

Regression and Analysis of Variance

Karen Keating

Kansas State University

Acknowledgements

Many of the examples and explanations provided in this text were provided by Dr. John Boyer, Dr. James Higgins and Dr. Nora Bello, all of the Statistics Department at Kansas State University. We greatly appreciate their willingness to share their expertise.

Karen Keating, 2019

Contents

Chapter 1: Simple Linear Regression	1
Section 1.1. Introduction	1
1.1.1. Scatterplots	2
1.1.2. Simple linear regression model and least squares	6
1.1.3. Line of “best” fit	7
Section 1.2. Model Assumptions	11
1.2.1. The assumptions	12
1.2.2. Minimum variance, unbiased estimators	13
Section 1.3. Inference	16
1.3.1. Point estimates	16
1.3.2. Mean and variance of the estimators.....	18
1.3.3. Confidence intervals for the slope and intercept	19
1.3.4. Hypothesis tests.....	21
1.3.5. Prediction and estimation.....	24
1.3.6. Summary.....	28
Section 1.4. Software and Diagnostic Plots	29
1.4.1. SAS code for the (Traffic, Lead) example	30
1.4.2. SAS output for the (Traffic, Lead) example	32
1.4.3. Using the SAS output to assess model assumptions	38
1.4.4. Summary.....	43
Section 1.5. NASA Rocket Propellant Example	43
Section 1.6. ANOVA Table, F and t tests	50
1.6.1. Partitioning the total sum of squares	50
1.6.2. The ANOVA table	52
1.6.3. ANOVA F test.....	54
1.6.4. Some additional considerations.....	57
Section 1.7. Model Diagnostics & Transformations.....	60
1.7.1. Scatterplots	61
1.7.2. Normal probability plots	61
1.7.3. Residual plots	63
1.7.4. Detecting nonconstant variance	65
1.7.5. Goodness of fit.....	68

1.7.6. Transformations.....	69
1.7.7. Summary.....	78
Section 1.8. Correlation Analysis	79
1.8.1. T test for correlation	80
1.8.2. Fisher's Z test for correlation	84
1.8.3. An application: consistency of judges.....	85
1.8.4. Summary.....	89
Chapter 2: Multiple Linear Regression	91
Section 2.1. Introduction	91
2.1.1. Multiple regression with two predictors	91
2.1.2. Least squares estimation	92
2.1.3. Correlation and scatterplot matrix	94
2.1.4. 3D scatterplots.....	96
2.1.5. Summary.....	97
Section 2.2. Body Fat Example.....	98
2.2.1. Examine the correlations	99
2.2.2. Check the assumptions	101
2.2.3. ANOVA table for multiple regression.....	102
2.2.4. Parameter Estimates table.....	104
2.2.5. Estimating the response	107
2.2.6. Summary.....	109
Section 2.3. More than Two Predictors	109
2.3.1. Multicollinearity and variance inflation.....	111
2.3.2. Criteria for comparing models	116
2.3.3. Compare models for the body fat data.....	117
2.3.4. Summary.....	120
Section 2.4. General Linear Regression Model.....	121
2.4.1. Response surface	121
2.4.2. Some terminology.....	122
2.4.3. "All models are wrong..."	123
2.4.4. Example: A second-order model.....	123
2.4.5. Nested model F test.....	126
2.4.6. SAS programming notes.....	128
2.4.7. Summary.....	130

Section 2.5. Qualitative Predictors	131
2.5.1. Indicator ('dummy') variables	131
2.5.2. Example: headache drugs	132
2.5.3. Additive model.....	133
2.5.4. Interaction model	134
2.5.5. SAS programming notes.....	135
2.5.6. Fit models to the headache data	135
2.5.7. Interpret the SAS output.....	146
2.5.8. Summary.....	150
Section 2.6. Influence and Outliers.....	151
2.6.1. Leverage.....	151
2.6.2. Outliers.....	152
2.6.3. Influence	153
2.6.4. Summary.....	156
Chapter 3: Model Building.....	157
Section 3.1. Introduction	157
3.1.1. Considerations	158
3.1.2. Variable Selection	159
3.1.3. Criteria for Model Selection.....	159
3.1.4. Summary.....	163
Section 3.2. Procedures for Model Selection.....	164
3.2.1. Automated Search Methods.....	164
3.2.2. Comparison of Methods	166
Section 3.3. Example using SENIC data.....	168
3.3.1. Fullest possible model.....	170
3.3.2. Forward selection	171
3.3.3. Backward elimination	174
3.3.4. Stepwise	175
3.3.5. Other model selection criteria	176
3.3.6. Choose the final model	178
Section 3.4. Prediction Models	181
3.4.1. Deleted Residuals.....	182
3.4.2. PRESS Statistic.....	182
3.4.3. K-fold Cross Validation.....	184

3.4.4. Mean Square for Prediction.....	184
Chapter 4: One-Way ANOVA	187
Section 4.1. Principles of Experimental Design.....	188
4.1.1. Experimental error.....	189
4.1.2. Randomization	190
4.1.3. Replication	191
4.1.4. Examples	191
4.1.5. Causation vs. association	192
4.1.6. Summary.....	193
Section 4.2. Single Factor Studies	193
4.2.1. Example: Effect of caffeine	194
4.2.2. Within- and between- group variability.....	195
4.2.3. Conduct the hypothesis test	200
Section 4.3. Linear Models.....	201
4.3.1. Cell means model.....	202
4.3.2. Effects model	205
4.3.3. SAS code for the effects model.....	206
4.3.4. SAS code and output for the caffeine data.....	208
4.3.5. Steps to interpret the SAS output.....	213
Section 4.4. Model Diagnostics.....	216
4.4.1. Violation of independence.....	217
4.4.2. Violation of normality	218
4.4.3. Violation of equal variance	219
4.4.4. SAS code for the caffeine example	220
4.4.5. Example: Check assumptions for a different dataset	221
4.4.6. Other transformations and back-transformations	223
4.4.7. SAS code for the simulated example	224
Section 4.5. Multiple Comparisons	225
4.5.1. Which means are different?	226
4.5.2. Example: Inflation of the Type I error rate	227
4.5.3. Controlling the Type I error rate	229
4.5.4. Comparison of methods.....	233
4.5.5. SAS code for multiple comparisons	234
Section 4.6. Contrasts	241

4.6.1. Contrast coefficients and tests	241
4.6.2. Linear and quadratic trends.....	244
4.6.3. Other polynomial contrasts	247
4.6.4. SAS code for contrasts	249
4.6.5. Examples of contrasts	251
4.6.6. Summary.....	252
Section 4.7. Power and Sample Size	253
4.7.1. Criteria for determining the sample size	253
4.7.2. Power of the ANOVA F test.....	256
Chapter 5: Two-Way ANOVA	261
Section 5.1. Definitions and models	261
5.1.1. Fabric data example.....	262
5.1.2. Notation for two-way ANOVA.....	264
5.1.3. Population main effects and interactions.....	265
5.1.4. Models for two-way ANOVA.....	267
5.1.5. Interactions in plain English	268
Section 5.2. Hypotheses for two-way ANOVA	270
5.2.1. Hypotheses for main effects.....	270
5.2.2. Hypotheses for interaction effects	271
5.2.3. An Agronomy Example.....	272
5.2.4. Interaction plots.....	275
Section 5.3. ANOVA Table and F-tests	276
5.3.1. Point estimates	276
5.3.2. Partitioning the total variation	277
5.3.3. Fabric data, re-visited	279
5.3.4. Nested model F test, re-visited.....	281
Section 5.4. t tests and contrasts.....	282
5.4.1. Confidence intervals for means	283
5.4.2. Hypothesis tests for the difference of two means	285
5.4.3. Contrasts	285
5.4.4. Summary.....	287
Section 5.5. SAS statements for two-way ANOVA	288
5.5.1. Contrasts in two-way ANOVA	298
Section 5.6. Examples of two-way ANOVA	300

5.6.1. Steel springs data	301
5.6.2. Preservative data	307
5.6.3. Average daily gain data	311
5.6.4. Summary	315
Chapter 6: Generalizations	317
Section 6.1. Three-Way ANOVA.....	317
6.1.1. Water heater example	318
6.1.2. Generic 3-way ANOVA table	320
6.1.3. SAS code for water heater data	321
6.1.4. Alternate analysis for water heater data	326
6.1.5. Summary	328
Section 6.2. Randomization and Blocking.....	331
6.2.1. Completely randomized design (CRD)	332
6.2.2. Randomized complete block (RCB) design	333
6.2.3. Analysis of the RCB design	336
6.2.4. Analysis of the agronomy data	337
6.2.5. Other cases of blocking.....	340
6.2.6. Example questions	341
6.2.7. Answers to example questions	343
Section 6.3. A Random Effects Model.....	347
6.3.1. Example: Corn chip data	348
6.3.2. SAS code for random effects.....	350
6.3.3. Example questions	352
6.3.4. Answers to example questions	353
Section 6.4. Mixed Effects Models.....	354
6.4.1. The mixed effects model.....	355
6.4.2. Example: Patient recovery times	356

List of Figures

Figure 1.1. Scatterplot for (Traffic, Lead) Data	2
Figure 1.2. Example scatterplot for a deterministic relationship	3
Figure 1.3. Positive and negative linear relationships	5
Figure 1.4. Some residuals for the (Traffic, Lead) example	8
Figure 1.5. Properties of estimators	14
Figure 1.6. Multiple confidence intervals	20
Figure 1.7. Normal probability plot for (Traffic, Lead).....	39
Figure 1.8. A variety of normal probability plots	39
Figure 1.9. Residual plot for (Traffic, Lead).....	41
Figure 1.10. A variety of residual plots	42
Figure 1.11. Partitioning the sum of squares.....	51
Figure 1.12. An F distribution.....	55
Figure 1.13. A nonlinear relationship.....	59
Figure 1.14. QQ plot and histogram #1.....	61
Figure 1.15. QQ plot and histogram #2.....	62
Figure 1.16. QQ plot and histogram #3.....	62
Figure 1.17 QQ plot and histogram #4.....	62
Figure 1.18 Suggested transformations based on scatterplot curves	69
Figure 1.19. Scatterplot of (Time, pH).....	70
Figure 1.20. Residual plot, original data	71
Figure 1.21. Goodness of fit, original data.....	71
Figure 1.22. Residual plot, using $1/X$	72
Figure 1.23. Goodness of fit, using $1/X$	72
Figure 1.24. Residual plot, using $\log(X)$	72
Figure 1.25. Goodness of fit, using $\log(X)$	72
Figure 1.26. Compare to Figure 1.23	73
Figure 1.27. Compare to Figure 1.25	73
Figure 1.28. Scatterplot for Example 2	74
Figure 1.29. Diagnostic plots for Model 1	74
Figure 1.30. Diagnostic plots for Model 2	75

Figure 1.31. Diagnostic plots for Model 3.....	75
Figure 1.32. Diagnostic plots for Model 4.....	76
Figure 1.33. Diagnostic plots for Model 5.....	76
Figure 1.34. Examples of correlation	79
Figure 1.35. Nonlinear association.....	80
Figure 1.36. A bivariate normal distribution	81
Figure 1.37. SAS output for PROC CORR	83
Figure 1.38. Line graph of five judges' scores.....	89
Figure 2.1. A regression plane.....	92
Figure 2.2. A scatterplot matrix	95
Figure 2.3. A 3D scatterplot	96
Figure 2.4. A regression plane.....	96
Figure 2.5. The regression plane, rotated.....	96
Figure 2.6. Residuals in 3D	97
Figure 2.7. Scatterplot matrix for body fat data	101
Figure 2.8. Diagnostic plots for the body fat data	102
Figure 2.9. Diagnostic plots for three-variable body fat model.....	110
Figure 2.10. Scatterplot matrix for the three-variable body fat model.....	113
Figure 2.11. Diagnostic plots for five candidate models.....	119
Figure 2.12. A variety of response surfaces	122
Figure 2.13. Diagnostic plots for simulated data	124
Figure 2.14. Partial residual plots	124
Figure 2.15. Partial residual plots with highlighted quadratic patterns	125
Figure 2.16. Diagnostic plots for second-order model	125
Figure 2.17. A point with high leverage	152
Figure 2.18. A potential outlier	152
Figure 2.19. An influential point	153
Figure 2.20. Graph for Cook's D	155
Figure 3.1. Which car is "best"?.....	157
Figure 4.1. Two populations for a t test.....	194
Figure 4.2. Multiple populations for an ANOVA F test	194
Figure 4.3. Scatterplot of caffeine data	195

Figure 4.4. Within-group variability for caffeine example.....	196
Figure 4.5. Between-group variability for caffeine example	197
Figure 4.6. Comparing large and small within-group variability.....	198
Figure 4.7. Scatterplot of simulated data	221
Figure 4.8. Diagnostic plots for simulated data	221
Figure 4.9. Diagnostic plots for the transformed data.....	222
Figure 4.10. Linear and quadratic trends.....	245
Figure 4.11. Data and graph for bean yields.....	246
Figure 4.12. Power curves for potato chip example	259
Figure 5.1. Mean profile plot for fabric data	263
Figure 5.2. Interaction plots for agronomy example	275
Figure 5.3. Interaction plot for the fabric data	297
Figure 5.4. Diagnostic plots for steel springs data.....	301
Figure 5.5. Interaction plot for steel springs data	303
Figure 5.6. Interaction plot for preservative data	308
Figure 5.7. Diagnostic plots for ADG data.....	312
Figure 5.8. Interaction plot for the ADG data	313
Figure 6.1. Result of LINES option for water heater data	327
Figure 6.2. Diagram of agricultural field divided into 12 plots	331
Figure 6.3. A non-random assignment.....	332
Figure 6.4. A “bad” spot in the field affects some treatments more than others.....	332
Figure 6.5. One possible random assignment.....	333
Figure 6.6. Diagram of blocking for agronomy study.....	334
Figure 6.7. Randomization for a block design.....	335
Figure 6.8. Hypothetical data for RCB agronomy example.....	338
Figure 6.9. Bags and batches of corn chips.....	348

Chapter 1: Simple Linear Regression

Section 1.1. Introduction

Regression analysis allows us to describe the relationship between two or more numeric variables. The response variable Y is the quantity that is being affected and the predictor variable X is the the variable that affects Y . To begin the discussion, we consider only one predictor variable. This is called simple linear regression. Later in the course we will consider multiple predictor variables, and this is called multiple linear regression. When there is only one predictor variable, the dataset will contain one column for X and one column for Y . This is sometimes called bivariate data (“bi” means two and “variate” mean variable).

For example, consider research conducted by an environmentalist that relates the concentration of pollutants (Y) to the traffic volume of a nearby highway (X). He hopes that the amount of traffic will help explain the pollutants, so the predictor variable (X) is traffic and the response variable (Y) is pollutants. To perform the regression analysis, he first needs to collect some data, so he selects several sites along the highway. At each site, he counted the traffic volume and measure the pollutants in nearby trees. The resulting data is shown in Table 1.1. The values for Traffic are in thousands of vehicles in a 24-hour period and. The pollutant he measured is the concentration of lead in nearby tree bark (measured in micrograms of lead per gram of bark).

Table 1.1 illustrates a typical layout for a regression dataset. The rows represent the sites, which are generically referred to as “observations”. The columns represent the variables, that is, the quantities that were measured at each site. Often, but not always, the first column consists of an identifier for each observation, which is used as a cross-reference to other information that is not part of the dataset. The identifier column is not part of the data we will analyze; it is strictly for identification purposes.

Site	Traffic	Lead
1	8.1	227
2	8.3	312
3	12.1	362
4	13.2	521
5	16.5	640
6	17.5	539
7	19.2	728
8	24.8	945
9	24.1	738
10	26.1	759
11	33.6	1263

Table 1.1. Bivariate Data

1.1.1. Scatterplots

When dealing with bivariate data, it is customary to treat the data as (X, Y) pairs. We can plot the data to visualize the relationship between X and Y . This type of plot is called a scatterplot. The scatterplot for the (Traffic, Lead) data is shown in Figure 1.1.

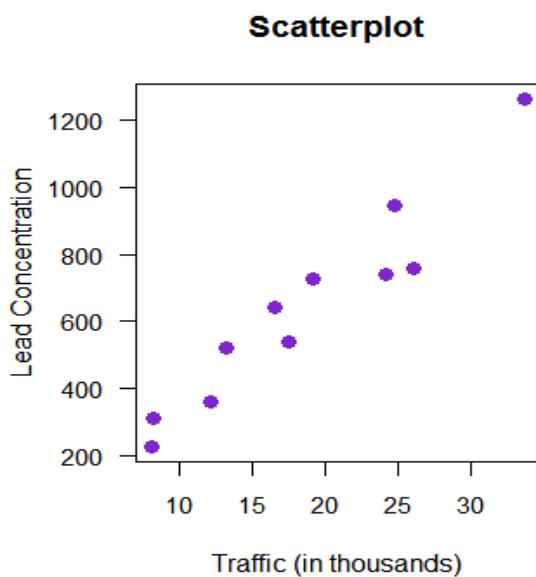


Figure 1.1. Scatterplot for (Traffic, Lead) Data

We can use the scatterplot to assess the relationship between X and Y. In particular, we are concerned with these characteristics.

- Is the relationship stochastic or deterministic?
- What is the shape of the relationship?
- What is the direction of the relationship?
- What is the strength of the relationship?

Stochastic vs. deterministic

In the Traffic/Lead example, the relationship is stochastic. In a deterministic relationship, each value of X produces exactly one value for Y, and this value for Y can be calculated with 100% accuracy. For example, suppose that C is the temperature in degrees Celsius and F is the temperature in degree Fahrenheit. If these two temperatures are measured at the same location at the same time, then there is a known relationship between C and F. This relationship can be expressed as either $C = (F - 32) * 5/9$... or ... $F = (C * 9/5) + 32$. If we recorded only the degrees Fahrenheit (at a particular location and time), then we could calculate the corresponding value for degrees Celsius. There would be no ambiguity in the value for C, in other words C could be calculated exactly. An example of a deterministic scatterplot is shown in Figure 1.2.

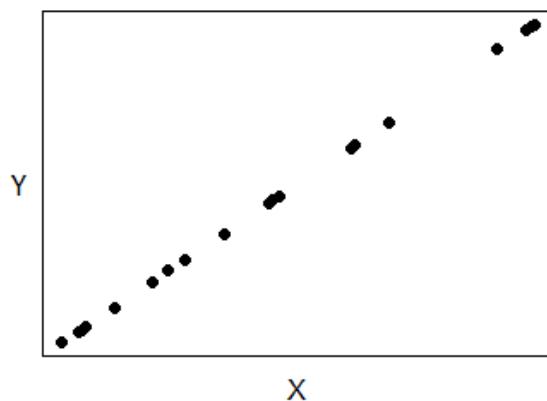


Figure 1.2. Example scatterplot for a deterministic relationship

If X and Y have a stochastic relationship, then each value for X has the potential to produce multiple values for Y. The variation in Y is sometimes called “noise” or “error”, but it simply represents the possibility that factors other than X can affect the value for Y. In the Traffic-Lead example, there are many other things (besides traffic) that can affect the concentration of lead. Other factors include

- The species of tree. Some species of trees may absorb lead more readily than other species.
- The distance between the tree and the highway. Trees that are closer to the highway are more likely to have higher concentrations of lead.
- The age of the tree. Younger trees may be more susceptible to lead.
- Atmospheric conditions. Prevailing wind patterns may affect the atmospheric lead concentration, which in turn could affect the concentration in the tree.
- Location characteristics. If some locations are adjacent to a manufacturing facility or if they are in a low-lying area, this could affect the lead concentration.
- Measurement “error”. The manner in which the bark was extracted from the tree, and slight changes in the chemical processes needed to process the bark, might produce different values for the lead concentration.

The measurement error, in conjunction with any of the other factors that have a minor affect on Y, will be automatically included in the simple linear regression analysis. If there are any factors that significantly affect the value for Y, then these factors should be included as part of the data collection process and they should also be included in the regression analysis. This would create multiple X variables in the regression analysis, and we will discuss that later.

For the topics we will cover in this course, we assume the relationship is stochastic. If the relationship is deterministic, statistical methods are not required to analyze the data.

Shape of the relationship

For simple linear regression, we want the overall shape in the scatterplot to be a straight line. There can be some variation (i.e., some “wiggle”), but the overall pattern should be a straight line. If the pattern is curved, we may be able to make some modifications in order to apply a simple linear regression analysis. We will discuss some options later in the course. If there is no pattern at all in the scatterplot (so that it looks like you dropped a bunch of marbles), then regression analysis is not an appropriate method for analyzing this dataset.

Direction of the relationship

If the shape of the relationship is linear, then we also want to consider the direction of the relationship. As X gets larger, what happens to Y? Does Y get bigger or does Y get smaller?

When larger values of X are associated with larger values for Y, then we say the relationship is positive (or increasing). On the scatterplot, most of the points will be in the upper right and lower left corners, with very few points in the opposing quadrants. An example of this is shown in Figure 1.3(a).

When larger values of X are associated with smaller values for Y, then we say the relationship is decreasing (or negative). On the scatterplot, most of the points will be in the lower right and in the upper left, with very few points in the opposing quadrants. An example of this is shown in Figure 1.3(b).

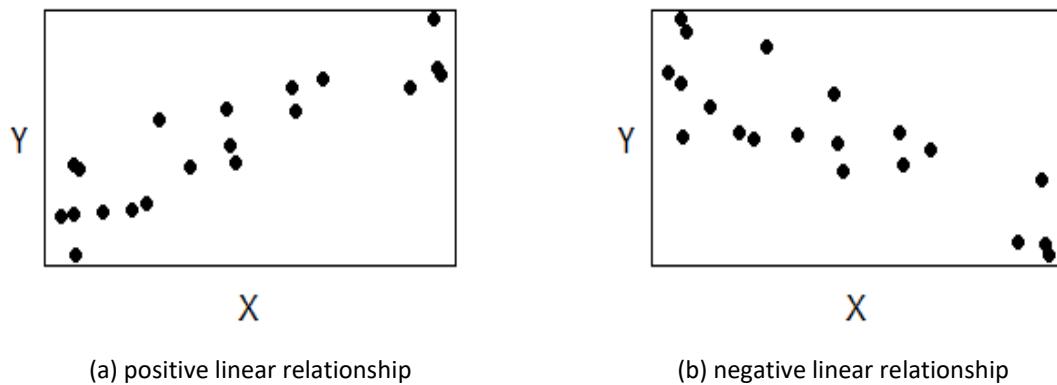


Figure 1.3. Positive and negative linear relationships

Strength of the relationship

If the shape of the relationship is linear, then we are also concerned with the strength of the relationship. If X and Y are strongly associated, then there will be only small amount of variation in Y. In other words, if we know the value for X we have a pretty good idea what the value for Y will be. We will not know the value for Y exactly, because this is a stochastic (not a deterministic) relationship. On the scatterplot, the (X, Y) pairs will form a fairly tight band in the graph. If you take a pencil or a ribbon (or any other straight, narrow object) and place it on the graph, the object will cover most of the points on the graph.

In contrast, if X and Y are weakly associated, then knowing the value for X does not provide much information regarding the value for Y. On the scatterplot, the points will still form a straight-line pattern (because this is a linear relationship), but the points will be more widespread than they would be if the relationship was stronger. You will still be able to take a straight object (like a ribbon) and place it over most of the points, but the ribbon will need to be wider for a weaker relationship, and narrower for a stronger relationship.

Now let us return to the (Traffic, Lead) example, and evaluate the scatterplot of the data in Figure 1.1. The relationship appears to be stochastic (not deterministic). It also appears to be linear (not curved) and fairly strong. All of these characteristics indicate that a simple linear regression analysis should be appropriate for this dataset. The next step is to develop a model (i.e., an equation) that will quantify the relationship between Traffic and Lead.

1.1.2. Simple linear regression model and least squares

Every regression analysis starts with a model. This is an equation that relates X and Y, and incorporates the stochastic variation. A simple linear regression model looks very similar to the equation of a straight line ($y = a + bx$), but it also needs to include the stochastic variation.

A simple linear regression model is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (1.1)$$

The terms in this model are defined as follows

Y_i is the observed value for the response variable for the i th observation

X_i is the observed value for predictor variable for the i th observation

β_0 is the population intercept

β_1 is the population slope

ε_i is the random “error” for the i th row of the data

The subscript “ i ” refers to the i^{th} row in the dataset. Note that β_0 and β_1 do not have the subscript “ i ”.

This means there is one β_0 and one β_1 for all the (X, Y) pairs in the data. β_0 and β_1 are called population parameters. We do not know the values for these two quantities, but we will use the results of the regression analysis to estimate their values. Since X, Y and ε each have the subscript “ i ”, they will have different values for different rows in the dataset.

The values for X_i and Y_i are in the dataset, but the value for ε_i is not known. Although this is called the “error” term, in reality it is not an error but instead it captures the effect of the other factors that have a (hopefully minor) effect on the response. After we perform a regression analysis, we will have an estimate for each of the ε_i and these are called the residuals. The residual measures how far “off” the model is from the observed value of the response.

1.1.3. Line of “best” fit

The main objective of simple linear regression analysis is to determine the values for the intercept and slope so that the straight line goes through the “middle” of all the (X, Y) pairs. There are many different ways to define “middle”, but we will use the method of least squares. This method has many nice mathematical and statistical properties, and is the customary method for performing regression.

The method of least squares is based on the principle of minimizing the total amount of deviation between the regression line and the points on the scatterplot. For each observation, the deviation is measured as the difference between the observed value for the response and estimated value for the response that is generated by the model. In other words, it is the residual, and it is calculated as

$$r_i = Y_i - (\beta_0 + \beta_1 X_i) \quad (1.2)$$

The residuals for a portion of the (Traffic, Lead) example are shown in Figure 1.4. Note that some of the residuals are positive and others are negative. For the least squares criterion, the positives will always equal the negatives, so that the sum of the residuals will always be equal to 0. To circumvent this difficulty, we square each residual (to make them all positive), and then we identify the values for β_0 and β_1 that will minimize the total squared residuals. In other words, the least squares criterion minimizes

$$\sum_{i=1}^n r_i^2 = \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 X_i))^2 \quad (1.3)$$

where n is the number of observations.

In order to find the values for β_0 and β_1 that minimize Equation 1.3, we would need to apply some techniques of calculus. Fortunately, we have software that will perform these calculations for us. It is important, however, to understand that there are specific formulas that are derived using calculus, and that these formulas will produce the same results if they are given the same data. It is also important to understand what these formulas are doing, in terms of the variability among the Y values in the data set and the variability among the corresponding X values.

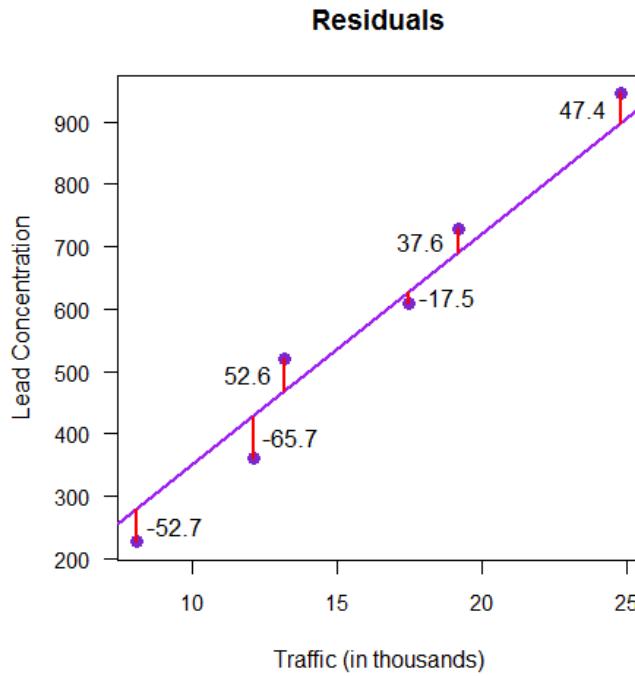


Figure 1.4. Some residuals for the (Traffic, Lead) example

Least squares estimation is based on the concept of the sum of squares, which is very closely related to the variance. From your earlier statistics course, you should recall that \bar{Y} is the sample mean for a variable Y and the sample variance is

$$\text{var}(Y) = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2 \quad (1.4)$$

The sample variance for Y is the average variability among the observed values for Y. It is an “average” because we divide by $n-1$. The total variability does not divide by $n-1$, and this is called the sum of squares. For least squares estimation, there are three sums of squares. These are

- the sum of squares for X: $SS_{XX} = \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - n(\bar{X})^2$
- the sum of squares for Y: $SS_{YY} = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n Y_i^2 - n(\bar{Y})^2$
- the sum of squares for XY: $SS_{XY} = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y}$

Note that there are two versions given for each of the sum of squares. The first version is the official definition, and the second version will always produce the same result. The second version is usually preferred if you are going to perform these calculations “by hand”. From these sums of squares, we can

calculate the least squares estimates for β_0 and β_1 . To indicate that these are estimates we use a “hat”, as in $\hat{\beta}_0$ and $\hat{\beta}_1$. These estimates are

$$\hat{\beta}_1 = \frac{SS_{XY}}{SS_{XX}} \quad (1.5)$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \quad (1.6)$$

Note that a bar on the top of a variable indicates the sample mean, while a caret (“hat”) indicates an estimate.

(Traffic, Lead) example, revisited

We will be using software to perform get these results, but in order to understand the complexity involved we will illustrate these calculations using the (Traffic, Lead) data. We first need to get the sums of squares, and this is best accomplished in a spreadsheet application (like Excel). We will use the second version of each equation for the sum of squares. Start with the (X, Y) pairs as separate columns in the spreadsheet, then create three new columns: one for X^2 , one for Y^2 , and one for $X*Y$. Then calculate the sum for each column. This is illustrated in Table 1.2

Site (i)	Traffic (X)	Lead (Y)	X^2	Y^2	$X*Y$
1	8.1	227	65.61	51,529	1,838.70
2	8.3	312	68.89	97,344	2,589.60
3	12.1	362	146.41	131,044	4,380.20
4	13.2	521	174.24	271,441	6,877.20
5	16.5	640	272.25	409,600	10,560.00
6	17.5	539	306.25	290,521	9,432.50
7	19.2	728	368.64	529,984	13,977.60
8	24.8	945	615.04	893,025	23,436.00
9	24.1	738	580.81	544,644	17,785.80
10	26.1	759	681.21	576,081	19,809.90
11	33.6	1263	1,128.96	1,595,169	42,436.80
Sums	203.5	7034	4,408.31	5,390,382	153,124.30

Table 1.2. Calculating sum of squares for (Traffic, Lead) example

The required sums are

$$\sum X_i = 203.5$$

$$\sum X_i^2 = 4,408.31$$

$$\sum Y_i = 7,034$$

$$\sum Y_i^2 = 5,390,382$$

$$\sum X_i Y_i = 153,124.3$$

The sample means are $\bar{X} = \frac{1}{11}(203.5) = 18.5$ and $\bar{Y} = \frac{1}{11}(7034) = 639.45$

Next, calculate the sums of squares.

$$SS_{XX} = 4,408.31 - (11)(18.5)^2 = 643.56$$

$$SS_{YY} = 5,390,382 - (11)(639.45)^2 = 892,522.67$$

$$SS_{XY} = 153,124.3 - (11)(18.5)(639.45) = 22,996.23$$

Now we can generate the estimated slope and intercept.

$$\text{estimated slope} = \hat{\beta}_1 = \frac{SS_{XY}}{SS_{XX}} = \frac{22,996.23}{643.56} = 35.7$$

$$\text{estimated intercept} = \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 639.45 - (35.7)(18.5) = -21$$

For the (Traffic, Lead) example, the estimated regression equation is Lead = -21 + 35.7*Traffic.

There are several ways that the estimated regression equation can be used. This will be explored in more detail later, but there are two examples that we provide now. First, we can use a value for X to generate an estimate for Y. For example, if a site has a traffic volume of 10,000 vehicles in a 24-hour period (so X = 10), we estimate the lead contamination to be Y = -21 + 35.7*10 = 336 micrograms of lead per gram of tree bark. We have used X = 10, and calculated Y = 336, so the point (10, 336) is a point on the regression line. Note that this value for Y is just an estimate. The accuracy of the estimate depends on numerous things, which we will discuss later.

Another use of the estimated regression equation is to interpret the slope. Since the equation is simply a straight line, we interpret the slope the same way we have always done: If X increases by 1, then Y changes by 'slope'. For an estimated regression equation, this becomes a little more complicated because we are dealing with a stochastic situation and not a deterministic one. For the (Traffic, Lead) example, the estimated slope is 35.7. This indicates that if the traffic volume increases by 1 thousand vehicles in a 24-hour period, then we expect the lead concentration in the tree bark to increase by 35.7 micrograms of lead per gram of tree bark. The "we expect" part of this sentence indicates that the

change may not be exactly 35.7, but that on average we expect it to be 35.7. This vagueness in the interpretation is necessary because we are dealing with a stochastic relationship.

Section 1.2. Model Assumptions

Every statistical procedure has assumptions, and regression is no exception. It is possible to generate estimates from a regression analysis, but gauging the accuracy of the estimates requires a probability distribution, which in turn requires some assumptions. For example, in the previous section we used the regression equation to estimate the amount of lead for for a site that has traffic 10 thousand, and we found that estimate to be 336 micrograms of lead per gram of bark. This tells us that 336 is reasonable number to expect for the lead concentration at that site, but we have no information regarding whether 335 or perhaps 300 would be reasonable numbers to expect. Similarly, we estimated the slope of the regression line to be 35.7, but would it be reasonable to believe that it could be 35 or perhaps 40? In order to answer these questions we need to perform a statistical analysis, and the type of analysis depends on the assumptions we make about the model.

To fully understand the model assumptions, we need to make a clear distinction between the regression model and the estimated regression equation. In its most general form, the simple linear regression model is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (1.7)$$

This model applies to the entire population, for example, all the sites that could possible be selected along the highway. β_0 and β_1 are population parameters that describe the “true” relationship between X_i and Y_i for *all* items in the population. When we select a sample of items from the population (e.g., a sample of sites along the highway), then we use the data collected at those sites to generate the estimated regression equation

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \quad (1.8)$$

Note that the two β ’s have been replaced by their estimates $\hat{\beta}$ ’s, and that the error term ε_i is absent. Also note that Y_i has been replaced with \hat{Y}_i , since this equation will be used to estimate values for Y . Equation 1.6 represents our estimate of the “true” relationship defined by equation (1.7).

The regression model is a conceptualization of a real process and, as such, it is a simplification of a much more complex phenomenon. The data we collect provides clues about the process, and we use those clues to decide whether or not our beliefs about the process are accurate. For some datasets, there may be many “good” models. For other datasets, finding even one good model can be difficult. George E.P. Box said it best: “All models are wrong, but some are useful.” We are simply looking for a “good” model.

1.2.1. The assumptions

For the simple linear regression model (equation 1.5) we implicitly assume that X and Y have a linear, stochastic relationship. This assumption can be evaluated via a scatterplot, as discussed in the previous section. If the relationship is clearly not linear, then a linear model is not appropriate unless the data can be transformed in order to make the relationship approximately linear. Transformations will be discussed in Section 1.7. When the relationship is linear, there are still other considerations before using a regression model. We assume that the two population parameters, β_0 and β_1 , are fixed, but unknown values. We also assume that the values for X_i are fixed. In contrast, the error term ε_i is random, that is, ε_i is influenced by some random phenomenon and this causes the values for ε_i to fluctuate according to some probability distribution. The values for Y_i incorporate the random values for ε_i , and this forces Y_i to be random as well. In summary, β_0 , β_1 , and X_i are fixed, while ε_i and Y_i are random.

The major assumptions for a simple linear regression model involve the random error term. Specifically, we assume that the values for ε_i follow a normal distribution that has mean 0 and a constant variance, and that the values for ε_i are all independent of each other. We abbreviate this as $\varepsilon \sim NIID(0, \sigma^2)$, where “NIID” is shorthand for Normal Independent and Identically Distributed.

By assuming that the distribution of the errors has mean 0, we are assuming that the error is just as likely to positive as it is to be negative (so that “on average” it is 0). This implies that the model is just as likely to over-estimate the value for Y as it is to under-estimate it. We are also assuming that the variance for the error term is a constant, and we designate this constant as σ^2 . The truly important part of this assumption is that the variance stays the same, regardless of the value for X .

The assumptions we make regarding the random variable ε have implications regarding the random variable Y . To understand these implications, consider equation (1.7). The only random component on the right hand side of this equation is ε_i , since β_0 , β_1 and X_i are all presumed to be fixed (not random). Since we assume that ε_i follows a normal distribution, this implies that Y_i also follows a normal distribution. Furthermore, the assumption that the ε_i are independent implies that the Y_i are independent. The close ties between ε_i and Y_i also extend to the values for their respective means and variances. From your earlier statistics course, recall that the mean of a random variable is also called the expected value of the random variable, denoted by a capital E . Also recall that the expected value of a fixed value is just that fixed value, and the variance of a fixed value is 0. When we take the expected value on both sides of equation (1.7), we obtain

$$\begin{aligned} E(Y_i) &= E(\beta_0 + \beta_1 X_i + \varepsilon_i) \\ E(Y_i) &= \beta_0 + \beta_1 X_i + 0 \\ E(Y_i) &= \beta_0 + \beta_1 X_i \end{aligned} \tag{1.9}$$

Thus the assumption that ε_i has mean 0 implies that Y_i has mean $\beta_0 + \beta_1 X_i$. Similarly, when we take the variance on both sides of equation (1.7), we see that the assumption $Var(\varepsilon_i) = \sigma^2$ implies that $Var(Y_i) = \sigma^2$. Note that the mean of Y depends on X , but the variance of Y does not depend on X . All of this can be summarized as follows: We assume that ε_i are independent, and follow a normal distribution with mean 0 and variance σ^2 . This implies that Y follows a normal distribution with mean $\beta_0 + \beta_1 X_i$ and variance σ^2 .

Every time we perform a regression analysis, we must first check the assumptions. We will not be able to determine if the assumptions are completely satisfied, but we will be able to determine if any of the assumptions appear to be grossly violated. Methods for verifying the assumptions are given in Section 1.4. If there are gross violations, potential remedies are presented in Section 1.7.

1.2.2. Minimum variance, unbiased estimators

When the assumptions are satisfied, the method of least squares produces the “best” estimates for the slope and intercept. This is given by the Gauss-Markov Theorem.

Gauss-Markov Theorem

Under the conditions of the linear regression model, the least squares estimators for β_0 and β_1 are unbiased and have minimum variance among all unbiased linear estimators.

To gauge the importance of minimum variance, unbiased estimators, consider Figure 1.5. The four targets represent four different estimators, that is, four different methods for generating the estimates. The bullseye on each target represents the true value of the population parameter we are trying to estimate, and the points on the target represent the values for the estimates that are obtained from different datasets (i.e., different random samples).

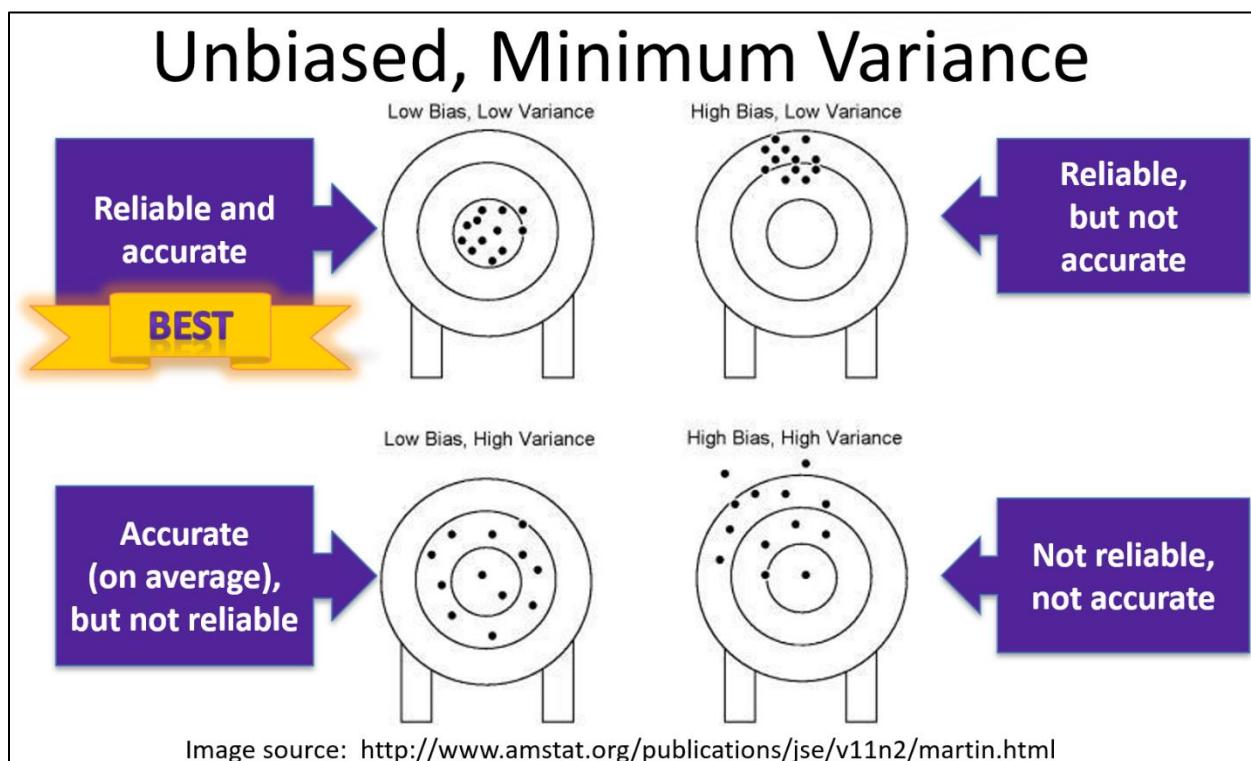


Figure 1.5. Properties of estimators

The accuracy of an estimator is based on the collection of values that could be generated by the estimator if the sampling procedure was repeated many times. In other words, if the data collection process is repeated, then it is likely that we would obtain different values for the estimated slope and intercept. Would the new estimates be similar to the original estimates or would they be very much different? It is desirable that the estimates be fairly close to each other, because this would indicate

that the estimator is reliable. In statistical terms, a reliable estimator has low variance. In contrast, estimators that have high variance generate estimates that vary dramatically, and these estimators are not reliable.

