
11	1962	3.20	2.80	0.40	3.60	-0.40	4.00	-0.80
12	1963	3.10	3.20	-0.10	3.00	0.10	3.47	-0.37
13	1964	3.10	3.10	0.00	3.15	-0.05	3.03	0.07
14	1965	2.60	3.10	-0.50	3.10	-0.50	3.13	-0.53
15	1966	2.00	2.60	-0.60	2.85	-0.85	2.93	-0.93
16	1967	1.60	2.00	-0.40	2.30	-0.70	2.57	-0.97
17	1968	1.30	1.60	-0.30	1.80	-0.50	2.07	-0.77
18	1969	1.20	1.30	-0.10	1.45	-0.25	1.63	-0.43
19	1970	1.20	1.20	0.00	1.25	-0.05	1.37	-0.17
20	1971	1.10	1.20	-0.10	1.20	-0.10	1.23	-0.13
21	1972	0.90	1.10	-0.20	1.15	-0.25	1.17	-0.27
22	1973	1.40	0.90	0.50	1.00	0.40	1.07	0.33
23	1974	2.00	1.40	0.60	1.15	0.85	1.13	0.87
24	1975	1.90	2.00	-0.10	1.70	0.20	1.43	0.47
25	1976	2.30	1.90	0.40	1.95	0.35	1.77	0.53
26	1977	3.10	2.30	0.80	2.10	1.00	2.07	1.03
27	1978	3.50	3.10	0.40	2.70	0.80	2.43	1.07
28	1979	3.80	3.50	0.30	3.30	0.50	2.97	0.83
29	1980	3.70	3.80	-0.10	3.65	0.05	3.47	0.23
30	1981	3.10	3.70	-0.60	3.75	-0.65	3.67	-0.57
31	1982	2.60	3.10	-0.50	3.40	-0.80	3.53	-0.93
32	1983	2.40	2.60	-0.20	2.85	-0.45	3.13	-0.73
33	1984	3.00	2.40	0.60	2.50	0.50	2.70	0.30
34	1985	2.40	3.00	-0.60	2.70	-0.30	2.67	-0.27
35	1986	1.80	2.40	-0.60	2.70	-0.90	2.60	-0.80
36	1987	1.70	1.80	-0.10	2.10	-0.40	2.40	-0.70
37	1988	2.20	1.70	0.50	1.75	0.45	1.97	0.23
38	1989	2.10	2.20	-0.10	1.95	0.15	1.90	0.20
39	1990	2.40	2.10	0.30	2.15	0.25	2.00	0.40
40	1991	2.10	2.40	-0.30	2.25	-0.15	2.23	-0.13
41	1992	2.20	2.10	0.10	2.25	-0.05	2.20	0.00
42	1993	2.70	2.20	0.50	2.15	0.55	2.23	0.47
43	1994	3.00	2.70	0.30	2.45	0.55	2.33	0.67
44	1995	2.80	3.00	-0.20	2.85	-0.05	2.63	0.17

Table 37: Dividend yields, Forecasts, errors — 1, 2, and 3 year moving Averages

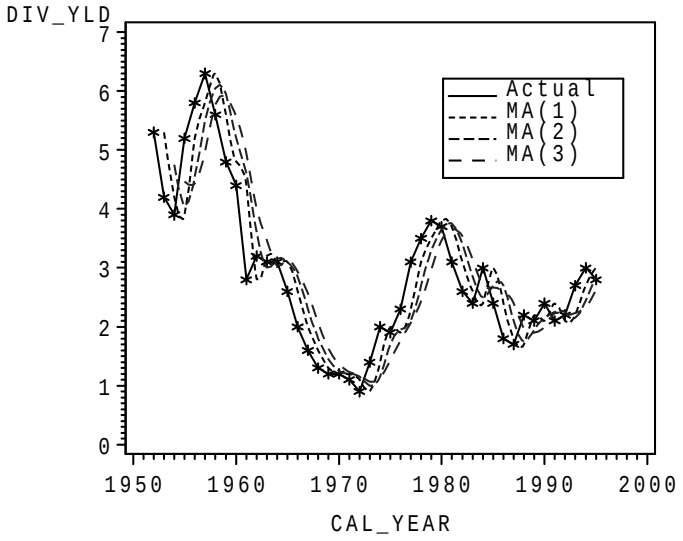

Figure 26: Plot of the data moving average forecast for Anheuser–Busch dividend data

