Notes on Applied Linear Regression

Jamie DeCoster

Department of Social Psychology
Free University Amsterdam
Van der Boechorststraat 1
1081 BT Amsterdam
The Netherlands

phone: +31 (0)20 444-8935
email: j.decoster@psy.vu.nl

April 10, 2003

These were compiled from notes made by Jamie DeCoster and Julie Gephart from Dr. Rebecca Doerge's class on applied linear regression at Purdue University. Textbook references refer to Neter, Kutner, Nachtsheim, & Wasserman's *Applied Linear Statistical Models*, fourth edition.

For help with data analysis visit
http://www.stat-help.com

Contents

Chapter 1

Linear Regression with One Independent Variable

1.1 General information about regression

- Regression analysis is a statistical tool that utilizes the relation between two or more quantitative variables so that one variable can predicted from another.
- The variable we are trying to predict is called the response or dependent variable. The variable predicting this is called the explanatory or independent variable.
- Relationships between variables can be either functional or statistical. A functional relationship is exact, while a statistical relationship has error associated with it.
- Regression analysis serves three purposes: Description, Control, and Prediction.
- Whenever reporting results, be sure to use at least four decimal places.

1.2 Simple linear regression model

- The simple linear regression function is

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i. \tag{1.1}$$

  In this model β_0, β_1, and ϵ_i are parameters and Y_i and X_i are measured values. This is called a "first order model" because it is linear in both the parameters and the independent variable.

- For every level of X, there is a probability distribution for Y having mean $E(Y_i) = \beta_0 + \beta_1 X_i$ and variance $\sigma^2(Y_i) = \sigma^2$, where σ^2 is the variance of the entire population.
- β_0 is called the intercept and β_1 is called the slope of the regression line.
- Data for the regression analysis may be either observational or experimental. Observational data is simply recorded from naturally occurring phenomena while experimental data is the result of some manipulation by the experimenter.

1.3 Estimated regression function

- The estimated regression function is

$$Y_i = b_0 + b_1 X_i. \tag{1.2}$$

- We calculate b_0 and b_1 using the methods of least squares. This chooses estimates that minimizes the sum of squared errors. These estimates can be calculated as

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{\sum X_i Y_i - \frac{\sum X_i \sum Y_i}{n}}{\sum {X_i}^2 - \frac{(\sum X_i)^2}{n}} \tag{1.3}$$

and

$$b_0 = \bar{Y} - b_1 \bar{X} = \frac{1}{n}\left(\sum Y_i - b_1 \sum X_i\right). \tag{1.4}$$

- Under the conditions of the regression model given above, the least squares estimates are unbiased and have minimum variance among all unbiased linear estimators. This means that the estimates get us as close to the true unknown parameter values as we can get.

- $E(Y)$ is referred to as the mean response or the "expected value" of Y. It is the center of the probability distribution of Y.

- The least squares regression line always passes through the point $(\bar{X}, \bar{Y})$.

- The predicted or fitted values of the regression equation calculated as

$$\hat{Y}_i = b_0 + b_1 X_i. \tag{1.5}$$

- Residuals are the difference between the response and the fitted value,

$$e_i = Y_i - \hat{Y}_i. \tag{1.6}$$

  We often examine the distribution of the residuals to check the fit of our model. Some general properties of the residuals are:

  - $\sum e_i = 0$
  - $\sum e_i^2$ is minimum when using least squares estimates
  - $\sum Y_i = \sum \hat{Y}_i$
  - $\sum X_i e_i = 0$
  - $\sum \hat{Y}_i e_i = 0$

- The sample variance s^2 estimates the population variance σ^2. This estimate is calculated as

$$s^2 = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2}, \tag{1.7}$$

  and is also referred to as the Mean Square Error (MSE).

- Note that estimated values are always represented by English letters while parameter values are always represented by Greek letters.

1.4 Normal error regression model

- The normal error regression model is

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \tag{1.8}$$

  where $i = 1 \ldots n$ are the trials of the experiment, Y_i is the observed response on trial i, X_i is the level of the independent variable on trial i, β_0 and β_1 are parameters, and the residuals ϵ_i are all normally distributed with a mean of 0 and a variance of σ^2.

- In this model Y_i and X_i are known values while β_0, β_1, and ϵ_i are parameters.

- The additional assumptions in this model allow us to draw inferences from our data. This means that we can test our parameter estimates and build confidence intervals.

Chapter 2

Inferences in Regression Analysis

2.1 General information about statistical inferences

- Every point estimate has a sampling distribution that describes the behavior of that estimate. By knowing the sampling distribution you know the center and the spread (expected value and variance, respectively) of the distribution of the estimate. Recall that a sampling distribution is also a probability distribution.
- We use two different ways to estimate parameters: point estimates and confidence intervals.

2.2 Hypothesis tests

- Specifically we are often interested in building hypothesis tests pitting a null hypothesis versus an alternative hypothesis. One hypothesis test commonly performed in simple linear regression is

$$\begin{aligned} H_0: &\quad \beta_1 = 0 \\ H_a: &\quad \beta_1 \neq 0. \end{aligned} \tag{2.1}$$

- There are four possible results to a hypothesis test. You can:
  1. Fail to reject the null when the null is true (correct).
  2. Reject the null when the alternative is true (correct).
  3. Reject the null when the null is true (Type I error).
  4. Fail to reject the null when the alternative is true (Type II error).
- The seven general steps to presenting hypothesis tests are:
  1. State the parameters of interest
  2. State the null hypothesis
  3. State the alternative hypothesis
  4. State the test statistic and its distribution
  5. State the rejection region
  6. Perform necessary calculations
  7. State your conclusion

2.3 Test statistics

- Knowing the sampling distribution of an estimate allows us to form test statistics and to make inferences about the corresponding population parameter.
- The general form of a test statistic is

$$\frac{\text{point estimate} - E(\text{point estimate})}{\text{standard deviation of point estimate}}, \tag{2.2}$$

  where E(point estimate) is the expected value of our point estimate under H_0.

2.4 Determining the significance of a test statistic

- Once we calculate a test statistic we can talk about how likely the results were due to chance if we know the underlying distribution.
- The "p-value" of our test statistic is a measure of how likely our observed results are just due to chance.
- To get the p-value we must know the distribution from which our test statistic was drawn. Typically this involves knowing the general form of the distribution (t, F, etc.) and any specific parameters (such as the degrees of freedom).
- Once we know the distribution we look up our calculated test statistic in the appropriate table. If the specific value of our calculated test statistic is not in the table we use the p-value associated with the largest test statistic that is smaller than the one we calculated. This gives us a conservative estimate. Similarly, if there is no table with the exact degrees of freedom that we have in our statistic we use the table with largest degrees of freedom that is less than the degrees of freedom associated with our calculated statistic.
- Before calculating our test statistic we establish a "significance level" α. We reject the null hypothesis if we obtain a p-value less than α and fail to reject the null when we obtain a p-value greater than α. The significance level is the probability that we make a Type I error. Many fields use a standard $\alpha = .05$.

2.5 Testing β_1 and β_0

- To test whether there is a linear relationship between two variables we can perform a hypothesis test on the slope parameter in the corresponding simple linear regression model.
- β_0 is the expected value when $X = 0$. If X cannot take on the value 0 then β_0 has no meaningful interpretation.
- When testing β_1:
  - Point estimate = least squares estimate b_1
  - E(point estimate) = β_1 under H_0 (This is typically = 0)
  - Standard deviation = $\sqrt{\sigma^2\{b_1\}}$
- We estimate $\sigma^2\{b_1\}$ using the following equation:

$$s^2\{b_1\} = \frac{\text{MSE from model}}{\sum(X_i - \bar{X})^2} \tag{2.3}$$

- The resulting test statistic

$$t^* = \frac{b_1 - \beta_1}{s\{b_1\}} \tag{2.4}$$

  follows a t distribution with $n - 2$ degrees of freedom.

- To make tests about the intercept β_0 we use
  - Point estimate = least squares estimate b_0
  - E(point estimate) = β_0 under H_0
  - Standard deviation = $\sqrt{\sigma^2\{b_0\}}$
- We estimate $\sigma^2\{b_0\}$ using the following equation:

$$s^2\{b_0\} = \text{MSE}\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right) \tag{2.5}$$

- The resulting test statistic

$$t^* = \frac{b_0 - \beta_0}{s\{b_0\}} \tag{2.6}$$

  follows a t distribution with $n - 2$ degrees of freedom.

2.6 Violations of normality

- If our Y's are truly normally distributed then we can be confident about our estimates of β_0 and β_1.
- If our Y's are approximately normal then our estimates are approximately correct.
- If our Y's are distinctly not normal then our estimates are likely to be incorrect, but they will approach the true parameters as we increase our sample size.

2.7 Confidence intervals

- Point estimates tell us about the central tendency of a distribution while confidence intervals tell us about both the central tendency and the spread of a distribution.
- The general form of a confidence interval is

$$\text{point estimate} \pm \text{(critical value)(standard deviation of point estimate)} \tag{2.7}$$

- We often generate a $1 - \alpha$ confidence interval for β_1:

$$b_1 \pm (t_{1-\frac{\alpha}{2}, n-2}) s\{b_1\}. \tag{2.8}$$

- When constructing confidence intervals it's important to remember to divide your significance level by 2 when determining your critical value. This is because half of your probability is going into each side.
- You can perform significance tests using confidence intervals. If your interval contains the value of the parameter under the null hypothesis you fail to reject the null. If the value under the null hypothesis falls outside of your confidence interval then you reject the null.

2.8 Predicting new observations

- Once we have calculated a regression equation we can predict a new observation at X_h by substituting in this value and calculating

$$\hat{Y}_h = b_0 + b_1 X_h. \tag{2.9}$$

- It inappropriate to predict new values when X_h is outside the range of X's used to build the regression equation. This is called extrapolation.

- We can build several different types of intervals around $\hat{Y}_h$ since we know its sampling distribution. In each case we have $(n-2)$ degrees of freedom and the $(1-\alpha)$ prediction interval takes on the form

$$\hat{Y}_h \pm (t_{1-\frac{\alpha}{2},n-2})s\{\text{pred}\}, \tag{2.10}$$

  where n is the number of observations to build the regression equation and $s\{\text{pred}\}$ is the standard deviation of our prediction.

- The exact value of $s\{\text{pred}\}$ depends on what we are specifically trying to predict. It is made up of the variability we have in our estimate and the variability we have in what we're trying to predict. This last part is different in each case described below.
  - To predict the mean response at X_h we use

$$s\{\text{pred}\} = \sqrt{\text{MSE}\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right)}, \tag{2.11}$$

  - To predict a single new observation we use

$$s\{\text{pred}\} = \sqrt{\text{MSE}\left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right)}, \tag{2.12}$$

  - To predict the mean of m new observations we use

$$s\{\text{pred}\} = \sqrt{\text{MSE}\left(\frac{1}{m} + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right)}, \tag{2.13}$$

  - These equations are related. Equation 2.11 is equal to equation 2.13 where $m = \infty$. Similarly, equation 2.12 is equal to equation 2.13 where $m = 1$.