In addition to the reliability of an estimator, we must also consider its accuracy. If an estimator routinely generates estimates that are too high, then we say the estimator is biased. It is also possible for a biased estimator to routinely generate estimates that are too low. In either case, a biased estimator routinely “misses the mark”, and it is therefore considered a poor estimator. Instead, we prefer an estimator that is unbiased, that is, it is just as likely to over-estimate as it is to under-estimate so that, on average, an unbiased estimator “hits the mark”.

In Figure 1.5, the target in the upper left represents an estimator that is both reliable and unbiased. It is reliable because the points (i.e., the estimates) are tightly clustered together, so this estimator has low variance. It is unbiased because the points are centered on the bullseye. Now consider the target in the upper right. It represents an estimator that is reliable, but not accurate. The points on this target are clustered together (so it has low variance), but the points are not centered on the bullseye (so it is biased). The target in the lower right represents an estimator that is neither reliable nor accurate. The points on this target are widely scattered (high variance, not reliable) and they are not centered on the bullseye (biased). The target in the lower left illustrates an estimator that is unbiased (points centered on bullseye) and has high variance (widely scattered).

Of these four targets, the best one is in the upper left. This represents an estimator that is unbiased and has low variance. The Gauss-Markov Theorem tells us that, if the regression model assumptions are satisfied, the estimators defined by the method of least squares are unbiased and they have the smallest variance of all unbiased estimators. This is the statistical “gold standard” for all estimators, and this is why the method of least squares is so widely used to perform regression analysis.

Section 1.3. Inference

1.3.1. Point estimates

The main reason for performing a regression analysis is to establish a relationship between the response variable (Y) and the predictor variable (X). This involves obtaining estimates for three population parameters:

- the population intercept β_0 is estimated by $\hat{\beta}_0$
- the population slope β_1 is estimated by $\hat{\beta}_1$
- the error variance σ^2 is estimated by $\hat{\sigma}^2$

Formulas for calculating $\hat{\beta}_0$ and $\hat{\beta}_1$ have already been provided (equations (1.5) and (1.6)), so we turn our attention to $\hat{\sigma}^2$. This is an estimate for the variance of the normal distribution. The error variance captures the variability in the response variable that is NOT explained by the regression equation, that is, it measures the variability of the points *around the regression line*. To measure this “excess” variability, we use the difference between the observed value for the response variable and the value indicated by the regression equation. In other words, we use the residuals (see equation (1.2)). As we have already noted, the sum of the residuals is always equal to 0, so we square them to make them all positive. Then the total variability around the line is the sum of squares due to error, abbreviated SSE, and it is calculated as

$$\text{Sum of Squares due to Error} = \text{SSE} = \sum_{i=1}^n r_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (1.10)$$

The estimate for the error variance is the *mean square for error*, abbreviated MSE.

$$\text{Mean Square for Error} = \text{MSE} = \frac{\text{SSE}}{n-2} \quad (1.11)$$

The estimate for σ^2 is $\hat{\sigma}^2 = \text{MSE}$.

For the (Traffic, Lead) example, the calculations for MSE are shown in Table 1.3. The columns for Site, Traffic and Lead are taken directly from Table 1.1. The column for estimated lead uses the regression equation $\hat{Y} = -21 + 35.7X$ and the residual is the difference between the original value and the estimated value for Lead. The last column is the square of the residual. The sum of the last column is SSE = 70,805.1, and this produces MSE = 70,805.1/(11-2) = 7867.2. Thus the estimate for the error variance is $\hat{\sigma}^2 = \text{MSE} = 7867.2$.

Site (i)	Traffic (X)	Lead (Y)	Est'd Lead	Residual	Resid ²
1	8.1	227	268.17	-41.17	1695
2	8.3	312	275.31	36.69	1346.2
3	12.1	362	410.97	-48.97	2398.1
4	13.2	521	450.24	70.76	5007
5	16.5	640	568.05	71.95	5176.8
6	17.5	539	603.75	-64.75	4192.6
7	19.2	728	664.44	63.56	4039.9
8	24.8	945	864.36	80.64	6502.8
9	24.1	738	839.37	-101.37	10275.9
10	26.1	759	910.77	-151.77	23034.1
11	33.6	1263	1178.52	84.48	7136.9
Sum				70805.1	

Table 1.3. Calculating SSE for (Traffic, Lead) example

The main objective of this section is to understand the assumptions of the linear regression model. It is also important to understand the difference between the regression model and the estimated regression equation and to understand the difference between a population parameter (β_0 , β_1 and σ^2)

and the estimates for these parameters ($\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}^2$). The precise manner in which these estimates are calculated are of secondary importance, since we will be using software to perform the calculations.

All three of these estimates ($\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}^2$) are considered *point estimates* because they consist of a single value that approximates the value of an unknown population parameter. The value for \hat{Y} , as generated by Equation (1.8), is a point estimate for Y . It is also possible to generate interval estimates, to provide a range of plausible values for these unknown quantities. These topics, as well as hypothesis tests, are discussed in Section 1.3.

The estimates for the slope and intercept and slope are obtained by using the formulas $\hat{\beta}_1 = \frac{SS_{XY}}{SS_{XX}}$ and

$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$. The formulas are often called the estimators. Once a sample of data has been collected, the estimators are used to generate the actual numeric estimates. If the sample changes, then the estimates may change but estimators remain the same. This implies that each estimator is a random variable, so the values it generates follow a probability distribution. Furthermore, this probability distribution has a mean and a standard deviation. When the assumptions of the regression model are satisfied, the mean and the standard deviation of both $\hat{\beta}_0$ and $\hat{\beta}_1$ can be calculated from the data.

1.3.2. Mean and variance of the estimators

For the intercept, the expected value of $\hat{\beta}_0$ is $E(\hat{\beta}_0) = \beta_0$. This tells us that $\hat{\beta}_0$ is an unbiased estimator of β_0 . For the slope, the expected value of $\hat{\beta}_1$ is $E(\hat{\beta}_1) = \beta_1$. This tells us that $\hat{\beta}_1$ is an unbiased estimator of β_1 . As discussed in Section 1.2., it is desirable to have an unbiased estimator, but it is also desirable for an estimator to have low variance. The variance for the intercept and the slope are given by

$$Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{SS_{xx}} \right) \quad \text{and} \quad Var(\hat{\beta}_1) = \frac{\sigma^2}{SS_{xx}} \quad (1.12)$$

For both $\hat{\beta}_0$ and $\hat{\beta}_1$, the standard deviation is the square root of the variance. These standard deviations are often called the *standard error* of the estimates, in order to distinguish these standard deviations from the standard deviation of the error term. Note that the formulas for both of the variances involve σ^2 , but this value is not known. We substitute the estimate for σ^2 , which is $\hat{\sigma}^2 = \text{MSE}$. Then $\hat{\beta}_0$ and $\hat{\beta}_1$ each follow a *t* distribution with $n - 2$ degrees of freedom.

To illustrate, we calculate the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$ for the (Traffic, Lead) example. Recall that $n = 11$, $\bar{X} = 18.5$, $SS_{xx} = 643.56$, and $\text{MSE} = 7867.2$.

With these values, the variance and the standard error of $\hat{\beta}_0$ are

$$Var(\hat{\beta}_0) = 7867.2 \left(\frac{1}{11} + \frac{18.5^2}{643.56} \right) = 4899.69$$

$$se(\hat{\beta}_0) = \sqrt{4899.69} = 70.0.$$

The corresponding values for the slope are

$$Var(\hat{\beta}_1) = \frac{7867.2}{643.56} = 12.22$$

$$se(\hat{\beta}_1) = \sqrt{12.22} = 3.5$$

The standard errors are used to create confidence intervals for β_0 and β_1 , and to perform hypothesis tests for these parameters. We consider confidence intervals first.

1.3.3. Confidence intervals for the slope and intercept

The general form of a confidence interval is

$$(\text{point estimate}) \pm (\text{critical value}) * (\text{standard error}) \quad (1.13)$$

For estimating the slope and intercept in a simple linear regression model, the critical value is derived from the t distribution with $n - 2$ degrees of freedom, where n is the number of (X, Y) pairs in the data. The critical value also depends on the confidence level, which we will assume is 95% unless indicated otherwise. For a 95% confidence interval, the significance level is $\alpha = 0.05$ and we split this in half to obtain $\alpha/2 = 0.025$. We find the critical value in the probability table for the t distribution using $\alpha/2 = 0.025$ and degrees of freedom $df = n - 2$.

For the (Traffic, Lead) example, the critical value is 2.262. The 95% confidence intervals are

- for β_0 : $-21 \pm (2.262)(70.0)$, or $(-179.3, 137.3)$
- for β_1 : $35.7 \pm (2.262)(3.5)$, or $(27.8, 43.6)$

These intervals provide a range of plausible values for their respective population parameters.

A more precise interpretation for these intervals can be a little tricky, because they are based on the potential estimates we could generate if we collected additional samples of data. It is very easy to mis-interpret a confidence interval. We have found a 95% confidence interval for β_1 to be $(27.8, 43.6)$. One common misinterpretation is to say “The probability is 95% that β_1 is between 27.8 and 43.6.” This is incorrect because the value for β_1 is a fixed population parameter. We do not know the value, but we know (or at least we have assumed) that it is not random. Since it is not random, it does not have a probability distribution.

A confidence interval is customarily reported in one of two ways.

- We are 95% confident that β_1 is between 27.8 and 43.6.
- A 95% confidence interval for β_1 is $(27.8, 43.6)$.

Unless you are well-versed in statistical theory, it is not clear what either of the statements is attempting to convey. The correct interpretation of a confidence interval is based on the concept of repeated sampling, that is, using multiple random samples from the same population. Each random sample produces a different dataset, and these datasets represent different “snapshots” of the

population. Each dataset generates a different confidence interval, but all of these confidence intervals are for the same population parameter. If all of these confidence intervals have confidence level 95%, then we expect that about 95% of these confidence intervals will contain the true value of the population parameter.

The concept of repeated sampling is illustrated in Figure 1.6. The vertical dashed line is the true value of the population parameter and each horizontal line segment represents the confidence interval that was obtained from a single random sample (i.e., a single dataset). The top four confidence intervals include the true value, since these line segments overlap the vertical dashed line. The fifth line segment does not cover the true value, so this confidence interval does not contain the true value.

When we generate a single 95% confidence interval from a single random sample, we do not know if we have generated one of the confidence intervals that covers the true value, or if the confidence interval we have generated is one of the 5% that “misses the mark”. This is the uncertainty we are attempting to describe when we make the statement: “We are 95% confident that our confidence interval contains the true value.”

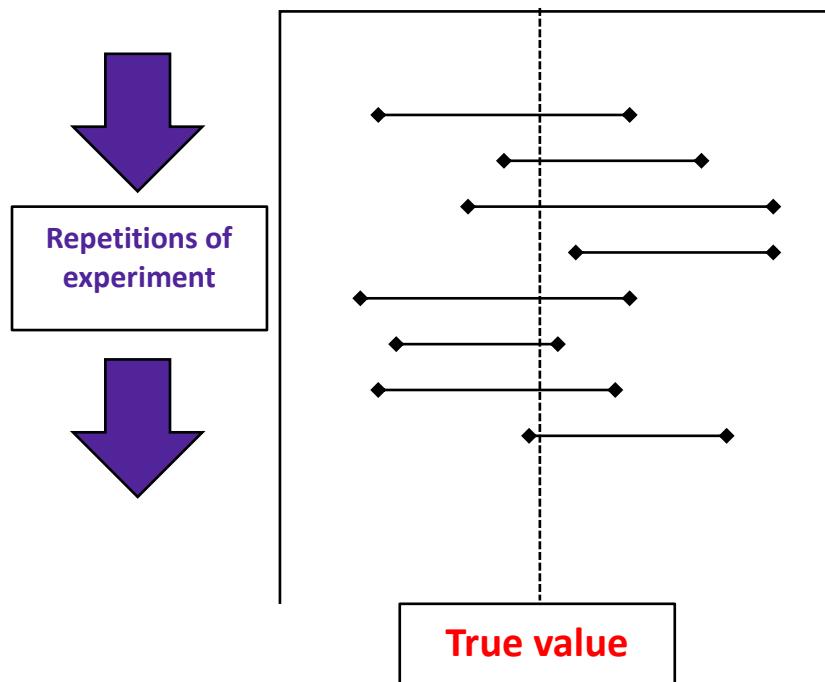


Figure 1.6. Multiple confidence intervals

1.3.4. Hypothesis tests

Recall the basic elements of a hypothesis test. Every hypothesis test has two competing hypotheses: the null hypothesis (H_0) and the alternative hypothesis (H_a , or sometimes H_1). Exactly one of these hypotheses is true, and we want to use the information in the dataset to decide which one we believe is true. Every test has a significance level, denoted by α , and we will assume $\alpha = 0.05$ unless it is specified otherwise. The values in the dataset are used to calculate a test statistic. There are different formulas for the test statistic, depending on what the hypotheses are and what the assumptions are. (Remember that every statistical model, including hypothesis tests, have certain assumptions that must be satisfied before the model is valid.) The test statistic will follow some probability distribution, which we call a reference distribution. The reference distribution is used to determine either (1) a critical value or (2) a p-value. For the critical value approach, we decide to reject H_0 if the absolute value of the test statistic is greater than the critical value. For the p-value approach, we decide to reject H_0 if the p-value is smaller than α . When we are conducting tests by hand, we usually use the critical value approach. If we are using software, we will use the p-value generated by the software to decide the test.

Hypothesis test for the population intercept

Suppose c is some constant (usually $c = 0$).

	In General	(Traffic, Lead) Example
Hypotheses	$H_0 : \beta_0 = c$ vs. $H_a : \beta_0 \neq c$	$H_0 : \beta_0 = 0$ vs. $H_a : \beta_0 \neq 0$
Test statistic	$\frac{\hat{\beta}_0 - c}{se(\hat{\beta}_0)}$	$\frac{-21 - 0}{70} = -0.3$
Reference distribution	t , with $df = n - 2$	t , with $df = 9$
Critical value	$t(\%, df = n - 2)$	2.262

- Decision: Since $|-0.3|$ is not greater than 2.262, we do not reject H_0
- Conclusion: It is reasonable to believe that the population intercept could be 0.

Although it is perfectly valid to perform a this hypothesis test for the population intercept, this test is rarely done because it is usually not of much interest. The population intercept is the value for Y when X

equals 0. In the (Traffic, Lead) example, we are testing whether or not the amount of Lead could be 0 at a site that has Traffic = 0. It stretches the imagination to think of a site along a highway that has 0 traffic, so this test is usually not considered as part of a regression analysis. It is included here solely for completeness. The real focus lies with the population slope.

Hypothesis test for the population slope

Suppose c is some constant (usually $c = 0$).

	In General	(Traffic, Lead) Example
Hypotheses	$H_0 : \beta_1 = c$ vs. $H_a : \beta_1 \neq c$	$H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$
Test statistic	$\frac{\hat{\beta}_1 - c}{se(\hat{\beta}_1)}$	$\frac{35.7 - 0}{3.5} = 10.2$
Reference distribution	t , with $df = n - 2$	t , with $df = 9$
Critical value	$t(\%, df = n - 2)$	2.262

- Decision: $|10.2| > 2.262$, so we reject H_0
- Conclusion: The sample provides evidence that the slope is not 0.

Since we have made the decision that the slope is not 0, this implies two things:

- (1) The predictor (Traffic) DOES help explain the values for the response variable (Lead).
- (2) The predictor should be kept in the model.

If the conclusion had been opposite (i.e., we had decided not to reject H_0), then any of the following could be plausible

- the population slope β_1 really is 0
- the ‘true’ regression model does not include the predictor
- a reduced model ($Y = \beta_0 + \varepsilon$) may be adequate
- the ‘true’ regression model is not linear

The first three items in this list are related to each other. If the true value for β_1 is 0, then the value for X in Equation (1.7) gets multiplied by 0. This implies that X can be removed from the model, which

indicates that the reduced model (excluding X) is just as valid as the simple linear regression model. This, in turn, indicates that there is not a linear relationship between X and Y , so that the ‘true’ regression model does not involve this X variable.

The last item in the list is different than the others, and it bears a little more investigating. The fact that we have concluded the slope could be 0 merely indicates that there is not a linear relationship between the response and the predictor. It is still entirely possible that there is some relationship between these two variables, but the relationship may be nonlinear. This is why it is important to look at a scatterplot of the data before attempting a regression analysis. If the scatterplot shows a curved pattern, then a linear model is not appropriate. It may be possible to transform either the predictor or the response (or both). Typical transformations are discussed in Section 1.7.

Relation between tests and confidence intervals

Two-sided hypothesis tests are directly related to confidence intervals, provided the level of significance of the test matches the confidence level of the confidence interval. For example, a test conducted at significance level 5% ($\alpha = 0.05$) is related to a 95% confidence interval, and a test conducted at significance level 0.01 is related to a 99% confidence interval.

For the (Traffic, Lead) data, we constructed a 95% confidence interval for β_1 to be (27.8, 43.6). This confidence interval does not contain 0, so 0 is not a plausible value for β_1 . Since 0 is not a plausible value, we should reject any hypothesis that claims it is equal to 0. This is precisely the result of our hypothesis test, when we rejected $H_0 : \beta_1 = 0$ at significance level 0.05. Based on the confidence interval for β_1 , we could perform numerous hypothesis. In addition to rejecting $H_0 : \beta_1 = 0$, we would also reject $H_0 : \beta_1 = 10$ and $H_0 : \beta_1 = 20$ and any hypothesized value for β_1 that is not in the confidence interval. We would fail to reject $H_0 : \beta_1 = 30$ and $H_0 : \beta_1 = 40$, since both 30 and 40 are in the confidence interval.

Since confidence intervals provide more information than a single hypothesis test, it has become routine for researchers to rely more on confidence intervals. While published research still relies heavily on hypothesis tests, using confidence intervals as a substitute is becoming more commonplace.

Coefficient of Determination

Also known as “R-squared”, the coefficient of determination provides the proportion of the variability in the response that is explained by the regression model. For simple linear regression, the coefficient of determination is directly related to the correlation between X and Y , which is discussed in more detail in Section 1.8. Because it is a proportion, the coefficient of determination is always between 0 and 1, but it is often expressed as a percentage between 0 and 100%. The value for the coefficient of determination is automatically generated by software as standard output of a regression analysis, or you can use the formula

$$R^2 = 1 - \frac{SSE}{SS_{yy}} \quad (1.14)$$

Higher values for R^2 indicates the model is a better fit to the data. For the (Traffic, Lead) example,

$$R^2 = 1 - \frac{70,805.1}{892,522.67} = 0.921.$$
 Approximately 92.1% of the variability in lead concentration can be

explained by the linear model that includes traffic volume.

It should be noted that 92.1% is an extraordinarily large value for R^2 , especially considering the fact that there is only one predictor variable in this model. Since the model explains 92.1% of the variation in Y then only 7.9% remains unexplained, and this is the part that gets incorporated into the error term. The decision regarding whether or not a value for R^2 is “large” is completely subjective, and it depends greatly on the subject matter. If the process for collecting the (X, Y) values is tightly controlled and the mechanism for measuring the values is very precise, then larger values of R^2 might be attained. In many situations, however, methods for measuring X and Y are imprecise or not well defined, and this tends to produce lower values for R^2 . One example might be an attempt to relate intelligence to academic achievement. We could use grade point average to measure academic achievement, but measuring intelligence is tricky business. Various IQ tests could produce different values for intelligence, and there is much debate as to whether any of the tests actually measure overall intelligence. When the variables are imprecisely defined, it is more likely to have a lower value for R^2 .

1.3.5. Prediction and estimation

In many circumstances, the primary purpose of regression analysis is to develop a model that will accurately predict or estimate a specific value for the response. “Estimate” and “predict” may sound like the same thing, but they represent two distinct goals in statistics. To understand the difference, let

us again consider the (Traffic, Lead) example, and suppose we are interested in the lead concentration when Traffic = 22. Suppose that you live along this highway, and proposed changes to the adjacent roadways is likely to increase the local traffic volume to 22. You are concerned about how this might change the lead concentration in your neighborhood. For this scenario, you are interested in only one site, so this is a prediction problem -- predicting the value for an individual Y that is derived from a specified value for X. On the other hand, suppose you are a highway maintenance worker (or perhaps an environmentalist), and you are interested in all the sites along the highway that have Traffic = 22. In this scenario, you would be interested in the average lead concentration at all the sites that have Traffic = 22. This would be an estimation problem -- estimating the mean value for Y for a specified X.

Estimation and prediction are very similar, but they are not the same. We can obtain point estimates and interval estimates for both individual Y values (prediction) and the mean Y values (estimation). The point estimates are the same for both prediction and estimation, but the confidence intervals are different. This is because the variability surrounding an individual Y value is greater than the variability surrounding a mean Y value.

Here is an example that may help to solidify this concept. Think back to some class that you took in which there were a lot of assignments. The students in this class completed the assignments and, at the end of the semester, an average score was computed for each student. When we consider only the averages, we find that most students scored in the B range, a few in the A range and a few in the C range, but hardly anyone scored D or F for the overall homework average. Now let's consider the scores on the first homework assignment. We will probably still see A's, B's and C', but we will probably also see several D's and F's. This is because the variability associated with an individual assignment is greater than the variability associated with the average all assignments. For an average, the high values tend to cancel out the low values, and the average ends up somewhere in the middle. This is exactly what is happening with the variability associated with a individual Y value and the variability associated with a mean Y value – the variability is greater for an individual value.

Table 1.4 shows the formulas that are used to calculate both the point estimate and the variance of the point estimate for estimation (an individual Y) and prediction (a mean Y). We will not use these formulas to calculate these by hand, we will use software instead. One thing to notice about these formulas is that the point estimates for estimation and prediction are exactly the same, and they are equal to the Y value on the the regression line corresponding the the X value x_0 .

The variances of the two point estimates are different, but only in the fact that the variance for prediction has an extra “1 +” in the formula. This is consistent with our notion that the variability associated with prediction will always be greater than the variability associated with estimation. (The assumes that the values for x_0 are the same for both estimation and prediction.)

Another thing to notice about the variance formulas is that both of them involve the square of $x_0 - \bar{X}$. If the value for x_0 is close to \bar{X} , then this difference will be small, which causes the variance of the point estimate to be small. If x_0 is far away from \bar{X} , then the difference will be greater and this causes the variance of the point estimate to be greater. This implies that interval estimates are narrower when x_0 is closer to \bar{X} , and the intervals are wider when x_0 is farther away from \bar{X} .

	Quantity to evaluate	Point estimate	Variance of the point estimate
Estimation	$E(Y X = x_0)$	$\hat{\beta}_0 + \hat{\beta}_1 x_0$	$\sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{SS_{XX}} \right)$
Prediction	$Y X = x_0$	$\hat{\beta}_0 + \hat{\beta}_1 x_0$	$\sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{X})^2}{SS_{XX}} \right)$

Table 1.4. Formulas for prediction and estimation

The final thing to notice about Table 1.4 is that the formulas for the variances both include σ^2 . Since the value for σ^2 is not known, we substitute the point estimate for σ^2 , which is $\hat{\sigma}^2 = \text{MSE}$.

To illustrate these concepts, we turn again to the (Traffic, Lead) example, and we use $x_0 = 22$. The point estimate is $\hat{Y} = -21 + (35.7)(22) = 764.4$ micrograms of lead per gram of bark. For a single site that has Traffic = 22, a point estimate for the lead concentration at that site is 764.4. For a collection of all sites that have Traffic = 22, a point estimate for the mean lead concentration for all these sites is also 764.4.

The variance for the mean estimate (estimation) is

$$\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{SS_{XX}} \right) = 7867.2 \left(\frac{1}{11} + \frac{(22 - 18.5)^2}{643.56} \right) = 864.95$$

The variance for the individual predicted value is

$$\hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{X})^2}{SS_{xx}} \right) = 7867.2 \left(1 + \frac{1}{11} + \frac{(22 - 18.5)^2}{643.56} \right) = 8732.15$$

This is not a typographical error. Simply including an extra “1 +” in the formula has created a variance for an individual value that is approximately 10 times larger than the variance for a mean value.

Prediction intervals vs. confidence intervals

For both prediction and confidence intervals, we use the basic form for an interval estimate

$$(\text{point estimate}) \pm (\text{critical value}) * (\text{standard error}) \quad (1.15)$$

The standard error is still the square root of the variance. The critical value is the same as it was when we were performing hypothesis tests. It is derived from the t distribution with degrees of freedom $n - 2$. For a 95% interval estimate, the critical value is 2.262.

A 95% confidence interval for the mean lead concentration across all sites with traffic = 22 is given by

$$(\text{pt. est.}) \pm (\text{critical value}) * (\text{std err}) = 764.4 \pm (2.262)\sqrt{864.95}, \text{ or } (697.9, 860.9)$$

A 95% prediction interval for the lead concentration at a single site that has traffic = 22 is given by

$$(\text{pt. est.}) \pm (\text{critical value}) * (\text{std err}) = 764.4 \pm (2.262)\sqrt{8732.15}, \text{ or } (553.0, 975.8)$$

Note that the prediction interval is wider than the confidence interval, and that both of these intervals are based on the same value for x_0 .

Now consider the prediction interval and the confidence interval when $x_0 = 18.5$. This also happens to be the sample mean for X , so these intervals should be narrower than the intervals when $x_0 = 22$.

The both intervals, the point estimate is $-21 + (35.7)(18.5) = 639.45$ and the critical value is still 2.262.

For the mean estimate,

- the variance is $7867.2 \left(\frac{1}{11} + \frac{(18.5 - 18.5)^2}{643.56} \right) = \frac{7867.2}{11} = 715.2$
- the confidence interval is $639.45 \pm (2.262)\sqrt{715.2}$, or (579, 700)

- note that the width of this interval is $700 - 579 = 121$, which is smaller than the width of the confidence interval when $x_0 = 22$ (that width is $860.9 - 697.9 = 163$)

For the individual predicted value,

- the variance is $7867.2 \left(1 + \frac{1}{11} + \frac{(18.5 - 18.5)^2}{643.56}\right) = 8582.4$
- the prediction interval is $639.45 \pm (2.262)\sqrt{8582.4}$, or $(429.9, 849.0)$
- note that the width of this interval is $849 - 429.9 = 419.1$, which is smaller than the width when when $x_0 = 22$ (that width is $975.8 - 553.0 = 422.8$)

The results of these calculations are summarized in Table 1.5. Note the following characteristics.

- The width of any of the intervals depends on the value of x_0
- Both types of intervals are narrower when x_0 is closer to the sample mean ($\bar{X} = 18.5$) than when x_0 is farther away
- For any given x_0 , confidence intervals are narrower than prediction intervals

	Confidence interval	Prediction interval
Traffic = 22,000 ($x_0 = 22$)	698 to 831 (width = 133)	553 to 976 (width = 423)
Traffic = 18,500 ($x_0 = 18.5$)	579 to 700 (width = 121)	430 to 849 (width = 419)

Table 1.5. Summary of confidence intervals and predictions intervals for (Traffic, Lead) data

1.3.6. Summary

These are the main topics of this section.

- Confidence intervals for the population intercept and slope.
- Hypothesis tests for the population intercept and slope.
- Point estimates and interval estimates for individual Y values and for a mean of Y values.

You will not be required to perform these calculations by hand, but you do need to understand what is involved in the calculations. You definitely need to know how to interpret the results of these calculations, and how these quantities are related to each other. You should also remember that all of hypothesis tests and interval estimates in this section utilize the t distribution with degrees of freedom $n - 2$.

Section 1.4. Software and Diagnostic Plots

Although there have been a lot of calculations in the preceding sections, we will rely on software to crunch the numbers, and use hand calculations only when necessary. There are many excellent statistical software systems available, but we will use SAS. This is an industry standard for high-end statistical computations. It is also free for academic users.

SAS is an actual programming system in which commands are typed into a SAS program and the program is executed in order to generate the output. SAS is not a menu-based point-and-click system. Complete SAS code will be provided for the many examples throughout this course. You are welcome to use this code and modify it for your own purposes. Please remember that software is simply a tool that we use so that we do not have to perform complicated calculations by hand. As with any software system, SAS will only do what it is programmed to do. It does not make any decisions and it cannot interpret any results. One small mistake in a SAS program can have dramatic results in the output, and a simple mistake can easily cause a SAS program to crash and not generate any output.

Every SAS program has a DATA step, which creates the dataset that SAS will use in its calculations. Every DATA step begins with the word DATA, has several lines of code to define the data values and the names of variables, and the DATA step ends with a semicolon. It is possible to read existing data files (like Excel files) into a SAS program, but all of our examples will contain the data directly in the DATA step.

After the DATA step, the SAS program needs to include one or more built-in SAS procedures (called PROC's) that can calculate summaries of the data (such as means and standard deviations), other PROC's can perform statistical analysis (such as t tests, or regression analysis), and still other PROC's can generate graphs. There are thousands of built-in procedures in SAS, and we will use only a handful of them.

An example SAS program is shown on the next page. It consists of one DATA step that creates a dataset for the (Traffic, Lead) example. After the DATA step, there is one PROC to generate a scatterplot and a second PROC to calculate the regression equation, and the point estimates, interval estimates, and other statistics that we have calculated by hand. We will then compare the output of the SAS program to the results we obtain by hand.

1.4.1. SAS code for the (Traffic, Lead) example

```
DATA example;
INPUT traffic lead;
DATALINES;
  8.1  227
  8.3  312
  12.1 362
  13.2 521
  16.5 640
  17.5 539
  19.2 728
  24.8 945
  24.1 738
  26.1 759
  33.6 1263
  22      .
;

PROC SGPLOT DATA=example;
  SCATTER X=traffic Y=lead;
  RUN;

PROC REG DATA=example;
  MODEL lead=traffic / P CLM CLI;
  RUN;
```

The diagram consists of three rectangular callout boxes with blue borders and white backgrounds, each containing a descriptive text. A blue bracket on the left side of the code points to the first three lines (DATA, INPUT, DATALINES). Another blue bracket on the right side points to the PROC SGPLOT and PROC REG statements. A third blue bracket at the bottom points to the entire PROC REG statement.

- The DATA step creates the SAS dataset
- Use PROC SGPLOT for a scatterplot
- Use PROC REG to perform regression

SAS keywords are in upper case letters. These must appear exactly as they are in the program. Words in lower case are names for datasets, variables, etc. You can choose whatever names you want, except that you cannot use any of the SAS keywords. Names in SAS are not case-sensitive, so `INPUT` is the same as `Input`, which is the same `input`.

The DATA step

The first line of every data step must include the key word `DATA` and end with a semicolon. The name after `DATA` defines the dataset name ('example'). You can choose a different name for your dataset. The second line is the `INPUT` statement. It defines the variable names ('traffic' and 'lead'), and ends with a semicolon. The third line is the `DATALINES` statement. There is nothing else on this line, except that it must end with a semicolon.

The data must be immediately after the `DATALINES` statement. There are no blank lines and no variable names after the `DATALINES` statement. There are no semicolons until you get to the end of the data. The data in this example is the (Traffic, Lead) data from Table 1.1.

The last data line (22 and a period) is used to get a prediction interval and a confidence interval of the mean when Traffic = 22.

Don't forget to put a semicolon on a new line at the end of the data.

PROC SGPLOT

This procedure is used to generate a variety of graphs. We will primarily use it to generate scatterplots. The first line begins with PROC SGPLOT and ends with a semicolon. The 'DATA = example' portion of the line tells SAS to use the dataset 'example' that you just created via the DATA step. If the 'DATA =' portion of the line is omitted, SAS will use the most recently created dataset. This may not be the data in the previous DATA step, so it is always advisable to include the 'DATA =' portion of the PROC SGPLOT statement.

The only graph we want to generate is a scatterplot of the (Traffic, Lead) pairs, so the only statement within PROC SGPLOT is the SCATTER statement. Following the key word SCATTER, use 'X =' and 'Y =' to define which variable is X and which variable is Y.

Every procedure, including PROC SGPLOT, ends with a RUN statement, and this statement must end with a semicolon.

PROC REG

This is the main procedure for performing regression analysis. The first line must contain the key words PROC REG and it must end with a semicolon. As with PROC SGPLOT, the 'DATA =' portion of this line is optional, but highly recommended.

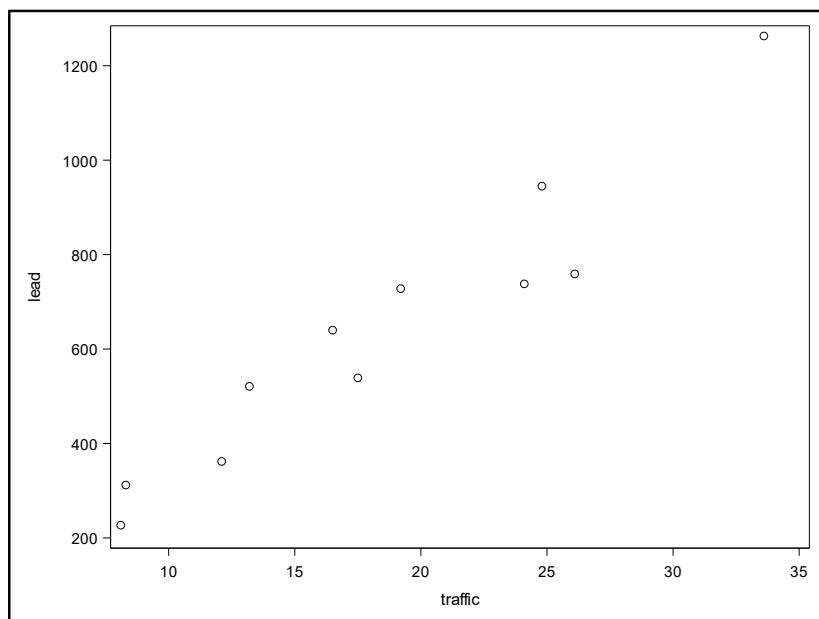
We want to generate a model in which the response variable (Y) is Lead and the predictor variable (X) is Traffic. The MODEL statement tells SAS which variable is X and which is Y. SAS does not remember that these were defined in PROC SGPLOT, so they must be specified again. The format for the MODEL statement is always "MODEL Y = X", where Y is on the left side of the equals and X is on the right. In the example code, the MODEL statement also includes a forward slash followed by P CLM CLI. Anything after the forward slash is an optional keyword to instruct SAS to calculate something additional.

- The keyword P tells SAS to calculate and print the predicted values. These are the point estimates for \hat{Y} .
- The keyword CLM is an abbreviation for confidence limits for the mean. This is what we have been calling the confidence interval for the mean.
- The keyword CLI is an abbreviation for confidence limits for individual, and this is what we have been calling a prediction interval.

Don't forget to include a RUN statement at the end of PROC REG, and the RUN statement needs to end with a semicolon.

1.4.2. SAS output for the (Traffic, Lead) example

The next several pages show the SAS output for the example program. The first graph is the results of PROC SGPlot with the SCATTER statement. It is simply a scatterplot of the (Traffic, Lead) pairs. SAS is notorious for printing a lot of output. It is not uncommon to have dozens of pages output for a very simple program. The next page of the SAS output begins the results from PROC REG.



The REG Procedure
Model: MODEL 1
Dependent Variable: lead

Number of Observations Read	12
Number of Observations Used	11
Number of Observations with Missing Values	1

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	821654	821654	104.44	<.0001
Error	9	70804	7867.16190		
Corrected Total	10	892459			

Root MSE	88.69702	R-Square	0.9207
Dependent Mean	639.45455	Adj R-Sq	0.9118
Coeff Var	13.87073		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-21.57636	69.99294	-0.31	0.7649
traffic	1	35.73140	3.49635	10.22	<.0001

The top table gives the total number of observations in the dataset and the number of observations used in the regression analysis. These are not the same numbers because we added an extra line of line (22 and a period). The first value is for Traffic and the second value is for Lead. (This is because the INPUT statement had Traffic first, and then Lead. The period tells SAS that the data value is missing, and SAS does not use missing values in the regression analysis.

The second table will be discussed in more detail later. For now, the value 892459 on the line for Corrected Total is what we calculated as SS_{YY} , the sum of squares for Y. We calculated the value as 892522.67. The difference is due to roundoff error in our hand calculations, so the value generated by SAS is more accurate.

The third table will also be discussed in more detail later. For now, the value 0.9207 for R-Square matches the value we calculate by hand (0.921). The last table provides the point estimates and their standard errors for both the intercept and the slope, and provides the test statistics (t Value) and p-value ($Pr>|t|$). We did not calculate the p-values, but the other values are within roundoff error of what we calculated in Section 1.3.

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	227	267.8480	45.1375	165.7399	369.9561	42.7144	492.9815	-40.8480
2	312	274.9943	44.5761	174.1561	375.8324	50.4338	499.5547	37.0057
3	362	410.7736	34.8699	331.8924	489.6548	195.1784	626.3688	-48.7736
4	521	450.0781	32.5358	376.4769	523.6793	236.3582	663.7980	70.9219
5	640	567.9917	27.6423	505.4606	630.5229	357.8270	778.1564	72.0083
6	539	603.7231	26.9707	542.7111	664.7352	394.0054	813.4409	-64.7231
7	728	664.4665	26.8549	603.7165	725.2166	454.8249	874.1082	63.5335
8	945	864.5624	34.6466	786.1864	942.9383	649.1515	1080	80.4376
9	738	839.5504	33.1445	764.5724	914.5284	625.3524	1054	-101.5504
10	759	911.0132	37.6999	825.7302	996.2962	692.9943	1129	-152.0132
11	1263	1179	59.1819	1045	1313	937.7881	1420	84.0013
12	.	764.5144	29.4100	697.9845	831.0444	553.1255	975.9034	.

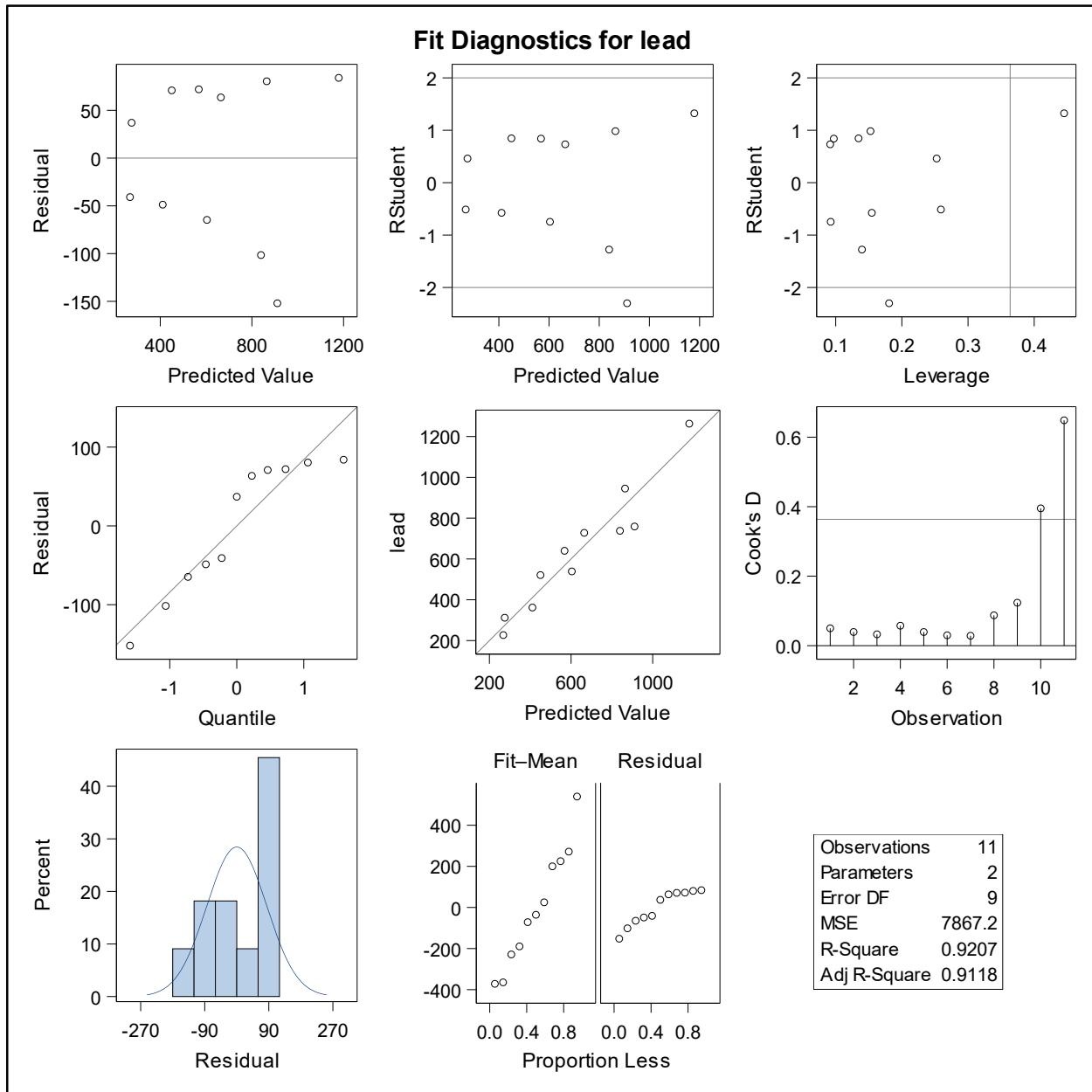
Sum of Residuals	0
Sum of Squared Residuals	70804
Predicted Residual SS (PRESS)	112161

This table is not part of the default output for PROC GLM. It was generated because we included the options P CLM CLI on the MODEL statement. These options, in conjunction with the extra line of data we added in the DATA step (22 and a period), provide the point estimate for Y and the confidence interval and prediction interval when Traffic = 22. All of the estimates related to Traffic = 22 are shown in the last line of this table, because the extra line of data we added was the last line in the DATA step. It is sometimes easier to put the extra line of data at the top in the DATA step (immediately below the DATALINES statement), so that it will appear on the top in this table. The point estimate is the “Predicted Value”, the confidence interval is “95% CL Mean” and the prediction interval is “95% CL Predict”. In addition to the values we wanted to calculate (which are in the last line of the table), SAS automatically generates these estimates for every observation in the dataset. This table can get quite long if there are very many observations.