7.5.2 Exponential Smoothing

Exponential smoothing is a method of forecasting that weights data from previous time periods with exponentially decreasing magnitudes. Forecasts can be written as follows, where the forecast for period 2 is traditionally (but not always) simply the outcome from period 1:

$$F_{t+1} = S_t = wy_t + (1-w)S_{t-1} = wy_t + (1-w)F_t$$

where :

- F_{t+1} is the forecast for period $t+1$
- y_t is the outcome at t
- S_t is the smoothed value for period t ($S_{t-1} = F_t$)
- w is the smoothing constant ($0 \leq w \leq 1$)

Forecasts are "smoother" than the raw data and weights of previous observations decline exponentially with time.

Example 7.4 (Continued)
3 smoothing constants (allowing decreasing amounts of smoothness) for illustration:

- $w = 0.2$ — $F_{t+1} = 0.2y_t + 0.8S_{t-1} = 0.2y_t + 0.8F_t$
- $w = 0.5$ — $F_{t+1} = 0.5y_t + 0.5S_{t-1} = 0.5y_t + 0.5F_t$
- $w = 0.8$ — $F_{t+1} = 0.8y_t + 0.2S_{t-1} = 0.8y_t + 0.2F_t$

Year 2 (1953) — set $F_{1953} = X_{1952}$, then cycle from there.

t	Year	y_t	$F_{w=.2,t}$	$e_{w=.2,t}$	$F_{w=.5,t}$	$e_{w=.5,t}$	$F_{w=.8,t}$	$e_{w=.8,t}$
1	1952	5.30	.	.	.	.	.	.
2	1953	4.20	5.30	-1.10	5.30	-1.10	5.30	-1.10
3	1954	3.90	5.08	-1.18	4.75	-0.85	4.42	-0.52
4	1955	5.20	4.84	0.36	4.33	0.88	4.00	1.20
5	1956	5.80	4.92	0.88	4.76	1.04	4.96	0.84
6	1957	6.30	5.09	1.21	5.28	1.02	5.63	0.67
7	1958	5.60	5.33	0.27	5.79	-0.19	6.17	-0.57
8	1959	4.80	5.39	-0.59	5.70	-0.90	5.71	-0.91
9	1960	4.40	5.27	-0.87	5.25	-0.85	4.98	-0.58
10	1961	2.80	5.10	-2.30	4.82	-2.02	4.52	-1.72
11	1962	3.20	4.64	-1.44	3.81	-0.61	3.14	0.06
12	1963	3.10	4.35	-1.25	3.51	-0.41	3.19	-0.09
13	1964	3.10	4.10	-1.00	3.30	-0.20	3.12	-0.02
14	1965	2.60	3.90	-1.30	3.20	-0.60	3.10	-0.50
15	1966	2.00	3.64	-1.64	2.90	-0.90	2.70	-0.70
16	1967	1.60	3.31	-1.71	2.45	-0.85	2.14	-0.54
17	1968	1.30	2.97	-1.67	2.03	-0.73	1.71	-0.41
18	1969	1.20	2.64	-1.44	1.66	-0.46	1.38	-0.18
19	1970	1.20	2.35	-1.15	1.43	-0.23	1.24	-0.04
20	1971	1.10	2.12	-1.02	1.32	-0.22	1.21	-0.11
21	1972	0.90	1.91	-1.01	1.21	-0.31	1.12	-0.22
22	1973	1.40	1.71	-0.31	1.05	0.35	0.94	0.46
23	1974	2.00	1.65	0.35	1.23	0.77	1.31	0.69
24	1975	1.90	1.72	0.18	1.61	0.29	1.86	0.04
25	1976	2.30	1.76	0.54	1.76	0.54	1.89	0.41
26	1977	3.10	1.86	1.24	2.03	1.07	2.22	0.88
27	1978	3.50	2.11	1.39	2.56	0.94	2.92	0.58
28	1979	3.80	2.39	1.41	3.03	0.77	3.38	0.42
29	1980	3.70	2.67	1.03	3.42	0.28	3.72	-0.02
30	1981	3.10	2.88	0.22	3.56	-0.46	3.70	-0.60
31	1982	2.60	2.92	-0.32	3.33	-0.73	3.22	-0.62
32	1983	2.40	2.86	-0.46	2.96	-0.56	2.72	-0.32
33	1984	3.00	2.77	0.23	2.68	0.32	2.46	0.54
34	1985	2.40	2.81	-0.41	2.84	-0.44	2.89	-0.49
35	1986	1.80	2.73	-0.93	2.62	-0.82	2.50	-0.70
36	1987	1.70	2.54	-0.84	2.21	-0.51	1.94	-0.24
37	1988	2.20	2.38	-0.18	1.96	0.24	1.75	0.45
38	1989	2.10	2.34	-0.24	2.08	0.02	2.11	-0.01
39	1990	2.40	2.29	0.11	2.09	0.31	2.10	0.30
40	1991	2.10	2.31	-0.21	2.24	-0.14	2.34	-0.24
41	1992	2.20	2.27	-0.07	2.17	0.03	2.15	0.05
42	1993	2.70	2.26	0.44	2.19	0.51	2.19	0.51
43	1994	3.00	2.35	0.65	2.44	0.56	2.60	0.40
44	1995	2.80	2.48	0.32	2.72	0.08	2.92	-0.12