2.9 ANOVA approach to regression

- Our observed variable Y will always have some variability associated with it. We can break this into the variability related to our predictor and the variability unrelated to our predictor. This is called an "Analysis of Variance."

- SSTO = Total sums of squares = $\sum(Y_i - \bar{Y})^2$ is the total variability in Y. It has $n-1$ degrees of freedom associated with it.

- SSR = Regression sums of squares = $\sum(\hat{Y}_i - \bar{Y})^2$ is the variability in Y accounted for by our regression model. Since we are using one predictor (X) it has 1 degree of freedom associated with it.

- SSE = Error sums of squares = $\sum(Y_i - \hat{Y}_i)^2 = \sum {e_i}^2$ is the variability in Y that is not accounted for by our regression model. It has $n-2$ degrees of freedom associated with it.

- These are all related because

$$\text{SSTO} = \text{SSR} + \text{SSE} \tag{2.14}$$

  and

$$\text{df}_{total} = \text{df}_{model} + \text{df}_{error} \tag{2.15}$$

- The MSR (regression mean square) is equal to the SSR divided by its degrees of freedom. Similarly, the MSE (error mean square) is equal to the SSE divided by its degrees of freedom.

- All of this is summarized in the following ANOVA table:

Source of variation	SS	df	MS
Model (R)	$\sum(\hat{Y}_i - \bar{Y})^2$	1	$\frac{\text{SSR}}{1}$
Error (E)	$\sum(Y_i - \hat{Y}_i)^2$	$n-2$	$\frac{\text{SSE}}{n-2}$
Total	$\sum(Y_i - \bar{Y})^2$	$n-1$	

2.10 The General Linear Test

- We can use the information from the ANOVA table to perform a general linear test of the slope parameter. In this we basically try to see if a "full" model predicts significantly more variation in our dependent variable than a "reduced" model.
- The full model is the model corresponding to your alternative hypothesis while your reduced model is the model corresponding to your null hypothesis.
- The basic procedure is to fit both your full and your reduced model and then record the SSE and DF from each. You then calculate the following test statistic:

$$F^* = \frac{\frac{\text{SSE}_{reduced} - \text{SSE}_{full}}{\text{df}_{reduced} - \text{df}_{full}}}{\frac{\text{SSE}_{full}}{\text{df}_{full}}}, \tag{2.16}$$

  which follows an F distribution with $(\text{df}_{full} - \text{df}_{reduced})$ numerator and df_{full} denominator degrees of freedom.
- When testing the slope in simple linear regression, F^* is equal to $\frac{\text{SSTO}-\text{SSE}}{\text{MSE}} = \frac{\text{MSR}}{\text{MSE}}$.
- $F_{1,n-2,1-\alpha} = (t_{n-2,1-\alpha})^2$.

2.11 Descriptive measures

- These measure the degree of association between X and Y.
- Coefficient of determination: r^2
  - This is the effect of X in reducing the variation in Y.
  - $r^2 = \frac{\text{SSR}}{\text{SSTO}}$
  - $0 \leq r^2 \leq 1$
- Coefficient of correlation: r
  - This is a measure of the linear association between X and Y.
  - $r = \pm\sqrt{r^2}$, where the sign indicates the slope direction
  - $-1 \leq r \leq 1$

2.12 Using SAS

- Something like the following commands typically appear at the top of any SAS program you run:

  ```
  options ls=72;
  title1 'Computer program 1';
  data one;
  infile 'mydata.dat';
     input Y X;
  ```

  The first line formats the output for a letter-sized page. The second line puts a title at the top of each page. The third establishes the name of the SAS data set you are working on as “one”. The fourth line tells the computer that you want to read in some data from a text file called “mydata.dat”. The fifth tells SAS that the first column in the dataset contains the variable “Y” and the second column contains the variable “X”. Each column in the text file is separated by one or more spaces.

- To print the contents of a dataset use the following code:

  ```
  proc print data=one;
     var X Y;
  ```

  where the variables after the **var** statement are the ones you want to see. You can omit the **var** line altogether if you want to see all the variables in the dataset.

- To plot one variable against another use the following code:

  ```
  proc plot data=one;
     plot Y*X = 'A';
  ```

  In this case variable Y will appear on the Y-axis, variable X will appear on the X-axis, and each point will be marked by the letter “A”. If you substitute the following for the second line:

  ```
  plot Y*X;
  ```

  then SAS will mark each point with a character corresponding to how many times that observation is repeated.

  You can have SAS do multiple plots within the same **proc plot** statement by listing multiple combinations on the **plot** line:

  ```
  plot Y*X='A' Y*Z='B' A*B;
  ```

  as long as all of the variables are in the same dataset. Each plot will produce its own graph. If you want them printed on the same axes you must also enter the **/overlay** switch on the plot line:

  ```
  plot Y*X Z*X /overlay;
  ```

- To have SAS generate a least-square regression line predicting Y by X you would use the following code:

  ```
  proc reg data=one;
     model Y = X;
  ```

This the output from this code would provide you with the regression coefficients and an ANOVA table. The coefficients appear at the bottom of the output: b_0 would be located across from "INTERCEPT" and b_1 would be across from "X."

The predicted values and residuals may be obtained by adding the switch **/p** to the model statement as in:

```
model Y = X /p;
```

You can also store the residuals and the predicted values in a new dataset (along with all of the other variables) by adding the following line after the **model** statement:

```
output out=two p=yhat r=resids;
```

where "two" is the name of the new dataset, "yhat" is the name of the variable to contain the predicted values, and "resids" is the name of the variable to contain the residuals.

To have SAS generate prediction intervals for a single new observation at each level of your predictor you add the line

```
print cli;
```

after the **model** statement. To have SAS generate prediction intervals for the mean response at each level of your predictor you add the line

```
print clm;
```

after the **model** statement. You can have both in the same **reg** procedure.

If you want prediction intervals at levels not in your data set the easiest way is to modify your data set so that the levels are present. You add an entry with the value you want in the predictor variable column and then a "." in the response variable column. The entry will not affect how SAS builds the regression equation, but SAS will calculate a predicted value for it.

- To have SAS calculate the correlations between variables you use the following code:

  ```
  proc corr;
    var X Y Z;
  ```

  The result is a matrix containing the correlations of the listed variables with each other.

- You can insert comments into your program that will be ignored by SAS. You place a /* at the beginning and a */ at the end of the comment. When searching for errors you might find it useful to comment out parts of your program and run it one piece at a time.

- Always remember to include a

  ```
  run;
  ```

  statement at the end of your program. Otherwise SAS may not execute all of your commands.

Chapter 3

Diagnostic and Remedial Measures I

3.1 General information about diagnostics

- It is important that we make sure that our data fits the assumptions of the normal error regression model. Violations of the assumptions lessen the validity of our hypothesis tests.
- Typical violations of the simple linear regression model are:
  1. Regression function is not linear
  2. Error terms do not have constant variance
  3. Error terms are not independent
  4. One or more observations are outliers
  5. Error terms are not normally distributed
  6. One or more important predictors have been omitted from the model

3.2 Examining the distribution of the residuals

- We don't want to examine the raw Y_i's because they are dependent on the X_i's. Therefore we examine the e_i's.
- We will often examine the standardized residuals instead because they are easier to interpret. Standardized residuals are calculated by the following formula:

$$e_i{}^* = \frac{e_i}{\sqrt{\text{MSE}}}. \tag{3.1}$$

- To test whether the residuals are normally distributed we can examine a normal probability plot. This is a plot of the residuals against the normal order statistic. Plots from a normal distribution will look approximately like straight lines.

  If your Y_i's are non-normal you should transform them into normal data.
- To test whether the variance is constant across the error terms we plot the residuals against $\hat{Y}_i$. The graph should look like a random scattering.

  If the variance is not constant we should either transform Y or use weighted least squares estimators. If we transform Y we may also have to transform X to keep the same basic relationship.
- To test whether the error terms are independent we plot the residuals against potential variables not included in the model. The graph should look like a random scattering.

  If our terms are not independent we should add the appropriate terms to our model to account for the dependence.

- We should also check for the possibility that the passage of time might have influenced our results. We should therefore plot our residuals versus time to see if there is any dependency.

3.3 Checking for a linear relationship

- To see if there is a linear relationship between X and Y you can examine a graph of Y_i versus X_i. We would expect the data to look roughly linear, though with error. If the graph appears to have any systematic non-linear component we would suspect a non-linear relationship.
- You can also examine a graph of e_i versus X_i. In this case we would expect the graph to look completely random. Any sort of systematic variation would be an indicator of a non-linear relationship.
- If there is evidence of a non-linear relationship and the variance is not constant then we want to transform our Y_i's. If there is evidence of a non-linear relationship but the variance is constant then we want to transform our X_i's.
- The exact nature of the transformation would depend on the observed type of non-linearity. Common transformations are logarithms and square roots.

3.4 Outliers

- To test for outliers we plot X against Y or X against the residuals. You can also check univariate plots of X and Y individually. There should not be any points that look like they are far away from the others.
- We should check any outlying data points to make sure that they are not the result of a typing error. True outliers should be dropped from the analysis.
- Dealing with outliers will be discussed more thoroughly in later chapters.

3.5 Lack of Fit Test

- In general, this tests whether a categorical model fits the data significantly better than a linear model.
- The assumptions of the test are that the Y_i's are independent, normally distributed, and have constant variance.
- In order to do a lack of fit test you need to have repeated observations at one or more levels of X.
- We perform the following hypothesis test:

$$\begin{aligned} H_0 &: Y_i = \beta_0 + \beta_1 X_i \\ H_a &: Y_i \neq \beta_0 + \beta_1 X_i \end{aligned} \tag{3.2}$$

  Our null hypothesis is a linear model while our alternative hypothesis is a categorical model. The categorical model uses the mean response at each level of X to predict Y. This model will always fit the data better than the linear model because the linear model has the additional constraint that its predicted values must all fall on the same line. When performing this test you typically hope to show that the categorical model fits no better than the linear model. This test is a little different than the others we've done because we don't want to reject the null.

- We perform a general linear test. Our full model is

$$Y_{ij} = \mu_j + \epsilon_{ij} \quad \begin{aligned} i &= 1 \dots n_j \\ j &= 1 \dots c \end{aligned} \tag{3.3}$$

where c is the number of different levels of X, n_j is the number of observations at level X_j, and μ_j is the expected value of Y_{ij}.

From our full model we calculate

$$\text{SSE}_{full} = \sum_j \sum_i (Y_{ij} - \hat{\mu}_j)^2, \tag{3.4}$$

where

$$\hat{\mu}_j = \frac{1}{c} \sum_i Y_i = \bar{Y}_j. \tag{3.5}$$

There are $n - c$ degrees of freedom associated with the full model.