The values in the last row are within roundoff error of the values previously calculated by hand.

The REG Procedure

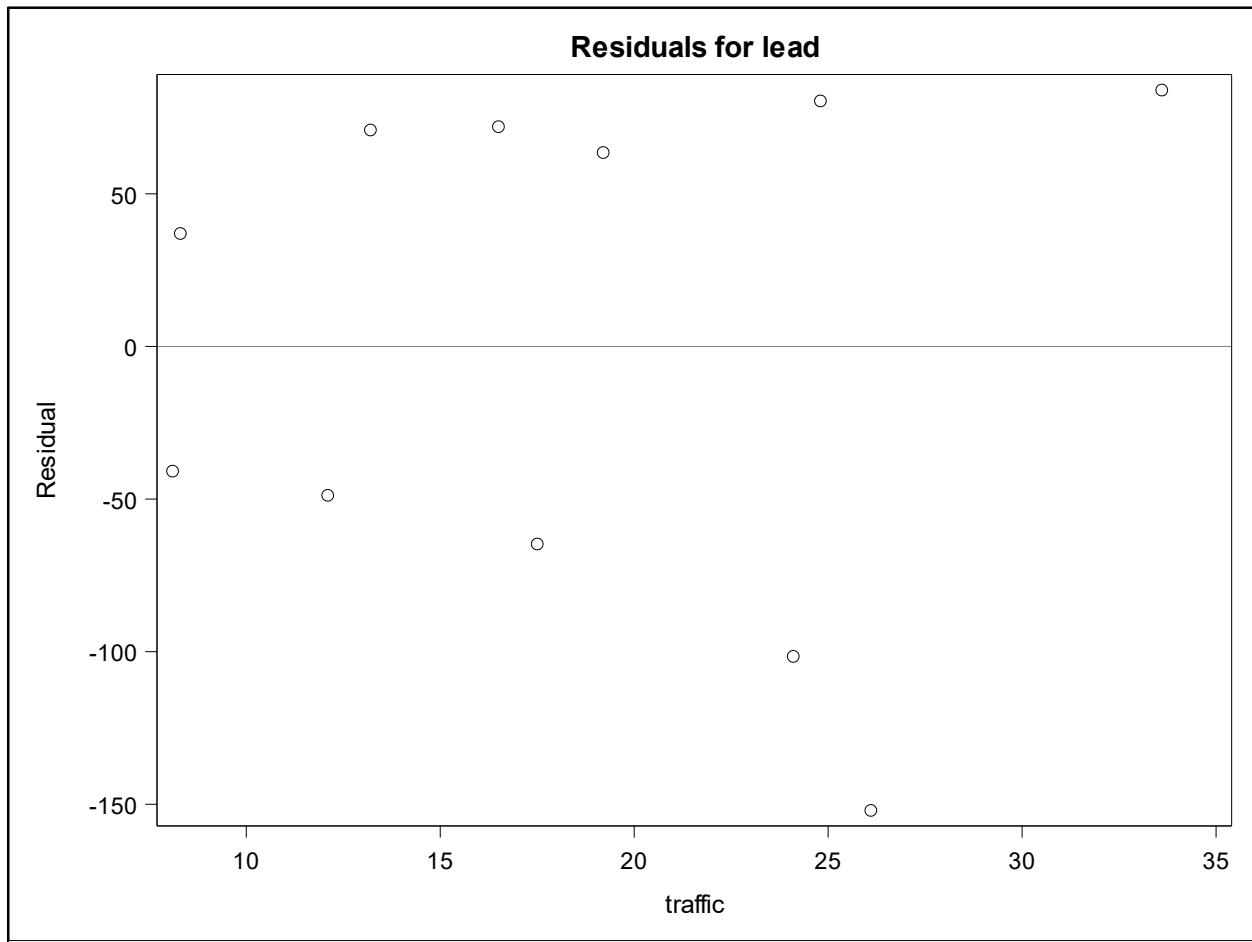
Model: MODEL1



This panel of plots is extremely important for assessing whether or not the model assumptions are satisfied. The residual plot is in the upper left corner and the normal probability plot is immediately below it. These will be discussed in more detail later.

The REG Procedure

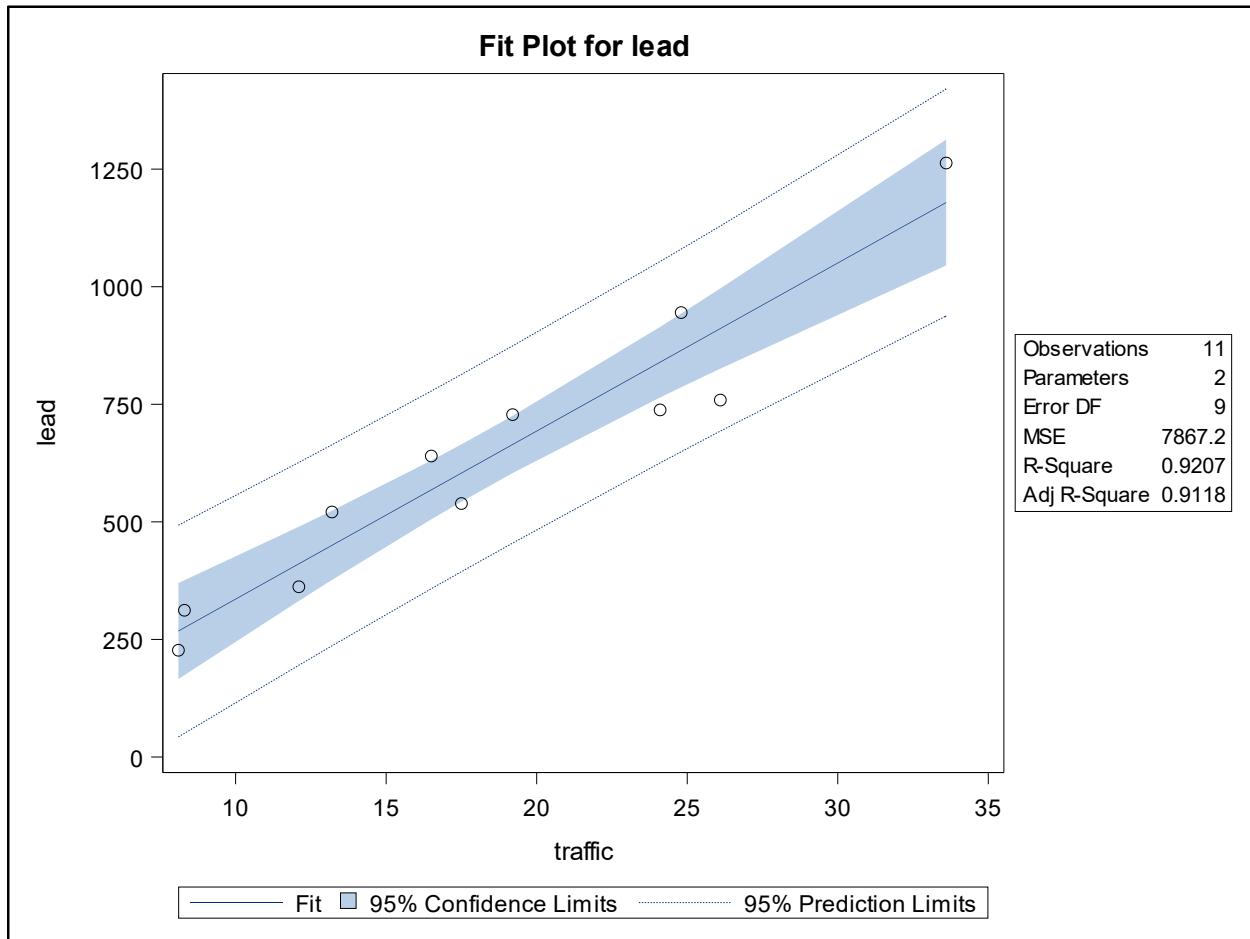
Model: MODEL1



This is another residual plot. It is similar to the graph in the upper left corner on the preceding page, in that they both have the residuals on the y axis. The difference between these two residual plots is that this graph has the original X values (for Traffic) on the x axis, while the residual plot on the preceding page has the fitted values (the \hat{Y} 's) on the x axis.

The REG Procedure

Model: MODEL1



This is a very interesting plot, and it is automatically generated by PROC REG. The points are simply a scatterplot, but the blue shaded area provides the confidence interval for the mean Y . Note that the shaded area narrows slightly near middle values for X . This is because the confidence interval is narrower when X is close to \bar{X} . The dotted lines represent the prediction intervals. Since the prediction intervals are always wider than the confidence intervals, the dotted lines are always outside of the blue shaded region. The dotted lines also become slightly more constricted near the center of the graph, because prediction intervals are also become more narrow when X is close to \bar{X} .

This ends the annotated SAS output for our example program.

1.4.3. Using the SAS output to assess model assumptions

The assumptions of a regression model can be stated succinctly as

$$\varepsilon_i \sim N(0, \sigma^2) \quad (1.16)$$

This indicates three basic assumptions:

1. The errors are follow a normal distribution
2. The errors are independent
3. The errors have constant variance (i.e., the variance does not depend on X)

If any of the assumptions are grossly violated, then the estimates and tests that are generated by the model may not be valid. For this reason, the assumptions need to be checked for every regression analysis and they need to be checked before any other results are interpreted.

Since all of the assumptions involve the error term, we must examine the residuals in order to check the assumptions. The only way we can get the residuals is to fit the model. We will be using SAS to fit the model, and SAS will automatically produce more output than we need. The output will not be in the order we need to examine it, so we will need to “jump around” in the output to perform a proper regression analysis.

Assess normality

To assess the assumption of normality, we will use the normal probability plot. In the SAS output, this graph is in the middle left in the panel of plots shown on page 36. This graph is also reproduced in Figure 1.7. The x-axis contains the quantiles from a normal distribution and y-axis contains the quantiles for the residuals. This type of graph is more commonly called a Q-Q plot (sometimes written QQ, without the hyphen), but it serves the same purpose as a normal probability plot. If the errors follow a normal distribution, then the points on this plot will follow the line. If there are “substantial” deviations from the line, then we would conclude that the errors do not follow a normal distribution. This would be a violation of an assumption, which would cause all the other results generated by SAS to be invalid.

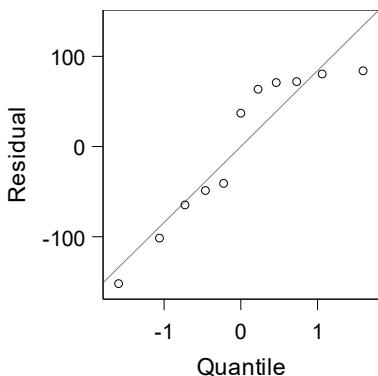


Figure 1.7. Normal probability plot for (Traffic, Lead)

The diagonal line on the QQ plot is not the regression line. It is merely a reference line so that we can how well the two sets of quantiles match. If they match perfectly, so that all the points are exactly on the line, the the obvious conclusion is that the residuals do follow a normal distribution. In most cases, there will be some deviation, and determining whether or not there are “substantial” deviations from the line is a subjective decision. It is possible for two very knowledgeable people to look at the same graph and arrive at two different conclusions.

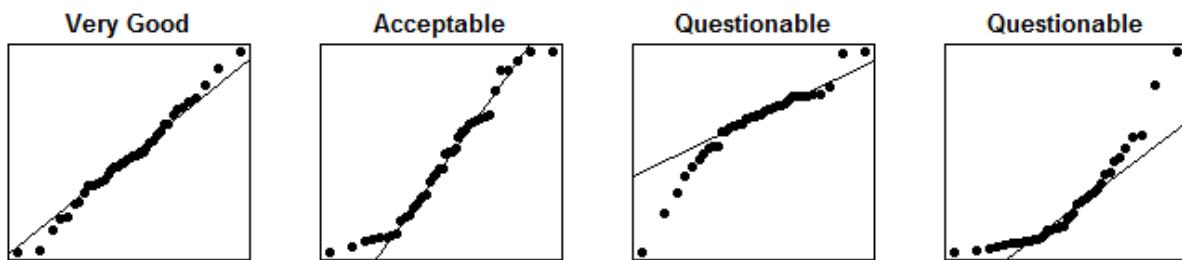


Figure 1.8. A variety of normal probability plots

Figure 1.8 shows a variety of normal probability plots. The graph on the left is excellent. Anyone who looks at this graph should conclude that normality is satisfied. The second graph shows more deviations, but this graph should also produce the conclusion that normality has not been violated. The two graphs on the right become much more questionable. There are obvious departures from the line in both of these graphs, so it quite possible that different conclusions could be drawn.

The least squares method for fitting a regression model is robust with respect to departures from normality. In other words, results from a regression model are still valid even when there are “minor” departures from normality. For this reason, we are looking for clear violations in the normal probability

plot. Although the plot in Figure 1.7 shows some deviations, we also need to consider the small sample size. It is more difficult to assess normality when the sample size is small, so we generally permit more erratic patterns in the QQ plot. For our example, we will conclude that the deviations are not severe enough for us to decide that normality has been violated.

There are many formal hypothesis tests to assess normality. These include Shapiro-Wilks and Anderson-Darling. Most tests for normality are sensitive to the sample size. For small samples, the tests usually fail to detect departures from normality. For large samples, the tests usually reject normality. For this reason, we will rely on a subjective interpretation of the normal probability plot.

Assess independence

There is nothing in the SAS output that allows us to assess independence. In fact, there is nothing in the data that provides information regarding independence. Instead, independence relies entirely on the manner in which the data was collected. This is why we insist that the data come from a *random* sample from the population, because the randomness ensures independence. Since we will not be collecting data as part of this course, we will assume that all the datasets we encounter were collected “properly”, so that independence is satisfied.

To get an understanding of how independence could be violated, consider this example. A researcher wants to understand the relationship between the height of an adult woman and the height of her mother. The dataset contains numerous (X, Y) pairs, where X is the mother’s height and Y is the woman’s height. To collect the data, the researcher went to the local mall and measured volunteers as they entered the food court. It is possible that some people arrived in family groups, and that some of the mothers had more than one daughter. Thus the dataset would contain more than one row of data for some of the mothers. This is an example of convenience sampling – an item is selected for inclusion in the sample merely because it is convenient. Since some of the mother’s values are repeated, these observations are not independent.

Here is another example that violates independence. This example comes from a recent student who was performing research on solar panels. A collection of panels was arranged in a grid and exposed to direct sunlight. Every minute, for approximately two hours, an electronic sensor recorded the amount of energy stored in each solar panel. This dataset was massive (tens of thousands of rows), but the measurements taken at any one time point would all be related (since they are all being exposed to the same amount of sunlight) and the measurements taken over time for one of the panels would also be

related. This is an example of time series data, in which one item is measured at several time points. In general, special statistical techniques are required for time series data.

Assess equality of variance

As with the normality assumption, we will use a graph in the SAS output to assess equality of variance. For regression analysis, there are no formal hypothesis tests for this. Recall that the variance we are referring to is the variance of the *errors* in the regression model. There is nothing in the data set that allows us to examine the errors. We must fit the model and generate the residuals, because the residuals are estimates for the errors. When we say “the variances are equal”, we mean that the variance is the same, regardless of the value for X, or the value for Y, or the predicted value for Y (\hat{Y}). If the variances are equal, we say that the data are homoscedastic. If the variances are not equal, we say the data are heteroscedastic.

To evaluate equality of variance, we use the residual plot, which is in the top left corner in the panel of plots produced by PROC REG. It is reproduced in Figure 1.9. A residual plot has the residuals on the y-axis. The x-axis could be the values for X, or the values for Y, or the predicted value for Y (\hat{Y}). The graph automatically generated by PROC REG uses the predicted value on the x-axis.

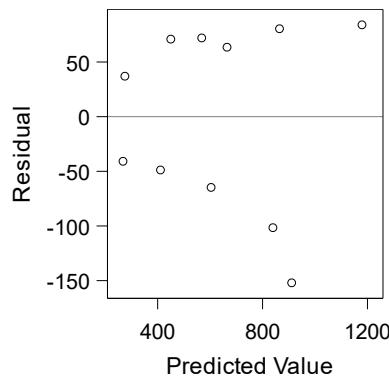


Figure 1.9. Residual plot for (Traffic, Lead)

If equality of variance is true, then the vertical spread of points in this plot should be roughly the same, regardless of whether you are looking on the left side, on the right side, and in the middle of the graph. The easiest way to evaluate this is to imagine a small picture frame. The height of the frame is exactly the height of the graph, and width of the frame is approximately 1/4 to 1/5 the width of the plot. Put the frame on the far left side of the graph, and note the vertical spread of points that fall within the

frame. Now shift the frame slowly to the right. As the frame moves, continue to note the vertical spread of points inside the frame. If, at any time, the vertical spread changes dramatically, then this is evidence that the assumption of equal variance has been violated.

This can sometimes be a subjective decision, so additional (fictitious) residual plots are shown in Figure 1.10. For the two graphs on the left, we have absolutely no concern about the equality of variance. The third graph (labeled “Unacceptable”), does not indicate a problem with equal variance – the vertical spread of points remains roughly the same no matter which part of the graph you focus on. The difficulty with this graph is that the residuals form a quadratic pattern (one that looks like the graph of $Y = X^2$). This is a problem, but it is not a problem regarding the variance. Potential remedies for this type of problem will be discussed in more detail in Section 1.7. The graph on the far right illustrates an obvious violation of the equal variance assumption. The points on the left have a very small vertical spread, while the points on the right have a large vertical spread. This is manifested by a wedge-shaped pattern in the residual plot. Some people see this as a funnel shape, laying on its side. This particular wedge shape is taller on the right, but it is also possible for the wedge (or funnel) to be taller on the left. It is also possible for the tallest part to be in the middle, and then the pattern would look more like a football.

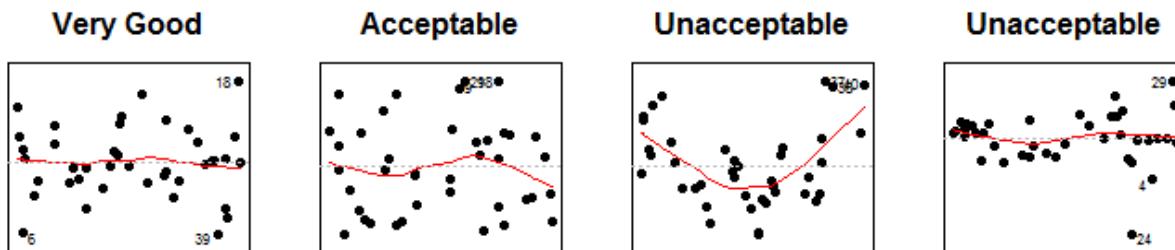


Figure 1.10. A variety of residual plots

We now return to the residual plot for the (Traffic, Lead) example, as shown in Figure 1.9. This graph shows a distinct wedge shape, which indicates that the assumption of equal variance has been violated. This implies that the confidence intervals and hypothesis tests that we conducted on this dataset should have never been examined. For this reason, we will no longer consider this dataset.

1.4.4. Summary

There is a lot of crucially important information in this section. You should be able to run the SAS code for the (Traffic, Lead) example and generate the output that was given. You should also be able to generate and interpret both the normal probability plots and the residual plots.

Section 1.5. NASA Rocket Propellant Example

We now consider a new example dataset, and apply the techniques we have learned thus far.

NASA is interested in the shear strength of the bond between propellants in a rocket motor and whether it may be related to age of the propellant batch. A sample of 20 propellant batches is collected and their shear strengths (in pounds per square inch) and ages (in weeks) are recorded. The data are shown in Table 1.6.

Obs (i)	Shear Strength (Y)	Age (X)
1	2158.7	15.5
2	1678.15	23.75
3	2316	8
4	2061.3	17
5	2207.5	5.5
6	1708.3	19
7	1784.7	24
8	2575	2.5
9	2357.9	7.5
10	2256.7	11
11	2165.2	13
12	2399.55	3.75
13	1779.8	25
14	2336.75	9.75
15	1765.3	22
16	2053.5	18
17	2414.4	6
18	2200.5	12.5
19	2654.2	2
20	1753.7	21.5

Table 1.6. Data for NASA rocket propellant example

Data source: Montgomery, et.al., 2006

We want to use this dataset to answer three questions of interest:

1. Is the shear strength linearly related to the age of the propellant? If so, quantify the relationship.
2. Estimate the mean shear strength for the propellants that are 10 weeks old.
3. Predict the shear strength for a propellant that is 10 weeks old.

Before we can answer these questions we need define and fit a linear regression model and assess the validity of this model. We will use SAS to generate the results that we need.

The first part of the SAS code is the DATA step. We create a temporary SAS dataset called 'nasa'. It contains all of the data in the dataset, and we will add one line of data that has Age = 10 and a missing value for shear strength, so that we can obtain the confidence interval (to answer question 2) and the prediction interval (to answer question 3). Note that we will not look at these intervals until after we have examined the model assumptions.

Here is the code for the DATA step.

It is not necessary to line up all the data values, or to force the same number of decimals, but this makes it easier to find any mistakes there might be in the data.

We have also included a PROC PRINT at the end of the DATA step. It is strongly recommended to print the dataset immediately after it is created. If there are any mistakes in the data, they need to be corrected before analysis continues.

The last line of data (for observation number 21) was not in the original dataset. This observation has Age = 10 and missing value (a period) for shear strength. Note that the period is in the middle column, while in the (Traffic, Lead) example, the period was in the last column. The location of the period must match the location of the variable name in the INPUT statement. For the NASA example, ShearStrength is the second variable in the INPUT

```
DATA nasa;
INPUT Obs ShearStrength Age;
DATALINES;
  1 2158.70 15.50
  2 1678.15 23.75
  3 2316.00 8.00
  4 2061.30 17.00
  5 2207.50 5.50
  6 1708.30 19.00
  7 1784.70 24.00
  8 2575.00 2.50
  9 2357.90 7.50
  10 2256.70 11.00
  11 2165.20 13.00
  12 2399.55 3.75
  13 1779.80 25.00
  14 2336.75 9.75
  15 1765.30 22.00
  16 2053.50 18.00
  17 2414.40 6.00
  18 2200.50 12.50
  19 2654.20 2.00
  20 1753.70 21.50
  21   .    10.00
;
RUN;

PROC PRINT DATA=nasa;
run;
```

statement (after Obs), so the missing value for ShearStrength must be in the second position in the data.

Once the dataset has been created, we can instruct SAS to perform the necessary calculations. This is shown below.

```
PROC MEANS DATA=nasa N SUM MEAN STD MIN MAX;
   VAR ShearStrength Age;
   RUN;

PROC SGPlot DATA=nasa;
   SCATTER X=Age Y=ShearStrength;
   RUN;

PROC REG DATA=nasa;
   MODEL ShearStrength = Age / P CLM CLI;
   RUN;
```

As with the (Traffic, Lead) example, the main part of the NASA analysis will consist of a scatterplot and a regression analysis. Whenever you are unfamiliar with a dataset, it is also advisable to generate some summary statistics of the data. If there was, for example, a missing or misplaced decimal place in one of the values for shear strength, SAS will not recognize that there is a mistake. The DATA step will execute correctly, and so will PROC PRINT. A mistake like this can sometimes be detected by looking at the sample means, etc.

In SAS, summary statistics are calculated via PROC MEANS. The additional letters at the end of the PROC MEANS statements are optional. They tell SAS to calculate the number of observation (N), the sum of the values (SUM), the mean value (MEAN), the standard deviation (STD), the minimum (MIN) and the maximum (MAX).

The VAR statement within PROC MEANS identifies the variables to use. The NASA dataset contains the variable Obs (short for observation), which is an identification variable and it is not part of the data we want to analyze. To avoid generating additional unnecessary output, we tell SAS to calculate the specified summary statistics only for the variables ShearStrength and Age.

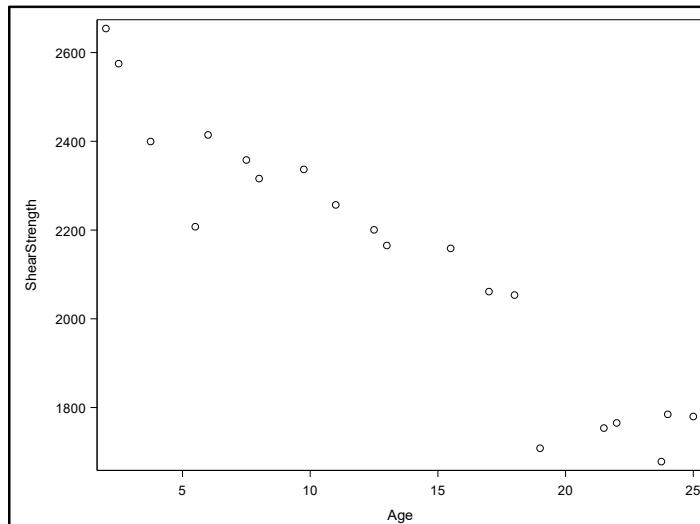
The rest of this code matches what we did with the (Traffic, Lead) example.

When this SAS program is executed, the first thing to examine is the table generated by PROC PRINT. If no mistakes are detected, proceed to the output generated by PROC MEANS.

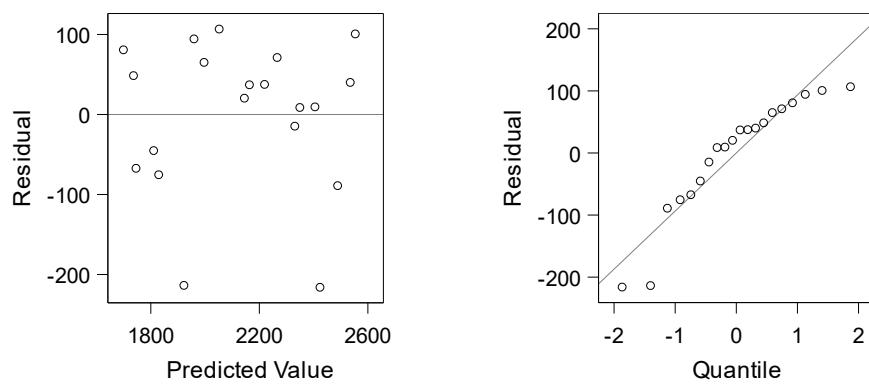
The MEANS Procedure

Variable	N	Sum	Mean	Std Dev	Minimum	Maximum
ShearStrength	20	42627.15	2131.36	298.5700660	1678.15	2654.20
Age	21	277.2500000	13.2023810	7.4743808	2.000000	25.000000

There is nothing obviously incorrect in these summaries, so the next step is to look at the scatterplot.



The scatterplot shows a linear trend, so a simple linear regression model seems appropriate. Next we look at the two diagnostic plots (the residual plot and the normal probability plot) in order to assess the assumptions.



The residual plot (on the left) shows no discernible pattern. In the normal probability plot (on the right), there is some deviation from the line, but it is not severe enough for us to terminate the analysis.

Overall, there is little evidence that the model assumptions have been violated. We can therefore proceed with the regression analysis and look at the output generated by PROC REG.

The REG Procedure
Model: MODEL1
Dependent Variable: ShearStrength

Number of Observations Read	21
Number of Observations Used	20
Number of Observations with Missing Values	1

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1527483	1527483	165.38	<.0001
Error	18	166255	9236.38100		
Corrected Total	19	1693738			

Root MSE	96.10609	R-Square	0.9018
Dependent Mean	2131.35750	Adj R-Sq	0.8964
Coeff Var	4.50915		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	2627.82236	44.18391	59.47	<.0001
Age	1	-37.15359	2.88911	-12.86	<.0001

The first table indicates that the dataset has 21 rows, but one of the rows has a missing value. This is consistent with the data we are using. The extra line of data (with a missing value) was inserted by us to generate the confidence interval and prediction interval for Age = 10. The second table contains the sums of squares, but we are not interested in those right now. The third table tells us that the value for R-squared is 0.9018, so that 90.18% of the variability in Shear Strength can be explained by this regression model. This is an extraordinarily high value.

The last table provides the estimated slope and intercept, the standard errors of these estimates, and the two hypothesis tests that we use to decide if either of the two population parameters is equal to 0. The line labeled 'Intercept' is testing the hypotheses $H_0 : \beta_0 = 0$ vs. $H_a : \beta_0 \neq 0$. The test statistic is

$t = 59.47$, with $p\text{-value} < 0.0001$. Since the $p\text{-value}$ is smaller than α (where $\alpha = 0.05$), we reject the null hypothesis and conclude the population intercept is not 0.

The line labeled 'Age' in the last table is providing the estimates and hypothesis test for the slope. The hypotheses are $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$. The test statistic is $t = -12.86$, with $p\text{-value} < 0.0001$. We reject the null hypothesis and conclude the population slope is not 0.

The estimates for the intercept and slope are, respectively 2627.82236 and -37.15359. These are used to construct the estimated regression equation. We will round these values to make the discussion more understandable. The estimated regression equation is

$$\text{ShearStrength} = 2627.8 - 37.15 \times \text{Age}$$

We are now prepared to answer Question 1: Is the shear strength linearly related to the age of the propellant? If so, quantify the relationship.

Yes, they are linearly related. This conclusion is justified by the results of the hypothesis test on the slope. The test statistic is $t = -12.86$, with $p\text{-value} < 0.0001$. To quantify the relationship, we interpret the estimated slope. For each additional week that the propellant ages, the shear strength of the bond between propellants is reduced, on average, by 37.15 pounds per square inch.

To answer the last two questions, we need to see the results of the CLM and CLI options on the MODEL statement in PROC REG. These options produce a fairly long table, with one row for each observation in the dataset. A portion of the table is shown below.

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	2159	2052	22.3597	2005	2099	1845	2259	106.7583
2	1678	1745	36.9114	1668	1823	1529	1962	-67.2746
... some rows have been removed ...								
19	2654	2554	39.2360	2471	2636	2335	2772	100.6848
20	1754	1829	31.8519	1762	1896	1616	2042	-75.3202
21	.	2256	23.5837	2207	2306	2048	2464	.

To answer question 2, we need to estimate the mean shear strength for the propellants that are 10 weeks old. This is asking for an interval estimate for the mean of a group of propellants, so we use the

confidence interval (95% CL Mean) generated by the option ‘CLM’. We use the last row in the this table, because we put the extra line of data at the bottom (in the SAS DATA step). The interval is (2207, 2306). For propellants that are 10 weeks old, we are 95% confident that the mean shear strength will be between 2207 and 2306 pounds per square inch.

To answer question 3, we need a prediction interval for a single propellant that is 10 weeks old. This is also on the last line in the table above, but now we use the columns 95% CL Predict (generated by the ‘CLI’ option). The prediction interval is (2048, 2464). For a single propellant that is 10 weeks old, we are 95% confident that its shear strength will be between 2048 and 2464 pounds per square inch.

This ends our analysis of the rocket propellant data. We presented an analysis of a simple linear regression model and answered three questions typically asked regarding regression models. There are many other questions that could have been asked, and these can be answered by either reading the SAS output or using values from the output to perform hand calculations. For example

- Find a 95% confidence interval for the population slope.
- Find a 95% interval estimate for the shear strength of a new propellant (age = 0 weeks).

Every statistical analysis is different, but there are certain patterns in the process that should be followed. Analyzing a dataset is not a “point-and-click” activity. It is a process that can take many different paths. Decisions that are made at one point in the process can dictate what path needs to be followed.

For any statistical analysis, we need to make sure the data values are correct. Identifying and correcting mistakes in the data often take as much time as actually analyzing the data. The datasets that are supplied as part of a textbook generally do not contain any errors, so the focus is on the analysis. If a dataset is procured from another source, take the time to make sure the data is accurate. If the data has mistakes, the results will also have mistakes.

Every statistical procedure has some underlying assumptions. Always check the assumptions before interpreting the results of a statistical analysis. If the assumptions are violated, the results are not valid. One customary assumption is that the data (or in the case of regression analysis, the errors) follow a normal distribution. If this assumption is violated, it may be possible to use an equivalent nonparametric statistical method that does not require normality. Nonparametric statistical methods are generally covered in a separate graduate-level class and are beyond the scope of what this book will cover.

Section 1.6. ANOVA Table, F and t tests

Some of the hypothesis tests required for a regression analysis have already been discussed. These include the test for deciding whether or not the population intercept is equal 0, and the the test for deciding whether or not the population slope is equal to 0. These are both t tests, that is, the t distribution is used to obtain the critical value and the p-value. There are many other tests that can be applied to a regression analysis, and some of these are automatically generated by PROC REG. Every hypothesis test in a regression analysis is defined in terms of the model parameters (β_0 and β_1), and we are usually more interested in the slope (β_1) than in the intercept (β_0). For simple linear regression there is only one slope, so much of the following discussion may seem obvious. In the next chapter, we will discuss multiple linear regression in which there are multiple slopes. That is when the analysis can become more complicated.

1.6.1. Partitioning the total sum of squares

The concept underlying a regression analysis involves identifying the sources of the variation in the values for the response variable (Y). For simple linear regression, we consider only one source and that is the value for the predictor variable (X). If we ignore this source, that is, if we remove the predictor from the model, then the best estimate we have for Y is \bar{Y} , the sample mean for Y . The total variation around this mean is $\sum(Y_i - \bar{Y})^2$. You may recognize this as the total sum of squares, SS_{YY} , as discussed in Section 1.1.

When we use the value for X to help identify the value for Y , some of the total variation in Y can be attributed to (or explained by) the value for X . The portion that remains unexplained is $\sum(Y_i - \hat{Y}_i)^2$. You may recognize this as the sum of squares due to error, SSE, as discussed in Section 1.2. Note that this sum is not the same as the total sum of squares, because overall mean \bar{Y} has been replaced by the predicted value for each observation \hat{Y}_i .

The difference between these two sums of squares is the amount of total variability that can be explained by the regression model. This sum of squares is designed SSReg, or sometimes SSModel. To see how the value for SSReg can be calculated, we partition the total sum of squares.

Start with the total sum of squares: $SSTot = SS_{YY} = \sum(Y_i - \bar{Y})^2$

Add and subtract the predicted \hat{Y}_i : $SSTot = \sum(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2$

Separate the terms: $SSTot = \sum(Y_i - \hat{Y}_i)^2 + \sum(\hat{Y}_i - \bar{Y})^2$

Partitioned Sum of Squares: $SSTot = SSE + SSReg$

The partitions can be visualized in Figure 1.11. This plot is a scatterplot, with X on the x-axis and Y on the y-axis. The diagonal line is the regression line and there is a horizontal line drawn at \bar{Y} . Each red triangle is an (X, Y) pair in the dataset. If we ignore the value for X (so that we ignore the regression line), then the total sum of squares is based on the vertical distance between each point and \bar{Y} . These are shown as gray vertical lines on the graph. (Some parts of the gray lines are obscured by the red lines. This is not avoidable.) This vertical distance can be split into two parts: (1) the distance between the point and the regression line, and (2) the distance between the regression line and \bar{Y} . The red lines illustrate the first part, and these are the distances that are used to calculate SSE. Distances in part (2) are used to calculate SSReg.

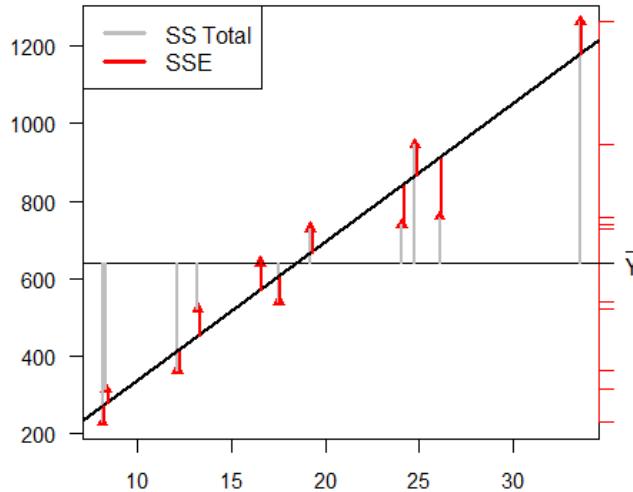


Figure 1.11. Partitioning the sum of squares

The concept behind partitioning the total sum of squares is relatively straightforward. If SSReg is “large” relative to SSE, then the regression model is useful. In other words, incorporating the values for X reduces the unexplained variability.

1.6.2. The ANOVA table

Calculations for the sums of squares are automatically performed by SAS for every regression analysis.

The results are reports in an ANOVA table, which is a standard way of reporting the results of many types of statistical analyses. (ANOVA is an abbreviation for Analysis of Variance.) The layout of an ANOVA table is shown in Table 1.7. This table also shows the formulas that are used to perform the calculations.

ANOVA Table					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	k-1	SSReg	SSReg / dfReg	MSReg/MSE	p-value
Error	n-k	SSE	SSE / dfE		
Corrected Total	n-1	SSTot			

Table 1.7. ANOVA table calculations for simple linear regression

An ANOVA table identifies the sources of variation, and these are listed in the first column. For a simple linear regression model, there is only one source of variation, which is the model. The next-to-last line (labeled “Error”) represents the remaining unexplained variation and the last line (“Corrected Total”) is reserved for the total variation. The column labeled DF stores the degrees of freedom for each source. You can think of one degree of freedom as one piece of “information” available in the dataset. Every time the data is used to estimate a parameter, the degrees of freedom is reduced by 1. The dataset contains a total of n (X, Y) pairs, so there are a total of n degrees of freedom. The mean for Y must be calculated (because the total sum of squares requires \bar{Y}), and this uses one degree of freedom. So the “Corrected Total” degrees of freedom is always $n - 1$. The degrees of freedom for Model is the number of slopes that must be estimated. In Table 1.7, the value k represents the number of parameters in the model. This is the number of slopes plus 1 for the intercept. For a simple linear regression model, the value for k is 2 (for the intercept and slope), so the degrees of freedom for Model will be equal to 1 (for one slope). Once the degrees of freedom for Model and Corrected Total have been identified, the degrees of freedom for Error can be calculated as a simple subtraction, because

$$\text{DF}(\text{Corrected Total}) = \text{DF}(\text{Model}) + \text{DF}(\text{Error})$$

The next column in the ANOVA table gives the sum of squares for each component. As with the degrees of freedom, the total sum of squares can also be calculated by adding the other sums of squares in the table.

$$SSTot = SSReg + SSE$$

The next column in the ANOVA table contains the mean squares. A mean square is an “average” of the sum of squares, but the divisor is the degrees of freedom. So the mean square for Model is calculated as the sum of squares for Model divided by the degrees of freedom for Model. Similarly, the mean square for error is the sum of squares for error divided by the degrees of freedom for error. There is no mean square for Corrected Total, so the last entry in this column will always be blank.

The next column is labeled “F Value”. The last two entries in this column will always be blank. The only entry in this column is a test statistic for a hypothesis test that we have not yet discussed. The hypotheses for this test can be stated in words better than symbols:

H_0 : None of the predictor variables in the model are useful for estimating Y

H_a : At least one of the predictors in the model is useful for estimating Y

The last column in the table, labeled “Pr > F” is the p-value for the test. The last two entries in this column will always be blank.

There will be a lot of ANOVA tables throughout this course. The ANOVA table for the NASA rocket propellant example is shown in Table 1.8.

ANOVA Table					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1,527,483	1,527,483.000	165.38	<.0001
Error	18	166,255	9236.381		
Corrected Total	19	1,693,738			

Table 1.8. ANOVA Table for the NASA rocket propellant example

1.6.3. ANOVA F test

The test reported on the “Model” line in the ANOVA table is testing the hypotheses

H_0 : None of the predictor variables in the model are useful for estimating Y

H_a : At least one of the predictors in the model is useful for estimating Y

For simple linear regression there is only predictor, so these hypotheses are equivalent to

H_0 : The predictor is not useful for estimating Y

H_a : The predictor is useful for estimating Y

We can also use the parameters in the model to specify these hypotheses.

$H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0$

There is yet another way to think about these hypotheses. If H_0 is true (so that the slope really is 0), then there is no reason to keep X in the model because it is just going to get multiplied by 0. If we remove X from the model, then we have what is called the reduced model: $Y_i = \beta_0 + \varepsilon_i$. The full (original) model is $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$. With these definitions, we can re-state the hypotheses

H_0 : The reduced model adequately fit the data, so the full model is not needed

H_a : The full model is needed to adequately fit the data

When the hypotheses are written this way, we think of this test as a comparison-of-models test.

No matter which set of hypotheses you prefer, the test statistic for this test is $F = \frac{\text{MSReg}}{\text{MSE}}$, where MSReg

and MSE are the two entries in the “Mean Square” column in the ANOVA table. The probability distribution for this test statistic is the F distribution. It is entirely plausible that you have never seen an F distribution, so we will digress for a moment and provide some characteristics of this distribution.

F distribution

An F distribution has two parameters, the numerator degrees of freedom and the denominator degrees of freedom. The numerator degrees of freedom is $k - 1$, where k is the number of parameters in the model. For a simple linear regression model, $k = 2$ (for the intercept and slope), so the numerator

degrees of freedom is 1. The denominator degrees of freedom is $n - k$, and for a simple linear regression model this is $n - 2$. These two parameters control the proportions of the F distribution.

The overall shape of the F distribution is shown in Figure 1.12. This distribution is very different than either a normal distribution or a t distribution. Values for the F distribution are always positive, so the curve starts at 0. There is no limit to how large the values can get, so the graph continues forever to the right. Any hypothesis test that uses the F distribution will be a right-tailed test. The p-value is calculated as the area under the curve to the right of the test statistic. We will not get the p-value by hand – we will rely on software. We use the p-value exactly the way would for any other test: We reject H_0 if the p-value is smaller than the significance level of the test.