Table 38: Dividend yields, Forecasts, and errors based on exponential smoothing with $w = 0.2, 0.5, 0.8$

Table 38 gives average dividend yields for Anheuser–Busch for the years 1952–1995 (Source: *Value Line*), forecasts and errors based on exponential smoothing based on lags of 1, 2, and 3.

Here we obtain Forecasts based on Exponential Smoothing, beginning with year 2 (1953):
1953: $F_{w=.2,1953} = y_{1952} = 5.30 \quad F_{w=.5,1952} = y_{1952} = 5.30 \quad F_{w=.8,1952} = y_{1952} = 5.30$

1954 ($w = 0.2$): $F_{w=.2,1954} = .2y_{1953} + .8F_{w=.2,1953} = .2(4.20) + .8(5.30) = 5.08$

1954 ($w = 0.5$): $F_{w=.5,1954} = .5y_{1953} + .5F_{w=.5,1953} = .5(4.20) + .5(5.30) = 4.75$

1954 ($w = 0.8$): $F_{w=.8,1954} = .8y_{1953} + .2F_{w=.5,1953} = .8(4.20) + .2(5.30) = 4.42$

Which level of w appears to be "discounting" more distant observations at a quicker rate? What would happen if $w = 1$? If $w = 0$? Figure 27 gives raw data and exponential smoothing forecasts.

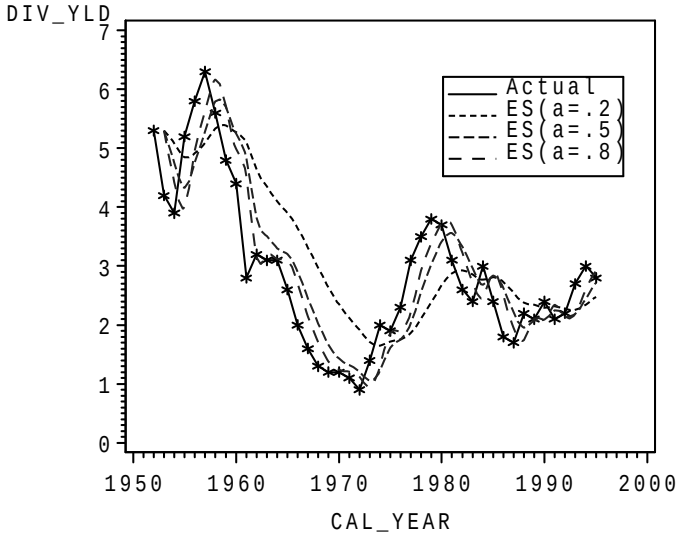

Figure 27: Plot of the data and Exponential Smoothing forecasts for Anheuser–Busch dividend data

Table 39 gives measures of forecast errors for three moving average, and three exponential smoothing methods.