- Our reduced model is a simple linear regression, namely that

  $$Y_{ij} = \beta_0 + \beta_1 X_i + \epsilon_{ij} \quad \begin{array}{l} i = 1 \ldots n_j \\ j = 1 \ldots c \end{array} \tag{3.6}$$

  The $\text{SSE}_{reduced}$ is the standard SSE from our model, equal to

  $$\sum_j \sum_i [Y_{ij} - (b_0 + b_1 X_i)]^2. \tag{3.7}$$

  There are $n - 2$ degrees of freedom associated with the reduced model.

- When performing a lack of fit test we often talk about two specific components of our variability. The sum of squares from pure error (SSPE) measures the variability within each level of X. The sum of squares from lack of fit (SSLF) measures the non-linear variability in our data. These are defined as follows:

  $$\text{SSLF} = \text{SSE}_{reduced} - \text{SSE}_{full} = \sum_j \sum_i (\bar{Y}_j - \hat{Y}_{ij})^2 \tag{3.8}$$

  $$\text{SSPE} = \text{SSE}_{full} = \sum_j \sum_i (Y_{ij} - \bar{Y}_j)^2. \tag{3.9}$$

  The SSPE has $n - c$ degrees of freedom associated with it while the SSLF has $c - 2$ degrees of freedom associated with it.

- These components are related because

  $$\text{SSE}_{reduced} = \text{SSPE} + \text{SSLF}. \tag{3.10}$$

  and

  $$\text{df}_{\text{SSE}_{reduced}} = \text{df}_{\text{SSPE}} + \text{df}_{\text{SSLF}}. \tag{3.11}$$

- From this we can construct a more detailed ANOVA table:

Source of variation	SS	df	MS
Model (R)	$\sum\sum(\hat{Y}_{ij} - \bar{Y})^2$	1	$\frac{\text{SSR}}{1}$
Error (E)	$\sum(Y_{ij} - \hat{Y}_{ij})^2$	$n-2$	$\frac{\text{SSE}}{n-2}$
LF	$\sum\sum(\bar{Y}_j - \hat{Y}_{ij})^2$	$c-2$	$\frac{\text{SSLF}}{c-2}$
PE	$\sum\sum(Y_{ij} - \bar{Y}_j)^2$	$n-c$	$\frac{\text{SSPE}}{n-c}$
Total	$\sum(Y_{ij} - \bar{Y})^2$	$n-1$	

- We calculate the following test statistic:

$$F^* = \frac{\frac{\text{SSE}_{reduced} - \text{SSE}_{full}}{\text{df}_{reduced} - \text{df}_{full}}}{\frac{\text{SSE}_{full}}{\text{df}_{full}}} = \frac{\frac{\text{SSE - SSPE}}{c-2}}{\frac{\text{SSPE}}{n-c}} = \frac{\frac{\text{SSLF}}{c-2}}{\frac{\text{SSPE}}{n-c}} \tag{3.12}$$

  which follows an F distribution with $c - 2$ numerator and $n - c$ denominator degrees of freedom.

- Our test statistic is a measure of how much our data deviate from a simple linear model. We must reject our linear model if it is statistically significant.

3.6 Using SAS

- To get boxplots and stem and leaf graphs of a variable you use the **plot** option of **proc univariate**:

```
proc univariate plot;
   var X;
```

- There are two different ways to generate a normal probability plot. One way is to use the **plot** option of **proc univariate**:

```
proc univariate plot;
   var Y;
```

  While this is the easiest method, the plot is small and difficult to read. The following code will generate a full-sized normal probability plot:

```
proc rank data=one normal=blom;
var Y;
ranks ynorm;
proc plot;
plot Y*ynorm;
```

- SAS refers to standardized residuals as "studentized residuals." You can store these in an output data set using the following code:

```
proc reg data=one;
model Y=X;
output out=two student=sresid;
```

- To transform data in SAS all you need to do is create a new variable as a function of the old one, and then perform your analysis with the new variable. You must create the variable in the data step, which means that it must be done ahead of any procedure statements. For example, to do a log base 10 transform on your Y you would use the code:

```
ytrans = log10(Y);
proc reg data=one;
   model ytrans = X;
```

  SAS contains funcions for performing square roots, logarithms, exponentials, and just about any other sort of transformation you might possibly want.

Chapter 4

Simultaneous Inferences and Other Topics

4.1 General information about simultaneous inferences

- When you perform multiple hypothesis tests on the same set of data you inadvertently inflate your chances of getting a type I error. For example, if you perform two tests each at a .05 significance level, your overall chance of having a false positive are actually $(1 - .95^2) = .0975$, which is greater than the .05 we expect.

- We therefore establish a "family confidence coefficient" α, specifying how certain we want to be that none of our results are due to chance. We modify the significance levels of our individual tests so that put together our total probability of getting a type I error is less than this value.

4.2 Joint estimates of β_0 and β_1

- Our tests of β_0 and β_1 are drawn from the same data, so we should present a family confidence coefficient when discussing our results.

- We describe the Bonferroni method, although there are others available. In this method we divide α by the number of tests we're performing to get the significance level for each individual test.

- For example, we would generate the following confidence intervals for β_0 and β_1 using a family confidence coefficient of α:

$$\begin{aligned} \beta_0 &: b_0 \pm (t_{n-2,1-\frac{\alpha}{4}})s\{b_0\} \\ \beta_1 &: b_1 \pm (t_{n-2,1-\frac{\alpha}{4}})s\{b_1\} \end{aligned} \tag{4.1}$$

  While we usually select our critical value at $(1 - \frac{\alpha}{2})$ in confidence intervals (because it is a two-tailed test), we use $(1 - \frac{\alpha}{4})$ here because we are estimating two parameters simultaneously.

4.3 Regression through the origin

- Sometimes we have a theoretically meaningful (0,0) point in our data. For example, the gross profit of a company that produces no product would have to be zero. In this case we can build a regression model with β_0 set to zero:

$$Y_i = \beta_1 X_i + \epsilon_i, \tag{4.2}$$

  where $i = 1 \ldots n$.

- The least squares estimate of β_1 in this model is

$$b_1 = \frac{\sum X_i Y_i}{\sum {X_i}^2}. \tag{4.3}$$

- The residuals from this model are not guaranteed to sum to zero.
- We have one more degree of freedom in this model than in the simple linear regression model because we don't need to estimate β_0. Our estimate of the variance (MSE) is therefore

$$s^2 = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-1}. \tag{4.4}$$

  Similarly, $\text{df}_{error} = n - 1$.

4.4 Measurement error

- Error with respect to our Y_i's is accommodated by our model, so measurement error with regard to the response variable is not problematic as long as it is random. It does, however, reduce the power of our tests.
- Error with respect to our X_i's, however, is problematic. It can cause our error terms to be correlated with our explanatory variables which violates the assumptions of our model. In this case the least-squares estimates are no longer unbiased.

4.5 Inverse prediction

- We have already learned how to predict Y from X. Sometimes, however, we might want to predict X from Y. We can calculate an estimate with the following equation:

$$\hat{X}_h = \frac{Y_h - b_0}{b_1}, \tag{4.5}$$

  which has a variance of

$$s^2\{\text{pred}X\} = \frac{\text{MSE}}{{b_1}^2}\left(1 + \frac{1}{n} + \frac{(\hat{X}_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right). \tag{4.6}$$

- While this can be used for informative purposes, it should not be used to fill in missing data. That requires more complicated procedures.

4.6 Choosing X levels

- The variance of our parameter estimates are all divided by $\sum(X_i - \bar{X})^2$. To get more precise estimates we would want to maximize this. So you want to have a wide range to your X's to have precise parameter estimates.
- It is generally recommended that you use equal spacing between X levels.
- When attempting to predict Y at a single value of X, the variance of our predicted value related to $(X_h - \bar{X})^2$. To get more precise estimate we would want to minimize this. So you want to sample all of your points at X_h if you want a precise estimate of $\hat{Y}_h$.

Chapter 5

Matrix Approach to Simple Linear Regression

5.1 General information about matrices

- A matrix is a collection of numbers organized into rows and columns. Rows go horizontally across while columns go vertically down. The following is an example of a matrix:

$$\underset{3\times 4}{\mathbf{M}} = \begin{pmatrix} 2 & 7 & 6 & 12 \\ 9 & 5 & 1 & 11 \\ 4 & 3 & 8 & 10 \end{pmatrix}. \tag{5.1}$$

  The dimension of the matrix (the numbers in the subscript) describes the number of rows and columns.

- Non-matrix numbers (the type we deal with every day) are called scalars. Matrix variables are generally typed as capital, bold-faced letters, such as $\mathbf{A}$ to distinguish them from scalars. When writing matrix equations by hand, the name of the matrix is usually written over a tilde, such as $\underset{\sim}{\mathrm{A}}$.

- A single value inside the matrix is called an element. Elements are generally referred to by the lowercase letter of the matrix, with subscripts indicating the element's position. For example:

$$\underset{3\times 3}{\mathbf{A}} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}. \tag{5.2}$$

- We use matrices because they allow us to represent complicated calculations with very simple notation. Any moderately advanced discussion of regression typically uses matrix notation.

5.2 Special types of matrices

- A square matrix is a matrix with the same number of rows and columns:

$$\begin{pmatrix} 5 & -2 & 2 \\ 9 & -4 & 3 \\ 11 & 5 & 12 \end{pmatrix}. \tag{5.3}$$

- A column vector is a matrix with exactly one column:

$$\begin{pmatrix} 7 \\ 2 \\ -8 \end{pmatrix}. \tag{5.4}$$

- A row vector is a matrix with exactly one row:

$$\begin{pmatrix} 1 & 3 & 4 \end{pmatrix}. \tag{5.5}$$

- A symmetric matrix is a square matrix that reflects along the diagonal. This means that the element in row i column j is the same as the element in row j column i:

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 5 & 7 \\ 3 & 7 & 6 \end{pmatrix}. \tag{5.6}$$

- A diagonal matrix is a square matrix that has 0's everywhere except on the diagonal:

$$\begin{pmatrix} 22 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -6 \end{pmatrix}. \tag{5.7}$$

- An identity matrix is a square matrix with 1's on the diagonal and 0's everywhere else:

$$\underset{3\times3}{\mathbf{I}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{5.8}$$

  Multiplying a matrix by the identity always gives you back the original matrix. Note that we always use **I** to represent an identity matrix.

- A unity matrix is a square matrix where every element equal to 1:

$$\underset{3\times3}{\mathbf{J}} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \tag{5.9}$$

  We can also have unity row vectors and unity column vectors. We use unity matrices for summing. Note that we always use **J** to represent a unity matrix and **1** to represent a unity vector.

- A zero matrix is a square matrix where every element is 0:

$$\underset{3\times3}{\mathbf{0}} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{5.10}$$

  Note that we always use **0** to represent a zero matrix.