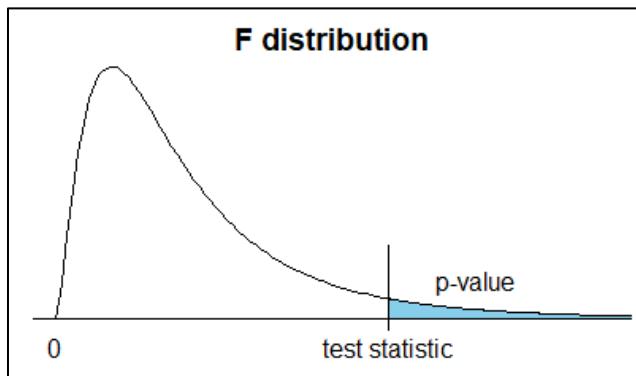


Figure 1.12. An F distribution

From NASA rocket propellant ANOVA table (Table 1.8), we see that the test statistic for the ANOVA F test is 165.38. The degrees of freedom for this test are also in the ANOVA table. The numerator degrees of freedom are the degrees of freedom associated with the numerator of F statistic, which is MSReg, and we have already determined that $df_{Reg} = 1$. (The labels “Reg” and “Model” are interchangeable for a regression analysis.) The denominator degrees of freedom for the F distribution are the degrees of freedom associated with denominator of the F statistic, which is MSE, and $df_E = 18$. When SAS calculates the p-value for this test, it uses an F distribution with degrees of freedom 1 and 18 (numerator and denominator, respectively), and the p-value is the probability of being greater than the test statistic. Since the p-value is $< .0001$, we reject H_0 (any version of these H_0 's) and conclude that Age is useful for estimating ShearStrength. The predictor Age should be kept in the model.

F test and t tests

For every regression analysis, SAS generates a parameter estimates table in addition to the ANOVA table. The parameter estimates table for the NASA example is shown in Table 1.9.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	2627.82236	44.18391	59.47	<.0001
Age	1	-37.15359	2.88911	-12.86	<.0001

Table 1.9. Parameter Estimates table for the NASA propellant data

The first line in this table is for the intercept. It provides the estimate for the intercept, $\hat{\beta}_0 = 2627.8$, the standard error of the estimate, $se(\hat{\beta}_0) = 44.18$. The columns labeled “t Value” and “Pr > |t|” are the test statistic and p-value, respectively, for testing $H_0: \beta_0 = 0$ vs. $H_a: \beta_0 \neq 0$. As mentioned earlier, this test is usually not very informative.

The second line in this table is providing the same information as the first line, but for the slope. The estimated slope is $\hat{\beta}_1 = -37.15$, with standard error $se(\hat{\beta}_1) = 2.889$, and the test statistic and p-value for testing $H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$. This test is usually of great importance, since it involves the slope. The p-value is <.0001, so we reject H_0 and conclude that the slope is not 0. The test statistic is $t = -12.86$, and the reference distribution is the t distribution. The degrees of freedom for the t distribution is NOT the degrees of freedom shown in the DF column. From earlier discussions, we know that the degrees of freedom for the t distribution is $n - 2$. The column labeled DF provides the number of parameters that are being tested, that is, the number of parameters that are “equal to 0” in the null hypothesis.

Since there is only one predictor in this model, there is a direct relationship between the t test for the slope and the F test for the model. In fact, they are exactly the same test. Mathematically, it can be shown that an F test with 1 and N degrees of freedom is the same as a t test with N degrees of freedom. Furthermore, the test statistics are related by $F = t^2$. For the NASA example, $t = -12.86$, so $t^2 = 165.38$, which is the value for F test statistic in the ANOVA table.

Since the F test and t test are performing the same test, the obvious question becomes: Why do we need both tests? A t test is testing exactly one parameter in the model, assuming all the other parameters remain in the model. The F test is testing all the parameters, assuming that only the intercept remains in the model. For a simple linear regression model, there is only one parameter in addition to the intercept, so the t test and the F test are the same. If there were 2 (or more) predictors in the model, then there will be multiple t tests and none of the t tests will be the same as the F test.

1.6.4. Some additional considerations

Observational vs. experimental data

The final interpretation of a statistical analysis is often restricted by how the data was collected. There are two basic ways in which data can be obtained: an observational study or an experiment.

Observational data is, as the name implies, simply observed or extracted, with no attempt to manipulate a situation in order to compel a response. Observational data include all types of surveys (e.g., election polls, opinion polls), but there are many other applications as well. The (Traffic, Lead) data that we studied in earlier sections is an example of observational data. The researcher simply counted the traffic at each location. There were no detours, no roadblocks, nothing that would disrupt the ordinary traffic volume. In addition, nothing was done to the trees that would affect the amount Lead. They were not watered, or fertilized, or pruned. Nothing was manipulated, it was simply recorded. This is observational data.

Another example of observational data is the NASA rocket propellant data. The researchers obtained a sample of 20 propellant batches, and recorded the age and shear strength of each batch. They did not subject the batches to specific temperatures. They did not add any chemical reagents to the batches. Nothing was manipulated. Data was simply measured and recorded.

With experimental data, there is a clear difference. It is called experimental data because it is usually the result of a scientific experiment. In advance of the data collection process, the researcher defines a narrow scope of interest and devises a plan to manipulate the situation in order to see if the response variables changes. For example, suppose a commercial baker is interested in changing a recipe for one of products. He believes he can reduce the sugar content of his chocolate chip cookies, and that this will not affect the flavor. He develops plan to test his theory. His plan includes 3 phases. In the first phase nothing changes. He keeps the sugar content the same as it has always been. In the second phase, he

reduces the sugar content by $\frac{1}{2}$ pound per batch of cookie dough. In the third phase, he reduces the sugar content by a full pound. For each phase, he makes several batches of cookies and recruits tasters to provide scores that indicate the quality of the cookies. Since the baker is manipulating the sugar content, this is experimental data.

With observational data, we can explore whether or not variables are associated, but we can NOT establish a cause-and-effect relationship between the variables. There could be other variables, that were not measured or recorded, that affect both the predictor and the response. For example, in the NASA example we estimated shear strength of a rocket propellant based its age. We did not consider any other variables that might affect the strength. Perhaps the propellant was stored in a container that somehow contaminated the propellant. The amount of contamination might increase over time, and that is what causes the propellant to lose strength. Based on the available data, we cannot say the increase in ages CAUSES the strength to decrease. We can only say that when the age increases, the strength decreases. In order to establish a cause-and-effect relationship, we need to have experimental data.

Interpreting a hypothesis test

When we conduct any hypothesis test (for regression or anything else), we are deciding which of two competing hypotheses (H_0 or H_a) is consistent with the data. We never “prove” that H_0 is true, or that H_a is true. We also never prove that they are false. Every hypothesis test begins with the assumption that H_0 is true. We reject this notion only if the data provides clear and convincing evidence that H_0 is false. If we reject H_0 , it does not mean that H_0 is false. It merely means that the data provides convincing evidence for us to believe that H_0 is false. If we do not reject H_0 , it does not mean that H_0 is true. It merely means that the data did not provide enough evidence for us to believe that H_0 is false.

When performing any statistical analysis, remain mindful of the sample size. With small samples, there may not be enough information (or there could be too much information) to reliably decide a test. When we are testing the slope in regression analysis, there may be a “true” linear relationship between the two variables, but the sample may be too small for us to detect it. It can also happen that there is not a “true” relationship, but the sample size is so large that analysis detects a minuscule relationship and exaggerates its relevance.

For regression analysis, always keep in mind that we are dealing with linear regression. If we fail to reject the hypothesis that the slope is 0, it does not automatically mean that there is no relationship

between the predictor and the response. There could still be a relationship between these two variables, but the relationship may be nonlinear. Figure 1.13 provides an scatterplot in which there is fairly strong relationship between X and Y, but the relationship is not linear. If a regression analysis is performed on this data, we would not reject the hypothesis that the slope is 0.

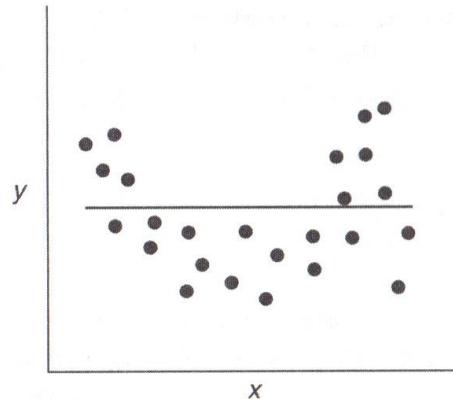


Figure 1.13. A nonlinear relationship

Extrapolation

In the NASA example, the estimated regression equation is $\text{ShearStrength} = 2627.8 - 37.15 * \text{Age}$. If we use this equation to estimate the shear strength of a propellant that is 75 weeks old, we obtain

$$\text{ShearStrength} = 2627.8 - 37.15 * 75 = -158.45 \text{ psi}$$

It is not possible to have a negative value for strength, so what went wrong?

The dataset contains values for Age that are between 2 and 24 weeks, so the regression analysis is limited to that range of values. It is not clear what happens to the strength when the age is greater than 24 weeks. It is possible that the strength would continue to decline, but it is also possible that the loss of strength would taper off and eventually stabilize. We don't have enough data to make an informed decision.

Extending inference beyond the scope of the data is called extrapolation, and it is not valid.

Section 1.7. Model Diagnostics & Transformations

To assess the adequacy of a model, we use diagnostic plots, such as a normal probability plot and various residual plots. For simple linear regression, an ordinary scatterplot of the (X, Y) pairs can also provide valuable information regarding model adequacy, but this is of limited value when there is more than one predictor in a model. The adequacy of a model is based on the model assumptions (that the errors follow a normal distribution with constant variance), but we must also evaluate whether or not the relationship between X and Y is linear. As discussed in earlier sections, a regression model is considered invalid if any of the assumptions are violated, and it is poor fit to the data if the relationship is not linear. In this section, we explore methods to detect model inadequacies, consequences of ignoring the inadequacies, and present possible remedies.

If the model assumptions are violated, then the F statistic (in the ANOVA table) may not follow an F distribution, and the test statistics for $\hat{\beta}_0$ and $\hat{\beta}_1$ may not follow t distributions. This means that all hypothesis tests, confidence intervals, and prediction intervals are not valid. The adequacy of the model must be examined before interpreting the results of a regression analysis.

To detect model inadequacies, we can use both graphical methods and formal hypothesis tests. The hypothesis tests are limited in scope, primarily because they each have their own set of assumptions that must be verified. We will concentrate on graphical methods, but they also have drawbacks. Interpreting diagnostic plots is a subjective process, and two reasonable analysts may have two different conclusions. In all of these plots, we are looking for extreme inadequacies, that is, obvious signs that something is “not right”. Minor deviations could be the result of sampling variability, and can be dismissed.

We will concentrate on three diagnostic plots. These are

- scatterplot of Y vs. X
- normal probability plot
- residual plot

1.7.1. Scatterplots

For simple linear regression, we use a scatterplot to assess whether or not it is reasonable to believe that there is a linear relationship between X and Y. The scatterplot should be examined before fitting the model. If the relationship between X and Y appears to be nonlinear, it may be possible to correct this by transforming one or both of the variables. Typical shapes and suggested transformation are discussed in Subsection 1.7.6.

1.7.2. Normal probability plots

The normal probability plot is used to decide if it is reasonable to believe that the errors follow a normal distribution. A normal probability plot is closely related to a QQ plot, which has quantiles from a normal distribution on the x-axis and quantiles from the residuals on the y-axis. The graph we use in the SAS output is actually a QQ plot, but it is interpreted the same way as a normal probability plot. If the points on the QQ plot follow the line, then the residuals follow a normal distribution. Note that it is highly unlikely that the points will follow the line exactly. There will (almost) always be minor deviations. This is especially true for small samples, which can produce very erratic QQ plots simply because there is so little data.

Example QQ plots are shown in the following figures. In addition to the QQ plot, each figure also contains the corresponding histogram of the residuals. The purple dashed curve on each histogram represents a normal curve. Histograms are generated by SAS, and are located in the lower left corner in the panel of plots generated by PROC REG.

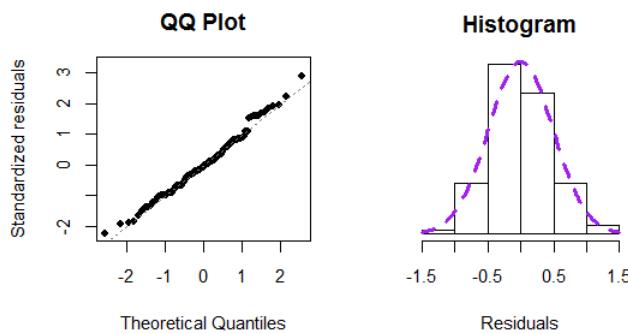


Figure 1.14. QQ plot and histogram #1

Figure 1.14 illustrates an “ideal” QQ plot and histogram. These data were simulated; it is unlikely to have such a perfect QQ plot when working with “real” data

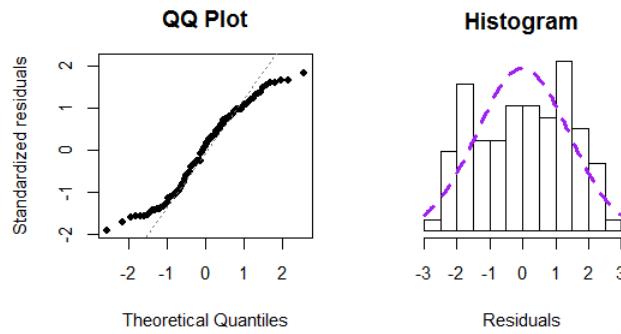


Figure 1.15. QQ plot and histogram #2

Figure 1.15 shows a distinct “S” shape in the QQ plot. This indicates a heavy-tailed distribution. In the histogram, a heavy-tailed distribution has taller than expected bars on both the left and right sides. However, the distribution is symmetric and centered at 0, so inference (using the F and t distributions) will be approximate. It is not necessary to abandon the analysis based on this QQ plot.

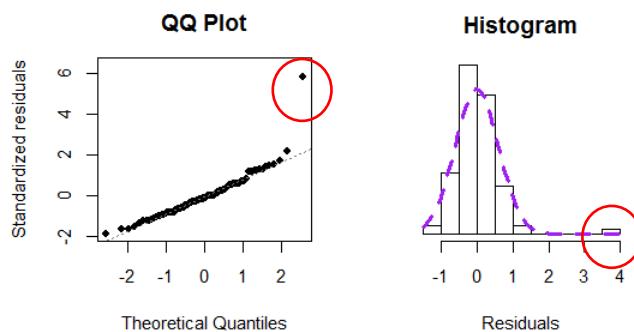


Figure 1.16. QQ plot and histogram #3

Figure 1.16 indicates a potential outlier, circled in red. This could be a mistake in the data, or just an unusual (X, Y) pair. This point should be investigated before proceeding with the analysis.

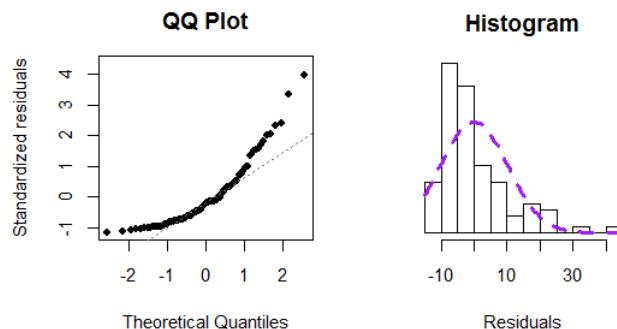


Figure 1.17 QQ plot and histogram #4

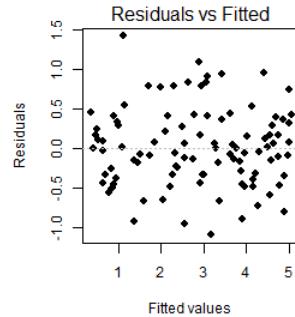
Figure 1.17 shows an extreme example of non-normal residuals. The QQ plot shows a “U” shape and the histogram is highly skewed. This is a clear violation of the normality assumption, and analysis should not proceed until this issue is resolved.

1.7.3. Residual plots

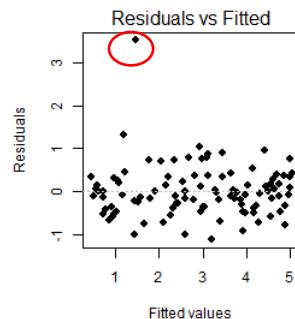
The residual plot contains a wealth of information about the adequacy of the model. It is possible for a residual plot to reveal a nonlinear relationship between X and Y that is not apparent in the scatterplot. Nonlinear relationships can often be remedied by a transformation, as discussed later in this section.

The residual plot can also be used to detect outliers. Outliers are specific (X, Y) pairs in the data that have an unusually high (or low) value for Y. Sometimes, an outlier can be detected in a scatterplot, since it would be a point away from the others. However, it is easier to detect outliers from a residual plot. An outlier is often simply a mistake in the data; once the mistake is corrected the outlier may disappear. If an outlier is not a mistake, it should not be removed from the data simply because it is an outlier. In many instances, outliers provide the impetus for important scientific discoveries.

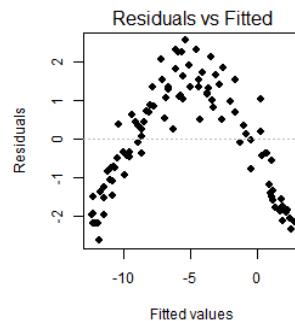
Example residual plots



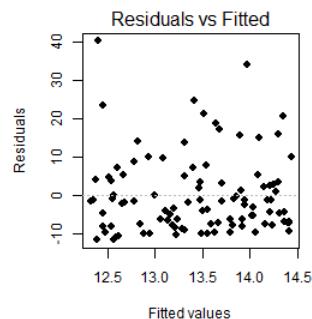
This is what the residual plot should look like. The points are scattered, with no obvious pattern.



There is a potential outlier, circled in red. This could be a mistake in the data, or just an unusual (X, Y) pair. This point should be investigated before proceeding with analysis.



This is absolutely clear and unambiguous. The quadratic shape indicates that including an X^2 term in the model may improve the fit.



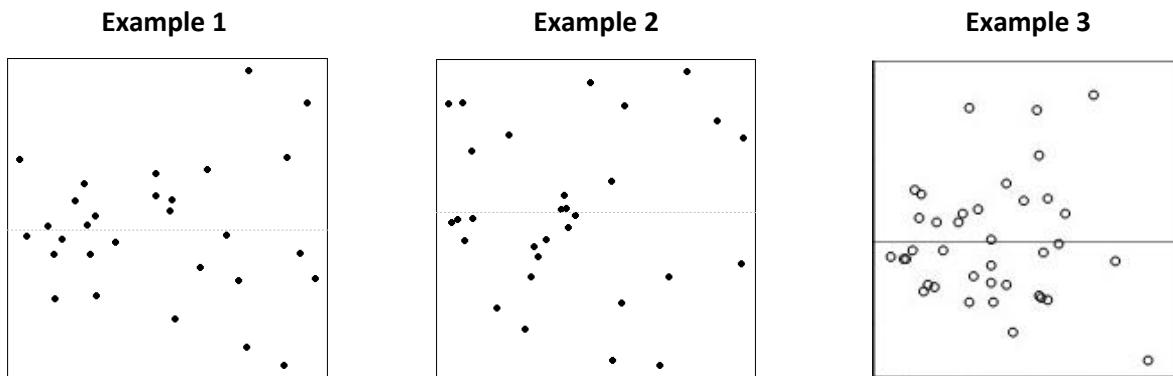
At first glance, this residual plot may appear to be okay. But on the y-axis 0 is near the bottom and it should be in the middle. The points above 0 are much more dispersed than points below 0. This pattern indicates the distribution of residuals is not symmetric, so they are unlikely to follow a normal distribution. When you encounter this type of residual plot, make sure to check the QQ plot and its histogram.

1.7.4. Detecting nonconstant variance

Residual plots can be used to detect violations of the assumption that the errors have constant variance. This was first presented in Section 1.4., and additional details are provided now. We present three examples of residual plots.

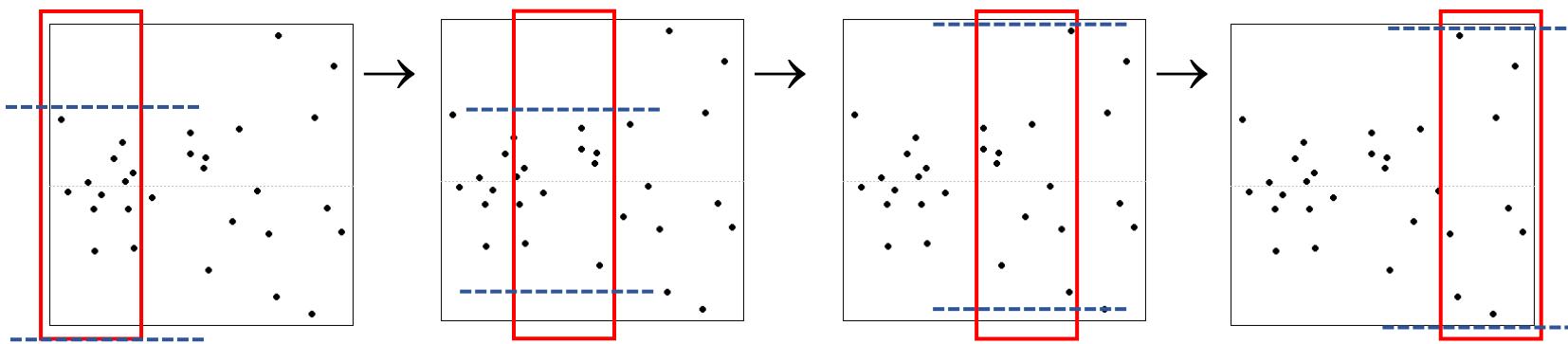
This is the general procedure. For each residual plot, draw a rectangle on the top of the plot. The width of the rectangle should be roughly one-fourth to one-fifth the width of the plot. Put the left edge of the box on the left edge of the plot, and notice the vertical spread of the points that are inside the box. Now move the box a little to the right. Do not change the width of the box. Again, notice the vertical spread of the points that are inside the box. Continue moving the box a little to the right and noting the vertical spread of the points inside the box. Does the vertical spread change dramatically as the box moves across the plot? If your answer is NO, then we have no evidence of non-constant variance, so there is no evidence that this model assumption has been violated. If your answer is YES, then there is evidence that the assumption of constant variance has been violated, and we need to consider modifying this model.

On the following pages, we perform this procedure on these three example residual plots.



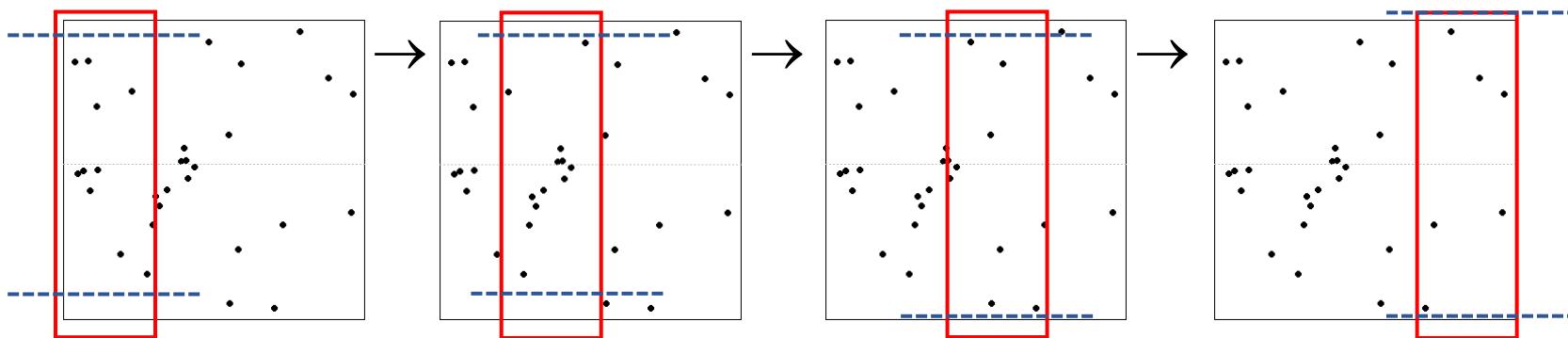
(These graphs have different formats because the first two were generated in R and the third one was generated in SAS.)

Example 1



Notice how the horizontal dotted lines get farther apart as the box moves from the left to the right. This is evidence that the variance is NOT constant, so that this model assumption appears to be violated. We need to modify this model before proceeding.

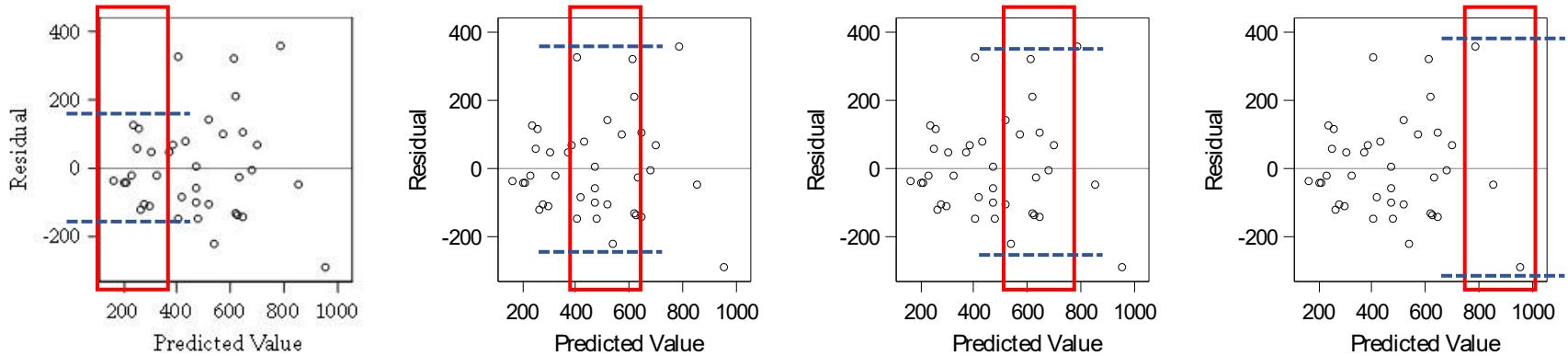
Example 2



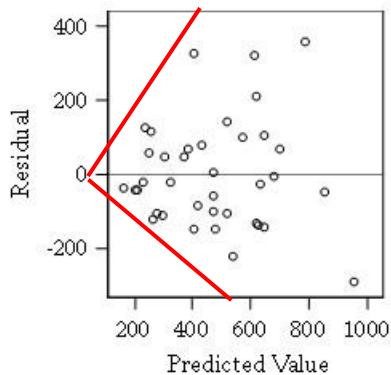
Notice how the horizontal dotted lines stay roughly the same distance apart as the box moves from the left to the right. This indicates that the variance remains fairly constant, so the assumption of constant variance appears to be satisfied.

Example 3

Notice how the vertical spread on the left-most graph is considerably smaller than the spread on the right three graphs. This forces us to consider the possibility that the variance may not be a constant.



A closer look at Example 3



This is sometimes called a “wedge” or “funnel” shape because we can draw two diagonal lines that define a triangular open space, as shown by the red lines. The funnel shape may open to the left instead of the right. It is also possible for the vertical spread to be greater in the middle of graph, in which case the pattern would like a football.

Other forms of residual plots

The residual plots we have been using have the fitted values (the \hat{Y} 's) on the x-axis and the residuals on the y-axis. There are other types of residual plots, but they are all interpreted the same way.

Sometimes, the x-axis contains the observed values for X, that is, the values for X that are in the dataset. This type of residual plot is rarely used, primarily because it is appropriate only for datasets that have a single predictor variable. If there are multiple predictors, it would be unclear which X variable to use in this plot.

Another type of residual plot uses the fitted value on the x-axis and the studentized residuals on the y-axis. Studentized residuals are standardized versions of the residuals. They have been re-scaled so that they follow a t distribution. Studentized residuals are compared to a t distribution in order to detect outliers in the data. Studentized residuals greater than 2 (or less than -2) are considered mild outliers, and those greater than 3 (or less than -3) are considered extreme outliers. SAS automatically generates this type of residual plot. It is in the middle of the top row in the panel of plots (immediately to the right of the original residual plot). The studentized residual plot generated by SAS also contains horizontal lines at 2 and -2, to better facilitate outlier detection.

1.7.5. Goodness of fit

Even when all the model assumptions appear to be satisfied, it is still possible for the model to be a poor fit to the data. There are several ways we can measure goodness of fit, and they all rely on how close the fitted Y values are to the observed Y values.

One measure of goodness of fit is the coefficient of determination, R-square. It is the proportion of the variability in Y that is explained by the regression model. It is a proportion, so it is always between 0 and 1, but it is often expressed as a percentage. Higher values for R-square indicate a better-fitting model.

Another measure of goodness of fit is the square root of the MSE, abbreviated RMSE. Recall that MSE is a point estimate for the error variance. When we take the square root, we have a point estimate for the error standard deviation. This measures the amount of variability around the regression line. Smaller values for RMSE indicate a better-fitting model.

1.7.6. Transformations

In some cases, a better-fitting model can be obtained by transforming either X or Y (or both). Certain shapes in the scatterplot and/or residual plot can provide clues as to what transformation might be appropriate. Typical transformations include squares, reciprocals and logarithms. For example, changing X to X^2 , or changing X to $1/X$, or changing Y to $\log(Y)$. Whenever transformed variables are used to fit a new model, the new model must still be checked for adequacy by examining its diagnostic plots. If the Y variable is transformed, it needs to be back-transformed before reporting the results of the analysis. There is no need for back-transformation when X is transformed, because the original X values are still in the dataset.

We first consider four curved patterns that can appear in a scatterplot, as shown in Figure 1.18. Other types of curves can be present in a scatterplot, but they are not as common. The suggested transformations for each curve are also shown in Figure 1.18. These are not the only transformations that could be applied, but they are the most likely to succeed.

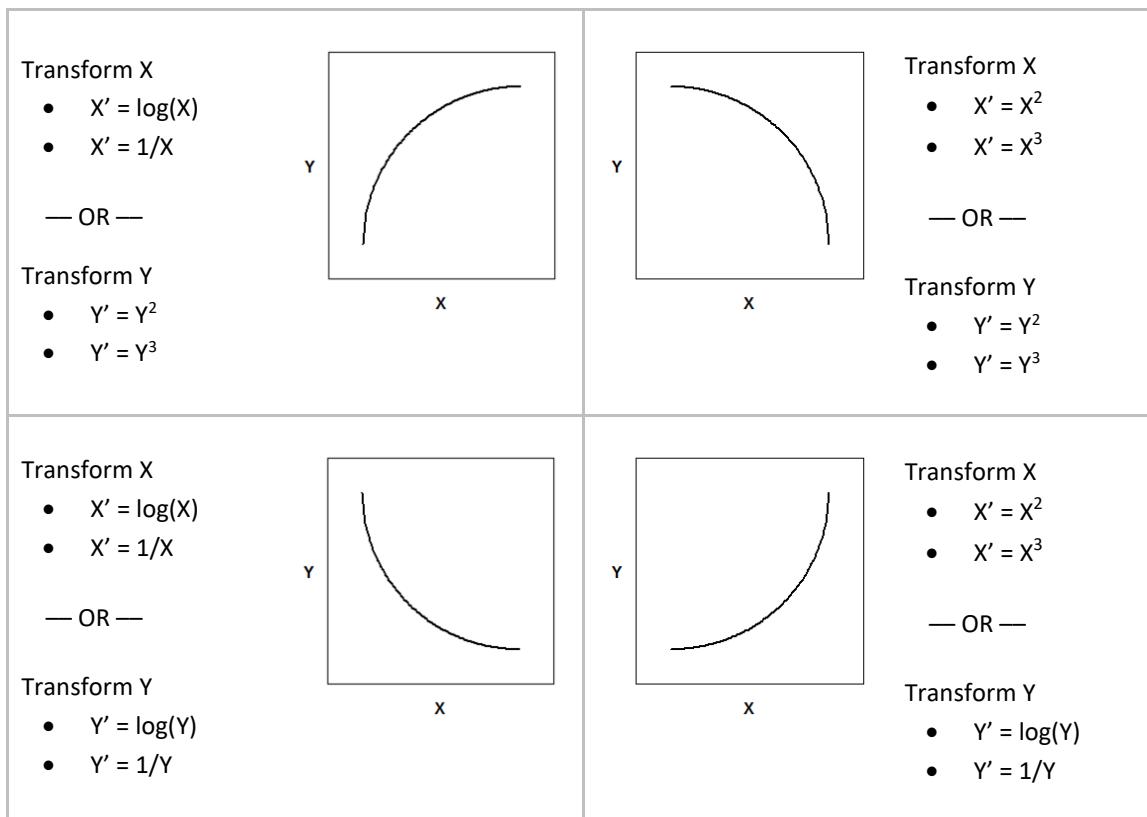


Figure 1.18 Suggested transformations based on scatterplot curves

Example 1

Following a hazardous waste accident, soil at was periodically tested to determine its pH level. We want to model the change in pH over time. The data are shown in Table 1.10. The predictor variable (X) is the time since the accident and the response variable (Y) is the measured pH in the soil.

We begin by generating a scatterplot, as shown in Figure 1.19. The points in this plot appear to follow a curved pattern, so it is unlikely that a linear model will fit these data very well. Comparing this scatterplot to the graphs in Figure 1.18, it seems that there are two reasonable choices for a transformation, and both of these involve replacing X (i.e., time) in the model. We could replace X with either $\log(X)$ or with $1/X$. It is not clear which of these transformations will generate a more suitable model, and it is entirely possible that neither of these transformation will work. The only way to know if a transformation is successful is to fit each model and evaluate the diagnostic plots.

time	pH
1	7.02
1	6.93
2	6.42
2	6.51
4	6.07
4	5.98
6	5.59
6	5.80
8	5.51
8	5.36

Table 1.10

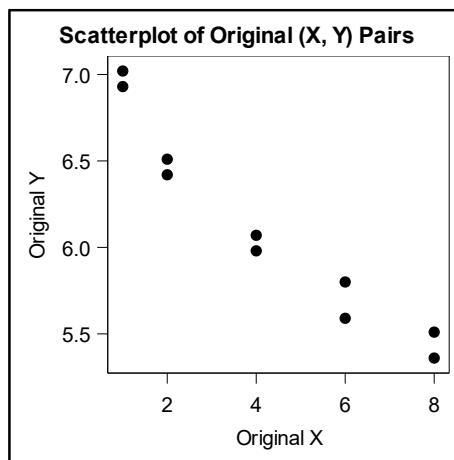


Figure 1.19. Scatterplot of (Time, pH)

We will consider these three models:

- Original data: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
- Using $1/X$: $Y_i = \beta_0 + \beta_1 \left(\frac{1}{X_i}\right) + \varepsilon_i$
- Using $\log(X)$: $Y_i = \beta_0 + \beta_1 \log(X_i) + \varepsilon_i$

Although we believe it will be a poor model, we fit a linear model to the original (X, Y) pairs. The residual plot in Figure 1.20 clearly shows a non-random pattern and the curvature indicates that the relationship between X and Y is not linear. This is further reinforced by the graph in Figure 1.21, which is a direct comparison of the original (observed) Y values and the Y values that are predicted by the model. In other words, Figure 1.21 is simply a scatterplot with additional points (red stars) to indicate the predicted Y for each X. Notice that the red stars form a straight line (because we fit a linear model), while the original data points (black circles) form a curve. This model is a poor fit to the data.

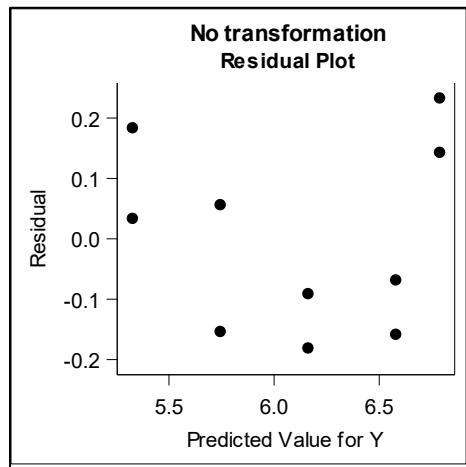


Figure 1.20. Residual plot, original data

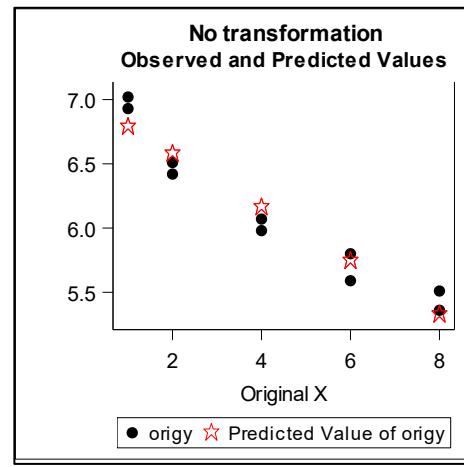


Figure 1.21. Goodness of fit, original data

We now consider a model in which X is replaced by the reciprocal of X . The residual plot, shown in Figure 1.22, still exhibits a curved pattern. Although the residual plot is not acceptable, the graph of observed and predicted values (Figure 1.23), shows that the “linear” model is now a curve. If the x-axis in Figure 1.23 had used the values $1/X$ instead of X , then the red stars would have formed a straight line.

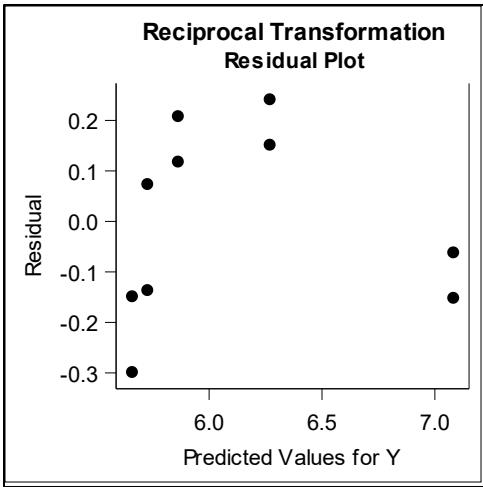


Figure 1.22. Residual plot, using $1/X$

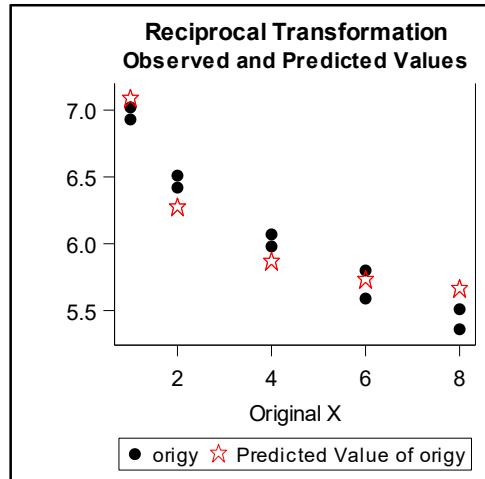


Figure 1.23. Goodness of fit, using $1/X$

The third model we consider uses $\log(X)$ instead of X . The residual plot (Figure 1.24) is a clear improvement from the two earlier models, and note how closely the observed Y and predicted Y coincide in Figure 1.25. The model that uses $\log(X)$ is clearly superior to either of the two other models.

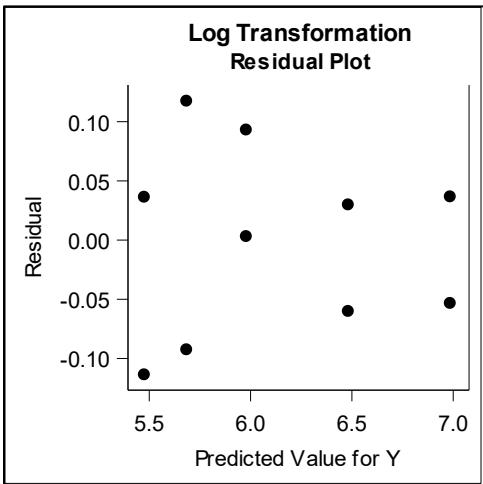


Figure 1.24. Residual plot, using $\log(X)$

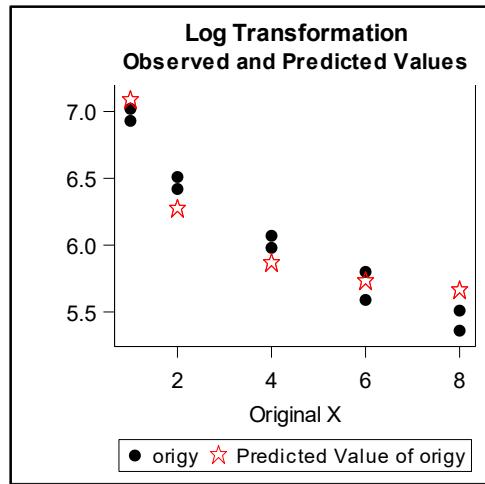


Figure 1.25. Goodness of fit, using $\log(X)$

For this particular example, using the transformation $\log(X)$ was obviously better than using either the original data or using $1/X$. In some cases, however, there may not be a clear “winner”. It is possible to have more than one transformation that generates acceptable diagnostic plots. To compare two or more “acceptable” models, we can use R-square (higher is better) or RMSE (root mean square error, smaller is better). These values can be obtained directly from the SAS output, and are shown in Table 1.11. The model that uses $\log(X)$ has the highest R-square and the lowest RMSE. This information, taken

in conjunction with the residual plots, indicates that the logarithmic model is the “best” of these three models.

Transformation	R-square	RMSE
Original data	0.9239	0.16092
Reciprocal	0.9019	0.19385
Logarithmic	0.9824	0.08214

Table 1.11. Goodness of fit measures for the three models

How do we get a curved linear model?

All three of the models we examined in the previous example are consider “linear” models even though two of them generated curved patterns between X and Y. When using transformed variables, the model will generate a straight line. When we revert back to the original (untransformed) variable, the pattern will become curved. For example, if we re-generate Figure 1.23, but use the values of $1/X$ on the x-axis, we would get the graph shown in Figure 1.26. If we re-generate the graph in Figure 1.25, but use $\log(X)$ on the x-axis, we would get Figure 1.27. Using the transformed values of X on the x-axis shows that the relationship is linear.