	Moving Average			Exponential Smoothing		
Measure	1–Period	2–Period	3–Period	$w = 0.2$	$w = 0.5$	$w = 0.8$
MAE	0.43	0.53	0.62	0.82	0.58	0.47
MSE	0.30	0.43	0.57	0.97	0.48	0.34

Table 39: Relative performances of 6 forecasting methods — Anheuser–Busch data

Note that MSE is SSE/n where n is the number of forecasts.

7.6 Forecasting with Seasonal Indexes

After trend and seasonal indexes have been estimated, future outcomes can be forecast by the equation:

$$F_t = [b_0 + b_1 t] \times SI_t$$

where $b_0 + b_1 t$ is the linear trend and SI_t is the seasonal index for period t.

Example 7.2 (Continued)

For the Texas FIRE gross sales data, we have:

$$b_0 = 1.6376 \quad b_1 = .08753 \qquad SI_{Q1} = .838 \quad SI_{Q2} = .901 \quad SI_{Q3} = .870 \quad SI_{Q4} = 1.390$$

Thus for the 4 quarters of 2003 ($t = 57, 58, 59, 60$), we have:

$$Q1 : F_{57} = [1.6376 + 0.08753(57)](.838) = 5.553 \quad Q2 : F_{58} = [1.6376 + 0.08753(58)](.901) = 6.050$$

$$Q3 : F_{59} = [1.6376 + 0.08753(59)](.870) = 5.918 \quad Q4 : F_{60} = [1.6376 + 0.08753(60)](1.390) = 9.576$$

7.6.1 Autoregression

Sometimes regression is run on past or "lagged" values of the dependent variable (and possibly other variables). An Autoregressive model with independent variables corresponding to k periods can be written as follows:

$$\hat{y}_t = b_0 + b_1 y_{t-1} + b_2 y_{t-2} + \cdots + b_k y_{t-k}$$

Note that the regression cannot be run for the first k responses in the series. Technically forecasts can be given for only periods after the regression has been fit, since the regression model depends on all periods used to fit it.

Example 7.4 (Continued)

From Computer software, autoregressions based on lags of 1, 2, and 3 periods are fit:

1–Period: $\hat{y}_t = 0.29 + 0.88 y_{t-1}$

2–Period: $\hat{y}_t = 0.29 + 1.18 y_{t-1} - 0.29 y_{t-2}$

3–Period: $\hat{y}_t = 0.28 + 1.21 y_{t-1} - 0.37 y_{t-2} + 0.05 y_{t-3}$

Table 40 gives raw data and forecasts based on three autoregression models. Figure 28 displays the actual outcomes and predictions.