5.3 Basic matrix operations

- Two matrices are considered equal if they are equal in dimension and if all their corresponding elements are equal.

- To transpose a matrix you switch the rows and the columns. For example, the transpose of **M** in equation 5.1 above would be

$$\underset{4\times3}{\mathbf{M}'} = \begin{pmatrix} 2 & 9 & 4 \\ 7 & 5 & 3 \\ 6 & 1 & 8 \\ 12 & 11 & 10 \end{pmatrix}. \tag{5.11}$$

  The transpose of a symmetric matrix is the same as the original matrix.

- To add two matrices you simply add together the corresponding elements of each matrix. You can only add matricies with the same dimension. For example,

$$\begin{pmatrix} 5 & 12 \\ 10 & 18 \end{pmatrix} + \begin{pmatrix} 6 & 9 \\ 4 & 15 \end{pmatrix} = \begin{pmatrix} 11 & 21 \\ 14 & 33 \end{pmatrix}. \tag{5.12}$$

- To subtract one matrix from another you subtract the corresponding element from the second matrix from the element of the first matrix. Like addition, you can only subtract a matrix from one of the same dimension. For example,

$$\begin{pmatrix} 5 & 12 \\ 10 & 18 \end{pmatrix} - \begin{pmatrix} 6 & 9 \\ 4 & 15 \end{pmatrix} = \begin{pmatrix} -1 & 3 \\ 6 & 3 \end{pmatrix}. \tag{5.13}$$

- You can multiply a matrix by a scalar. To obtain the result you just multiply each element in the matrix by the scalar. For example:

$$5\begin{pmatrix} -2 & 5 \\ 11 & 3 \end{pmatrix} = \begin{pmatrix} -10 & 25 \\ 55 & 15 \end{pmatrix}. \tag{5.14}$$

  Similarly, you can factor a scalar out of a matrix.

5.4 Matrix multiplication

- You can multiply two matrices together, but only under certain conditions. The number of columns in the first matrix has to be equal to the number of rows in the second matrix. The resulting matrix will have the same number of rows as the first matrix and the same number of columns as the second matrix. For example:

$$\underset{2\times3}{\mathbf{A}}\,\underset{3\times4}{\mathbf{B}} = \underset{2\times4}{\mathbf{C}}. \tag{5.15}$$

  Basically, the two dimensions on the inside have to be the same to multiply matrices ($3 = 3$) and the resulting matrix has the dimensions on the outside (2×4).

- An important fact to remember is that matrix multiplication is not commutative: $(\mathbf{AB}) \neq (\mathbf{BA})$. You always need to keep track of which matrix is on the left and which is on the right.

- Always multiply from left to right, unless parentheses indicate otherwise.

- Each element of the product matrix is calculated by taking the cross product of a row of the first matrix with a column of the second matrix. Consider the following example:

$$\begin{pmatrix} 3 & 6 & 1 \\ 4 & 10 & 12 \end{pmatrix}\begin{pmatrix} 8 & 11 \\ 7 & 2 \\ 5 & 9 \end{pmatrix} = \begin{pmatrix} 3(8)+6(7)+1(5) & 3(11)+6(2)+1(9) \\ 4(8)+10(7)+12(5) & 4(11)+10(2)+12(9) \end{pmatrix} = \begin{pmatrix} 71 & 54 \\ 162 & 172 \end{pmatrix} \tag{5.16}$$

  Basically, to calculate the element in cell ij of the resulting matrix you multiply the elements in row i of the first matrix with the corresponding elements of column j of the second matrix and then add the products.

5.5 Linear dependence

- Two variables are linearly dependent if one can be written as a linear function of the other. For example, if $x = 4y$ then x and y are linearly dependent. If no linear function exists between two variables then they are linearly independent.

- We are often interested in determining whether the columns of a matrix are linearly dependent. Let $\mathbf{c}$ be the column vectors composing a matrix. If any set of λ's exists (where at least one λ is nonzero) such that
$$\lambda_1 \mathbf{c_1} + \lambda_2 \mathbf{c_2} + \cdots + \lambda_i \mathbf{c_i} = 0, \tag{5.17}$$
(where there are i columns) then we say that the columns are linearly dependent. If equation 5.17 can be satisfied it means that at least one column of the matrix can be written as a linear function of the others. If no set of λ's exists satisfying equation 5.17 then your columns are linearlly independent.

- The rank of a matrix is the maximum number of linearly independent columns. If all the columns in a matrix are linearly independent then we say that the matrix is full rank.

5.6 Inverting a matrix

- If we have a square matrix $\mathbf{S}$, then its inverse $\mathbf{S}^{-1}$ is a matrix such that
$$\mathbf{S}\mathbf{S}^{-1} = \mathbf{I}. \tag{5.18}$$

- The inverse of a matrix always has the same dimensions as the original matrix.

- You cannot calculate the inverse of a matrix that is not full rank.

- There is a simple procedure to calculate the inverse of a 2 × 2 matrix. If
$$\underset{2\times 2}{\mathbf{A}} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \tag{5.19}$$
then
$$\underset{2\times 2}{\mathbf{A}}^{-1} = \frac{1}{ad - cb} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}. \tag{5.20}$$

- The quantity $ad - cb$ is called the determinant. If the matrix is not full rank this is equal to 0.

- There are ways to calculate more complicated inverses by hand, but people typically just have computers do the work.

5.7 Simple linear regression model in matrix terms

- We can represent our simple linear model and its least squares estimates in matrix notation.

- Consider our model
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \tag{5.21}$$
where $i = 1 \ldots n$. We can define the following matrices using different parts of this model:
$$\underset{n\times 1}{\mathbf{Y}} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad \underset{n\times 2}{\mathbf{X}} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} \quad \underset{2\times 1}{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \underset{n\times 1}{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \tag{5.22}$$

- We can write the simple linear regression model in terms of these matrices:
$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times 2}{\mathbf{X}} \underset{2\times 1}{\beta} + \underset{n\times 1}{\epsilon}. \tag{5.23}$$

- The matrix equation for the least squares estimates is
$$\underset{2\times 1}{\mathbf{b}} = (\mathbf{X'X})^{\mathbf{-1}}\mathbf{X'Y}. \tag{5.24}$$

- The matrix equation for the predicted values is

$$\underset{n\times 1}{\hat{\mathbf{Y}}} = \mathbf{Xb} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{5.25}$$

  To simplify this we define the hat matrix $\mathbf{H}$ such that

$$\underset{n\times n}{\mathbf{H}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \tag{5.26}$$

  and

$$\hat{\mathbf{Y}} = \mathbf{HY}. \tag{5.27}$$

  We shall see later that there are some important properties of the hat matrix.

- The matrix equation for the residuals is

$$\underset{n\times 1}{\mathbf{e}} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}. \tag{5.28}$$

5.8 Covariance matrices

- Whenever you are working with more than one random variable at the same time you can consider the covariances between the variables.

- The covariance between two variables is a measure of how much the varibles tend to be high and low at the same time. It is related to the correlation between the variables. The covariance between variables Y_1 and Y_2 would be represented as

$$\sigma\{Y_1, Y_2\} = \sigma\{Y_2, Y_1\}. \tag{5.29}$$

- We can display the covariances between a set of variables with a variance-covariance matrix. This matrix is always symmetric. For example, if we had k different variables we could build the following matrix:

$$\begin{pmatrix} {\sigma_1}^2 & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{12} & {\sigma_2}^2 & \cdots & \sigma_{2k} \\ \vdots & \vdots & & \vdots \\ \sigma_{1k} & \sigma_{2k} & \cdots & {\sigma_k}^2 \end{pmatrix}. \tag{5.30}$$

5.9 Using SAS

- SAS does have an operating system for performing matrix computations called IML. However, we will not need to perform any complex matrix operations for this class so it will not be discussed in detail.

- You can get $\mathbf{X}'\mathbf{X}$, $\mathbf{X}'\mathbf{Y}$, and $\mathbf{Y}'\mathbf{Y}$ when you use **proc reg** by adding a **print xpx** statement:

```
proc reg;
   model Y=X;
   print xpx;
```

  You can get $(\mathbf{X}'\mathbf{X})^{-1}$, the parameter estimates, and the SSE when you use **proc reg** by adding a **print i** statement:

```
proc reg;
   model Y=X;
   print i;
```

  If you want both you can include them in the same **print** statement.

Chapter 6

Multiple Regression I

6.1 General information about multiple regression

- We use multiple regression when we want to relate the variation in our dependent variable to several different independent variables.
- The simple linear regression model (which we've used up to this point) can be done using multiple regression methodology.

6.2 First order model with two independent variables

- The general form of this model is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \tag{6.1}$$

  where $i = 1 \ldots n$ and the ϵ_i's are normally distributed with a mean of 0 and a variance of σ^2.

- We interpret the parameters in this model as follows:
  - ○ β_0 : If X_1 and X_2 have meaningful zero points then β_0 is the mean response at $X_1 = 0$ and $X_2 = 0$. Otherwise there is no useful interpretation.
  - ○ β_1 : This is the change in the mean response per unit increase in X_1 when X_2 is held constant.
  - ○ β_2 : This is the change in the mean response per unit increase in X_2 when X_1 is held constant.
- β_1 and β_2 are called partial regression coefficients.
- If X_1 and X_2 are independent then they are called additive effects.

6.3 The General linear model (GLM)

- This is the generalization of the first-order model with two predictor variables. In this model we can have as many predictors as we like.
- The general form of this model is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i, \tag{6.2}$$

  where $i = 1 \ldots n$ and ϵ_i follows a normal distribution with mean 0 and variance σ^2.

- This model can be used for many different types of regression. For example, you can:

- ○ Have qualitative variables, where the variable codes the observation as being in a particular category. You can then test how this categorization predicts some response variable. This will be discussed in detail in chapter 11.
- ○ Perform polynomial regression, where your response variable is related to some power of your predictor. You could then explain quadratic, cubic, and other such relationships. This will be discussed in detail in chapter 7.
- ○ Consider interactions between variables. In this case one of your predictors is actually the product of two other variables in your model. This lets you see if the effect of one predictor is dependent on the level of another. This will be discussed in detail in chapter 7.