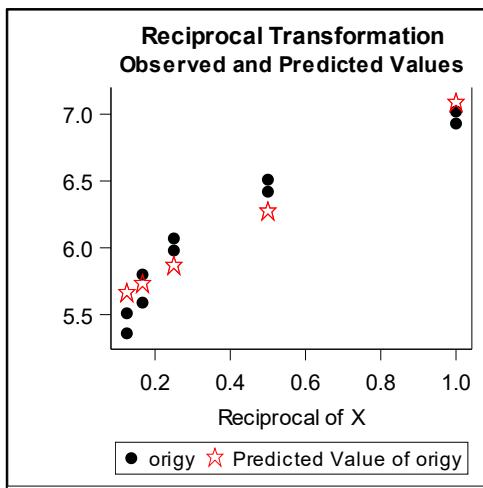


Figure 1.26. Compare to Figure 1.23

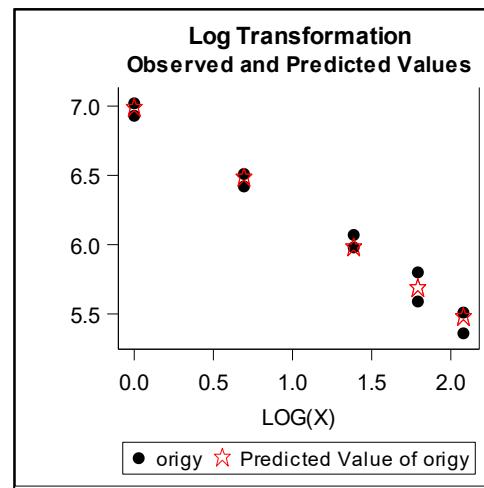


Figure 1.27. Compare to Figure 1.25

Example 2

A random sample of (X, Y) pairs are shown in the scatterplot, Figure 1.28. We want to develop a linear regression model to fit these data, but the relationship appears to be curved. The pattern is similar to the graph in lower right of Figure 1.18, and we will consider the suggested transformations. In addition, we will also fit a model to the original data and we will consider a model that uses a square root transformation.

In total, we will consider five models. We use SAS to fit each model and superimpose the regression line on a scatterplot of the transformed data. We also generate a residual plot and a QQ plot for each model. We will use these plots to compare and contrast the models.

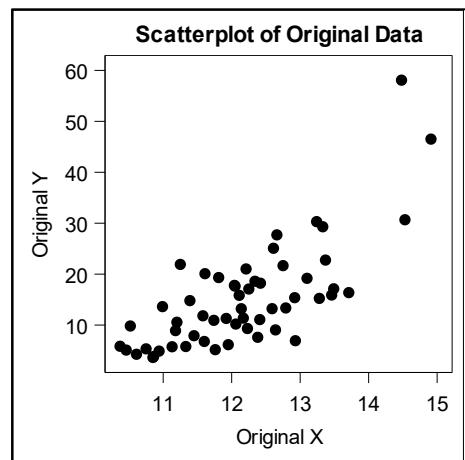


Figure 1.28. Scatterplot for Example 2

Model 1 (Original data): $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

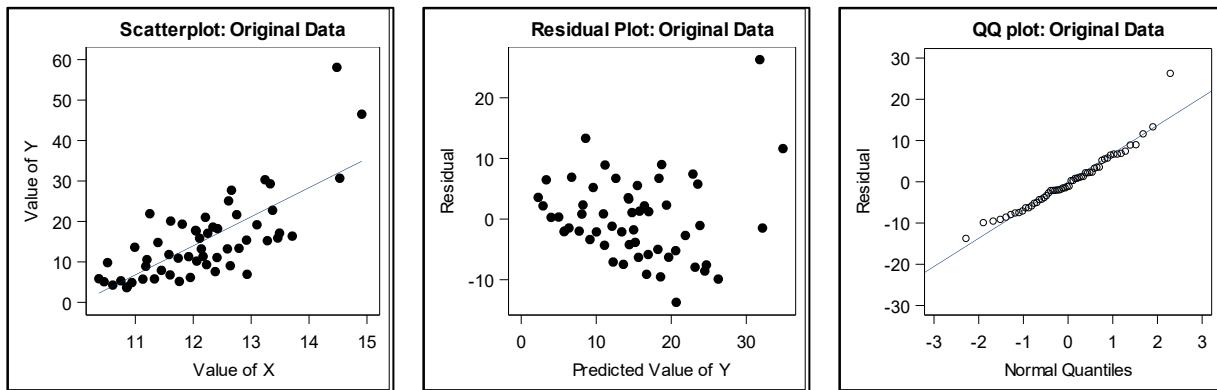


Figure 1.29. Diagnostic plots for Model 1

The scatterplot indicates a definite relationship between X and Y , but the relationship appears to be nonlinear. In addition, there is a wedge shape in the residual plot; the points on the left side of the residual plot have much smaller variation than the points on the right. This indicates that the assumption of equal variances may be violated. Except for one stray point, the QQ plot is excellent (points are following the line), but the pattern in the residual plot indicates we should look for another model.

Model 2 (Using X^2): $Y_i = \beta_0 + \beta_1 X_i^2 + \varepsilon_i$

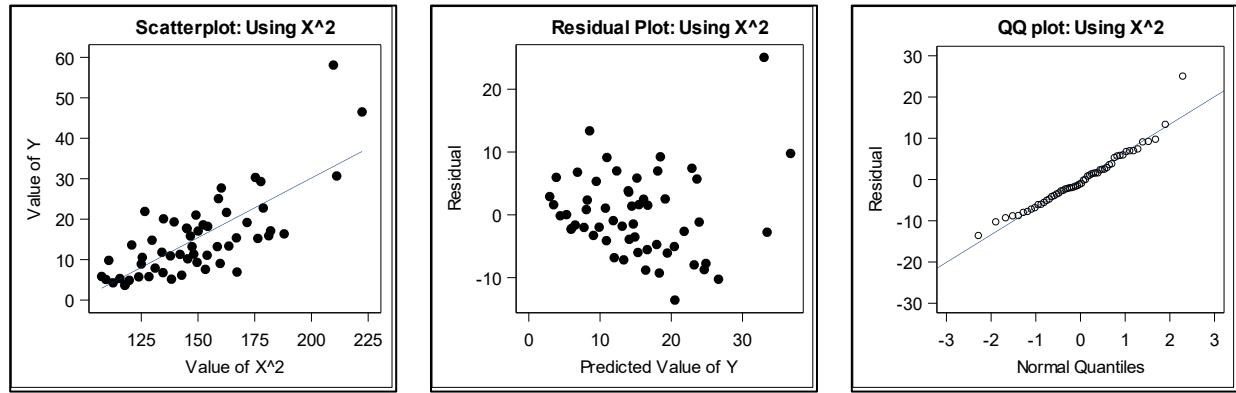


Figure 1.30. Diagnostic plots for Model 2

Although using X^2 was one of the suggested transformations, it does not seem to affect these plots other than to change the scale of the x-axis on the scatterplot. The scatterplot still shows a curved pattern, the residual plot still shows a wedge, and the QQ plot is fine. The pattern in the residual plot indicates we should look for another model.

Model 3 (Using $1/Y$): $\frac{1}{Y_i} = \beta_0 + \beta_1 X_i + \varepsilon_i$

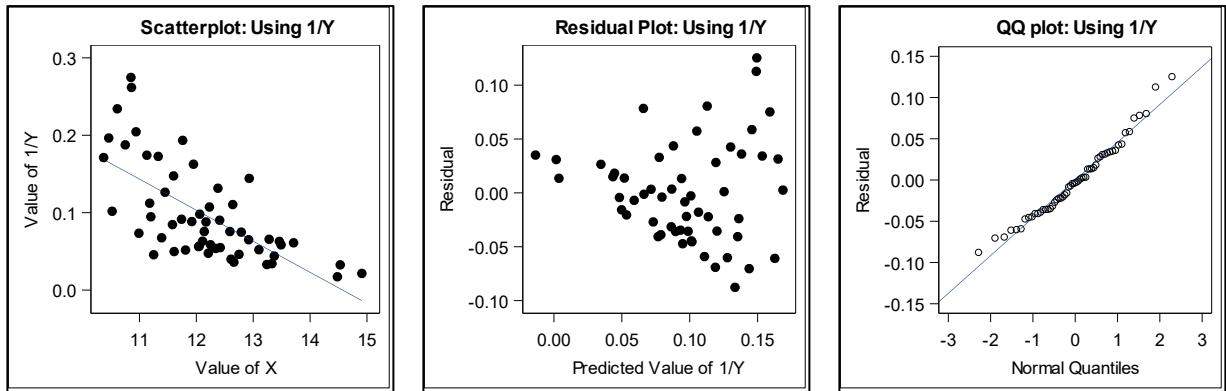


Figure 1.31. Diagnostic plots for Model 3

There is still a curve in the scatterplot, and the wedge shape in the residual plot has become more exaggerated. This is not a good model, even though the QQ plot looks fine.

Model 4 (Using $\log(Y)$): $\log(Y_i) = \beta_0 + \beta_1 X_i + \varepsilon_i$

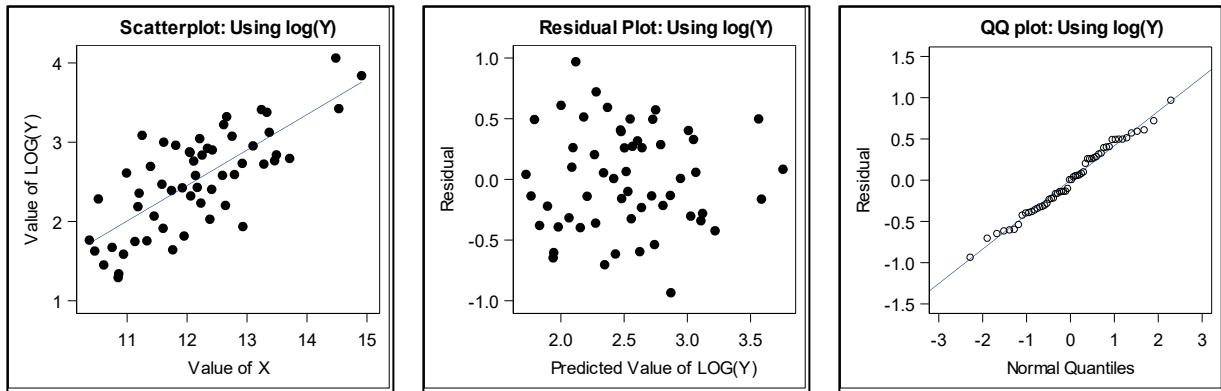


Figure 1.32. Diagnostic plots for Model 4

The log transformation has generated a better model. There are no major concerns with any of these plots. Using $\log(Y)$ instead of Y has straightened the curve in the scatterplot and dissipated the wedge shape in the residual plot. The QQ plot actually looks better than the other models, since the points at the end no longer stray away from the line. There is nothing in any of the these plots that give us cause for concern about this model.

Model 5 (Using \sqrt{Y}): $\sqrt{Y_i} = \beta_0 + \beta_1 X_i + \varepsilon_i$

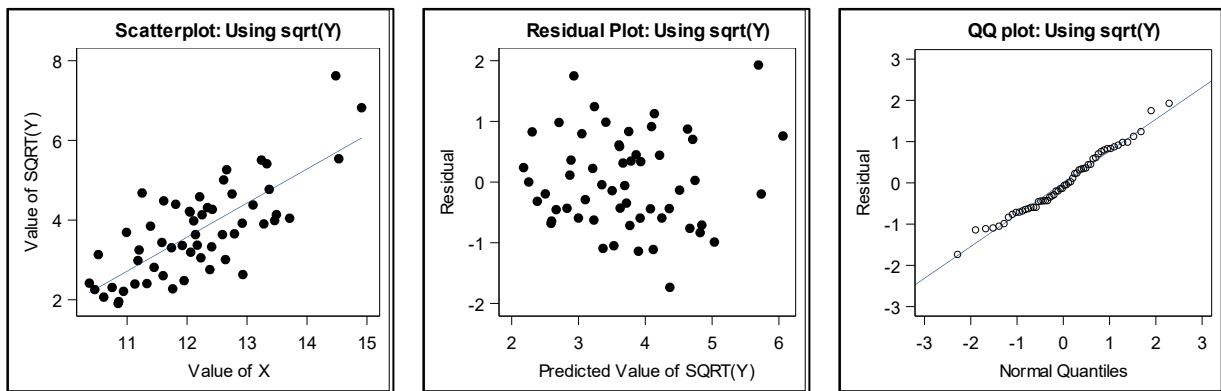


Figure 1.33. Diagnostic plots for Model 5

The square root transformation has also generated a perfectly acceptable model. The scatterplot shows a linear pattern, the residual plot has no pattern, and the points on the QQ plot are following the line.

Special considerations when Y is transformed

Examination of the diagnostic plots in Example 2 has revealed two acceptable models: one that uses a logarithmic transformation of Y and one that uses a square root transformation of Y. Choosing between these two models is a subjective decision, based on personal preference and the ease with which subsequent analysis can be interpreted. **We cannot use R-square or RMSE to compare these two models because they have different response variables.**

Since both of these models involve a transformation of the response variable, special care must be taken when reporting the results of either model. Specifically, any estimates for the response variable will be based on the transformed values. These include values for \hat{Y} and any confidence intervals or prediction intervals for \hat{Y} . When the results of the analysis are reported, these estimates need to be back-transformed so that they are on the same scale as the original data. It is not necessary to back-transform values for X , because estimates for X are not generated as part of a regression analysis.

Suppose, for example, that we want a point estimate for Y when $X = 12$. The log(Y) model from Example 2 generates the estimate 2.4534 and the square root model generates the estimate 3.5721. From the scatterplot of the original data (the graph on the left in Figure 1.29), we see that when $X = 12$, the Y values range from about 5 to about 20. How can the two models we generated produce estimates that are so much lower? The reason is that the estimates for "Y" generated by the log(Y) model are really for the natural log of Y, and the estimates for "Y" generated by the square root model are really for the square root of Y. The estimates need to be back-transformed to put them on the same scale as the original data.

To perform a back-transformation, use the inverse function of the original transformation. Some of the more common transformations are shown in Table 1.12.

- Using the log(Y) model, with $X = 12$
 - the estimate for $\log(Y)$ is 2.4534
 - the estimate for Y is $\exp(2.4534) = 11.63$.
- Using the square root model, with $X = 12$
 - the estimate for \sqrt{Y} is 3.5721
 - the estimate for Y is $(3.5721)^2 = 12.76$

If we had obtained confidence interval or prediction intervals estimates for Y from either of these two models, the endpoints of these intervals would need to be back-transformed as well.

Transformation	Back-transformation
$Y' = \log(Y)$	$Y = \exp(Y')$
$Y' = \sqrt{Y}$	$Y = (Y')^2$
$Y' = 1/Y$	$Y = 1/Y'$
$Y' = (Y)^2$	$Y = \sqrt{Y'}$
$Y' = (Y)^3$	$Y = \sqrt[3]{Y'} = (Y')^{1/3}$

Table 1.12. Common transformations and back-transformations

1.7.7. Summary

In this section, we identified some of the methods that are commonly used to detect inadequacy of a linear model. The inadequacy could be due to a violation of model assumptions, or it could be simply that a linear model is a poor fit to the data. In some cases, model inadequacies can be remedied by transforming one or both of the variables in the model. There is no guarantee that using transformed values will alleviate the inadequacy, so it imperative that the diagnostic plots for the new model be examined. For some datasets, it may be difficult to find even one “good” model, but for other datasets there might be many “good” models. If we want to compare two “good” models (that are fit to the same data) we can use the value of R-square (higher is better) or RMSE (lower is better), but these can only be used when the two model use the same Y values. If one model uses the original Y values and another model uses transformed Y values, then neither R-square nor RMSE can be used to compare these models.

Whenever transformed values for the response variable (Y) are used in a model, the estimates for Y generated by the model will be based on the transformed values. These estimates need to be back-transformed so that the values will be consistent with the original Y values that are in the data.

Always remember that we are simply looking for a “good” model. There is no such thing as a “perfect” model. George E.P. Box, a famous statistician, said it best: “All models are wrong, but some are useful.” We are simply looking for a useful model.

Section 1.8. Correlation Analysis

The correlation coefficient, denoted by r , measures the strength of the linear association between two numeric variables. The official name is “Pearson’s product-moment coefficient of correlation”, but it is usually simply called the correlation. There are other types of correlation coefficients, including Spearman’s and Kendall’s tau, but using the word “correlation” implies Pearson’s correlation unless specifically stated otherwise. For this course, we will always use Pearson’s correlation. The value for the correlation is always between -1 and 1. Values closer to 0 indicate weaker correlation and values closer to 1 or -1 indicate stronger correlation.

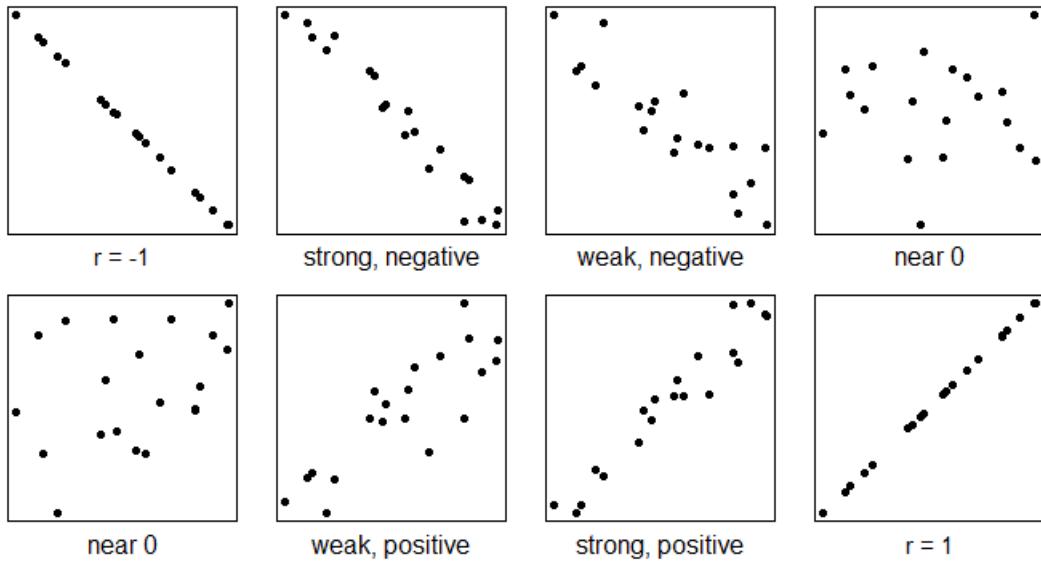


Figure 1.34. Examples of correlation

To visualize the correlation, consider the scatterplots in Figure 1.34. The graph in the upper left has perfect correlation because the points follow a line, with no deviation. The slope of the line is negative, so the correlation is $r = -1$. The next graph (to the right) also shows negative correlation, but there are minor deviations from a linear pattern, so this is strong negative correlation (perhaps $r = -0.8$). The next graph shows weaker, but still negative, correlation (perhaps $r = -0.4$). The last graph on the top does not show a linear pattern, so the correlation is near 0. The graphs in the bottom row show a similar pattern, from almost no correlation on the left to perfect correlation on the right. Since the slopes are positive, the correlations are also positive.

It is important to remember that the correlation is measuring the strength of the linear association between X and Y. It is possible for X and Y to be strongly related, but if the relationship is not linear then the correlation could be near 0. This is illustrated in Figure 1.35.

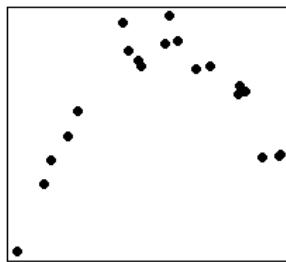


Figure 1.35. Nonlinear association

For simple linear regression, there are 3 basic ways to measure the strength of the linear relationship between X and Y.

$$1) \text{ coefficient of determination: } R^2 = \frac{SSTot - SSE}{SSTot} = \frac{SSReg}{SSTot}$$

$$2) \text{ correlation coefficient: } r = \frac{SS_{yy}}{\sqrt{SS_{xx} \cdot SS_{yy}}} = \hat{\beta}_1 \sqrt{\frac{SS_{xx}}{SS_{yy}}}$$

$$3) \text{ estimated slope: } \hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$$

Recall that R^2 measures the proportion of the variability in Y that is explained by the regression on X. For simple linear regression, $R^2 = r^2$.

1.8.1. T test for correlation

When we use a dataset to calculate the value for r , we are calculating the sample correlation. This is a point estimate for the “true” correlation that exists in the population, which is denoted by the Greek letter ρ (rho). Often, we want to perform a hypothesis test to decide if there is a linear association between X and Y. The hypotheses are

$$H_0 : \rho = 0 \quad (\text{There is not a linear association between X and Y.})$$

$$H_a : \rho \neq 0 \quad (\text{There is a linear association between X and Y.})$$

It is also possible to perform right-tailed tests (in which the alternative hypothesis is $\rho > 0$) or a left-tailed test (in which the alternative hypothesis is $\rho < 0$), but we will consider only two-tailed tests.

Every statistical test has certain assumptions that must be satisfied in order for the test to be valid. For the t test for correlation, we must assume that both X and Y are random variables and that the joint distribution for (X, Y) is bivariate normal. A typical bivariate normal distribution is shown in Figure 1.36. In essence, it is two normal distributions that are combined to form a 3D surface. If you make a vertical slice through the surface that is parallel to either of the axes, the result will be a normal curve. If the assumptions for the regression model are satisfied, then we can assume that the assumptions for this t test are also satisfied.

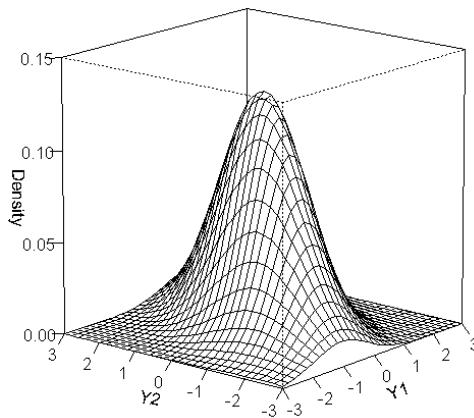


Figure 1.36. A bivariate normal distribution

To perform the hypothesis test, we need the test statistic and the critical value. The test statistic is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad (1.17)$$

When the assumptions are satisfied, the test statistic follows a t distribution with $n - 2$ degrees of freedom, where n is the number of (X, Y) pairs in the data. We illustrate the calculations using the NASA rocket propellant data. There are $n = 20$ (X, Y) pairs in the data, and the required sums are

$$SS_{XY} = -41,112.654, SS_{XX} = 1106.559, \text{ and } SS_{YY} = 1,693,737.601.$$

This produces the following correlation coefficient and test statistic.

$$\text{correlation coefficient: } r = \frac{SS_{XY}}{\sqrt{SS_{XX} \cdot SS_{YY}}} = \frac{-41,112.654}{\sqrt{1106.559 \times 1,693,737.601}} = -0.94965$$

$$\text{test statistic: } t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0.94965\sqrt{20-2}}{\sqrt{1-(-0.94965)^2}} = -12.86$$

To complete this test, we need to know the critical value. Since we are performing a two-tailed test at significance level $\alpha = 0.05$, we split α in half and use 0.025. From the t table with degrees of freedom 18 and $\alpha/2 = 0.025$, the critical value is 2.101. The absolute value of the test statistic is greater than the critical value, ($12.86 > 2.101$) so we reject H_0 . The sample data provides convincing evidence that there is a linear relationship between X and Y.

We can also use SAS to perform this test. In addition, SAS can generate a scatterplot matrix, which is a group of scatterplots (arranged row-by-column, like a matrix) for several numeric variables. To instruct SAS to produce this information, we use PROC CORR. Once the SAS dataset has been created, the SAS code is

```
PROC CORR DATA=nasa PLOTS=MATRIX;  
  VAR ShearStrength Age;  
  RUN;
```

This generates the output shown in Figure 1.37. The table of simple statistics provides a summary of each variable in the VAR statement, but we are most interested in the next table which is the correlation matrix. There is one row and one column for each variable in the VAR statement. Each cell contains three numbers: (1) the correlation, (2) the p-value for the t test, and (3) the number of (X, Y) pairs that were used to perform the test. The diagonal entries (from the upper left to the lower right) will always have correlation equal to 1, since this is the correlation between a variable and itself. The entries in the upper right cell will always be the mirror image of the entries in the lower left cell, since the correlation between X and Y is the same as the correlation between Y and X. The t test that we just conducted by hand is reported in the upper right cell (or the lower left cell). The correlation is -0.94965, which matches exactly what we calculated by hand. The p-value for this test is $< .0001$, and $n = 20$ pairs of (X, Y) values were used. Since the p-value is less than α , we reject H_0 and arrive at the same conclusion we obtained by hand.

The additional option PLOTS=MATRIX on the PROC CORR statement generates the scatterplot matrix. The plot in the upper right corner has Age on the x-axis and ShearStrength on the y-axis. The plot in the lower left corner has ShearStrength on the x-axis and Age on the y-axis. When there are only two variables, the scatterplot matrix does not provide any more information than a regular scatterplot. But when there are multiple X variables (in Chapter 2), these plots will be more important.

The CORR Procedure

2 Variables: ShearStrength Age

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
ShearStrength	20	2131	298.57007	42627	1678	2654
Age	21	13.20238	7.47438	277.25000	2.00000	25.00000

Pearson Correlation Coefficients		
Prob > r under H0: Rho=0		
Number of Observations		
	ShearStrength	Age
ShearStrength	1.00000	-0.94965 <.0001 20
Age	-0.94965 <.0001 20	1.00000 21

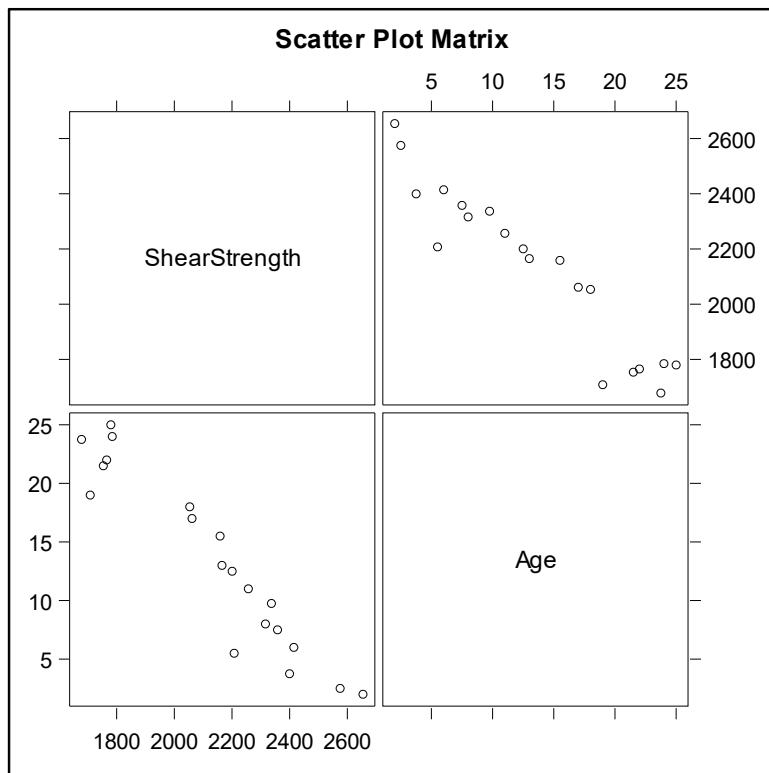


Figure 1.37. SAS output for PROC CORR

1.8.2. Fisher's Z test for correlation

The t test for correlation can be used only when the null hypothesis is $\rho = 0$. If we want to test whether or not the correlation is equal to a specific value, say 0.5, then we cannot use the t test. Instead, we can use Fisher's Z test. For this test, the hypotheses are

$$H_0 : \rho = \rho_0 \text{ vs. } H_a : \rho \neq \rho_0$$

where ρ_0 is a known value between -1 and 1. To perform this test, we need to transform both ρ_0 and the sample correlation (r), then calculate a test statistic using the transformed values.

- Transform r : $r' = \frac{1}{2} \log_e \left(\frac{1+r}{1-r} \right)$
- Transform ρ : $\rho'_0 = \frac{1}{2} \log_e \left(\frac{1+\rho_0}{1-\rho_0} \right)$
- Test statistic: $Z = \sqrt{n-3} (r' - \rho'_0)$

The test statistic follows an approximate standard normal distribution, and we reject H_0 if the absolute value of the test statistic is greater than the critical value. The critical value is obtained from the normal probability table, using $\alpha/2 = 0.025$. (We split α in half because this is a two-tailed test.)

Example: Fisher's Z test

The body weights of 100 fathers and their first-born sons are measured, resulting in a sample correlation $r = 0.38$. Is this compatible with an underlying correlation of 0.5 that might be expected under genetic theory (i.e., Mendelian sampling)?

We are testing the hypotheses

$$H_0 : \rho = 0.5 \quad (\text{Weights of fathers and first-born sons follow genetic theory.})$$

$$H_a : \rho \neq 0.5 \quad (\text{Weights of fathers and first-born sons do not follow genetic theory.})$$

We are using $\rho_0 = 0.5$, with $r = 0.38$ and $n = 100$, so we perform the following calculations

- Transform r : $r' = \frac{1}{2} \log_e \left(\frac{1+0.38}{1-0.38} \right) = 0.40$
- Transform ρ : $\rho'_0 = \frac{1}{2} \log_e \left(\frac{1+0.5}{1-0.5} \right) = 0.549$
- Test statistic: $Z = \sqrt{100-3} (0.400 - 0.549) = -1.47$

The critical value is 1.96, which is obtained from the normal probability table with $\alpha/2 = 0.025$. The absolute value of the test statistic is 1.47, and this is not greater than the critical value, so we do not reject H_0 . We conclude that the sample does not provide enough evidence to suspect that the body weight of fathers and their first-born sons do not follow genetic theory.

1.8.3. An application: consistency of judges

Suppose there are several raters (judges) who will evaluate a number of objects. For example, the objects could be

- contestants in a beauty contest
- figure skaters in the Olympics
- flowers in a garden show
- cookies made from a new recipe

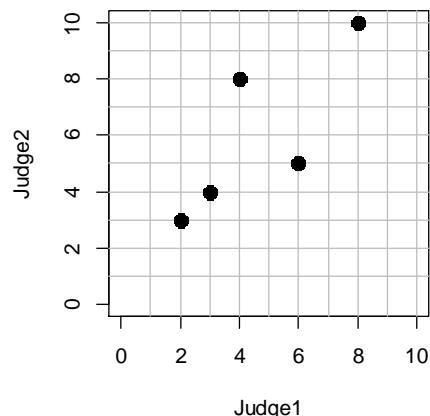
Many of these types of evaluations are subjective, and we are interested in assessing whether the judges' ratings are consistent, or if there are one (or more) judges whose evaluations are somehow different from the others.

To illustrate, suppose there are only two judges, and they each evaluate 5 objects. For each object, each judge assigns a number between 1 and 10, where 1 is least desirable and 10 is most desirable. The judges ratings are given in the following table.

	Object 1	Object 2	Object 3	Object 4	Object 5
Judge 1	2	6	3	4	8
Judge 2	3	5	4	8	10

A graph may help put this data in perspective.

The (X, Y) pairs are (Judge1, Judge2) for each object: (2, 3), (6, 5), (3, 4), (4, 8) and (8, 10). If the two judges have similar rating for the objects, then these points should be (approximately) on a line. We can use correlation to measure the strength of the linear relationship.



To decide if there is a linear relationship between these two sets of scores, we can use the t test automatically generated by PROC CORR. The code and relevant output are shown below.

```
DATA example;
INPUT one two @@;
TITLE 'Two Judges';
DATALINES;
2 3      6 5      3 4      4 8      8 10
;
PROC CORR DATA=example;
RUN;
```

Pearson Correlation Coefficients, N = 5		
Prob > r under H0: Rho=0		
	one	two
one	1.00000	0.78332 0.1171
two	0.78332 0.1171	1.00000

The correlation is 0.78332, and the t test has p-value 0.1171. Since the p-value is larger than 0.05, we do not reject the null hypothesis. We conclude that there is not a linear relationship between the scores of these two judges.

The default in SAS is to calculate Pearson's product-moment correlation coefficient, which measures the strength of the linear relationship between the judge's scores. If one judge gives a wider range of score than the other judge, then the relationship might not be linear. In this case, it would be advantageous to use a different method for calculating the correlation. The nonparametric alternative to Pearson's method is based on ranks, and it called Spearman's correlation coefficient.

To calculate Spearman's correlation, the actual scores are converted to ranks. For each judge, the smallest value has rank 1, the second smallest value has rank 2, etc. The ranks for our example data are shown in the square brackets in the table below.

	Object 1	Object 2	Object 3	Object 4	Object 5
Judge 1	2 [1]	6 [4]	3 [2]	4 [3]	8 [5]
Judge 2	3 [1]	5 [3]	4 [2]	8 [4]	10 [5]

Spearman's correlation uses the same formula as Pearson's correlation, except that the actual values are replaced by the ranks. To get SAS to calculate Spearman's correlation, use the option SPEARMAN on the PROC CORR statement. For example,

```
PROC CORR DATA=example SPEARMAN;
RUN;
```

Spearman Correlation Coefficients, N = 5		
Prob > r under H0: Rho=0		
	one	two
one	1.00000	0.90000 0.0374
two	0.90000 0.0374	1.00000

Spearman's correlation measures the monotone relationship between the judges' scores. (A monotone relationship means that as one value goes up, the other value also goes up. This is not as strict as a linear relationship.) For the example data, Spearman's correlation is 0.90000, and the p-value is 0.0374. Since the p-value is smaller than 0.05, we reject H_0 and conclude that there is a monotone relationship between the scores of these two judges. In other words, the judges' scores are consistent.

Example: Consistency of Five Judges

We now consider a more complicated situation. Suppose that there are 5 judges, and each judge evaluates the 6 projects (the same 6 projects for each judge). We want to assess the consistency of the judges. The data are shown in Table 1.13.

	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5
Project 1	5.0	3.2	4.0	5.4	3.8
Project 2	4.0	3.6	5.6	5.1	5.8
Project 3	4.4	4.2	5.8	5.8	6.4
Project 4	5.0	6.0	6.2	6.2	6.0
Project 5	5.2	6.4	6.8	7.0	7.2
Project 6	6.2	8.0	7.4	7.4	7.6

Table 1.13. Ratings for five judges

Since correlation is only defined for pairs of variables, we will need to calculate the correlation between each pair of judges. In other words, we need the correlation matrix. The code to read this data into SAS and generate both Pearson's and Spearman's correlations is shown below. Note that PROC CORR can calculate both Pearson's and Spearman's correlations at the same time.

```

DATA pairwise;
INPUT project judge1 judge2 judge3 judge4 judge5;
DATALINES;
1 5.0 3.2 4.0 5.4 3.8
2 4.0 3.6 5.6 5.4 5.8
3 4.4 4.2 5.8 5.8 6.4
4 5.0 6.0 6.2 6.2 6.0
5 5.2 6.4 6.8 7.0 7.2
6 6.2 8.0 7.4 7.4 7.6
;
PROC CORR DATA=pairwise PEARSON SPEARMAN;
VAR judge1 -- judge5;
RUN;

```

Pearson Correlation Coefficients, N = 6 Prob > r under H0: Rho=0					
	judge1	judge2	judge3	judge4	judge5
judge1	1.00000 0.0431	0.82526 0.2742	0.53483 0.0132	0.83596 0.0205	0.42669 0.0008
judge2	0.82526 0.0431	1.00000 0.0132	0.90479 0.0014	0.96907 0.0205	0.81747 0.0008
judge3	0.53483 0.2742	0.90479 0.0132	1.00000 0.0008	0.88084 0.0205	0.97614 0.0008
judge4	0.83596 0.0382	0.96907 0.0014	0.88084 0.0205	1.00000 0.0008	0.82745 0.0421
judge5	0.42669 0.3988	0.81747 0.0469	0.97614 0.0008	0.82745 0.0421	1.00000 0.0008

Table 1.14. Pearson correlation for five judges

Spearman Correlation Coefficients, N = 6 Prob > r under H0: Rho=0					
	judge1	judge2	judge3	judge4	judge5
judge1	1.00000 0.0835	0.75370 0.0835	0.75370 0.0835	0.83824 0.0371	0.66674 0.1481
judge2	0.75370 0.0835	1.00000 0.0001	1.00000 <.0001	0.98561 0.0003	0.94286 0.0048
judge3	0.75370 0.0835	1.00000 <.0001	1.00000 0.0003	0.98561 0.0003	0.94286 0.0048
judge4	0.83824 0.0371	0.98561 0.0003	0.98561 0.0003	1.00000 0.0003	0.92763 0.0077
judge5	0.66674 0.1481	0.94286 0.0048	0.94286 0.0048	0.92763 0.0077	1.00000 0.0008

Table 1.15. Spearman correlation for five judges

Two judges are consistent if the correlation between their ratings is close to 1. This is true for all judges except judge 1. If we use Pearson's correlation (in Table 1.14), notice that all correlations are above 0.8 except for judges 1 and 3 and for judges 1 and 5. So judge 1 seems to be different than judges 3 and 5, but the other four judges seem to be consistent. If we use Spearman's correlation (in Table 1.15), we arrive at a similar conclusion. All of the Spearman correlations are above 0.8 except for judge 1 with judge 2, 3 or 5. Again, it seems that judge 1 assigns different ratings than the other judges.

We can also arrive at the same conclusion by looking at the p-values provided in the correlation matrix. These p-values are for testing $H_0: \rho = 0$ vs. $H_a: \rho \neq 0$. Consistent judges will have p close to 1, if the two

judges are providing consistent ratings we expect to reject the null hypothesis. Using Pearson's correlation, this is not the case for judge 1 and judge 3 ($p = 0.2742$) or for judge 1 and judge 5 ($p = 0.3988$). In both of these pairings, we fail to reject the null hypothesis that the correlation is zero.

For this particular data set, it seems that judge 1 is not consistent with the other judges.

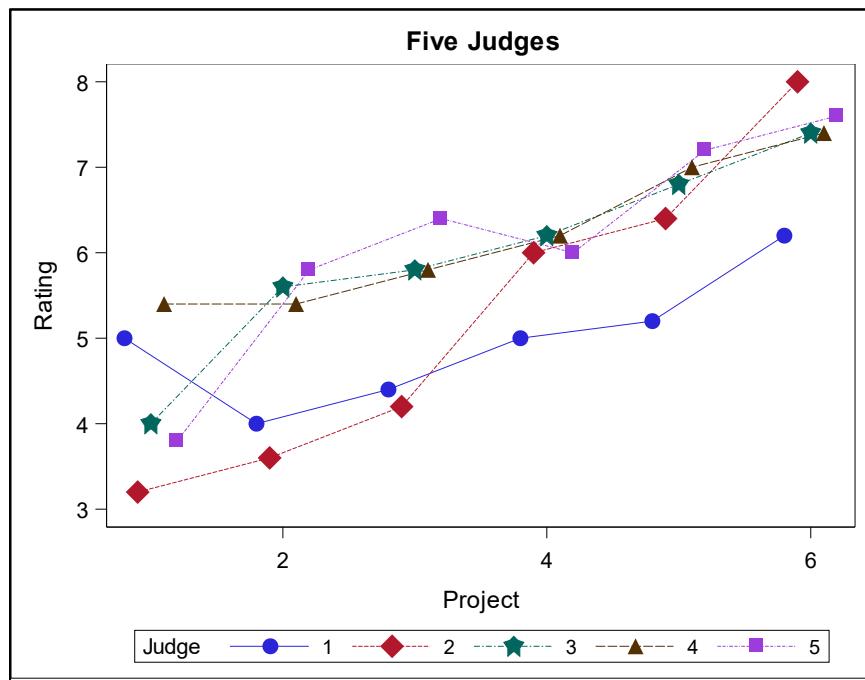


Figure 1.38. Line graph of five judges' scores

A plot of the data, shown in Figure 1.38, may help us understand why judge 1 is not consistent. Judge 1 (whom we have already identified as being “different”) has lower ratings for most projects except for project 1. All of the other judges have increasing ratings across the projects (except for a slip dip for project 4 and judge 5). Judge 1 is different from the other judges primarily because of the rating for project 1, which appears to be higher than we would expect if all the judges were consistent.

1.8.4. Summary

In this section, we have explored using correlation as a mechanism for assessing the relationship between two numeric variables. There are two types of hypothesis tests for correlation: one that uses a t distribution and one that uses the standard normal (Z) distribution. The test utilizing the t distribution can be generated by SAS, but the other cannot. These tests are closely related to the tests for the slope parameter in a simple linear regression model. The t test for correlation requires that the (X, Y) pairs

follow a bivariate normal distribution. If this is not the case, then the nonparametric correlation (Spearman's) should be used.

Correlation is not causation. Simply because two variables X and Y have strong correlation, this does not mean that X causes Y , nor does it mean that Y causes X . Consider this example. The per capita consumption of soft drinks in the U.S. has increased steadily since 1950. The rate of obesity in the U.S. has also increased steadily since 1950. These two variables are highly correlated. This does not imply that soft drink consumption causes obesity, nor does it imply that obesity causes soft drink consumption. In order to determine causation, we need to isolate other factors that might be affecting both of these two variables. Such factors might include consumption of fast food, which might be positively correlated with both of these variables, or perhaps the amount of daily exercise, which might be negatively correlated with both of these variables. In order to determine causation, these "other" factors must be controlled, and this occurs when the data are generated as a result of an experimental study. Causation cannot be determined when the data are derived from an observational study.