t	Year	y_t	$F_{AR(1),t}$	$e_{AR(1),t}$	$F_{AR(2),t}$	$e_{(AR(2),t}$	$F_{AR(3),t}$	$e_{AR(3),t}$
1	1952	5.3	.	.	.	.	.	.
2	1953	4.2	4.96	-0.76	.	.	.	.
3	1954	3.9	3.99	-0.09	3.72	0.18	.	.
4	1955	5.2	3.72	1.48	3.68	1.52	3.72	1.48
5	1956	5.8	4.87	0.93	5.30	0.50	5.35	0.45
6	1957	6.3	5.40	0.90	5.64	0.66	5.58	0.72
7	1958	5.6	5.84	-0.24	6.06	-0.46	6.03	-0.43
8	1959	4.8	5.22	-0.42	5.09	-0.29	5.03	-0.23
9	1960	4.4	4.52	-0.12	4.34	0.06	4.35	0.05
10	1961	2.8	4.16	-1.36	4.10	-1.30	4.12	-1.32
11	1962	3.2	2.75	0.45	2.33	0.87	2.29	0.91
12	1963	3.1	3.11	-0.01	3.26	-0.16	3.35	-0.25
13	1964	3.1	3.02	0.08	3.03	0.07	3.00	0.10
14	1965	2.6	3.02	-0.42	3.06	-0.46	3.05	-0.45
15	1966	2	2.58	-0.58	2.47	-0.47	2.44	-0.44
16	1967	1.6	2.05	-0.45	1.90	-0.30	1.90	-0.30
17	1968	1.3	1.70	-0.40	1.60	-0.30	1.61	-0.31
18	1969	1.2	1.43	-0.23	1.36	-0.16	1.37	-0.17
19	1970	1.2	1.35	-0.15	1.33	-0.13	1.34	-0.14
20	1971	1.1	1.35	-0.25	1.36	-0.26	1.36	-0.26
21	1972	0.9	1.26	-0.36	1.24	-0.34	1.23	-0.33
22	1973	1.4	1.08	0.32	1.03	0.37	1.03	0.37
23	1974	2	1.52	0.48	1.68	0.32	1.70	0.30
24	1975	1.9	2.05	-0.15	2.25	-0.35	2.23	-0.33
25	1976	2.3	1.96	0.34	1.96	0.34	1.92	0.38
26	1977	3.1	2.31	0.79	2.46	0.64	2.47	0.63
27	1978	3.5	3.02	0.48	3.29	0.21	3.28	0.22
28	1979	3.8	3.37	0.43	3.53	0.27	3.49	0.31
29	1980	3.7	3.64	0.06	3.77	-0.07	3.75	-0.05
30	1981	3.1	3.55	-0.45	3.56	-0.46	3.54	-0.44
31	1982	2.6	3.02	-0.42	2.88	-0.28	2.86	-0.26
32	1983	2.4	2.58	-0.18	2.47	-0.07	2.47	-0.07
33	1984	3	2.40	0.60	2.37	0.63	2.39	0.61
34	1985	2.4	2.93	-0.53	3.14	-0.74	3.16	-0.76
35	1986	1.8	2.40	-0.60	2.26	-0.46	2.20	-0.40
36	1987	1.7	1.87	-0.17	1.72	-0.02	1.73	-0.03
37	1988	2.2	1.79	0.41	1.78	0.42	1.80	0.40
38	1989	2.1	2.23	-0.13	2.40	-0.30	2.41	-0.31
39	1990	2.4	2.14	0.26	2.13	0.27	2.10	0.30
40	1991	2.1	2.40	-0.30	2.52	-0.42	2.53	-0.43
41	1992	2.2	2.14	0.06	2.08	0.12	2.05	0.15
42	1993	2.7	2.23	0.47	2.28	0.42	2.29	0.41
43	1994	3	2.67	0.33	2.84	0.16	2.85	0.15
44	1995	2.8	2.93	-0.13	3.05	-0.25	3.03	-0.23

Table 40: Average dividend yields and Forecasts/errors based on autoregression with lags of 1, 2, and 3 periods

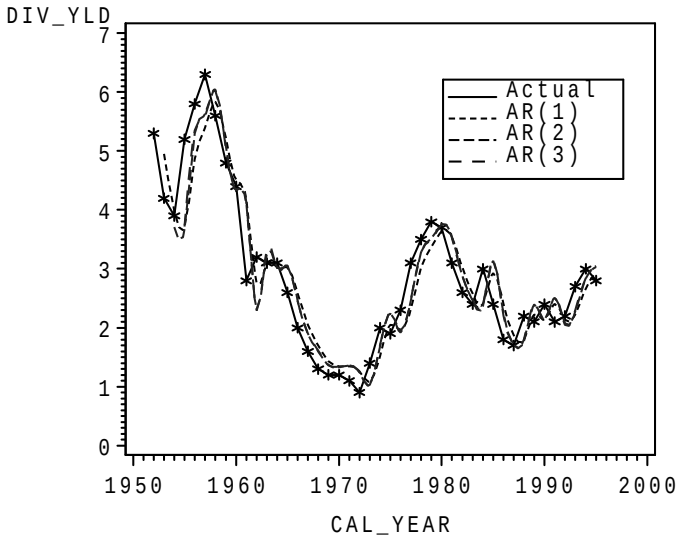

Figure 28: Plot of the data and Autoregressive forecasts for Anheuser–Busch dividend data