6.4 The matrix form of the GLM

- The GLM can be expressed in matrix terms:

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times p}{\mathbf{X}}\,\underset{p\times 1}{\beta} + \underset{n\times 1}{\epsilon}, \tag{6.3}$$

where

$$\underset{n\times 1}{\mathbf{Y}} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \qquad \underset{n\times p}{\mathbf{X}} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{pmatrix}$$
$$\underset{p\times 1}{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} \qquad \underset{n\times 1}{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}. \tag{6.4}$$

- As in simple linear regression, the least squares estimates may be calculated as

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, \tag{6.5}$$

where

$$\mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{pmatrix}. \tag{6.6}$$

- Using this information you can write an ANOVA table:

Source of variation	SS	df	MS
Model (R)	$\mathbf{b}'\mathbf{X}'\mathbf{Y} - \frac{1}{n}\mathbf{Y}'\mathbf{J}\mathbf{Y}$	$p-1$	$\frac{\text{SSR}}{p-1}$
Error (E)	$\mathbf{e}'\mathbf{e} = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}$	$n-p$	$\frac{\text{SSE}}{n-p}$
Total	$\mathbf{Y}'\mathbf{Y} - \frac{1}{n}\mathbf{Y}'\mathbf{J}\mathbf{Y}$	$n-1$	

6.5 Statistical inferences

- You can test whether there is a significant relationship between your response variable and any of your predictors:

$$\begin{array}{ll} H_0: & \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0 \\ H_a: & \text{at least one } \beta_k \neq 0. \end{array} \tag{6.7}$$

Note that you are not testing β_0 because it is not associated with a predictor variable. You perform a general linear test decide between these hypotheses. You calculate

$$F^* = \frac{\text{MSR}}{\text{MSE}}, \tag{6.8}$$

which follows an F distribution with $p-1$ numerator and $n-p$ denominator degrees of freedom.

- You can calculate a coefficient of multiple determination for your model:

$$R^2 = \frac{\text{SSR}}{\text{SSTO}} = 1 - \frac{\text{SSE}}{\text{SSTO}}. \tag{6.9}$$

This is the proportion of the variation in Y that can be explained by your set of predictors. A large R^2 does not necessarily mean that you have a good model. Adding parameters always increases R^2, but might not increase its significance. Also, your regression function was designed to work with your data set, so your R^2 could be inflated. You should therefore be cautious until you've replicated your findings.

- You can also test each β_k individually:

$$\begin{array}{ll} H_0: & \beta_k = 0 \\ H_a: & \beta_k \neq 0. \end{array} \tag{6.10}$$

To do this you calculate

$$t^* = \frac{b_k - \beta_k}{s\{b_k\}} = \frac{b_k}{s\{b_k\}}, \tag{6.11}$$

which follows a t distribution with $n-p$ degrees of freedom. You can obtain $s\{b_k\}$ by taking the square root of the kth diagonal element of

$$\text{MSE}\,(\mathbf{X}'\,\mathbf{X})^{-1}, \tag{6.12}$$

which is the variance/covariance matrix of your model $\mathbf{s^2}\{\mathbf{b}\}$.

Remember to use a family confidence coefficient if you decide to perform several of these test simultaneously.

- You can also build prediction intervals for new observations. You first must define the vector

$$\mathbf{X_h} = \begin{pmatrix} X_{h1} \\ X_{h2} \\ \vdots \\ X_{h,p-1} \end{pmatrix}, \tag{6.13}$$

where $X_{h1} \ldots X_{h,p-1}$ is the set of values at which you want to predict Y. You may then calculate your predicted value

$$\hat{Y}_h = \mathbf{X_h}'\,\mathbf{b}\,. \tag{6.14}$$

It is inappropriate to select combinations of X's that are very different from those used to build the regression model.

As in simple linear regression, we can build several different types of intervals around $\hat{Y}_h$ depending on exactly what we want to predict. In each case we have $(n-p)$ degrees of freedom and the $(1-\alpha)$ prediction interval takes on the form

$$\hat{Y}_h \pm (t_{1-\frac{\alpha}{2},n-p})s\{\text{pred}\}, \tag{6.15}$$

where n is the number of observations to build the regression equation, and $s\{\text{pred}\}$ depends on exactly what you want to predict.

  - ○ To predict the mean response at $\mathbf{X_h}$ you use

$$s\{\text{pred}\} = \sqrt{\text{MSE}\,(\mathbf{X_h}'(\mathbf{X}'\,\mathbf{X})^{-1}\,\mathbf{X_h})}. \tag{6.16}$$

- ◦ To predict a single new observation you use

$$s\{\text{pred}\} = \sqrt{\text{MSE}\,(1 + \mathbf{X_h}'(\mathbf{X}'\,\mathbf{X})^{-1}\,\mathbf{X_h})}. \tag{6.17}$$

- ◦ To predict the mean of m new observations you use

$$s\{\text{pred}\} = \sqrt{\text{MSE}\,\left(\frac{1}{m} + \mathbf{X_h}'(\mathbf{X}'\,\mathbf{X})^{-1}\,\mathbf{X_h}\right)}. \tag{6.18}$$

6.6 Diagnostics and remedial measures

- Locating outliers takes more effort in multiple regression. An observation could be an outlier because it has an unusually high or low value on a predictor. But it also could be an outlier if it simply represents an unusual combination of the predictor values, even if they all are within the normal range.

  You should therefore examine plots of the predictors against each other in addition to scatterplots and normal probability plots to locate outliers.

- To assess the appropriateness of the multiple regression model and the consistancy of variance you should plot the residuals against the fitted values. You should also plot the residuals against each predictor variable seperately.

- In addition to other important variables, you should also check to see if you should include any interactions between your predictors in your model.

6.7 Using SAS

- Running a multiple regression is SAS is very easy. The code looks just like that for simple linear regression, but you have more than one variable after the equals sign:

```
proc reg
   model Y = X1 X2 X3;
```

- Predicting new observations in multiple regression also works the same way. You just insert an observation in your data set (with a missing value for Y) and then add a **print cli** or **print clm** line after your **proc reg** statement.

Chapter 7

Multiple Regression II

7.1 General information about extra sums of squares

- Extra sums of squares measure the marginal reduction in the error sum of squares when one or more independent variables are added to a regression model.
- Extra sums of squares are also referred to as partial sums of squares.
- You can also view the extra sums of squares as the marginal increase in regression sum of squares when one or more independent variables are added to a regression model.
- In general, we use extra sums of squares to determine whether specific variables are making substantial contributions to our model.

7.2 Extra sums of squares notation

- The regression sum of squares due to a single factor X_1 is denoted by SSR(X_1). The corresponding error sum of squares is denoted by SSE(X_1). This is the SSR from the regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \tag{7.1}$$

  where $i = 1 \ldots n$.

- The sums of squares due to multiple factors is represented in a similar way. For example, the regression sum of squares due to factors X_1, X_2, and X_3 is denoted by SSR(X_1, X_2, X_3) and the error sum of squares by SSE(X_1, X_2, X_3). This is the SSR from the regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i, \tag{7.2}$$

  where $i = 1 \ldots n$.

- When we have more than one predictor variable we can talk about the extra sum of sqares associated with a factor or group of factors. In model 7.2 above, the extra sum of squares uniquely associated with X_1 is calculated as

$$\begin{aligned} \text{SSR}(X_1|X_2, X_3) &= \text{SSE}(X_2, X_3) - \text{SSE}(X_1, X_2, X_3) \\ &= \text{SSR}(X_1, X_2, X_3) - \text{SSR}(X_2, X_3). \end{aligned} \tag{7.3}$$

  As mentioned above, SSR($X_1|X_2, X_3$) is a measure of how much factor X_1 contributes to the model above and beyond the contribution of factors X_2 and X_3.

- The unique sums of squares associated with each factor are often referred to as Type II sums of squares.

7.3 Decomposing SSR into extra sums of squares

- Just as we can partition SSTO into SSR and SSE, we can partition SSR into the extra sums of squares associated with each of our variables. For example, the SSR from model 7.2 above can be broken down as follows:
$$\text{SSR}(X_1, X_2, X_3) = \text{SSR}(X_1) + \text{SSR}(X_2|X_1) + \text{SSR}(X_3|X_1, X_2). \tag{7.4}$$
These components are often referred to as the Type I sums of squares.
- We can represent the relationships between our sums of squares in an ANOVA table. For example, the table corresponding to model 7.2 would be:

Source of variation		SS	df	MS
Model (R)		$\text{SSR}(X_1, X_2, X_3)$	$(p-1) = 3$	$\text{MSR}(X_1, X_2, X_3)$
	X_1	$\text{SSR}(X_1)$	1	$\text{MSR}(X_1)$
	$X_2\|X_1$	$\text{SSR}(X_2\|X_1)$	1	$\text{MSR}(X_2\|X_1)$
	$X_3\|X_1, X_2$	$\text{SSR}(X_3\|X_1, X_2)$	1	$\text{MSR}(X_3\|X_1, X_2)$
Error		$\text{SSE}(X_1, X_2, X_3)$	$n-4$	$\text{MSE}(X_1, X_2, X_3)$
Total		SSTO	$n-1$	

- Note that putting our parameters in a different order would give us different Type I SS.

7.4 Comparing Type I and Type II sums of squares

- To review, to calculate the Type I SS you consider the variability accounted for by each predictor, conditioned on the the predictors that occur earlier in the model:
$$\begin{aligned} X_1 &: \text{SSR}(X_1) \\ X_2 &: \text{SSR}(X_2|X_1) \\ X_3 &: \text{SSR}(X_3|X_1, X_2). \end{aligned} \tag{7.5}$$
To calculate the Type II SS you consider the unique variablility accounted for by each predictor:
$$\begin{aligned} X_1 &: \text{SSR}(X_1|X_2, X_3) \\ X_2 &: \text{SSR}(X_2|X_1, X_3) \\ X_3 &: \text{SSR}(X_3|X_1, X_2). \end{aligned} \tag{7.6}$$
- The Type I SS are guaranteed to sum up to $\text{SSR}(X_1, X_2, X_3)$, while the Type II SS are not guaranteed to sum up to anything. We say that the Type I SS partition the SSR.
- The Type II SS are guaranteed to be the same no matter how you enter your variables in your model, while the Type I SS may change if you use different orders.
- Testing the significance of parameters is the same as testing the Type II SS.

7.5 Testing parameters using extra sums of squares

- To test whether a single $\beta_k = 0$ we can perform a general linear model test using the extra sums of squares associated with that parameter.

- For example, if we wanted to test parameter β_2 from model 7.2 our hypotheses would be

$$\begin{array}{ll} H_0: & \beta_2 = 0 \\ H_a: & \beta_2 \neq 0. \end{array} \tag{7.7}$$

  The full model is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i. \tag{7.8}$$

  Our reduced model is the same except without the variable we want to test:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_3 Xi3 + \epsilon_i. \tag{7.9}$$

  We then calculate the following test statistic:

$$F^* = \frac{\frac{\text{SSE}_{reduced} - \text{SSE}_{full}}{\text{df}_{reduced} - \text{df}_{full}}}{\frac{\text{SSE}_{full}}{\text{df}_{full}}} = \frac{\frac{\text{SSE}(X_1, X_3) - \text{SSE}(X_1, X_2, X_3)}{1}}{\frac{\text{SSE}(X_1, X_2, X_3)}{n-4}}, \tag{7.10}$$

  which follows an F distribution with 1 numerator and $n - 4$ denominator degrees of freedom. Testing parameters in this way is referred to as performing partial F-tests.