Chapter 2: Multiple Linear Regression

Section 2.1. Introduction

In the previous chapter, we were working with regression models that have exactly one predictor variable. These are called “simple” linear regression models. We now consider models that have two or more predictors. These are called “multiple” linear regression models. The obvious difficulty with a simple linear regression model is that it is likely to be too simplistic to realistically capture complex relationships, and therefore it can fail to provide an adequate description of the behavior of the response variable. Including additional predictor variables can mitigate this inadequacy. In addition, multiple linear regression models can

- model curved relationships between Y and X’s
- attain much more precise inference
- accommodate non-numeric (qualitative) predictors
- incorporate interactions between predictors

We begin our discussion with multiple linear regression models that contain exactly two predictors, then we generalize the concepts to include additional predictors.

2.1.1. Multiple regression with two predictors

When there are two predictor variables, the linear regression model is written

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \quad i = 1, \dots, n \quad (2.1)$$

where

Y_i = observed value of the response variable for observation i

X_{1i} = value of the first predictor variable recorded for observation i

X_{2i} = value of the second predictor variable recorded for observation i

β_0 = the intercept

β_1 = the slope on X_1

β_2 = the slope on X_2

ε_i = the error for observation i

As we did with simple linear regression, we assume that $\varepsilon_i \sim \text{NIID}(0, \sigma^2)$, that is, the errors are normal and independent, with constant variance σ^2 . Note that the β 's do not depend on i , because the intercept and both slopes apply to the entire population, not just to one observation in the dataset.

When there are two predictors, the regression LINE becomes a two dimensional regression PLANE, as illustrated in Figure 2.1Figure . One observation is one point in the three-dimensional space, and the predicted Y for this observation is the point on the plane directly above (or below) the observation. The residual for each observation is the vertical distance from the point to the plane.

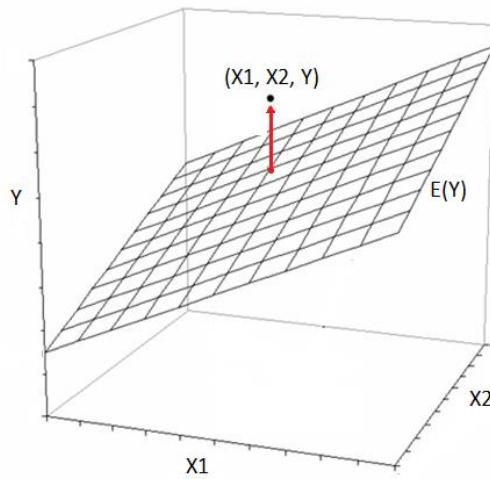


Figure 2.1. A regression plane

2.1.2. Least squares estimation

Least squares estimation of the β 's will be based on minimizing the squares of the vertical distances between the observations and the plane, i.e., the sum of the squared residuals. In mathematical notation, we want to minimize

$$Q = \sum_{i=1}^n r_i^2 = \sum_{i=1}^n \left\{ Y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}) \right\}^2 \quad (2.2)$$

In order to do this “by hand”, we would need to get the partial derivatives of Q with respect to each β , set each derivative equal to zero, then solve the simultaneous equations for the β 's. This should sound similar, because it is the same procedure as for simple linear regression. We will not be finding partial

derivatives or solving simultaneous equations in this course. We will rely on SAS to perform these calculations.

The estimators for the β 's are designated by $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ and the equation for the regression plane is

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2.$$

The population parameters (β_0 , β_1 , and β_2) are constants. They are not random variables and they do not have probability distributions. The specific values that are calculated from the random sample are the point estimates for the population parameters. The estimators ($\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$) are random variables, so they each follow a probability distribution.

Because the estimators are derived by solving equation (2.2), they are least squares estimators. From the Gauss-Markov Theorem (in Chapter 1), all least squares estimators possess two very desirable statistical properties. First, they are unbiased estimators. This means that, under repeated sampling, the expected value of the estimator is equal to the corresponding population parameter. Second, least squares estimators have the smallest variance among all unbiased estimators. Together, these two properties tell us that, on average, we are aiming at the right target (unbiased) and we should not miss it by much (low variance). This is the statistical “gold standard” of estimators.

The estimators are NOT independent because they are each calculated from the same sample data. The dependence among estimators is recorded in a variance-covariance matrix. Every variance-covariance matrix is symmetric, that is, the entries in the upper right are mirrored in the lower left, and the diagonal entries are the variances of the sampling distributions for each estimator. For an estimated model with two predictors (plus an intercept), the variance-covariance matrix has 3 rows and 3 columns.

$$\begin{bmatrix} \text{var}(\hat{\beta}_0) & \text{cov}(\hat{\beta}_0, \hat{\beta}_1) & \text{cov}(\hat{\beta}_0, \hat{\beta}_2) \\ \text{cov}(\hat{\beta}_0, \hat{\beta}_1) & \text{var}(\hat{\beta}_1) & \text{cov}(\hat{\beta}_1, \hat{\beta}_2) \\ \text{cov}(\hat{\beta}_0, \hat{\beta}_2) & \text{cov}(\hat{\beta}_1, \hat{\beta}_2) & \text{var}(\hat{\beta}_2) \end{bmatrix}$$

A covariance is similar to a correlation, but it is not restricted to be between -1 and 1. Larger values (either positive or negative) indicate a stronger relationship between the two $\hat{\beta}$'s, and values that are closer to 0 indicate a weaker relationship.

The variance-covariance matrix is important because hypothesis tests and confidence/prediction intervals require not only the point estimates, but they also require the standard errors of the estimates. The standard errors incorporate both the variances and covariances of the estimates. In addition, extremely large or extremely small values in this matrix can produce numerical instability (like dividing by 0). Ignoring the covariance leads to incorrect standard errors, and therefore faulty inference. The values for the standard errors that are reported by SAS will automatically incorporate the variances and covariances, so we will need to examine the variance-covariance matrix only if there is a problem.

2.1.3. Correlation and scatterplot matrix

When we had a single predictor variable, we began each regression analysis with a scatterplot of the data. When there are multiple predictors, we use a scatterplot matrix. This is simply a grid of scatterplots that plot each X against every other X and each X against Y. When there are two X's, we have a 3x3 grid of scatterplots, as shown in Figure 2.2. The three graphs in the upper right are mirror images of the three graphs in the lower left.

The labels along the diagonal of the matrix indicate which variable is being plotted on each axis in the corresponding scatterplots. The label for Y does not have to be in the lower right corner, and the X's do not have to be in order. These labels can be in any order along the diagonal.

Since the label X_1 is in row 1 and column 1, all the scatterplots in row 1 (to the right of the label) have X_1 plotted on the y-axis and all the scatterplots in column 1 (below the label) have X_1 plotted on the x-axis. The label X_2 is in row 2 and column 2, so all the scatterplots in row 2 (right and left of the label) have X_2 plotted on the y-axis and all the scatterplots in column 2 (above and below the label) have X_2 plotted on the x-axis. Similarly, all the scatterplots in row 3 have Y on the y-axis and those in column 3 have Y on the x-axis.

Since we are accustomed to having Y on the y-axis, we will concentrate on the three graphs in the lower left. The ones in the upper right are mirror images. For the two graphs in the bottom row, the first graph is plotting (X_1, Y) and the second one is plotting (X_2, Y) . Since each of these graphs is plotting a predictor against the response, we want to see a strong linear relationship in each of the graphs. It appears that the linear relationship between X_1 and Y is stronger than the relationship between X_2 and Y. This implies that X_1 may be a better predictor than X_2 . The graph in the middle on the left is plotting X_1 (on the x-axis) against X_2 (on the y-axis). We do not want to see a strong relationship between these

two variables because they are both predictors. In Figure, it appears that there is a moderate relationship (not very weak and not very strong).

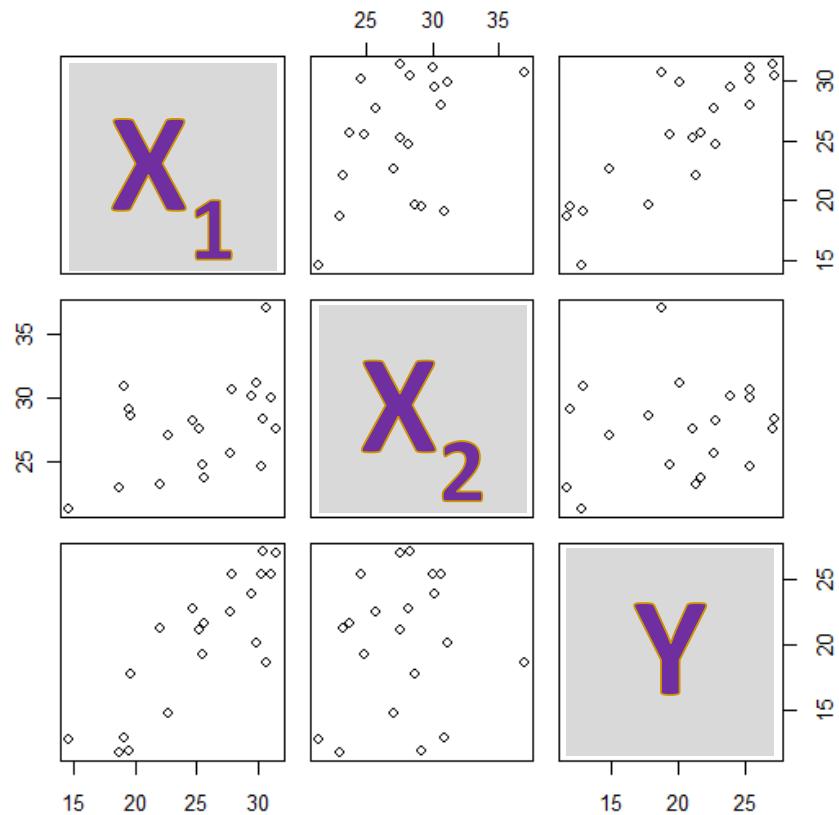


Figure 2.2. A scatterplot matrix

The information gleaned from the scatterplot matrix can also be obtained from a correlation matrix. When there are two predictors, the correlation matrix will contain 3 rows and 3 columns (like the scatterplot matrix). The format of the correlation matrix was discussed in Chapter 1 (Figure 1.37), and we will provide an example in the next section. In many cases, a scatterplot matrix is preferred over the correlation matrix because scatterplots can also reveal patterns (e.g., curves) that are not readily apparent by looking only at the correlation. For example, if the scatterplot of (X_1, Y) had a quadratic pattern, the correlation would be close to 0 (and this is not very informative). But seeing a quadratic pattern in the scatterplot would indicate that including an $(X_1)^2$ term might improve the model.

2.1.4. 3D scatterplots

When there are exactly two predictors, we can plot the data in three dimensions. This type of graph could be used instead of the scatterplot matrix, but it will not applicable if there are more than two predictors. A 3D scatterplot is shown in Figure 2.3. The vertical axis is the response and the two horizontal axes are the two predictors: X_1 and X_2 . Each point represents an (X_1, X_2, Y) triple from one row in the dataset.

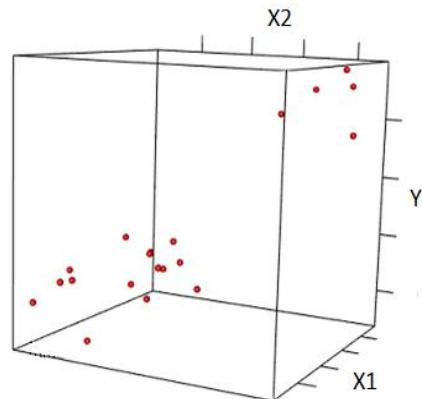


Figure 2.3. A 3D scatterplot

The least squares procedure fits a surface (a plane) through these points.

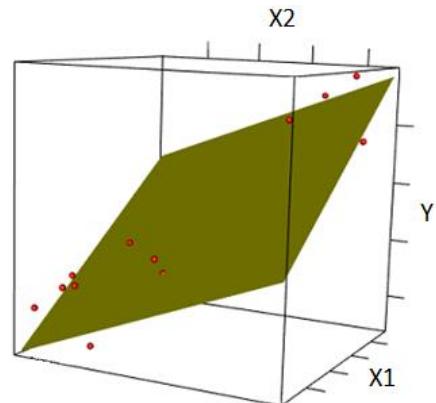


Figure 2.4. A regression plane

If the graph is rotated clockwise, points above and below the least squares plane are visible.

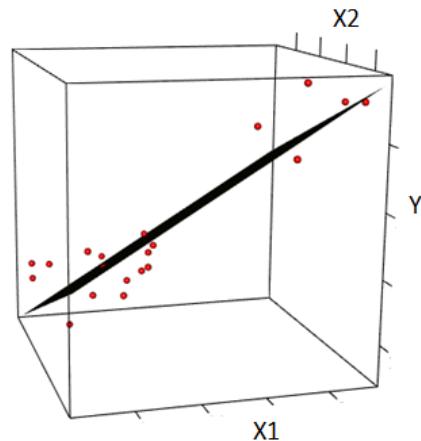


Figure 2.5. The regression plane, rotated

Residuals are the vertical distance from each point to the plane. There is one residual for each point (i.e. each observation, each row in the dataset). Points above the plane have positive residuals and points below the plane have negative residuals. The location of the least squares plane is guaranteed to minimize the sum of the squared residuals.

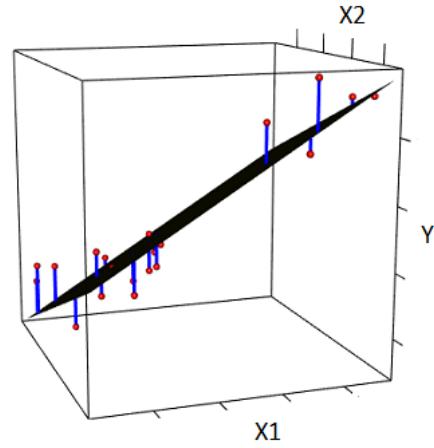


Figure 2.6. Residuals in 3D

There is no theoretical limit to the number of predictors that can be in a multiple regression model. The only limitations are the resources needed to collect the data and the amount of computer power needed to perform the calculations. If there are p predictors, the regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_p X_{pi} + \varepsilon_i \quad i = 1, \dots, n \quad (2.3)$$

The estimated equation can no longer be thought of as a regression plane because the data will not fit in three dimensions. We will call it a regression “surface”. Estimates for the parameters (the β ’s) are obtained by solving the least squares equation:

$$\text{minimize } Q = \sum_{i=1}^n r_i^2 = \sum_{i=1}^n \left\{ Y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_p X_{pi}) \right\}^2 \quad (2.4)$$

As with equation (2.2), solving this equation involves taking partial derivatives and solving simultaneous equations. We will use SAS to perform the calculations.

2.1.5. Summary

Multiple regression models are simply an extension of the simple linear regression models discussed in Chapter 1. The assumptions are the same, and the parameter estimates are still based on the least squares criteria. When the assumptions are satisfied, the least squares criteria still produce the “best” estimators (unbiased, with minimum variance). We can use the correlation matrix or a scatterplot matrix to help decide which predictors might be “important” in a model, but this is not the only consideration.

Section 2.2. Body Fat Example

In a long-term study on obesity, researchers are interested in the relationship between body fat and morphological measurements. Body fat is expensive to measure accurately, and it is hoped that there could be other, less expensive, measurements that can be used to reliably estimate body fat. Data was collected from a random sample of 20 healthy females, all 25 to 34 years of age.

The data, shown in Table 2.1, contains the following variables:

predictor #1: X_1 = triceps skinfold thickness (in mm)
predictor #2: X_2 = midarm circumference (in cm)
response: Y = Body fat (in points)

We want to use the data to fit the model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \quad i = 1, \dots, n$$

The SAS code is as follows:

```
/* ... data step goes here ... */  
  
TITLE 'Correlation Analysis';  
PROC CORR DATA=fat PLOTS=MATRIX;  
  VAR triceps midarm bodyfat;  
RUN;  
  
TITLE 'Predictors are triceps and midarm';  
PROC REG DATA=fat PLOTS=DIAGNOSTICS;  
  MODEL bodyfat = triceps midarm / CLB COVB P CLM CLI;  
RUN;
```

Triceps	Midarm	Body Fat
19.5	29.1	11.9
24.7	28.2	22.8
30.7	37	18.7
29.8	31.1	20.1
19.1	30.9	12.9
25.6	23.7	21.7
31.4	27.6	27.1
27.9	30.6	25.4
22.1	23.2	21.3
25.5	24.8	19.3
31.1	30	25.4
30.4	28.3	27.2
18.7	23	11.7
19.7	28.6	17.8
14.6	21.3	12.8
29.5	30.1	23.9
27.7	25.7	22.6
30.2	24.6	25.4
22.7	27.1	14.8
25.2	27.5	21.1

Table 2.1. Body fat data

2.2.1. Examine the correlations

We will use the output from PROC CORR to examine the strength of the relationship between the response and each of the two predictors, and to evaluate the strength of the relationship between the two predictors. In the SAS code, the option “PLOTS=MATRIX” will generate scatterplot matrix. The code also contains TITLE statements. These are optional, but recommended. They can serve as comments in the code, and they also appear in the printed output.

We have included several options on the MODEL statement (after the slash). We have already seen the options P, CLM and CLI, which produce estimates for, respectively, the predicted value for Y, the confidence limits for the mean of Y, and the prediction limits for an individual Y. There are also two new options on the MODEL statement. These are CLB and COVB, which provide confidence limits for the β 's and the variance-covariance matrix for the $\hat{\beta}$'s. A concise regression analysis does not always include these two options, but we are including them in our first example of multiple regression.

We begin the regression analysis by looking at the correlation matrix, shown in Table 2.2. We are concerned with three correlations in this table.

- (1) Between the response and one of the predictors (bodyfat and triceps).

This has the highest correlation at 0.84327. The p-value is <.0001, so we should reject the notion that this correlation could be 0. We conclude that there is a fairly strong linear relationship between the response and triceps.

- (2) Between the response and the other predictor (bodyfat and midarm).

This correlation is only 0.14244, which is pretty low. The p-value is 0.5491, so we would NOT reject the hypothesis that this correlation could be 0. This indicates that there may not be a linear relationship between the response and midarm.

- (3) Between two predictors (midarm and triceps).

This correlation is 0.50102 and the p-value (0.0207) indicates that this correlation is not 0. If the correlation between the predictors is greater than 0.7 (or so), we might have problems with the fitted model. (This will be discussed in more detail later.) The correlation between our two predictors is less than 0.7, so this should not be an issue with the current model.

Pearson Correlation Coefficients Prob > r under H0: Rho=0 Number of Observations			
	triceps	midarm	bodyfat
triceps	1.00000 21	0.50102 0.0207 21	0.84327 <.0001 20
midarm	0.50102 0.0207 21	1.00000 21	0.14244 0.5491 20
bodyfat	0.84327 <.0001 20	0.14244 0.5491 20	1.00000 20

Table 2.2. Correlation matrix for the body fat data

Based on the correlation matrix, we can infer that the predictor triceps is more strongly correlated with body fat than is midarm (0.84 vs. 0.14). This implies that triceps is more likely to be a better predictor than midarm. We should definitely include triceps in the model, but the inclusion of midarm is not as certain. Even though midarm is weakly correlated with body fat (0.14), it may still be useful in the model. This is consistent with information in the scatterplot matrix, which was first shown in Figure 2.2 and is reproduced in Figure 2.7 for convenience.

As with the correlation matrix, the graphs in the upper right are mirror images of the graphs in the lower left. We will consider only the graphs in the lower left.

The graph in the bottom left corner has BodyFat on the y-axis and Triceps on the x-axis. We can see a definite linear relationship in this graph, so Triceps should be a “good” predictor for BodyFat. This is consistent with the estimated correlation (0.84) between these two variables.

The middle graph in the bottom row has BodyFat on the y-axis and Midarm on the x-axis. These points are relatively dispersed, which is consistent with the relatively low correlation (0.14) between Midarm and BodyFat.

The third graph we need to consider is between the two predictors, Triceps and Midarm. This is the first graph in the middle row. We see a linear relationship, but it is not very strong. In fact, this relationship appears to be slightly curved. We are not concerned about this curvature because both of these variables are predictors. The correlation between these two variables is 0.50, which is consistent with the graph.

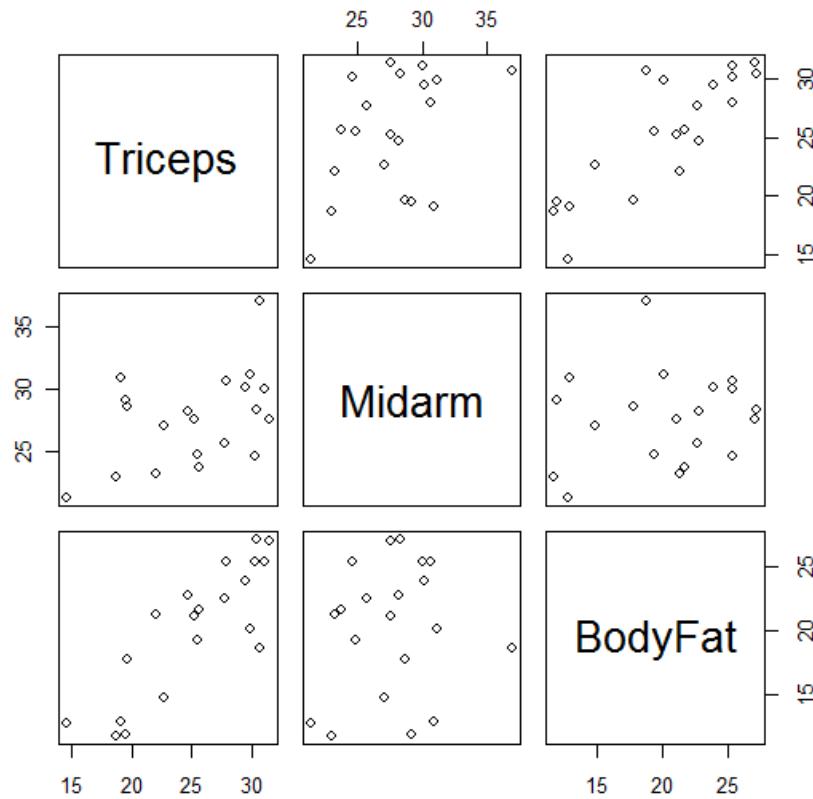


Figure 2.7. Scatterplot matrix for body fat data

It is not necessary to evaluate both the correlation matrix and scatterplot matrix, because they both convey the same information. The correlation matrix provides numeric summaries (point estimates and p-values for the correlations) while the scatterplot matrix provides a visual representation.

2.2.2. Check the assumptions

After examining the correlations, the next step in a regression analysis is to check the assumptions. The correlation analysis simply provides clues as to which predictors might be useful in the model. We use SAS to actually fit the model (via PROC REG), and then assess which predictors are needed. As with simple linear regression, we need to check the assumptions before we interpret any of the other results from the model. We are using the same two diagnostic plots that we used for simple linear regression, and we interpret these plots exactly the same way.

The diagnostic plots for the body fat example are shown in Figure 2.8. The residual plot (on the left) is perfectly acceptable because the points show not obvious pattern. The QQ plot (on the right) is also

perfectly fine because the points are following the line. Overall, there is no evidence that the assumptions have been violated, so we can proceed with the analysis.

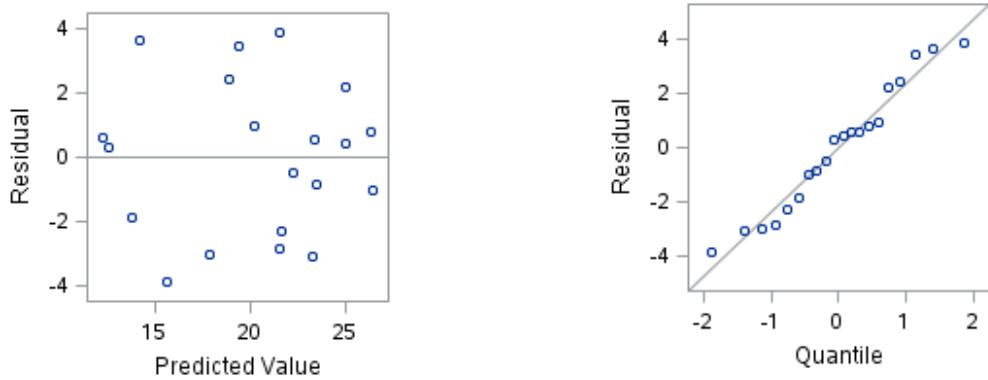


Figure 2.8. Diagnostic plots for the body fat data

2.2.3. ANOVA table for multiple regression

The ANOVA table for multiple regression is very similar to the ANOVA table for simple linear regression, which we examined in Chapter 1. A generic ANOVA table for multiple regression is shown in Table 2.3. In this table, the value for k is the number of parameters in the model. This is the number of predictor variables plus 1 (for the intercept).

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	$k - 1$	SSReg	SSReg / dfReg	MSReg / MSE	
Error	$n - k$	SSE	SSE / dfE		
Corrected Total	$n - 1$	SSTot			

Table 2.3. Generic ANOVA table

The relationships we established in simple linear regression are all still true:

- Total degrees of freedom is always one less than the sample size
- The degrees of freedom are additive: $df_{Reg} + df_E = df_{Total}$
- The sums of squares are additive: $SS_{Reg} + SSE = SSTot$
- Each mean square is “average”: $MS = SS / df$
- Test statistic: $F = MS_{Reg} / MSE$
- Total sum of squares is related to the variance of Y: $SSTot = (n - 1) Var(Y)$

The ANOVA table for the body fat example is shown in Table 2.4. Recall that the model for this example is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \quad i=1, \dots n \quad (2.5)$$

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	389.46	194.73	31.25	<.0001
Error	17	105.93	6.23		
Corrected Total	19	495.39			

Table 2.4. ANOVA table for body fat data

The “Corrected Total” degrees of freedom is 19, which is one less than our sample size of 20. The “Model” degrees of freedom is 2 because there are two predictors in the model. Note that the degrees of freedom (DF) are additive ($2 + 17 = 19$) and that the sum of squares are also additive ($389.46 + 105.93 = 495.39$),, but that the mean squares are NOT additive. (There is no mean square for “Corrected Total”).

The line for “Model” provides the test statistic ($F = 31.25$) and p-value ($p<.0001$) for a hypothesis test. This test is sometimes called the “overall” ANOVA F test. For the body fat example, the hypotheses for the overall ANOVA F test are

$$H_0 : \beta_1 = 0 \text{ and } \beta_2 = 0 \quad \text{vs. } H_a : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0$$

When there are more than two predictors in a model, it is easier to think of these hypotheses in words:

$$\begin{aligned} H_0 &: \text{all population slopes are equal to 0} \\ \text{vs. } H_a &: \text{at least one population slope is not 0} \end{aligned}$$

If H_0 is true, then every predictor can be removed from the model (because they are all going to get multiplied by 0). Then the model reduces to $Y_i = \beta_0 + \varepsilon_i$.

For the body fat model, we reject H_0 in the overall ANOVA F test. We conclude that either triceps or forearm (or both) is useful for estimating body fat. The next step is to determine if BOTH of these predictors are needed, or if we can simplify the model by removing one of them.

2.2.4. Parameter Estimates table

The Parameter Estimates table is automatically generated by PROC REG. It provides information about each parameter in the model, which includes the intercept and the slopes on all the predictors. The Parameter Estimates table for the body fat example is shown in Table 2.5. The columns for ‘95% Confidence Limits’ are provided in this table only because we included the option CLB on the MODEL statement in the SAS code. (CLB stands for confidence limits for the beta’s.)

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	95% Confidence Limits	
Intercept	1	6.79163	4.48829	1.51	0.1486	-2.67783	16.26109
triceps	1	1.00058	0.12823	7.80	<.0001	0.73004	1.27113
midarm	1	-0.43144	0.17662	-2.44	0.0258	-0.80407	-0.05882

Table 2.5. Parameter estimates table for the body fat data

Each row in the table is estimating and testing one parameter in the model.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \quad (2.6)$$

which is

$$(\text{BodyFat})_i = \beta_0 + \beta_1 (\text{Triceps})_i + \beta_2 (\text{Midarm})_i + \varepsilon_i \quad (2.7)$$

For the intercept

In the model the intercept is β_0 , so all the values on the line labeled ‘Intercept’ are related to β_0 . The point estimate for β_0 is $\hat{\beta}_0 = 6.79163$, and the standard error of this estimate is $se(\hat{\beta}_0) = 4.48829$. The 95% confidence interval for β_0 is (-2.67783, 16.2609). Since the confidence interval contains 0, we conclude that 0 is a plausible value for β_0 . This is confirmed by the test statistic ($t = 1.51$) and the p-value (0.1486) for testing $H_0 : \beta_0 = 0$ vs. $H_a : \beta_0 \neq 0$. We do not reject H_0 , so the test indicates that 0 is a plausible value for the intercept β_0 . (The conclusion derived from the confidence interval should always match the conclusion from the hypothesis test.)

Although the hypothesis test indicates that the intercept can be removed from the model, it is generally not a good idea to do so. The intercept by itself is usually not of particular importance, but keeping it in the model allows the other estimates to be more precise.

For Triceps

In the model (equation (2.7)), the slope on triceps is β_1 , so all the values on the line labeled ‘triceps’ are related to β_1 . The estimated slope is $\hat{\beta}_1 = 1.00058$ and the standard error of this estimate is $se(\hat{\beta}_1) = 0.12823$. Recall that $\hat{\beta}_1$ is a random variable, and the value we obtained from the current dataset is just one possible value for this random variable. The potential values for $\hat{\beta}_1$ follow a t distribution with degrees of freedom equal to the degrees of freedom for error, which is 17 for the body fat data. Using the probability table for the t distribution (with $df = 17$ and $\alpha/2 = 0.025$), the critical value is 2.110. Therefore, a 95% confidence interval for β_1 is $1.00 +/- (2.110)(0.128)$, which is $1.00 +/- 0.27$, or $(0.73, 1.27)$. These confidence limits are also provided in the parameter estimates table, but this is only because we included the CLB option on the MODEL statement in PROC REG. Since 0 is not in the confidence interval, we should conclude that 0 is not a plausible value for β_1 . The columns “t Value” and “Pr > |t|” are providing the test statistic ($t = 7.80$) and the p-value ($p < 0.0001$), respectively, for the hypotheses $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$. Since the p-value is smaller than 0.05, we reject H_0 and conclude that the slope on triceps is not equal to 0. The predictor triceps should be kept in the model.

For Midarm

In the model (equation (2.7)), the slope on midarm is β_2 , so all the values on the line labeled ‘midarm’ are related to β_2 . The estimated slope is $\hat{\beta}_2 = -0.43144$ and the standard error of this estimate is $se(\hat{\beta}_2) = 0.17662$. The potential values for $\hat{\beta}_2$ follow a t distribution with degrees of freedom equal to the degrees of freedom for error, which is 17 for the body fat data. (Note that the degrees of freedom for $\hat{\beta}_2$ is exactly the same as the degrees of freedom for $\hat{\beta}_1$.) The critical value is 2.110, which is the same as for $\hat{\beta}_1$. A 95% confidence interval for β_2 is $-0.431 +/- (2.110)(0.177)$, which is $-0.431 +/- 0.373$, or $(-0.804, -0.058)$. These confidence limits are within roundoff error of the values in the Parameter Estimates table. Since 0 is not in the confidence interval, we should conclude that 0 is not a plausible value for β_2 . The columns “t Value” and “Pr > |t|” are providing the test statistic ($t = -2.44$) and the

p-value ($p = 0.0258$), respectively, for the hypotheses $H_0 : \beta_2 = 0$ vs. $H_a : \beta_2 \neq 0$. Since the p-value is smaller than 0.05, we reject H_0 and conclude that the slope on midarm is not equal to 0. The predictor midarm should remain in the model.

Both the test for triceps and the test for midarm are called marginal tests. They are only testing one parameter in the model, and it is assumed that all the other parameters remain in the model. In essence, each of these t tests is evaluating the additional contribution of the variable whose slope is being tested, after the contributions from all the other predictors have been accounted for.

Interpreting the parameter estimates

Using the point estimates from the Parameter Estimates table, the estimated regression equation is

$$E(\text{Body Fat}) = 6.79 + 1.00 * (\text{Triceps}) - 0.43 * (\text{Midarm})$$

(Recall that E stands for expected value.)

The estimated intercept is $\hat{\beta}_0 = 6.79$, and it represents the expected body fat when both triceps and midarm are equal to 0. Obviously, this interpretation does not make any sense. It is not possible to have a value of 0 for either triceps or midarm. It is often the case that the value for the intercept is nonsensical, but we usually keep it in model unless there is a good reason to take it out.

The remaining model parameters are the slopes. The slope for the k^{th} predictor is the change in the expected value for the response for each unit increase in the predictor, **while keeping all the other predictors constant**. When we interpret the estimated slopes in multiple regression, we must include this qualification, and we should also include the units on each variable.

- The estimated slope on triceps is $\hat{\beta}_1 = 1.00$.

Interpretation: If the triceps measurement increases by 1 mm and the midarm measurement stays the same, then we expect the body fat to increase by 1.00 points.

- The estimated slope on midarm is $\hat{\beta}_2 = -0.43$.

Interpretation: If the midarm measurement increases by 1 cm and the triceps measurement stays the same, then we expect the body fat to decrease by 0.43 points. (Note that this is a decrease because the slope is negative.)

2.2.5. Estimating the response

So far, we have concluded two things: (1) the model assumptions do not appear to be violated , and (2) both predictors are significant. Now we can use the model for prediction and estimation.

There are two quantities of interest:

- The mean response is the average value of the response for all items in the population that have the specified values of the predictors.
- The individual predicted value is the value for one of the items in the population that has the specified values of the predictors.

Both quantities require that we specify a value for every predictor that is in the model.

Each of these two quantities can be estimated by either a point estimate or an interval estimate. As with simple linear regression, the point estimate for a mean response is the same as the point estimate for an individual response, but the interval estimates are different. The interval estimate for the mean response is called a confidence interval and the interval estimate for an individual value is called a prediction interval.

It is a fairly straightforward task to calculate a point estimate from the regression equation. For example, if a woman has triceps = 20 mm and midarm = 20 cm, then we estimate her body fat to be

$$\text{Body Fat} = 6.79 + 1.00*(20) - 0.43*(20) = 18.19 \text{ points.}$$

If we want the estimate the mean body fat for all women who have triceps = 20 mm and midarm = 20 cm, the calculation is exactly same. However, the interval estimates for these two values will be different. We are not showing the formulas needed to calculate the interval estimates. We will let SAS perform these calculations.

The options we include in the SAS code are exactly the same as they were with simple linear regression. We include an extra line of data in the DATA step, and on the MODEL statement we use the option CLM for the confidence interval and/or the option CLI for the prediction interval. For multiple regression, the extra line of data needs to include a value for every predictor in the model.

```

DATA fat;
INPUT triceps thigh midarm bodyfat;
DATALINES;
19.5 43.1 29.1 11.9
24.7 49.8 28.2 22.8
... more data lines here ...
25.2 51.0 27.5 21.1
20.0 50.0 20.0 .
;
PROC REG DATA=fat;
MODEL bodyfat = triceps midarm / CLM CLI;
RUN;

```



extra line of data

Note that the SAS code is reading the data for four variables (triceps, thigh, midarm, and bodyfat), but we are not using the values for thigh in our current model. It is perfectly acceptable to have unused variables in a dataset; there is no need to edit the dataset to remove the variable thigh. The last line in the DATA step assigns the value of 20 to both triceps and midarm, with a period (a missing value) for bodyfat. This line also assigns the value 50 to thigh, but this value will be ignored in our current analysis. The output generated by the CLM and CLI options is shown in Table 2.6. For a single woman who has midarm 20 cm and triceps 20 mm, we predict her body fat to be between 12.2150 and 24.1340 points, with probability 0.95. For all the women who have these measurements, we estimate their mean body fat to be between 15.3856 and 20.9634 points, with confidence 0.95.

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	11.9000	13.7481	1.0546	11.5231	15.9731	8.0307	19.4655	-1.8481
2	22.8000	19.3394	0.5791	18.1176	20.5612	13.9329	24.7460	3.4606
...
20	21.1000	20.1417	0.5585	18.9633	21.3201	14.7448	25.5386	0.9583
21	.	18.1745	1.3219	15.3856	20.9634	12.2150	24.1340	.

Table 2.6. Confidence and prediction limits for the body fat data

Note that, as with simple linear regression, the prediction interval is wider than the confidence interval, and this will always be true provided the values of the predictor variables remain unchanged. It is also true that both of these intervals will get narrower if the predictor values are closer to the mean, but this is harder to visualize now that there is more than one predictor.

2.2.6. Summary

The process for performing a multiple linear regression analysis with two predictors is basically the same as that for simple linear regression. The differences are:

1. We now have multiple scatterplots to examine, which we can see in a scatterplot matrix. Optionally, we can look at the correlation matrix.
2. We need to consider the relationship between the predictor variables.
3. The overall ANOVA F test is not the same as the t tests in the Parameter Estimates table. The overall ANOVA F test is testing whether or not all the slopes in the model are equal to 0. Each t test is testing whether or not one of slopes is equal to 0, assuming all the other predictors remain in the model.
4. To interpret a slope, we need to keep constant the values of the other predictors.
5. To generate estimates for the response variable, we need to specify values for all the predictors that are in the model.

Section 2.3. More than Two Predictors

For the body fat data, a third variable was collected for each woman in the dataset. This variable contains the measurement of the woman's thigh. We want to know if including this variable as a predictor in the model will improve the fit of the model. The new model we want to consider is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i \quad (2.8)$$

where Y_i , X_{1i} and X_{2i} are as defined in Section 2.2., and X_{3i} is the thigh measurement for the i^{th} woman.

This is an additive model with three predictors. The SAS code is as follows.

```
DATA fat;
  INPUT triceps thigh midarm bodyfat;
  DATALINES;
  19.5 43.1 29.1 11.9
  24.7 49.8 28.2 22.8
  . . . more data lines . .
;
PROC REG DATA=fat;
  MODEL bodyfat = triceps midarm thigh;
  RUN;
```

Note that we are not yet considering the correlation among these variables. We will look at the correlations at the end of this example. The diagnostic plots are shown in Figure 2.9. The residual plot (on the left) and the QQ plot (on the right) have no indication that the model assumptions have been violated, so we proceed to the ANOVA table.

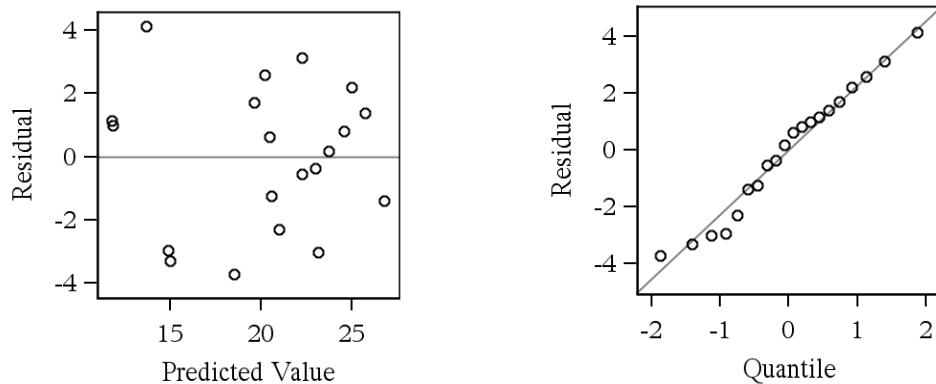


Figure 2.9. Diagnostic plots for three-variable body fat model

In the ANOVA table (shown in Table 2.7), the p-value for the overall ANOVA F test is less than 0.0001. We reject the hypothesis that all three slopes are equal to 0. Instead, we conclude that at least one slope is not 0. In other words, at least one of the three predictors is useful for estimating body fat.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	396.98461	132.32820	21.52	<.0001
Error	16	98.40489	6.15031		
Corrected Total	19	495.38950			

Table 2.7. ANOVA table for the three-variable body fat model

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	117.08469	99.78240	1.17	0.2578
triceps	1	4.33409	3.01551	1.44	0.1699
midarm	1	-2.18606	1.59550	-1.37	0.1896
thigh	1	-2.85685	2.58202	-1.11	0.2849

Table 2.8. Parameter estimates table for the three-variable body fat model

Next, we examine the parameter estimates table, shown in Table. Each line in the table is providing information about exactly one of the parameters in the model (equation (2.8)).

- The line labeled ‘intercept’ is for β_0 . The hypotheses are $H_0 : \beta_0 = 0$ vs. $H_a : \beta_0 \neq 0$. The p-value is 0.2578, which is greater than 0.05, so we do not reject H_0 . It is plausible that the intercept is equal to 0.
- The line labeled ‘triceps’ is for β_1 . The hypotheses are $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$. The p-value is 0.1699, which is greater than 0.05, so we do not reject H_0 . It is plausible that the slope on triceps is equal to 0.
- The line labeled ‘midarm’ is for β_2 . The hypotheses are $H_0 : \beta_2 = 0$ vs. $H_a : \beta_2 \neq 0$. The p-value is 0.1896, which is greater than 0.05, so we do not reject H_0 . It is plausible that the slope on midarm is equal to 0.
- The line labeled ‘thigh’ is for β_3 . The hypotheses are $H_0 : \beta_3 = 0$ vs. $H_a : \beta_3 \neq 0$. The p-value is 0.2849, which is greater than 0.05, so we do not reject H_0 . It is plausible that the slope on thigh is equal to 0.

For all three slopes, we FAIL TO REJECT H_0 , so it seems that NONE of these variables are useful predictors for body fat.

The overall ANOVA F test and the three individual t tests for the slopes appear to be providing conflicting results. The overall ANOVA F test has p-value less than 0.0001, which indicates at least one of the predictors is useful for estimating body fat. But the individual t tests indicate that none of the three predictors are useful. The answer to this conundrum is related to two important facts:

- 1) The individual t tests are testing one predictor at a time, assuming the other predictors remain in the model.
- 2) The relationship between the predictors cannot be ignored.