- We can also test the contribution of several of your variables simultaneously. For example, if we wanted to simultaneously test the contribution of X_1 and X_3 in model 7.2 above our hypotheses would be

$$\begin{array}{ll} H_0: & \beta_1 = \beta_3 = 0 \\ H_a: & \text{at least one} \neq 0. \end{array} \tag{7.11}$$

  The full model is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i \tag{7.12}$$

  and the reduced model is

$$Y_i = \beta_0 + \beta_2 X_{i2} + \epsilon_i. \tag{7.13}$$

  We then calculate the following test statistic:

$$F^* = \frac{\frac{\text{SSE}(X_2) - \text{SSE}(X_1, X_2, X_3)}{2}}{\frac{\text{SSE}(X_1, X_2, X_3)}{n-4}}, \tag{7.14}$$

  which follows an F distribution with 2 numerator and $n - 4$ denominator degrees of freedom.

- We can perform other tests as well. For example, if we wanted to test whether $\beta_1 = \beta_2$ our hypotheses would be

$$\begin{array}{ll} H_0: & \beta_1 = \beta_2 \\ H_a: & \beta_1 \neq \beta_2. \end{array} \tag{7.15}$$

  The full model would be

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i \tag{7.16}$$

  and the reduced model would be

$$Y_i = \beta_0 + \beta_c (X_{i1} + X_{i2}) + \beta_3 X_{i3} + \epsilon_i. \tag{7.17}$$

  To build the reduced model we could create a new variable

$$X_{ic} = X_{i1} + X_{i2} \tag{7.18}$$

  and then build a model predicting Y from X_c and X_3. Once we obtain the SSR from these models we perform a general linear test to make inferences.

7.6 Coefficients of partial determination

- Just as the coefficient of multiple determination R^2 measures the proportionate reduction in the variation of Y associated with your full model, coefficients of partial determination r^2 measure the proportionate reduction of error associated with specific parameters.

- The coefficient of partial determination is defined as the marginal reduction in error obtained by including a parameter in your model. For example, the coefficient of partial determination for variable X_1 in model 7.2 above is

$$r^2{}_{Y1.23} = \frac{\mathrm{SSR}(X_1|X_2, X_3)}{\mathrm{SSE}(X_2, X_3)}. \tag{7.19}$$

- You can take the square root of the coefficient of partial determination to get the coefficient of partial correlation. It is given the same sign as that of the corresponding regression coefficient.

7.7 Standardized multiple regression

- In multiple regression we sometimes work with standardized variables to control for roundoff errors and to permit comparisons between the regression coefficients.

- The first step to building a standardized model is to perform a correlation transformation on our variables as follows:

$$\begin{aligned} Y'_i &= \tfrac{1}{\sqrt{n-1}}\left(\tfrac{Y_i-\bar{Y}}{s\{Y\}}\right) \\ X'_{ik} &= \tfrac{1}{\sqrt{n-1}}\left(\tfrac{X_{ik}-\bar{X}_k}{s\{X_k\}}\right) \end{aligned} \tag{7.20}$$

where $k = 1 \ldots p-1$, and the standard deviations are defined as

$$\begin{aligned} s\{Y\} &= \sqrt{\tfrac{\sum(Y_i-\bar{Y})^2}{n-1}} \\ s\{X_k\} &= \sqrt{\tfrac{\sum(X_{ik}-\bar{X}_k)^2}{n-1}}. \end{aligned} \tag{7.21}$$

- We use these variables to generate a standardized multiple regression model:

$$Y'_i = \beta'_1 X'_{i1} + \beta'_2 X'_{i2} + \cdots + \beta'_{p-1} X'_{i,p-1} + \epsilon'_i, \tag{7.22}$$

where $i = 1 \ldots n$. Notice that we do not have an intercept parameter – our transformation always leads to an intercept of zero.

- The parameters from the standardized model are related to the original parameters as follows:

$$\begin{aligned} \beta_k &= \left(\tfrac{s\{Y\}}{s\{X_k\}}\right)\beta'_k \\ \beta_0 &= \bar{Y} - \beta_1\bar{X}_1 - \beta_2\bar{X}_2 - \cdots - \beta_{p-1}\bar{X}_{p-1}, \end{aligned} \tag{7.23}$$

where $k = 1 \ldots p-1$.

- An interesting result is that the $\mathbf{X'X}$ matrix of the transformed variables is the correlation matrix of the X variables $\mathbf{r_{XX}}$. Similarly, the $\mathbf{X'Y}$ matrix is the correlation matrix of the X's with Y $\mathbf{r_{YX}}$.

7.8 Multicollinearity

- You have multicollinearity when your predictor variables are correlated. This is bad because it reduces the precision of your parameter estimates. It is difficult to draw conclusions about a model with significant multicollinearity.
- While it affects the estimation of your individual parameters, multicollinearity does not affect inferences regarding the full model.
- If X_1 and X_2 are uncorrelated then $\mathrm{SSR}(X_1|X_2) = \mathrm{SSR}(X_1)$ and $\mathrm{SSR}(X_2|X_1) = \mathrm{SSR}(X_2)$. This also extends to more than two variables. This means that if you have no multicollinearity among your variables then your Type I SS are exactly equal to your Type II SS and are not dependent on the order of the variables in your model.
- The primary solution to multicollinearity is to drop variables from your model that are highly correlated with others. Using centered variables can help reduce some types of multicollinearity. Otherwise, there is not much else you can do.

7.9 Polynomial models

- Polynomial regression is actually just a type of general linear model where you take your terms to higher powers, such as squares and cubes.
- You need to center your predictors (subtract off the mean) before performing a polynomial regression. Using standardized variables works just as well.
- A simple type is the second order model with one predictor:

  $$Y_i = \beta_0 + \beta_1 X_i + \beta_{11} {X^2}_i + \epsilon_i, \tag{7.24}$$

  where $i = \ldots n$.
- Note that you always include the lower-order factors in your model. It is inappropriate to have a model with X^k as a factor unless you also include factors $X^1 \ldots X^{k-1}$.
- You can also have polynomial models with more than one independent variable, such as

  $$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_{11} {X^2}_{i1} + \beta_2 X_{i2} + \beta_{22} {X^2}_{i2} + \epsilon_i, \tag{7.25}$$

  where $i = 1 \ldots n$.
- A lower order predictor should always be entered in your model before a higher order predictor from the same independent variable.
- You can tell how high of an order function you should use by the shape of the data. If your data appears to reverse p times, you should use a function whose highest order parameter is at the $p - 1$ power.

7.10 Interaction models

- If the effect of one predictor variable appears to depend on the level of another variable in your model you should include an interaction term. You can also have interactions with more than two variables.
- You construct an interaction term by multiplying the variables involved in the interaction together.
- A model of the interaction between two variables would be

  $$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i, \tag{7.26}$$

  where $i = 1 \ldots n$.

- In this model the mean change in Y with a unit increase in X_1 when X_2 is held constant is $\beta_1 + \beta_3 X_2$. Similarly, the mean change in Y with a unit increase in X_2 when X_1 is held constant is $\beta_2 + \beta_3 X_1$.

- It is inappropriate to include an interaction term in a model without also including all of the variables involved in the interaction individually.

- It is possible to have interactions with higher-order terms, such as squares and cubes. However, these types of interactions are difficult to interpret.

7.11 Using SAS

- From here on out we will occasionally use the SAS **proc glm** to perform our analyses instead of **proc reg**. The basic syntax for **proc glm** is exactly the same as **proc reg**, but it has some additional capabilities.

- To get SAS to output the Type I or Type II SS you can just add the **ss1** or **ss2** toggle to your model statement in either **proc reg** or **proc glm**:

```
proc reg data=one;
   model Y=X1 X2 X3 / ss1 ss2;
```

- SAS does not have a procedure to perform a correlation transformation. However, you can use **proc standard** to standardize your variables and then divide each by $\sqrt{n-1}$. The following code would perform a correlation transformation on a dependent and three independent variables:

```
data one;
/* You would have code here to read in the variables Y, X1, X2, and X3 */
proc standard data=one mean=0 std=1 out=two;
   var Y X1 X2 X3;

data three;
set two;
/* You would have to specify the exact value of n-1 in these statements */
YC = Y / sqrt(n-1);
X1C = X1 / sqrt(n-1);
X2C = X2 / sqrt(n-1);
X3C = X3 / sqrt(n-1);

proc reg data=two;
   model YC = X1C X2C X3C;
```

You can have SAS test the significance of sets of parameters using the **partial** statement under **proc reg**. The following code would test whether both β_1 and β_3 are equal to zero:

```
proc reg data=one;
   model Y = X1 X2 X3;
   partial: test X1=0, X3=0;
```

Chapter 8

Building the Regression Model I: Selection of Predictor Variables

8.1 General information about model building

- A good model is one that accurately reflects the data uses a relatively small number of meaningful independent variables to predict a relatively large proportion of the variability in the dependent variable.

- There are four basic steps to building a good regression model:

  1. Collect the data appropriately.
  2. Decide on the essential predictor variables, including interactions and polynomial terms.
  3. Refine your model. Make sure that you account for all the appropriate factors without having extra variables in your model.
  4. Test your model. Your model was built for your specific data set so you shouldn't put too much faith in it until you've validated it with new data.

8.2 "All possible models" methods

- In this method you systematically test every possible combination of variables. You should be careful not to consider invalid models (like interaction models without main effects).

- For each model you calculate some measure of "fitness" and choose the one that fits best. The common measures are:

  R^2 criterion You calculate $R^2 = \frac{\text{SSR}}{\text{SST0}}$ for each model. Models with more parameters will have larger R^2's, so you need to take this into account. Generally you will see a point after which adding more variables makes little difference in the R^2.

  Adjusted R^2 criterion You calculate ${R^2}_a = 1 - \frac{\text{MSE}\,(n-1)}{\text{SSTO}}$ for each model. This puts models with different numbers of parameters on an equal footing, so you can just look for higher ${R^2}_a$'s.

  Mallows C_p criterion You calculate $C_p = \frac{\text{SSE}}{\text{MSE}} - (n - 2p)$ for each model. You look for a model whose C_p is close to the number of parameters used in the model.

  PRESS$_p$ criterion You calculate $\text{PRESS}_p = \sum(Y_i - \hat{Y}_{i(i)})^2$ for each model, where $\hat{Y}_{i(i)}$ is the predicted value of Y_i when the ith observation is not used to build the regression model. Models with small PRESS$_p$ values are better than those with large values.

8.3 Forward stepwise regression

- In forward stepwise regression we start with a simple model and gradually add parameters to it until we are unable to make any significant improvement. If at any time during the process a previously included variable is no longer predictive it is removed from the model.

- Stepwise regression is generally used when there are too many variables to use the "all possible models" procedure.