2.3.1. Multicollinearity and variance inflation

To investigate the apparent discrepancy between the F test and the three t tests, we now look at the correlations among all the variables in the model. The SAS code shown below produces the correlation matrix in Table 2.9 and the scatterplot matrix in Figure 2.10.

```
PROC CORR DATA=fat PLOTS=MATRIX;
  VAR triceps midarm thigh bodyfat;
  RUN;
```

Pearson Correlation Coefficients				
	triceps	midarm	thigh	bodyfat
triceps	1.00000	0.50102	0.90943	0.84327
midarm	0.50102	1.00000	0.09777	0.14244
thigh	0.90943	0.09777	1.00000	0.87809
bodyfat	0.84327	0.14244	0.87809	1.00000

Table 2.9. Correlation matrix for the three-variable body fat model

We want strong correlation between body fat (Y) and each of the X's. Based on the last column in the correlation matrix (Table 2.9) and bottom row in the scatterplot matrix (Figure 2.10), it seems that thigh and triceps would be very good predictors for body fat, but midarm would not be as good.

- Body fat and thigh: correlation = 0.878, very high
- Body fat and triceps: correlation = 0.843, very high
- Body fat and midarm: correlation = 0.142, weak

In addition to having strong correlation between the response and each of the predictors, we also want to have weak correlation among the predictors.

- Midarm and triceps: correlation = 0.501, moderate (acceptable)
- Midarm and thigh: correlation = 0.098, very low (good)
- Triceps and thigh: correlation = 0.909, very high (very bad)

When two or more predictors are highly correlated (greater than 0.7 or so), this is an indication the multicollinearity may be a problem. The correlation between triceps and thigh is 0.909, and this implies that these two variables are essentially measuring the same thing. If the value for triceps is known, then its value can be used to estimate the value for thigh with great accuracy. The reverse is also true: If the value for thigh is known, it can be used to estimate the value for triceps very accurately. Since these two variables are so closely related, it is not necessary to have them both in the model. In fact, it can be detrimental to have them both in the same model.

Recall that the least squares estimation procedure relies on minimizing the sum of the squared residuals (equation (2.4)). The mathematical methods that SAS (or any other software) uses perform this minimization can become unstable when some of the predictors are highly correlated. (In this sense, “unstable” is similar to dividing by 0 or almost 0.) It is entirely possible that SAS will not recognize that there is a problem, but the standard errors of the estimates can explode into very large numbers. This

affects all the inference (i.e., hypothesis tests and confidence/prediction intervals) that rely on the standard errors.

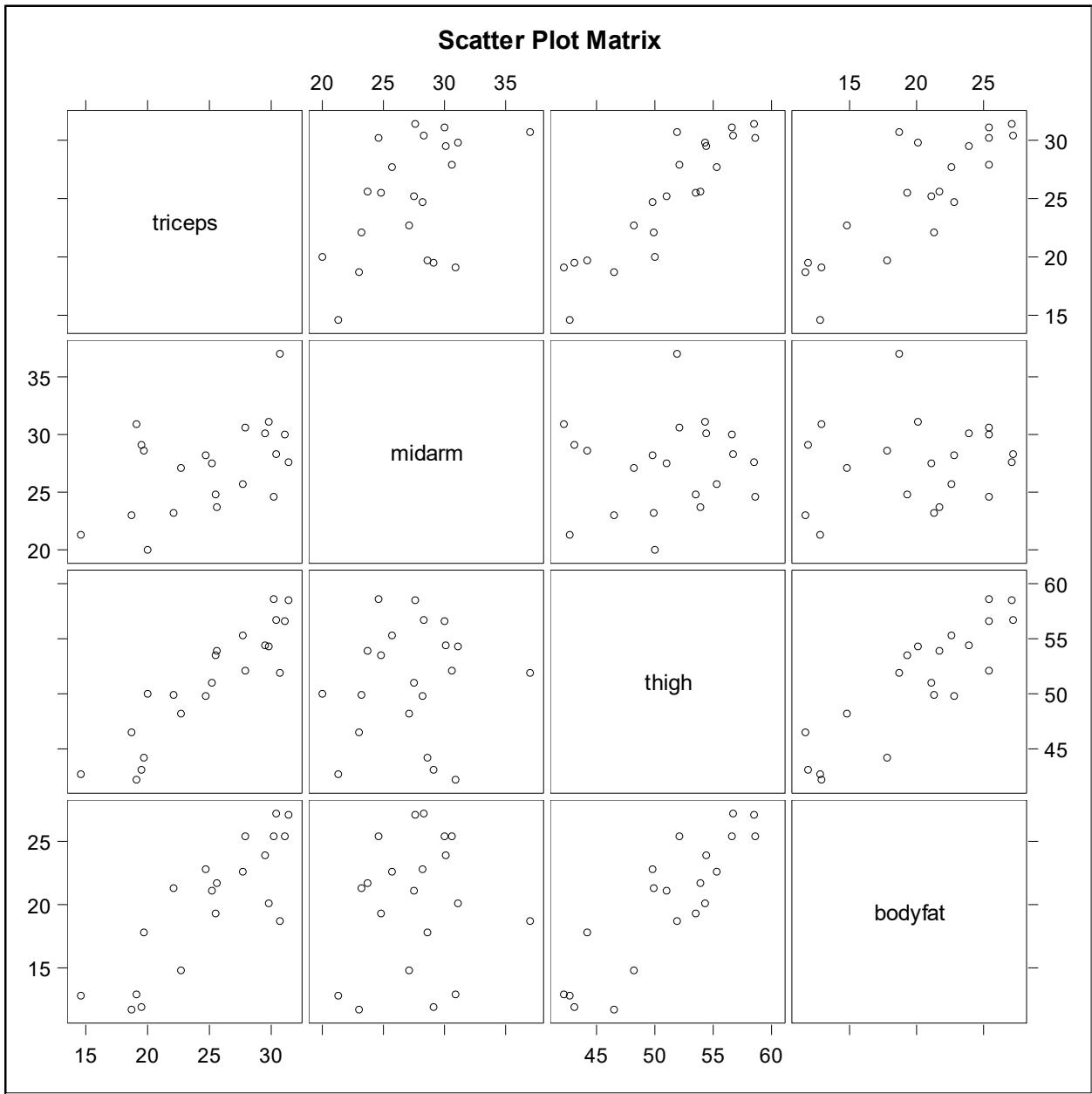


Figure 2.10. Scatterplot matrix for the three-variable body fat model

Having a strong correlation between two or predictors does not guarantee that the standard errors will “explode”, but it is a warning sign. It is also possible for the standard errors to become large even when the pairwise correlations are all weak. This is because the strong relationship may be between three or more predictors, and correlation only measures the relationship between two predictors. Consider a model in which there are three predictors. We will call them A, B and C so as not to confuse them with the body fat variables. Suppose that C is exactly equal to the sum of A and B (so $A + B = C$). The correlation between A and C might be weak and the correlation between B and C might also be weak, but the correlation between $A + B$ and C is exactly equal to 1 (i.e., perfect correlation). This would imply that if A and B are already in the model, then there is no need to include C in the model.

The linear relationships between combinations of predictor variables is called multicollinearity. It is measured by the variance inflation factors, abbreviated VIF. There is one VIF for each predictor in the model. For the k^{th} predictor X_k , the variance inflation factor is calculated as

$$\text{VIF}_k = \frac{1}{1 - R_k^2} \quad (2.9)$$

where R_k^2 is the coefficient of determination (R-square) when X_k is regressed on the remaining predictors. In other words, X_k is treated as the response variable and the other S's are the predictors. Note that the variance inflation factors depend only on the X values in the data set. The values for Y are irrelevant to the VIF.

The variance of the estimator $\hat{\beta}_k$ is a function of VIF_k .

- When $R_k^2 = 0$ then $\text{VIF}_k = 1$. This implies that X_k is not linearly related to other predictors in the model, so the variance is not inflated.
- When $R_k^2 > 0$ then $\text{VIF}_k > 1$. This implies that the variances are inflated due to correlation between predictors.
- As R_k^2 gets larger (i.e., closer to 1), then VIF_k also gets larger.

It is acceptable (in most cases, unavoidable) to have a small amount of variance inflation. There is some debate regarding how much variance inflation is acceptable. In some disciplines, having any variance inflation factor greater than 4 is unacceptable, but others consider this too strict. We will consider 10 to be the cutoff, so that any VIF greater than 10 implies serious multicollinearity issues.

To get SAS to calculate the variance inflation factors, use the option ‘VIF’ on the MODEL statement.

```
PROC REG DATA=fat;
  MODEL bodyfat = triceps midarm thigh / VIF;
  RUN;
```

This will produce an extra column in the Parameter Estimates table, appropriately labeled “Variance Inflation”. This is shown in Table 2.10.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Variance Inflation
Intercept	1	117.08469	99.78240	1.17	0.2578	0
triceps	1	4.33409	3.01551	1.44	0.1699	708.84291
midarm	1	-2.18606	1.59550	-1.37	0.1896	104.60601
thigh	1	-2.85685	2.58202	-1.11	0.2849	564.34339

Table 2.10. Variance inflation factors for the three-variable body fat model

Since all three of the predictors in the body fat model have variance inflation factors much greater than 10, there are serious multicollinearity issues in this data set. When all three predictors are used in the model, the inference is invalidated by multicollinearity.

To remedy multicollinearity issues, the predictors in the model must be modified. This may involve simply removing one or more of the predictors from the model. However, it may not be clear which predictor should be removed, because the predictor that has the high value for VIF may not be the predictor that is causing the problem. For the body fat model, we know that introducing the predictor thigh into the model created the multicollinearity problems, so if thigh is removed then the multicollinearity issues should dissipate. But it is entirely possible that we could remove triceps, and keep both thigh and midarm, and this might also alleviate the multicollinearity issues. It might also produce a better-fitting model.

Another potential remedy for multicollinearity issues is to use composite predictors. This can be achieved via Principal Component Analysis. The drawback of this approach is that the composite predictors can be difficult to interpret. This is a graduate-level statistical topic and is beyond the scope of our current work.

A final suggestion to alleviate multicollinearity problems is to center the predictors. For each predictor, subtract the mean from the observed value, then use the difference instead of the original value. Some

people recommend that the predictors be standardized. This would involve subtracting the mean and dividing by the standard deviation for each predictor, then using the transformed value instead of the original value. While this has potential to remove the multicollinearity problems, it also creates predictors that are very difficult to interpret.

While multicollinearity is fairly easy to detect, it is sometimes difficult to resolve. In most cases, it will be necessary to consider several different models and then choose the one that appears to be the “best”.

2.3.2. Criteria for comparing models

There is no such thing as a “best” model. We are looking for a model that fits the data well and does not violate the model assumptions. There are many different criteria we can use to compare models, and some of them might provide conflicting information. For example, one criterion might indicate Model 2 is better than Model 1, but a different criterion might indicate that Model 1 is better than Model 2. In the end, choosing a model is a subjective decision, but there are certain guidelines we should follow.

- Principle of parsimony

We want to use the simplest model that accurately describes the data. In other words, we want as few predictors as possible, and we want to use un-transformed predictors, if possible.

- Smaller MSE is better

Recall that the MSE is an estimate for the error variance σ^2 , which measures the amount of variability around the regression line. We could use the RMSE (square root of MSE) to get the same result. This criterion can only be used when the two models that are being compared have the same number of predictors.

- Larger R-square is better

Since R-square measures the proportion of variability in Y that is explained by the regression model, having a large R-square indicates the model fits the data well. This criterion can only be used when the two models that are being compared have the same number of predictors.

- Adjusted R-square (larger is better)

Every time we add a predictor to the model, then the value for R-square will go up and the value for MSE (and RMSE) will go down. If the two models we are comparing have different numbers of predictors, then we must account for the fact that the model with more predictors will

automatically appear to be better. One way to do this is to place a penalty on R-square that is proportional to the number of predictors in the model. This is called adjusted R-square. It appears in the SAS output as Adj R-Sq, and is located in the table between the ANOVA table and the Parameter Estimates table. The value for adjusted R-square is calculated via this formula.

$$R_{\text{adj}}^2 = 1 - \left(\frac{n-1}{n-p-1} \right) (1 - R^2), \text{ where } p \text{ is the number of predictors} \quad (2.10)$$

There are additional criteria that we will consider later.

2.3.3. Compare models for the body fat data

For the body fat data, there are several models we could create using the three predictors. Based on previous exploration and the subsequent multicollinearity issues, we should not include both thigh and triceps in the same model. We will compare and contrast these five candidate models:

- Model 1: Use only midarm
- Model 2: Use only thigh
- Model 3: Use only triceps
- Model 4: Use midarm and triceps
- Model 5: Use midarm and thigh

Here is the SAS code for the results we will need.

```
. . . DATA step goes here . . .

/* 5 candidate models      */
PROC REG DATA=fat;
  MODEL bodyfat = midarm;           * model 1;
  MODEL bodyfat = thigh;            * model 2;
  MODEL bodyfat = triceps;          * model 3;
  MODEL bodyfat = midarm triceps / VIF; * model 4;
  MODEL bodyfat = midarm thigh / VIF;  * model 5;
RUN;
```

Note that we have several MODEL statements within one PROC REG. SAS will automatically label them in the output as “Model 1”, “Model 2”, etc. We can provide our own names for these models, but this is optional. To assign a name to a model, insert the name and a colon before the word MODEL, as in

```
ModelA : MODEL bodyfat = midarm;
```

If the name of the model has any spaces, enclose the name in either single or double quotes, as in

```
'Model A' : MODEL bodyfat = midarm;
```

The name of the model will appear in the header of every page of SAS output for that model.

The SAS code provided above will generate many pages of results, but for the purpose of comparing these models, we only need to consider only the diagnostic plots, the MSE (or RMSE) , R-square and adjusted R-square. In the SAS output, MSE is in the ANOVA table (Mean Square, on the line labeled Error), while RMSE (“Root MSE”), R-square and adjusted R-square (“Adj R-Sq”) are all in the table immediately below the ANOVA table.

For our five candidate models, the values for the numeric criteria are given in Table 2.11 and the diagnostic plots are shown in **Error! Reference source not found.** The values in the table have been collected from scattered places in the SAS output. (RMSE is not listed because it is just the square root of MSE.)

	Predictors in the Model				
	Model 1	Model 2	Model 3	Model 4	Model 5
Criteria	midarm	thigh	triceps	midarm & triceps	midarm & thigh
MSE	26.96	6.301	7.951	6.231	6.536
R ²	0.0203	0.7710	0.7111	0.7862	0.7757
adj R ²	-0.0341	0.7583	0.6950	0.7610	0.7493

Table 2.11. Criteria for comparing five models

None of the diagnostic plots (**Error! Reference source not found.**) show any signs for concern, so it appears that none of the five models violate the assumptions. (If any model had obviously violated the assumptions, we would discard that model.) We now turn our attention to the numeric criteria in Table 2.11.

For Model 1, the value for R-square is much lower than the other models, and the value for adjusted R-square is actually negative. This is a very poor model, and it should be discarded. Of the remaining models, Model 3 (that uses only triceps) has the lowest R-square, the lowest adjusted R-square and the highest MSE, so this model appears to be inferior. However, we should not let a minor difference in these criteria to dictate which model we choose. The remaining three models are nearly identical in their values for the criteria.

To be absolutely thorough, we should examine the variance inflation factors for any model that has more than one predictor. Models 2 and 3 do not have VIFs because they each have only one predictor.

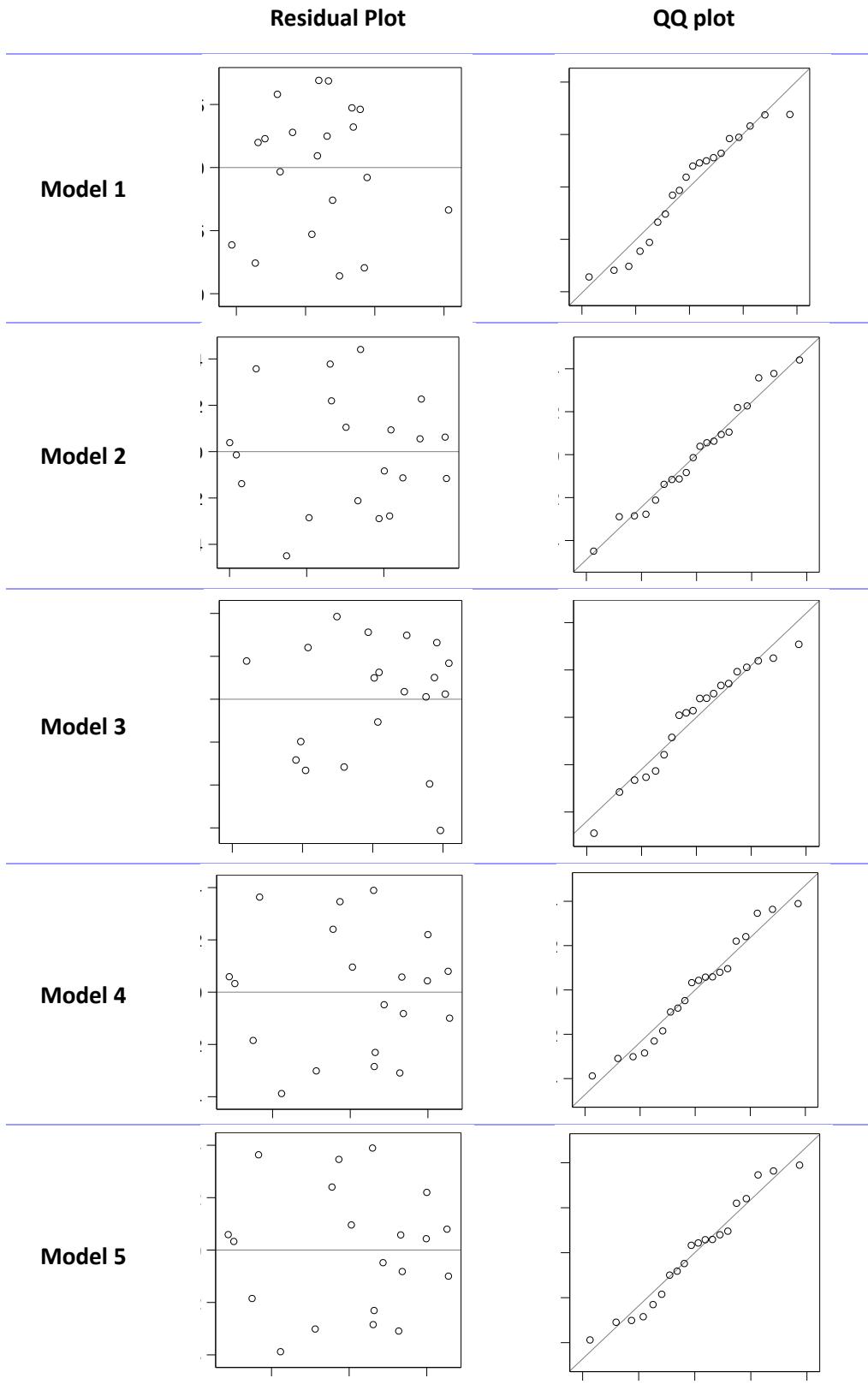


Figure 2.11. Diagnostic plots for five candidate models

The VIF for both predictors in Model 4 is 1.26512, and for Model 5 the VIF for both predictors is 1.00722. (Note that these values are included in the SAS output only because we include the option VIF on the corresponding MODEL statements in the SAS code.) Multicollinearity is not an issue with any of these models.

There is no clear “winner” among Models 2 through 5. We should choose the model that will be the easiest to implement and explain.

2.3.4. Summary

The basic steps for performing a multivariate regression analysis are very similar to simple linear regression analysis. The differences are

- The overall ANOVA F test and the individual t tests are now testing different hypotheses.
- When there are multiple predictors, interpretation of the estimated slopes must include references to the other predictors in the model.
- Multiple predictor variables can be correlated, so we have to be aware of potential multicollinearity issues. This includes recognizing when there is a multicollinearity issue, and how to mitigate it.
- If we want to compare two models that have the same number of predictors, we can use MSE, RMSE, R-square or adjusted R-square. If the two models have different numbers of predictors, the only criterion of these four that is applicable is adjusted R-square. (There will be other criteria later.)

Section 2.4. General Linear Regression Model

So far, we have used data values exactly as they appear in the dataset. Sometimes it is necessary to transform one or more of the variables, in order to obtain a model that satisfies the assumptions and fits the data well. When we use transformed values, the model is called a general linear regression model. The equation for a general linear regression model is

$$Y_i = \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2i} + \dots + \beta_p Z_{pi} + \varepsilon_i \quad i = 1, 2, \dots, n \quad (2.11)$$

This is called a general linear model because

- the Z 's can be X 's (i.e., the observed variables in the data set)
- the Z 's can be transformations of the X 's (e.g. X^2)
- the Z 's can represent non-numeric data (e.g., gender)

There can be any combination of observed/transformed/non-numeric predictors in a general linear regression model. The model is still “linear” because it is linear in the β 's.

To illustrate, suppose there are two measured variables (X_1 and X_2) in the data set. The Z 's could be $Z_1 = X_1$, $Z_2 = X_2$, $Z_3 = X_1^2$, $Z_4 = X_2^2$, and $Z_5 = X_1 \cdot X_2$. Then the general linear model is

$$Y_i = \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2i} + \beta_3 Z_{3i} + \beta_4 Z_{4i} + \beta_5 Z_{5i} + \varepsilon_i \quad i = 1, 2, \dots, n \quad (2.12)$$

which is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i}^2 + \beta_4 X_{2i}^2 + \beta_5 X_{1i} X_{2i} + \varepsilon_i \quad i = 1, 2, \dots, n \quad (2.13)$$

This model has six parameters (one intercept and five slopes), even though there are only two predictors in the dataset.

2.4.1. Response surface

The additional terms allow the least squares plane to become a least squares surface. Some typical response surfaces are shown in Figure 2.12. Although these surfaces appear complicated, the model is still linear in the β 's, and all our work with multiple regression models is still valid and applicable. We continue to use the least squares method to estimate the β 's. For specified values of the predictors, the expected value of the response is the corresponding point on the response surface.

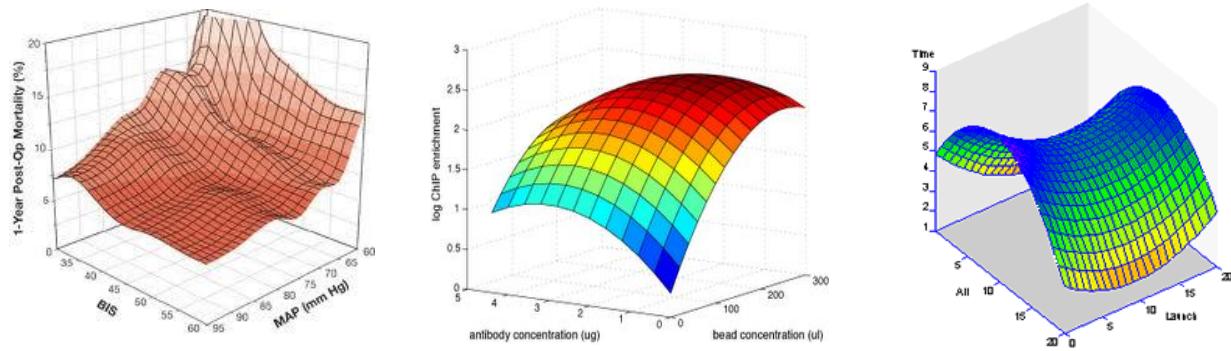


Figure 2.12. A variety of response surfaces

2.4.2. Some terminology

Some models for linear regression occur often enough that they are given special descriptors. It is not mandatory to use these labels to refer to these models, but they help to convey more information about the model than simply calling it “Model 1”.

- A ‘first order’ model has no exponents on the predictors.
For example, a first order model with two predictors is $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$.
- A ‘second order’ model has squared terms.
For example, a second order model with one predictor is $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \varepsilon_i$.
- An ‘interaction’ term is the product of two predictors.
For example, in equation (2.13) the term $\beta_5 X_{1i} X_{2i}$ is an interaction term. A model that contains an interaction term is an interaction model.
- An ‘additive’ model has no interaction terms.
- A ‘nonlinear’ model is not linear in the β ’s.
For example, $Y_i = \beta_0 \exp(\beta_1 X_i) + \varepsilon_i$ is a nonlinear model, as is $Y_i = \beta_0 X_i^{\beta_1} + \varepsilon_i$.

It is possible for one model to fit into more than one of these categories. For example, a model can be a first-order model with interaction or it can be a first-order additive model.

2.4.3. "All models are wrong..."

It can be quite tricky to decide which predictor variables should be considered for inclusion in the model. There is an unlimited number of potential predictors and an unlimited number of potential transformations of these predictors, so it is impossible to consider them all. We need to limit the scope to something that is reasonable. Knowledge of the subject matter is invaluable in deciding which predictors might be useful, and it can also provide clues as to which transformations might be appropriate. We can also use the output from one model to assist in deciding which transformations might be worthwhile. Remember the words of George E.P. Box: "All models are wrong, but some are useful." We are looking for a "useful" model.

*"All models are wrong,
but some are useful."*

George E. P. Box
(1921 – 2013)



Photo courtesy of [DavidMCEddy](#) at [en.wikipedia](#)

2.4.4. Example: A second-order model

A data set contains two measured predictors (X_1 and X_2) and a response (Y). This is simulated data, so there is no story to put the data in context.

We will fit a first-order model to these data.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \quad (2.14)$$

The diagnostic plots are shown in Figure 2.13. The QQ plot (on the right) could be okay, but the residual plot has a very distinct quadratic shape. We must modify this model, and the obvious choice is to add a squared term to the model. Since there are two predictors, we must decide which one (or both) we should square.

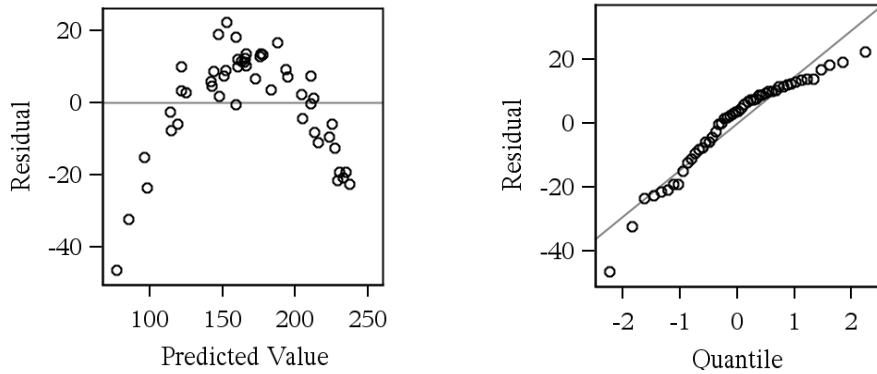


Figure 2.13. Diagnostic plots for simulated data

A new type of residual plot

When there are multiple predictors in a model, a new type of plot called partial residual plots can aid in assessing which (if any) transformations may improve the model. There is one partial residual plot for each predictor. The x-axis contains the values for the predictor and the y-axis contains the residuals from the fitted model. The plots for the current model are shown in Figure 2.. If there is a pattern in a partial residual plot, this may indicate what transformation should be considered for the corresponding predictor. Both graphs in Figure 2. show a quadratic pattern, as highlighted in Figure 2.. Although the quadratic pattern is more distinct for X_2 , we will include squared terms for both of these predictors.

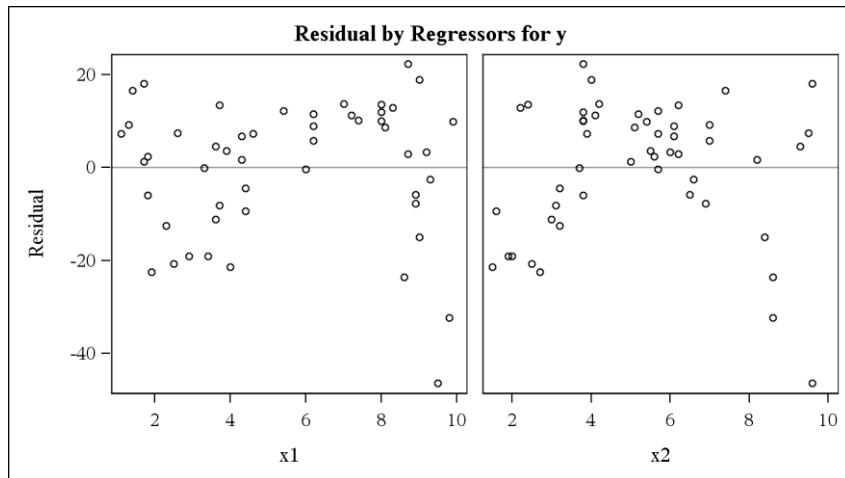


Figure 2.14. Partial residual plots

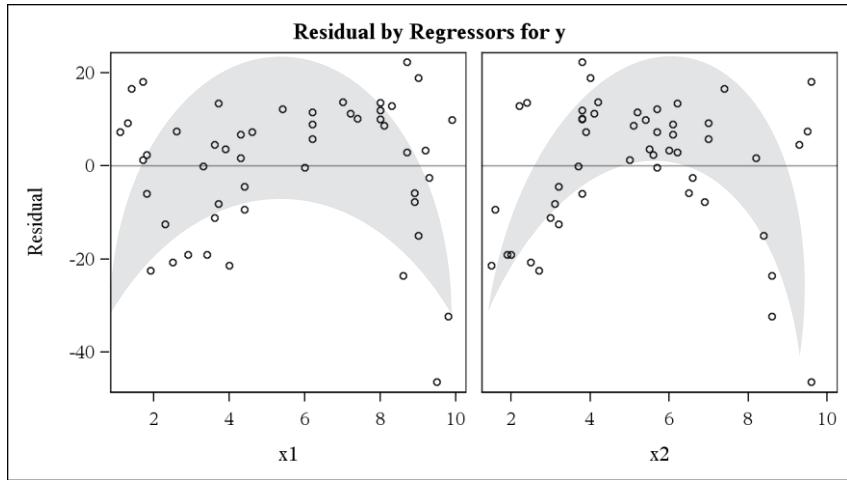


Figure 2.15. Partial residual plots with highlighted quadratic patterns

We now consider a second-order model for these data.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i}^2 + \beta_4 X_{2i}^2 + \varepsilon_i \quad (2.15)$$

The diagnostic plots for the second-order model, shown in Figure 2.16, look MUCH better than those for the first-order model (Figure 2.13). We will take a closer look at the second-order model.

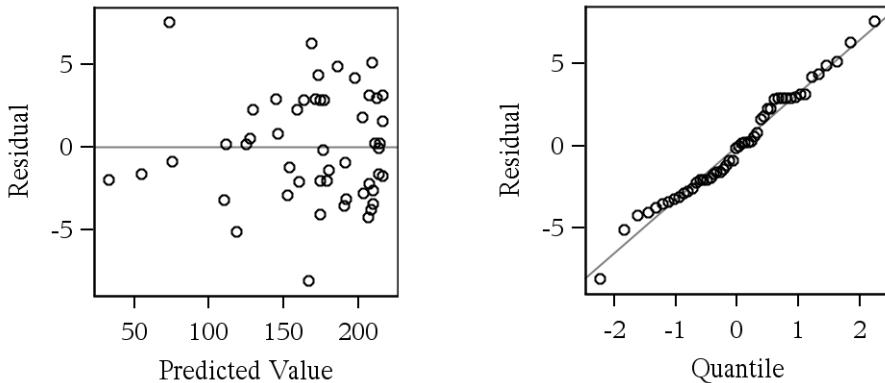


Figure 2.16. Diagnostic plots for second-order model

The ANOVA table and related output for the second-order model are shown in Table 2.12. Notice that the overall ANOVA F test is significant (test statistic is $F=139.09$, $p<.0001$), and the value for R-square is extraordinarily high (0.9252). The estimated model is

$$Y = 242.73 - 5.84X_1 - 4.00X_2 - 0.46X_1^2 - 1.36X_2^2 \quad (2.16)$$

From the parameter estimates table, only the intercept and X_2^2 are significant predictors, so we need to keep X_2^2 in the model. Whenever we keep a squared term in the model, it is customary to also keep the lower-order term in the model, so we should keep X_2 in the model as well. Now we need to decide if we can remove X_1 and X_1^2 . The results of the t tests in the parameter estimates table tell us that we can remove X_1 from the model, provided all the other terms remain in the model. We could also remove X_1^2 , provided all the other terms remain in the model. These t tests do not tell us if we can remove both X_1 and X_1^2 from the model. To make this decision, we need a different hypothesis test.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	91624	22906	139.09	<.0001
Error	45	7410.81752	164.68483		
Corrected Total	49	99035			

Root MSE	12.83296	R-Square	0.9252
Dependent Mean	169.76200	Adj R-Sq	0.9185
Coeff Var	7.55938		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	242.73456	13.36723	18.16	<.0001
x1	1	-5.84354	3.67487	-1.59	0.1188
x2	1	4.00068	3.96260	1.01	0.3181
x1sq	1	-0.46071	0.32827	-1.40	0.1673
x2sq	1	-1.36271	0.34632	-3.93	0.0003

Table 2.12. SAS output for second-order model

2.4.5. Nested model F test

The overall ANOVA F test is used to decide if we should remove ALL terms from the model. The individual t tests (in the Parameter Estimates table) are used to decide if we can remove ONE term from the model, assuming all the other terms remain in the model. To decide if we can remove SOME of the terms from a model, we need a new test. This is called a nested model F test. It is also sometimes called a comparison of models F test.

This test is comparing two models: the “full” model includes all the predictors of interest and the “reduced” model contains only some of the predictors. It is important that all of the predictors in the reduced model also be contained in the full model.

For the current example, the full model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i}^2 + \beta_4 X_{2i}^2 + \varepsilon_i \quad (2.17)$$

and the reduced model is

$$Y_i = \tau_0 + \tau_1 X_{2i} + \tau_2 X_{2i}^2 + \varepsilon_i \quad (2.18)$$

(The τ 's are simply the coefficients. We use a different Greek letter because they can have different values than the β 's in the full model.) Note that the reduced model is the same as the full model, except that the terms involving X_1 and X_1^2 have been removed.

To compare these two models, the hypotheses for the nested model F test are

$$H_0 : \beta_1 = 0 \text{ and } \beta_3 = 0 \quad \text{vs.} \quad H_a : \beta_1 \neq 0 \text{ or } \beta_3 \neq 0$$

If H_0 is true, then we can remove both X_1 and X_1^2 from the full model, i.e., the reduced model adequately fits the data. If H_a is true, then we cannot remove these two terms, i.e., the full model is needed.

The test statistic for the nested model F test requires four numbers:

- $SSE(\text{Red})$ = sum of squares for error in the reduced model
- $SSE(\text{Full})$ = sum of squares for error in the full model
- $dfE(\text{Red})$ = degrees of freedom for error in the reduced model
- $dfE(\text{Full})$ = degrees of freedom for error in the full model

All of these values are obtained from the ANOVA tables for the two models.

The test statistic is

$$F = \frac{\frac{SSE(\text{Red}) - SSE(\text{Full})}{dfE(\text{Red}) - dfE(\text{Full})}}{\frac{SSE(\text{Full})}{dfE(\text{Full})}} \quad (2.19)$$

The test statistic follows an F distribution, with numerator degrees of freedom $dfE(\text{Red}) - dfE(\text{Full})$ and denominator degrees of freedom $dfE(\text{Full})$.

For the current example, we have

- $SSE(\text{Red}) = 52633$
- $SSE(\text{Full}) = 7410.81752$
- $dfE(\text{Red}) = 47$
- $dfE(\text{Full}) = 45$

The test statistic is

$$F = \frac{\frac{52633 - 7410.81752}{47 - 45}}{\frac{7410.81752}{45}} = \frac{\frac{45222.18}{2}}{\frac{7410.81752}{45}} = 137.3$$

To find the critical value, we use the probability table for the F distribution, with numerator degrees of freedom 2 and denominator degrees of freedom 45. The critical value is somewhere between 3.23 and 3.15. (The critical value is actually 3.204, but we would need to use software to determine this, since 45 denominator degrees of freedom is not in our F probability table.) Since the test statistic is MUCH greater than the critical value, we STRONGLY reject H_0 . We conclude that the reduced model does NOT adequately fit the data, and that we need to use the full model. In other words, we should not remove these two terms from the model.

2.4.6. SAS programming notes

It is possible, even desirable, to get SAS to perform the transformations and other calculations described in this section. All of the transformations of the predictor variables occur in the DATA step, and various options and additional statements in PROC REG can perform the other calculations. Since it will not be known which transformations might be necessary, the first part of the SAS code should read the data and generate the original (untransformed) model. This is done the same way we have been doing it.

```
DATA simulated;
INPUT x1 x2 y;
DATALINES;
. . . data goes here . . .
;
PROC REG DATA=simulated;
MODEL y = x1 x2;
RUN;
```

From the output of this model, we realized that including both the square of X_1 and the square of X_2 might improve the model. This requires that we create two new variables, which we will call $x1sq$ and $x2sq$, and this is done in another DATA step.

```
DATA transformed;
SET simulated;
x1sq = x1**2;          * square of x1;
x2sq = x2**2;          * square of x2;
RUN;
```

The SET statement makes a copy of the dataset ‘simulated’ (which was created in the first DATA step), and the additional statements create the two transformed variables that we need. If we had wanted to include an interaction term in the model we are about to create, we would need to include this statement in the DATA step:

```
interact = x1*x2;  * interaction;
```

Note that this DATA step does not need a DATALINES statement, nor does it need to have all the original data values repeated. The SET statement is doing that for us.

Now we use PROC REG to fit the new model. This is the second-order model defined (equation (2.17)). We will also use PROC REG to fit the reduced model (equation (2.18)).

```
PROC REG DATA=transformed;
'Second-Order Model': MODEL y = x1 x2 x1sq x2sq;
'Reduced Model':      MODEL y = x2 x2sq;
RUN;
```

Note that this PROC REG is using the dataset ‘transformed’, which we have just created. It cannot use the original dataset (‘simulated’) because that dataset does not contain the transformed variables that we are using in the MODEL statements. After viewing the results of the second-order model, we wanted to test if we could remove both X_1 and X_1^2 . This is the nested model F test. To get SAS to perform this test, we need to edit our previous section of code for PROC REG and include a TEST statement. This is shown below.

```
PROC REG DATA=transformed;
'Second-Order Model': MODEL y = x1 x2 x1sq x2sq;
TEST x1, x1sq;    * ----- THIS LINE IS ADDED ----- ;
'Reduced Model':      MODEL y = x2 x2sq;
RUN;
```

The TEST statement is performing a nested model F test. The full model is the model that is defined in the immediately preceding MODEL statement, and the variables listed in the TEST statement are the ones we want to remove from the full model. The variables listed in the TEST statement are separated by commas. It is important that the TEST statement be immediately after the MODEL statement for the full model. In the SAS output, the result of the TEST statement is reported immediately after the results from the full model. For our example, the output is shown Table 2.13.

Test 1 Results for Dependent Variable y				
Source	DF	Mean Square	F Value	Pr > F
Numerator	2	22611	137.30	<.0001
Denominator	45	164.68483		

Table 2.13. Results for nested model F test

Note that the test statistic is $F = 137.30$, the numerator degrees of freedom is 2 and the denominator degrees of freedom is 45. All of these values match what we calculated “by hand”. SAS does not report the critical value; it reports the p-value instead. The p-value for this test is $p < 0.0001$, so we would strongly reject the null hypothesis. This is the same conclusion we derived “by hand”.

2.4.7. Summary

In this section, we have expanded our ability to model more complex relationships by using transformed predictor variables. The essential part of this process lies in determining which variables should be transformed, and what transformation is appropriate. For this, we can use the partial residual plots as a guide, but we must still assess the accuracy and validity of the transformed model. This section also introduced a very important hypothesis test – the nested model F test. This test is used to determine if it is appropriate to remove two or more predictors from a model.

Section 2.5. Qualitative Predictors

In the previous section, we considered a general linear regression model in which some of the predictor variables were transformations of the original variables in the dataset. This is not the only way that a regression model can become “general”. It is also possible for a general linear regression model to incorporate non-numeric variables. Such variables are called qualitative variables. They are also called categorical variables, and SAS calls them classification variables. Examples include Gender (with values Male or Female) and Season (with values Spring, Summer, Fall, Winter). There are no units associated with qualitative predictors, but they must be converted to numbers before they can be used in a regression model.

Regression models that include both numeric predictors and qualitative predictors are called ANCOVA models. This is an acronym for Analysis of Covariance.

2.5.1. Indicator ('dummy') variables

Every qualitative variable must be converted to numeric before it can be used in a regression model. The possible values for a qualitative variable are called the “levels”. For a qualitative predictor with k levels, there will be $k-1$ variables in the model. These variables are called indicator (or dummy) variables. Every indicator variable will have the value either 0 or 1.

For example, if the qualitative variable is Gender (with levels Male and Female), there will be 1 indicator variable. This could be defined as

$$X_1 = \begin{cases} 1 & \text{if Gender is Male} \\ 0 & \text{otherwise} \end{cases}$$

or it could be defined as

$$X_1 = \begin{cases} 1 & \text{if Gender is Female} \\ 0 & \text{otherwise} \end{cases}$$

For another example, consider the qualitative variable Season, with levels Spring, Summer, Fall and Winter. Since there are 4 levels, there will be 3 indicator variables. These could be defined as

$$X_1 = \begin{cases} 1 & \text{if Season is Spring} \\ 0 & \text{otherwise} \end{cases} \quad X_2 = \begin{cases} 1 & \text{if Season is Summer} \\ 0 & \text{otherwise} \end{cases} \quad X_3 = \begin{cases} 1 & \text{if Season is Fall} \\ 0 & \text{otherwise} \end{cases}$$

One commonly asked question is: Why is there not an indicator variable for Winter? The short answer is: Because we don't need one.

- If Season = Spring, then . . . $X_1 = 1$, $X_2 = 0$ and $X_3 = 0$
- If Season = Summer, then . . $X_1 = 0$, $X_2 = 1$ and $X_3 = 0$
- If Season = Fall, then $X_1 = 0$, $X_2 = 0$ and $X_3 = 1$
- If Season = Winter, then . . . $X_1 = 0$, $X_2 = 0$ and $X_3 = 0$

If all three of the indicator variables are equal to 0, then Season must be Winter.