- The steps for performing forward stepwise regression are:

  1. Fit the simple linear regression models for each of the $p-1$ predictor variables. Select the one that accounts for the most variance as a candidate for the model.
  2. Test all the two-factor regression models that include the candidate chosen in step 1. Examine the two-factor model that accounts for the most variance. If the second factor makes a significant contribution we add it to our model. If not, we stop with the model we currently have.
  3. Check to see if you need to drop any of your included variables. Compute partial F-tests for each:

     $$F_a{}^* = \frac{\text{MSR}(X_a|X_b)}{\text{MSE}(X_a, X_b)} = \left(\frac{b_a}{s\{b_A\}}\right)^2. \tag{8.1}$$

     Remove the variable if this statistic is not significant.
  4. Check for more candidates. If you add new variables you must check to see if you need to drop older variables. Cycle between steps 2 and 3 until you have no more candidates.

8.4 Using SAS

- Model building procedures are available under **proc reg**. The general format is:

  ```
  proc reg data=one;
     model Y = X1 X2 X3 X4 /selection= *keyword* best=*number*;
  ```

  keyword defines the model selection procedure that you want to perform and is either **forward, backward, stepwise, rsquare, adjrsq,** or **cp**. To perform forward stepwise regression you use the **stepwise** keyword. ***number*** specifies the maximum number of subset models you want SAS to print.

  When using forward selection you can control the p-values at which variables enter and leave the model using the **slentry** and **slstay** keywords:

  ```
  proc reg data=one;
     model Y = X1 X2 X3 X4 /selection=forward slentry=.05 slstay=.05;
  ```

- SAS will not perform model selection using the PRESS$_p$ statistic, but you can store the PRESS values for a single model in an output data set using the following code:

  ```
  proc reg data=one;
     model Y = X1 X2 X3 X4;
     output out=two press=press;
  proc print data=two;
  ```

  The PRESS$_p$ for the model would be the sum of these values.

- SAS will not perform "all possible models" procedures when you have more than ten predictors.

Chapter 9

Building the Regression Model II: Diagnostics

9.1 Checking the model's predictors

- To make sure you are not leaving out any important predictor variables you should perform partial regression plots on any variables you've excluded.
- Consider the case where Y is your dependent variable, X_1 and X_2 are included in your model, and X_3 was excluded from your model. To create a partial regression plot for X_3 you:
  1. Regress Y on the terms in your model, X_1 and X_2.
  2. Calculate the residuals $= e_i(Y|X_1, X_2)$.
  3. Regress X_3 on the terms in your model, X_1 and X_2.
  4. Calculate the residuals $= e_i(X_3|X_1, X_2)$.
  5. Plot $e_i(Y|X_1, X_2)$ versus $e_i(X_3|X_1, X_2)$. If you see any pattern it indicates that you would benefit by adding some function of X_3 to your model. The nature of the pattern tells you whether the term should be linear, quadratic, etc.
- You should also check the partial regression plots of variables included in your model to see if you should add higher polynomial terms.

9.2 Improved methods for calculating residuals

- We often perform transformations on our residuals make it easier to identify outliers.
- *Studentized residuals* represent the magnitude of each residual relative to the estimated standard deviation. The studentized residual is calculated as

  $$r_i = \frac{e_i}{s\{e_i\}}, \tag{9.1}$$

  where

  $$s\{e_i\} = \sqrt{\text{MSE}\,(1 - h_{ii})}. \tag{9.2}$$

  The studentized residuals follow a t distribution with $n - p$ degrees of freedom.
- *Deleted residuals* measure how accurate your model is in predicting each observation when the observation under consideration is not used to build the regression model. The deleted residual is calculated as

  $$d_i = Y_i - \hat{Y}_{i(i)}, \tag{9.3}$$

  where $\hat{Y}_{i(i)}$ is the predicted value for observation i from the regression model built without using i.

- *Studentized deleted residuals* combine both the above methods, giving you the residual relative to its standard deviation when the observation from which the residual is calculated is not used to build the regression model. The studentized deleted residual is calculated as

$$t_i = \frac{d_i}{s\{d_i\}} = \frac{e_i}{\sqrt{\text{MSE}_{(i)}\,(1 - h_{i(i)})}}, \tag{9.4}$$

  where $\text{MSE}_{(i)}$ refers to the mean square error from the model built without observation i. The studentized deleted residuals follow a t distribution with $n - p - 1$ degrees of freedom.

9.3 Checking for outliers using leverage

- Recall that

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \tag{9.5}$$

  and

$$\hat{\mathbf{Y}} = \mathbf{HY}\,. \tag{9.6}$$

- We can examine the $\mathbf{H}$ matrix to look for outliers. For a single residual e_i,

$$\sigma^2\{e_i\} = \sigma^2(1 - h_{ii}), \tag{9.7}$$

  where h_{ii} is the ith diagonal element of the $\mathbf{H}$ matrix.

- The value h_{ii} is a measure of the distance between the predictor values of the ith observation and the mean predictor values and is called that observation's leverage.

- Two properties of the leverage values are that

$$0 \leq h_{ii} \leq 1 \tag{9.8}$$

  and

$$\sum h_{ii} = p. \tag{9.9}$$

- If an observation has a large leverage then we know that it is very different from the average. If

$$h_{ii} > (\frac{2p}{n} = 2\bar{h}) \tag{9.10}$$

  then observation i can be considered an outlier.

9.4 Identifying influential cases

- After we identify an outlying observation we need to determine if it is influential. That is, if it had undue influence on our parameter estimates.

- To detect influential cases you can examine:

  DFFITS - measures the influence that observation i has on the fitted value $\hat{Y}_i$. An observation is influential if

$$\begin{array}{ll} |\text{DFFITS}| > 1 & \text{when } n \leq 30 \\ |\text{DFFITS}| > 2\sqrt{\frac{p}{n}} & \text{when } n > 30. \end{array} \tag{9.11}$$

  Cook's D - measures the influence that observation i has on all of the fitted values. Cook's D follows an F distribution with p numerator and $n - p$ denominator degrees of freedom. A significant D indicates that the observation is influential.

  DFBETAS - measures the influence that observation i has on each regression coefficient. There is a DFBETA for each coefficient. An observation is influential if

$$\begin{array}{ll} |\text{DFBETA}_k| > 1 & \text{when} n \leq 30 \\ |\text{DFBETA}_k| > \frac{2}{\sqrt{n}} & \text{when} n > 30, \end{array} \tag{9.12}$$

  where $k = 1 \ldots p - 1$.

9.5 Dealing with outliers

- When you locate outliers you should first check to make sure that the data was entered correctly.
- After establishing that you have true outliers you should determine if they are influential or not.
- If your outliers are not influential then you don't need to worry about correcting for them. If they are influential you have to decide whether you want to delete them from your data. You might also want to consider if adding another predictor might be appropriate.
- If you don't want to delete the outliers (because either there are too many of them or your data set is too small) you can try transforming one or more of your variables. You can also try using least absolute deviations estimators instead of least squares estimators.

9.6 Multicollinearity

- There are a number of "symptoms" of multicollinearity that you may notice when working with regression:
  - ○ Your regression coefficients change significantly when you add or remove predictors.
  - ○ Established predictors of an effect turn up nonsignificant.
  - ○ The regression coefficient has a sign opposite to the known correlation between the predictor and the response variable.
  - ○ Your model is highly significant but none of your parameter estimates are significant.
- To formally diagnose multicollinearity we use Variance Inflation Factors (VIFs). These estimate how much of the variance in the estimated regression coefficients is caused by multicollinearity.
- For each coefficient β_k you calculate

  $$\mathrm{VIF}_k = (1 - {R_k}^2)^{-1}, \tag{9.13}$$

  where ${R_k}^2$ is the coefficient of multiple determination when X_k is regressed on the other $p-2$ predictors. If the largest VIF ≥ 10 then multicollinearity is a problem in your model.
- We talked about the effects of multicollinearity in chapter 7.

9.7 Using SAS

- To have SAS output the studentized residuals to a data set you use the **student** keyword in the **output** line. Similarly, to have SAS output the studentized deleted residuals you use the **rstudent** keyword. The following code would produce a data set called **two** which would have the studentized residuals under the variable **stud** and the studentized deleted residuals under the variable **studdel**.

  ```
  proc reg data=one;
     model Y=X1 X2 X3;
     output out=two student=stud rstudent=studdel;
  ```

- You can have SAS output the leverage values (h_{ii}) to a data set using the **h** keyword on the **output** line. The following code would create a data set called **two** with the leverage values in a variable called **lev**.

  ```
  proc reg data=one;
     model Y=X1 X2 X3;
     output out=two h=lev;
  ```

- To get the DFFITS and DFBETAS for your observations you add the **influence** switch to your **model** statement. To get Cook's D, however, you need to use the **cookd** keyword of the **output** statement and print it out separately. The following code would produce all three influence measures:

  ```
  proc reg data=one;
    model Y=X1 X2 X3 / influence;
    output out=two cookd=D;

  proc print data=two;
  ```

- To get SAS to calculate VIF's for your variables you need only add the **vif** switch to your model statement:

  ```
  proc reg data=one;
    model Y=X1 X2 X3 / vif;
  ```

Chapter 11

Qualitative Predictor Variables

11.1 General information about qualitative variables

- So far we have only dealt with quantitative predictor variables, where the variable is some sort of numerical measurement. In this chapter we learn how to use qualitative variables, where refer to how items are categorized.
- When we have multiple groups it is better to model them with a qualitative variable than to build several different regression lines. Using a qualitative variable provides a more accurate estimate of the MSE, plus you get more precise parameter estimates because you're using your entire sample.
- When you build a model you can have:
  - All quantitative variables. This is simple or multiple regression.
  - All qualitative variables. This is analysis of variance (ANOVA).
  - A mixture of quantitative and qualitative variables. This is analysis of covariance (ANCOVA).

11.2 Indicator variables

- We represent a qualitative predictor in our regression model with one or more indicator variables.
- Consider a qualitative variable with three levels, A, B, and C. We would create two indicator variables, X_1 and X_2, to represent this categorization. We define these for each observation in the following way:

$$X_1 = \begin{cases} 1 & \text{if in category A} \\ 0 & \text{if not in category A} \end{cases} \qquad X_2 = \begin{cases} 1 & \text{if in category B} \\ 0 & \text{if not in category B} \end{cases} \tag{11.1}$$

- We don't need an indicator variable for the third category because we know that if it's not in the first two it has to be in the third. If you include a third indicator variable your $\mathbf{X}$ matrix will not be full rank.
- In general, you use a number of indicator variables to represent a factor equal to one less than the number of different factor levels.
- The significance of a qualitative predictor is usually examined using the General Linear Test.

11.3 Using SAS

- To analyze qualitative variables using **proc reg** you must manually build your indicator variables in the data step of your program.

To define interaction terms with indicator variables you create a new set of indicators by multiplying together the variables used to represent the factors you want to interact. For example, say you wanted to interact the qualitative variable A with three levels (coded by indicator variables A_1 and A_2) with qualitative varible B with three levels (coded by indicator variables B_1 and B_2). The interaction between A and B would be coded by four indicator variables: A_1B_1, A_1B_2, A_2B_1, and A_2B_2.

- To analyze qualitative variables using **proc glm** you can either build your indicator variables in the data step or you can use the **class** statement.

  For example, if you wanted to predict college gpa (**Y**) by the high school that a student attended (coded by different values of **X**) you could use the following code:

  ```
  proc glm data=one;
    class X;
    model Y=X;
  ```

  You can define interaction terms with qualitative variables declared in the **class** statement the same way you define interactions between quantitative variables:

  ```
  proc glm data=one;
    class X1 X2;
    model Y=X1 X2 X1*X2;
  ```

Chapter 16

Single Factor ANOVA

16.1 General information about ANOVAs

- ANOVA differs from regression in two ways:

  1. The predictors may be qualitative.
  2. No assumptions are made about the nature of the relationship between the dependent variable and any qualitative variables.

- In ANOVA we refer to predictor variables as "factors." The different categories a composing a factor are called the "levels" of that factor.

- We can have either experimental factors, where the levels are assigned, or classification factors, where the levels are observed.

- We have the following assumptions when performing fixed-factor ANOVAs:

  1. For each factor level there is a probability distribution of responses that is normally distributed.
  2. Each probability distribution within a factor has the same variance.
  3. The observations at each factor level are randomly drawn from the probability distribution associated with that factor and are independent of each other.

16.2 Cell means model

- The cell means model is a way of representing a single-factor ANOVA. Its general form is

  $$Y_{ij} = \mu_i + \epsilon_{ij}, \tag{16.1}$$

  where $i = 1 \ldots r$, $j = 1 \ldots n_i$, r is the number of levels in the factor, and n_i is the number of observations in factor level i.

- This model assumes that the response variable has a different mean at each level of our factor.

- We often consider the following calculations with regard to the cell means model:

  - $n_t = \sum n_i$
  - $Y_{i.} = \sum_j Y_{ij}$
  - $\hat{Y}_{ij} = \frac{Y_{i.}}{n_i} = \bar{Y}_{i.}$
  - $\bar{Y}_{..} = \frac{\sum\sum Y_{ij}}{n_t}$
  - $e_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \bar{Y}_i$

16.3 Partitioning variance in ANOVA

- The deviation in Y may be broken down as follows:

$$\underset{\text{Total deviation}}{(Y_{ij} - \bar{Y}_{..})} = \underset{\text{Deviation caused by the treatment}}{(\bar{Y}_{i.} - \bar{Y}_{..})} + \underset{\text{Individual variation}}{(Y_{ij} - \bar{Y}_{i.})} \tag{16.2}$$

- This corresponds to the following ANOVA table:

Source of variation	SS	df
SSTR	$\sum n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2$	$r-1$
SSE	$\sum\sum(Y_{ij} - \bar{Y}_{i.})^2$	$n_t - r$
SSTO	$\sum\sum(Y_{ij} - \bar{Y}_{..})^2$	$n_t - 1$

- In ANOVA we assume that the variances within each factor level are the same. We refer to this as the expected mean square error:

$$\text{E(MSE)} = \sigma^2. \tag{16.3}$$

Consider the variance we would get if we considered the entire sample. Not only would we have variability due to error, we would also have the variability associated with the different levels of our factor. This would be the expected treatment mean square:

$$\text{E(MSTR)} = \sigma^2 + \frac{\sum n_i(\mu_i - \mu_.)}{r-1}. \tag{16.4}$$

- If the MSE and the MSTR are the same then we can infer that the means at each level of the factor are the same. We can use this to test whether our factor is creating meaningful divisions. To test

$$\begin{array}{rl} H_0: & \mu_1 = \mu_2 = \ldots = \mu_r \\ H_a: & \text{not all means equal} \end{array} \tag{16.5}$$

we can calculate the following statistic:

$$F^* = \frac{\text{MSTR}}{\text{MSE}}, \tag{16.6}$$

which follows an F distribution with $r-1$ numerator and $n_t - r$ denominator degrees of freedom.

16.4 Regression approach to ANOVA

- Consider the cell means model above:

$$Y_{ij} = \mu_i + \epsilon_{ij}. \tag{16.7}$$

Just as we did in partitioning variance, we can write this equation in terms of treatment deviations from the grand mean:

$$Y_{ij} = \mu_. + (\mu_i - \mu_.) + \epsilon_{ij}. \tag{16.8}$$

Let $\tau_i = \mu_i - \mu_.$ = treatment i's deviation from grand mean. We can then rewrite the cell means model as

$$Y_{ij} = \mu_. + \tau_i + \epsilon_{ij}. \tag{16.9}$$

- Note that the treatment deviations are dependent because $\sum \tau_i = 0$. This means that we can write the last treatment deviation as

$$\tau_r = -\tau_1 - \tau_2 - \ldots - \tau_{r-1}. \tag{16.10}$$

- Using these we can build the factor effects model:

$$Y_i = \mu_. + \tau_1 X_{i1} + \tau_2 X_{i2} + \ldots + \tau_{r-1} X_{i,r-1} + \epsilon_i, \tag{16.11}$$

  where

$$X_{ij} = \begin{cases} 1 & \text{if observation } i \text{ is a member of group } j \\ -1 & \text{if observation } i \text{ is a member of group } r \\ 0 & \text{otherwise} \end{cases} \tag{16.12}$$

  and $i = 1 \ldots n_t$. Setting up your X's in this way is called effect coding.

16.5 Using SAS

- If you use **proc glm** to regress a response variable on a categorical factor, you can use the **means** statement to have SAS report the mean of the response variable and each level of the factor:

```
proc glm data=one;
   class X;
   model Y=X;
   means X;
```

- We have already learned how to perform regression with qualitative variables using **proc reg** and **proc glm**. We can also use **proc anova**, which works in basically the same way as **proc glm**. It has some additional options for examining differences between factor level means.

Chapter 17

Analysis of Factor Level Effects

17.1 General information about factor level effects

- Using basic ANOVA techniques we can determine if at least one level of a categorical variable is different from the others. However, they cannot tell us which levels differ.
- This chapter presents two basic methods to examine factor level effects. One method allows us to test hypotheses related to specifc patterns among the factor level means. The other compares the means of all the different factor levels to determine which are significantly different.

17.2 Graphical methods

- To determine which factor levels are different we can examine a line plot. In a line plot we draw a horizontal line representing different levels of our response variable. We then draw dots to represent the means at each level of our categorical variable. We then look for a pattern of differences.
- To simply determine whether there are differences between any of our factor levels we can examine a normal probability plot. A significant departure from normality indicates that our factor levels are different.

17.3 Estimation of factor level effects

- The factor level mean μ_i is estimated by $\bar{Y}_{i.}$. To test

  $$\begin{array}{ll} H_0: & \mu_i = c \\ H_a: & \mu_i \neq c. \end{array} \tag{17.1}$$

  you calculate the statistic

  $$t^* = \frac{\bar{Y}_{i.} - c}{s\{Y_{i.}\}}, \tag{17.2}$$

  which follows a t distribution with $n_t - r$ degrees of freedom. The standard deviation of $Y_{i.}$ is calculated as

  $$s\{Y_{i.}\} = \sqrt{\frac{\sigma^2}{n_i}} \approx \sqrt{\frac{\text{MSE}}{n_i}}. \tag{17.3}$$

- You can estimate the difference between two factor level means $D = \mu_i - \mu_{i'}$ by calculating

  $$\hat{D} = \bar{Y}_{i.} - \bar{Y}_{i'.}. \tag{17.4}$$

To test

$$\begin{array}{ll} H_0: & D = c \\ H_a: & D \neq c \end{array} \tag{17.5}$$

you calculate the statistic

$$t^* = \frac{\hat{D} - c}{s\{D\}}, \tag{17.6}$$

which follows a t distribution with $n_t - r$ degrees of freedom. The standard deviation of D is calculated as

$$s\{D\} = \sqrt{\sigma^2 \left(\frac{1}{n_i} + \frac{1}{n_{i'}} \right)} \approx \sqrt{\text{MSE} \left(\frac{1}{n_i} + \frac{1}{n_{i'}} \right)}. \tag{17.7}$$

- You can also calculate linear contrasts. A contrast is a linear combination of the factor level means $L = \sum c_i \mu_i$, where the coefficients c_i always sum to zero. You estimate the contrast by

$$\hat{L} = \sum c_i \bar{Y}_{i.}. \tag{17.8}$$

To test

$$\begin{array}{ll} H_0: & L = L_c \\ H_a: & L \neq L_c \end{array} \tag{17.9}$$

you calculate the statistic

$$t^* = \frac{\hat{L} - L_c}{s\{L\}}, \tag{17.10}$$

which follows a t distribution with $n_t - r$ degrees of freedom. The standard deviation of L is calculated as

$$s\{L\} = \sqrt{\sigma^2 \sum \frac{{c_i}^2}{n_i}} \approx \sqrt{\text{MSE} \sum \frac{{c_i}^2}{n_i}}. \tag{17.11}$$

17.4 Multiple pairwise comparisons

- Often we want to test the means of all of our levels to see which are significantly different from each other. To do this we must imploy special multiple comparisons procedures.
- Tukey's method, based on confidence limits, is one of the most popular. To perform a Tukey test you:
  1. Calculate all possible comparisons D between the factor level means. For each pair of i and i' (where $i \neq i'$) you calculate the test statistic

     $$q^* = \sqrt{2} t^* = \sqrt{2} \left[\frac{\hat{D}}{s\{D\}} \right] = \sqrt{2} \left[\frac{\bar{Y}_i - \bar{Y}_{i'}}{\sqrt{\text{MSE} \left(\frac{1}{n_i} + \frac{1}{n_{i'}} \right)}} \right], \tag{17.12}$$

     which follows a q distribution (also called the studentized range distribution) with r numberator and $n_t - r$ denominator degress of freedom. The p-values for this distribution are found in table B9 of your textbook.
  2. Order the factor level means on a line.
  3. Underline the pairs of means that are not significant.
- The textbook also shows how to perform Scheffe's and Bonferroni's multiple comparisons tests.

17.5 Using SAS

- You can have SAS perform linear contrasts in **proc glm** using the **contrast** statement. You specify a label for the contrast, the categorical variable you're basing the contrast on, and the coefficients for the contrasts. For example, the following code would test whether the average of the first two groups is the same as the average of the last two groups:

  ```
  proc glm data=one;
     class X;
     model Y=X;
     contrast 'First two vs last two' X .5 .5 -.5 -.5;
  ```

- You can have SAS calculate Tukey's, Bonferroni's, or Scheffe's multiple comparisons groupings by using options on the **means** statement in either **proc glm** or **proc anova**. The following code would have SAS output all three:

  ```
  proc anova data=one;
     class X;
     model Y=X;
     means X /tukey bon scheffe;
  ```