Every qualitative variable will have one level that is absent from the indicator variables. The missing level is called the reference level.

2.5.2. Example: headache drugs

We want to compare two over-the-counter pain relievers: acetaminophen and ibuprofen. Specifically, we want to know which one works better on headaches. We recruit 50 human volunteers, all of whom routinely suffer from chronic headaches, and we randomly assigned 25 to take acetaminophen and the other 25 to take ibuprofen. The volunteers do not know which medication they were given. The next time a volunteer has a headache, he recorded the severity of the headache (on a scale from 1 to 10), then took the assigned medication and recorded how long (in minutes) it took to get pain relief. It was necessary to record the severity of the headache, because that could impact the time to get relief.

The response variable is time until pain relief. There are two predictor variables: Severity and Drug. Severity is a numeric variable and Drug is qualitative with two possible values: acetaminophen or ibuprofen. Since Drug has two levels, it requires one indicator variable. The model will have two predictor variables, defined by

$$X_1 = \text{Severity} \quad \text{and} \quad X_2 = \begin{cases} 1 & \text{if Drug is acetaminophen} \\ 0 & \text{otherwise} \end{cases} \quad (2.20)$$

Note that Severity is not using indicator variables because it is already numeric. The indicator variable for Drug is X_2 .

We will analyze a dataset corresponding to this example, but first we need to understand how to set up the model.

2.5.3. Additive model

Continuing with the headache example, consider the linear model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \quad (2.21)$$

Although this model looks just like the models we have been working with, this model is different because the value for indicator variable X_2 must be either 0 or 1.

If the drug is ibuprofen, then $X_2 = 0$ and the model becomes

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 \cdot 0 + \varepsilon_i \\ Y_i &= \beta_0 + \beta_1 X_{1i} + \varepsilon_i \end{aligned}$$

If the drug is acetaminophen, then $X_2 = 1$ and the model reduces to

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 \cdot 1 + \varepsilon_i \\ Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 + \varepsilon_i \\ Y_i &= (\beta_0 + \beta_2) + \beta_1 X_{1i} + \varepsilon_i \end{aligned}$$

Compare these two equations:

For ibuprofen: $Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$

For acetaminophen: $Y_i = (\beta_0 + \beta_2) + \beta_1 X_{1i} + \varepsilon_i$

Both of these are equations for straight lines. For ibuprofen, the slope is β_1 and the intercept is β_0 .

For acetaminophen, the slope is β_1 and the intercept is $\beta_0 + \beta_2$. These two lines have the same slope, but their intercepts could possibly be different. In order to determine if there is any difference between the two drugs, we need to test whether or not the intercepts are the same. In other words, we need to test whether or not $\beta_0 = \beta_0 + \beta_2$, i.e., whether or not $\beta_2 = 0$. The official hypotheses are

$$H_0 : \beta_2 = 0 \text{ vs. } H_a : \beta_2 \neq 0.$$

- If H_0 is true, the intercepts are the same. There is no difference between these two drugs.
- If H_a is true, the intercepts are not the same. There IS a difference between these two drugs.

The model defined by equation (2.21) is called an additive model, and it will produce one straight line for each level of the qualitative predictor. Because it is an additive model, all the lines will be parallel (i.e., they will all have the same slope), but their intercepts can be different.

2.5.4. Interaction model

An additive model forces the lines to be parallel, and this will be an inappropriate model if we want to consider the possibility that one drug works better for mild (low severity) headaches, while the other drug works better for extreme (high severity) headaches. To incorporate this possibility, we need to use an interaction model.

An interaction model uses the two predictor variables that are defined in equation (2.20), but it also includes a term that is the product of X_1 and X_2 , so that the model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i}X_{2i} + \varepsilon_i \quad (2.22)$$

The new term is called the interaction term.

If the drug is ibuprofen, then $X_2 = 0$ and the model becomes

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 \cdot 0 + \beta_3 X_{1i} \cdot 0 + \varepsilon_i \\ Y_i &= \beta_0 + \beta_1 X_{1i} + \varepsilon_i \end{aligned}$$

If the drug is acetaminophen, then $X_2 = 1$ and the model reduces to

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 \cdot 1 + \beta_3 X_{1i} \cdot 1 + \varepsilon_i \\ Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 + \beta_3 X_{1i} + \varepsilon_i \\ Y_i &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_{1i} + \varepsilon_i \end{aligned}$$

Compare these two equations:

$$\text{For ibuprofen: } Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

$$\text{For acetaminophen: } Y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_{1i} + \varepsilon_i$$

Like the additive model, both of these equations are lines. Also like the additive model, the intercepts of these two lines can be different (β_0 vs. $\beta_0 + \beta_2$). With the interaction model, however, the slopes are not forced to be equal because one slope is β_1 and the other slope is $\beta_1 + \beta_3$. If we want to test whether or not the slope for one drug is equal to the slope for the other drug, then we want to test whether or not $\beta_1 = \beta_1 + \beta_3$, i.e., whether or not $\beta_3 = 0$. If we want to test whether or not the intercept for one drug is equal to the intercept of the other drug, then we want to test whether or not $\beta_0 = \beta_0 + \beta_2$, i.e., whether or not $\beta_2 = 0$. In an interaction model, there is no direct hypothesis test for comparing acetaminophen to ibuprofen, because this could depend on the value of X_1 .

2.5.5. SAS programming notes

Whenever the model contains one or more qualitative/categorical/classification predictors, we cannot use PROC REG to fit the model. Instead, we must use PROC GLM. (GLM stands for General Linear Model.) The SAS statements that go into PROC GLM are, in many cases, similar to those in PROC REG. But there are some basic differences.

1. To read character data into a SAS dataset, put a dollar sign (\$) after the name of the variable in the INPUT statement in the DATA step. The dollar sign is not needed anywhere else in the program.
2. PROC GLM requires a CLASS statement that lists all the qualitative predictors that will be included in the model. SAS uses this list to create all the necessary indicator variables for all the qualitative predictors. By default, SAS uses the last level of qualitative predictor as the reference level. (In this sense, “last” is last alphabetically.) If the levels are Male and Female, SAS will choose Male as the reference level. If the levels are Spring, Summer, Fall and Winter, SAS will choose Winter as the reference level.
3. PROC GLM does not automatically generate the diagnostic plots. We need to include the option PLOTS=DIAGNOSTICS or the option PLOTS=ALL to get these graphs.
4. PROC GLM does not automatically generate the Parameter Estimates table. To get this table, we need to include the option SOLUTION on the MODEL statement.
5. PROC GLM can have only one MODEL statement. If two models are to be considered, then two PROC GLM's are needed.

2.5.6. Fit models to the headache data

The SAS code to fit both the additive and the interaction models is shown below. Note the dollar sign after the word Drug in the INPUT statement. This is necessary because the values for the variable Drug are character data. The CLASS statement in PROC GLM instructs SAS to define the indicator variable for Drug. Since Drug has 2 levels (acetaminophen and ibuprofen), there will be only one indicator variable and SAS will automatically make ibuprofen the reference level. The only difference between the two PROC GLM's is in the MODEL statement. The interaction model includes the term Drug*Severity, while the additive model excludes this term.

```

DATA headaches;
INPUT Subject Drug $ Severity Time; ; * use $ after character variable;
DATALINES;
1 acetaminophen      2 15.4
2 acetaminophen      3 8.4
... more datalines here ...
;

TITLE 'Additive Model';
PROC GLM DATA = headaches PLOTS=ALL;
CLASS Drug;
MODEL Time = Severity Drug / SOLUTION;
RUN;

TITLE 'Interaction Model';
PROC GLM DATA = headaches PLOTS=ALL;
CLASS Drug;
MODEL Time = Severity Drug Drug*Severity / SOLUTION;
RUN;

```

This is the beginning of the complete SAS output.

Additive Model

The GLM Procedure

Class Level Information		
Class	Levels	Values
Drug	2	acetamin ibuprofe

Number of Observations Read	50
Number of Observations Used	50

Comments:

- The Class Level Information table will have one row for each variable listed in the CLASS statement. The levels ("Values") are given in alphabetical order. The last one is the reference level. There is not an indicator variable for the reference level.
- Make sure the "Number of Observations Read" matches the number of rows in the dataset.

Additive Model

The GLM Procedure

Dependent Variable: Time

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1250.150116	625.075058	24.49	<.0001
Error	47	1199.758084	25.526768		
Corrected Total	49	2449.908200			

R-Square	Coeff Var	Root MSE	Time Mean
0.510284	26.05137	5.052402	19.39400

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Severity	1	1035.021642	1035.021642	40.55	<.0001
Drug	1	215.128474	215.128474	8.43	0.0056

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Severity	1	1158.484316	1158.484316	45.38	<.0001
Drug	1	215.128474	215.128474	8.43	0.0056

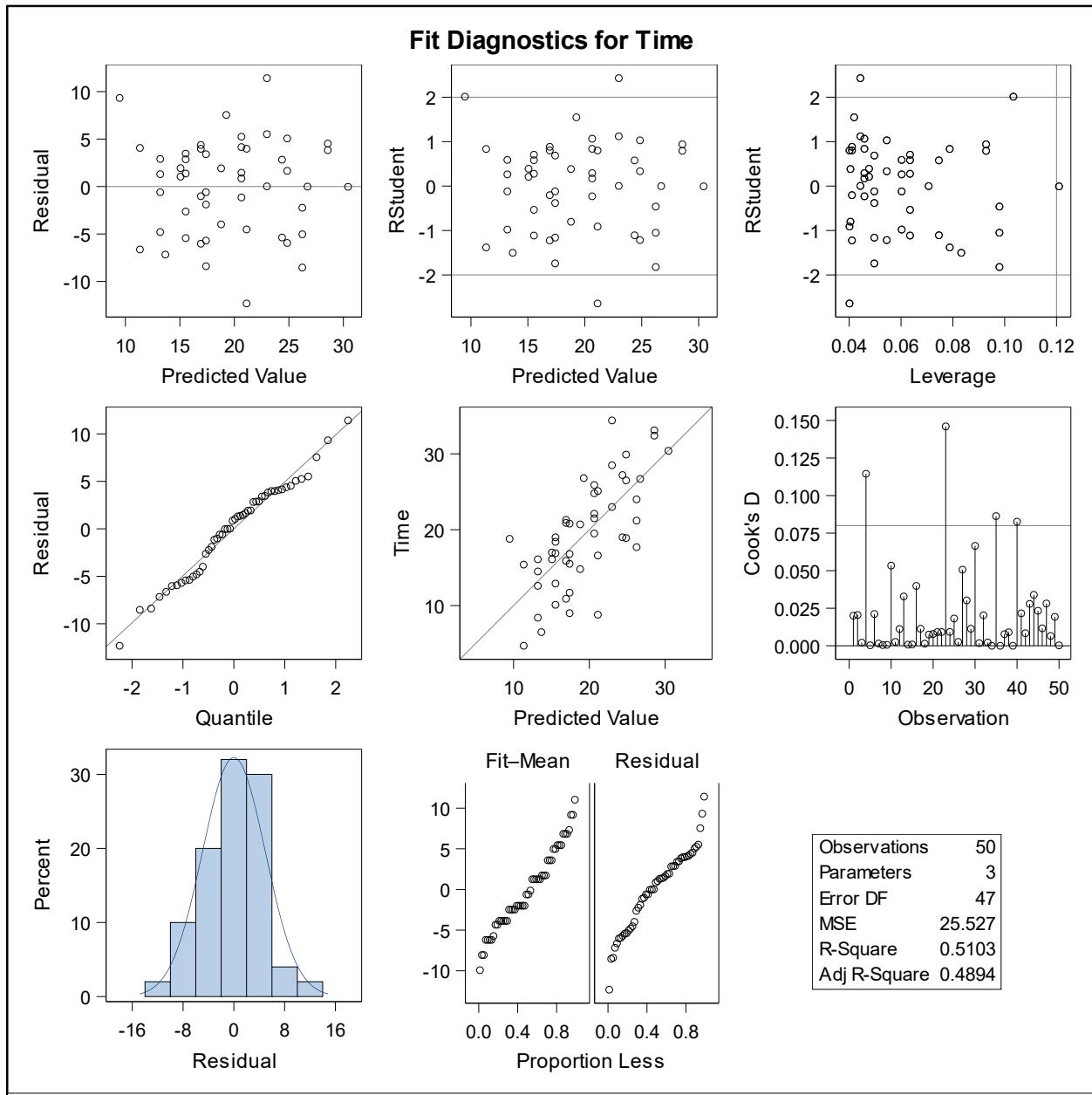
Parameter	Estimate	Standard Error		t Value	Pr > t
		B	t		
Intercept	11.80850299	B	1.66792060	7.08	<.0001
Severity	1.86239521		0.27645515	6.74	<.0001
Drug acetamin	-4.19791617	B	1.44604806	-2.90	0.0056
Drug ibuprofe	0.00000000	B	.	.	.

Note: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

Comments:

- The top table is the ANOVA table, and the second table gives R-square and RMSE.
- The third table is for Type I sums of squares. **IGNORE THIS TABLE.**
- The next table is the Type III sums of squares. We will interpret these tests.
- The last table is the result of the SOLUTION option.
- Ignore the note at the bottom of the page. This note will appear every time there is a classification variable.

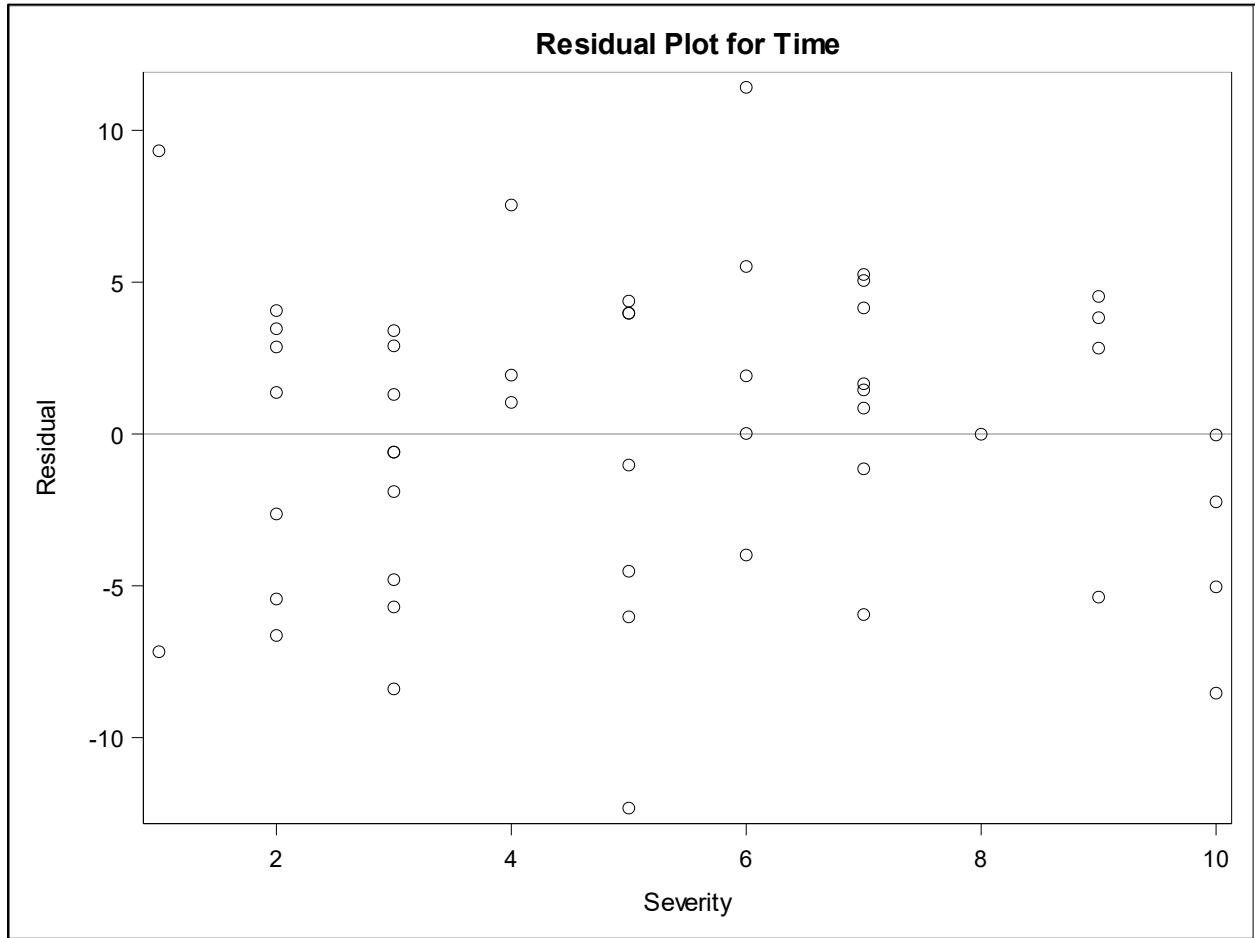
Additive Model
The GLM Procedure
Dependent Variable: Time



Comments:

This is the same panel of diagnostic plots as those produced by PROC REG. The residual plot is in the upper left corner, and the normal QQ plot is immediately below it.

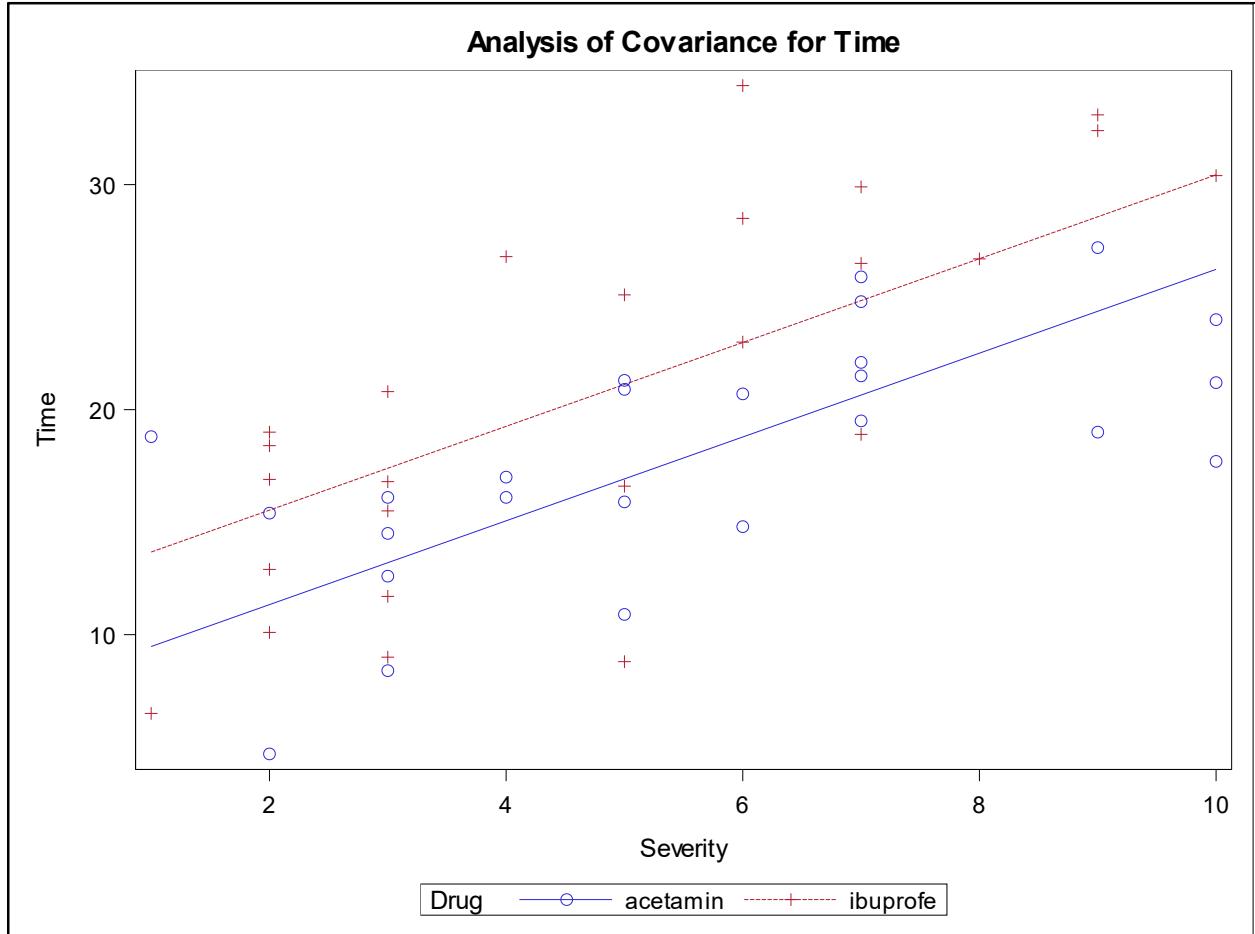
Additive Model
The GLM Procedure
Dependent Variable: Time



Comments:

This is another residual plot. It looks different than that one in the upper left corner in the panel of plots (on the previous page) because this one has the Severity values on the x-axis and the other graph has the fitted values (the \hat{Y} 's) on the x-axis.

Additive Model
The GLM Procedure
Dependent Variable: Time



Comments:

This is a scatterplot of the data, using different colors and plotting symbols for the two levels of Drug. The lines are the estimated regression lines. Note that the lines are parallel because this is an additive model.

This is the end of the output for the additive model. The interaction model starts on the next page.

Comment:

This is the beginning of the output for the interaction model.

All of the comments regarding the additive model also apply here, except for the graph on the last page of the output.

Interaction Model

The GLM Procedure

Class Level Information		
Class	Levels	Values
Drug	2	acetamin ibuprofe

Number of Observations Read	50
Number of Observations Used	50

Interaction Model

The GLM Procedure

Dependent Variable: Time

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1370.815233	456.938411	19.48	<.0001
Error	46	1079.092967	23.458543		
Corrected Total	49	2449.908200			

R-Square	Coeff Var	Root MSE	Time Mean
0.559537	24.97371	4.843402	19.39400

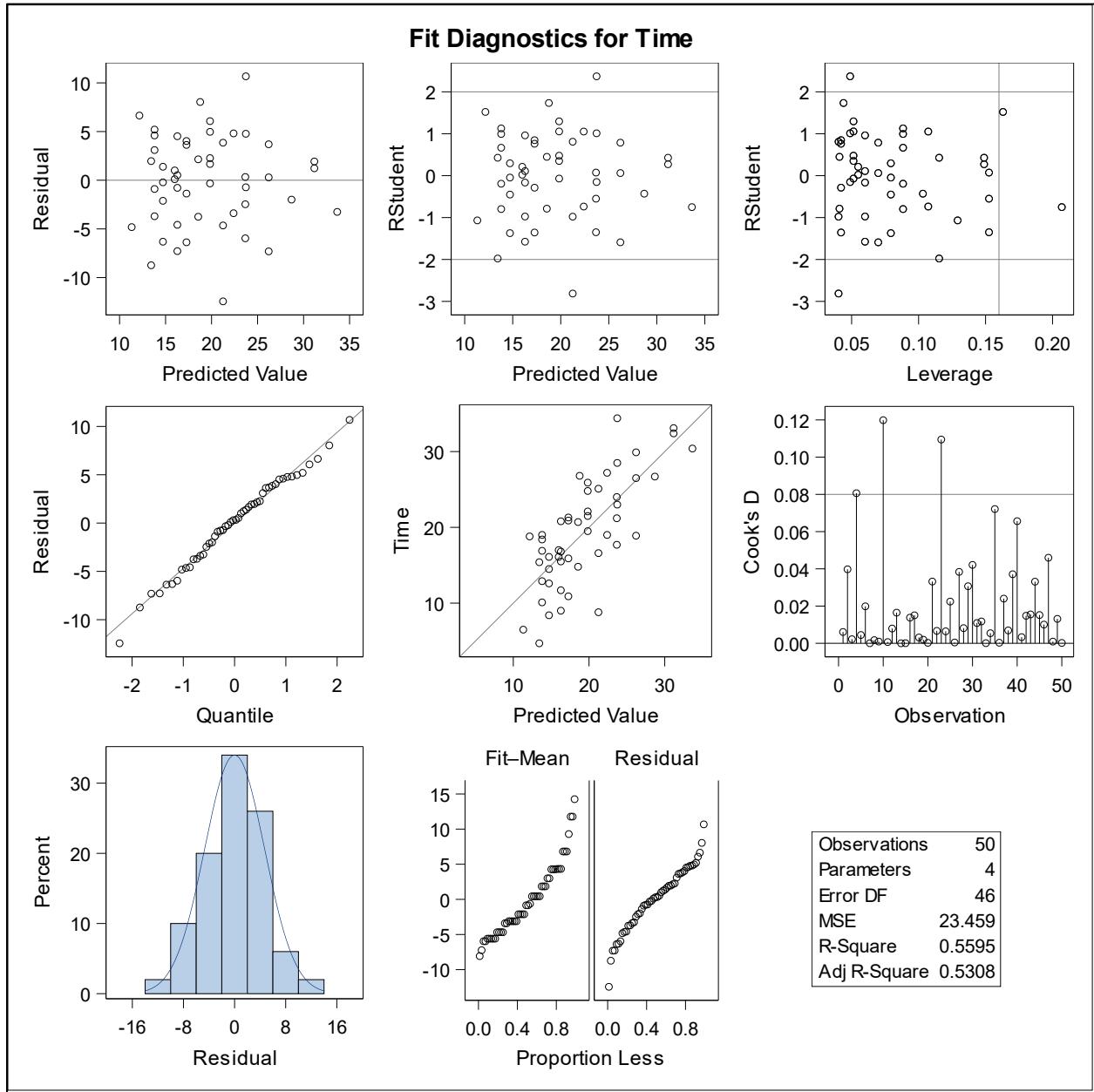
Source	DF	Type I SS	Mean Square	F Value	Pr > F
Severity	1	1035.021642	1035.021642	44.12	<.0001
Drug	1	215.128474	215.128474	9.17	0.0040
Severity*Drug	1	120.665117	120.665117	5.14	0.0281

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Severity	1	1179.932182	1179.932182	50.30	<.0001
Drug	1	10.302231	10.302231	0.44	0.5108
Severity*Drug	1	120.665117	120.665117	5.14	0.0281

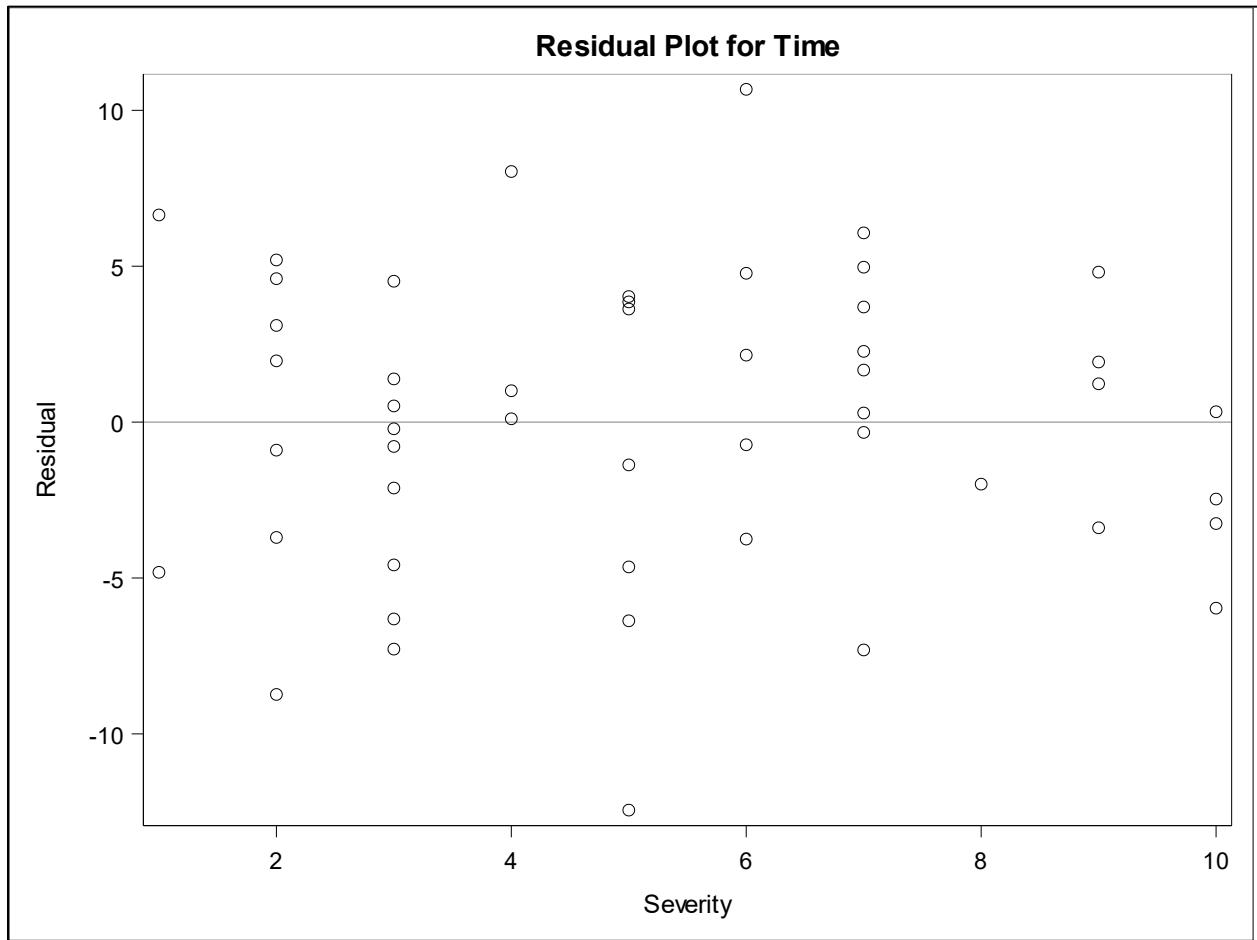
Parameter	Estimate		Standard Error	t Value	Pr > t
Intercept	8.835703704	B	2.06752698	4.27	<.0001
Severity	2.481728395	B	0.38053360	6.52	<.0001
Drug acetamin	2.041505599	B	3.08060091	0.66	0.5108
Drug ibuprofe	0.000000000	B	.	.	.
Severity*Drug acetamin	-1.202658628	B	0.53027606	-2.27	0.0281
Severity*Drug ibuprofe	0.000000000	B	.	.	.

Note: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

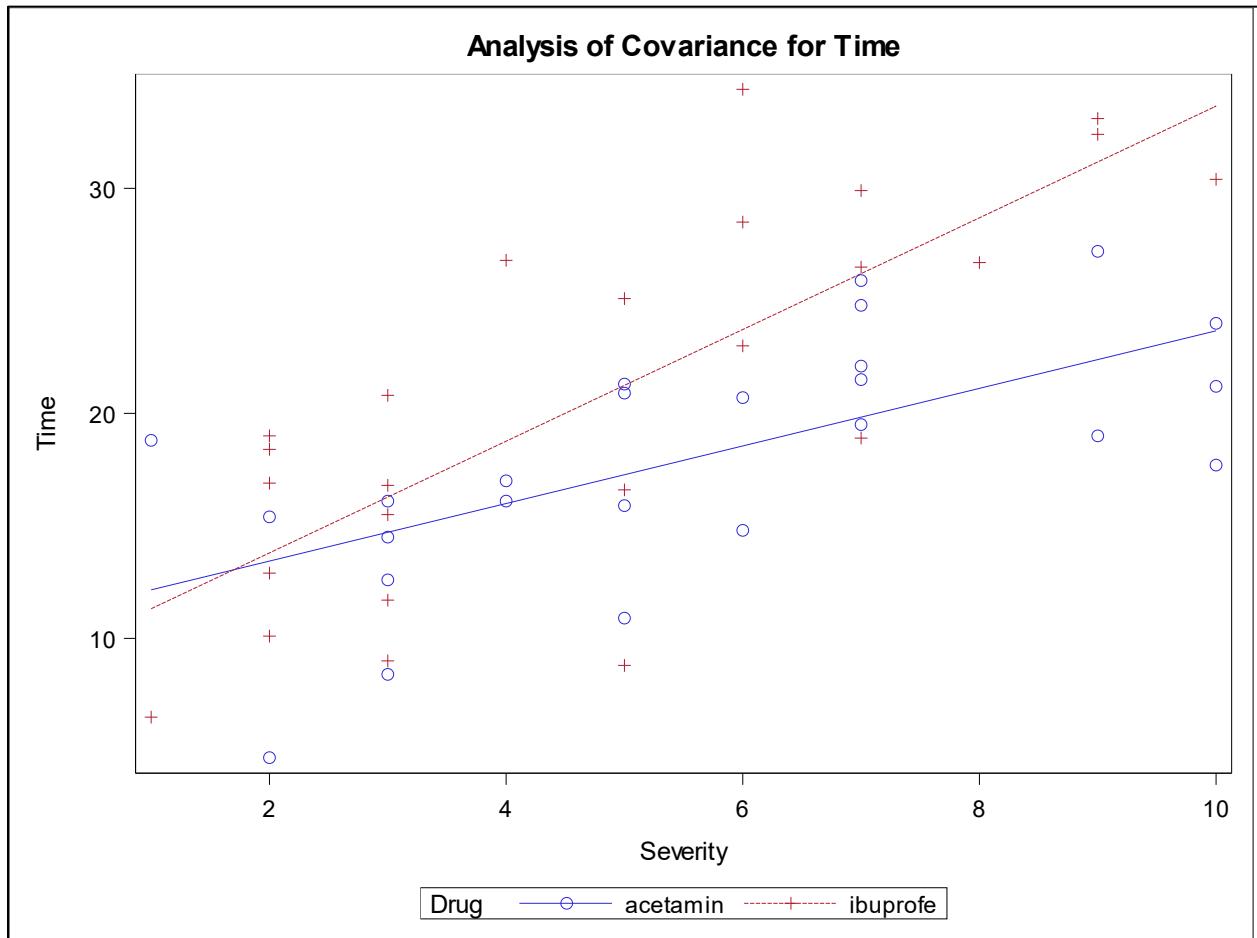
Interaction Model
The GLM Procedure
Dependent Variable: Time



Interaction Model
The GLM Procedure
Dependent Variable: Time



Interaction Model
The GLM Procedure
Dependent Variable: Time



Comments:

This is a scatterplot of the data, using different colors and plotting symbols for the two levels of Drug. The lines are the estimated regression lines. Note that the lines are **NOT** parallel because this is an interaction model.

This is the end of the complete SAS output.

2.5.7. Interpret the SAS output

The additive model

The equation for the model is $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$

where $X_1 = \text{Severity}$ and $X_2 = \begin{cases} 1 & \text{if Drug is acetaminophen} \\ 0 & \text{otherwise} \end{cases}$

Both the residual plot and the normal QQ plot look very good. There is no reason to suspect the assumptions have been violated.

The overall ANOVA F test is testing $H_0 : \beta_1 = 0$ and $\beta_2 = 0$ vs. $H_a : \beta_1 \neq 0$ or $\beta_2 \neq 0$. The test statistic is $F = 24.49$, with $p < .0001$. We reject H_0 . (Note: If we did not reject H_0 , the interpretation of this model would end here. We would not look at any more hypothesis tests for this model.)

Next, we interpret the test results given in the Type III sums of squares table.

- The line labeled ‘Severity’ is testing the hypotheses $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$.

Note that this is the slope on the numeric predictor Severity.

The test statistic is $F = 45.38$, with $p < .0001$.

We reject H_0 and conclude the slope on Severity is not 0.

- The line labeled ‘Drug’ is testing the hypotheses $H_0 : \beta_2 = 0$ vs. $H_a : \beta_2 \neq 0$.

Note that this is the multiplier on the indicator variable. It is also the parameter that will tell us if there is a difference between the two drugs (H_a) or if the two drugs are the same (H_0).

The test statistic is $F = 8.43$, with $p = 0.0056$.

We reject H_0 and conclude the two drugs are different.

The next step is to examine the Parameter Estimates table. Each line in this table contains a point estimate and the standard error of the estimate, as well as a test statistic and p-value for testing exactly one parameter in the model. Since the classification variable Drug has exactly two levels, the t tests given in the Parameter Estimates table are testing the same hypotheses as the F tests given in the Type III sums of squares table. The corresponding tests have the same p-value, and the test statistics are related by $F = t^2$. If the classification variable had more than two levels, these tests would be different.

The Parameter Estimates table also provides the point estimate for each parameter in the model. The line labeled ‘Drug ibuprofen’ has an estimate of 0 and dots for the remaining entries. This is because ibuprofen is the reference level for Drug, so it does not explicitly appear in the model. The least squares regression equation for the additive model is $Y = 11.8085 + 1.8624X_1 - 4.1979X_2$.

- For ibuprofen ($X_2 = 0$):
$$Y = 11.8085 + 1.8624X_1$$
- For acetaminophen ($X_2 = 1$):
$$Y = (11.8085 - 4.1979) + 1.8624X_1$$

which simplifies to
$$Y = 7.6106 + 1.8624X_1$$

These are the two lines in the last graph of the output for the additive model. They have the same slope, but different intercepts.

The final question: Is there a difference between ibuprofen and acetaminophen? Based on our analysis of the additive model, the answer is yes. This conclusion comes from the hypothesis test for β_2 , and we can use either the t test for ‘Drug ibuprofen’ in the Parameter Estimates table or the F test for ‘Drug’ in the Type III sums of squares table. The p-value is 0.0056.

The interaction model

The equation for the interaction model is $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i}X_{2i} + \varepsilon_i$

where $X_1 = \text{Severity}$ and $X_2 = \begin{cases} 1 & \text{if Drug is acetaminophen} \\ 0 & \text{otherwise} \end{cases}$

The additional term in the model ($\beta_3 X_{1i}X_{2i}$) is called the interaction term, and it allows for the possibility that the two levels of Drug might have different slopes.

Both the residual plot and the normal QQ plot look very good. There is no reason to suspect the assumptions have been violated.

The overall ANOVA F test is testing $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ vs. $H_a : \text{at least one of these } \beta\text{'s is not 0}$.

The test statistic is $F = 19.48$, with $p < .0001$, so we reject H_0 . (Note: If we did not reject H_0 , the interpretation of this model would end here. We would not look at any more hypothesis tests for this model.)

Next, we interpret the test results given in the Type III sums of squares table. Look at the interaction test first. This test is on the line labeled ‘Severity*Drug’. It is testing $H_0 : \beta_3 = 0$ vs. $H_a : \beta_3 \neq 0$. The test statistic is $F = 5.14$, with $p = 0.0281$. We reject H_0 and conclude that the interaction is significant.

By rejecting H_0 in the interaction test, we are concluding that the slopes of the lines are significantly different. Because the interaction is significant, there is no valid interpretation of the test for Severity or the test for Drug. The point estimates for the model parameters are still valid, but when the interaction is significant the hypothesis tests for two main parameters must be ignored.

The least squares regression equation for the interaction model is

$$Y = 8.8357 + 2.4817X_1 + 2.0415X_2 - 1.2027X_1X_2$$

- For ibuprofen ($X_2 = 0$):

$$\begin{aligned} Y &= 8.8357 + 2.4817X_1 + 2.0415(0) - 1.2027X_1(0) \\ Y &= 8.8357 + 2.4817X_1 \end{aligned}$$

- For acetaminophen ($X_2 = 1$):

$$\begin{aligned} Y &= 8.8357 + 2.4817X_1 + 2.0415(1) - 1.2027X_1(1) \\ Y &= 8.8357 + 2.4817X_1 + 2.0415 - 1.2027X_1 \\ Y &= 10.8772 + 1.2790X_1 \end{aligned}$$

These are the two lines that are graphed at the end of the output for the interaction model.

Since the interaction is significant, it is more difficult to answer the question: Is there a difference between ibuprofen and acetaminophen? From the graph (the very last graph in the output), it appears that there is not a difference when the severity is approximately 2, because that is where the two lines intersect. There appears to be a difference when the severity is higher, but it is not clear at what value of severity will the differences between the two drugs become significant. We could answer this question by generating confidence intervals for the mean for each drug, and at various values for Severity. We would accomplish this the same way we did with PROC REG. Include the CLM option on the MODEL statement, and add extra lines of data in the DATA step. For example, we could include the following lines of data. Each line has a value (or a dot) for each variable in the INPUT statement: Obs, Drug, Severity, and Time.

```

51 acetaminophen 1 .
52 ibuprofen 1 .
53 acetaminophen 2 .
54 ibuprofen 2 .
55 acetaminophen 3 .
56 ibuprofen 3 .
57 acetaminophen 4 .
58 ibuprofen 4 .
59 acetaminophen 5 .
60 ibuprofen 5 .
61 acetaminophen 6 .
62 ibuprofen 6 .

```

These extra lines of data (in conjunction with the CLM option on the MODEL statement) tells SAS to generate a confidence interval for the mean time to pain relief for each drug and each value of Severity between 1 and 6. For each of these Severity values, compare the confidence interval for acetaminophen to the confidence interval for ibuprofen. If the confidence intervals overlap, then there is not a significant difference between the two drugs for that particular value of Severity. If the confidence intervals do not overlap, then there is a significant difference. Repeat this comparison for every value of Severity. This is a bit more work than what we have been doing, but the statistical analysis typically becomes more complicated when there is a significant interaction.

For the headache data, the confidence intervals for the mean time to relief are summarized in Table 2.14. The confidence interval for acetaminophen overlaps the confidence interval for ibuprofen when the Severity value is 4 or less, but they do not overlap when the Severity is 5 or more. We would conclude that there is not a difference between these two drugs when the severity of the headache is 4 or less, but there is significant difference when the severity is 5 or more.

Severity	for acetaminophen	for ibuprofen
1	(8.22, 16.09)	(7.81, 14.82)
2	(10.12, 16.75)	(10.90, 16.70)
3	(11.97, 17.46)	(13.90, 18.67)
4	(13.71, 18.28)	(16.72, 20.81)
5	(15.27, 19.27)	(19.29, 23.20)
6	(16.58, 20.52)	(21.57, 25.88)

Table 2.14. Confidence Limits for Mean Time to Pain Relief

2.5.8. Summary

Non-numeric predictor variables are called by several different names, including classification variables, qualitative variables, and categorical variables. A general linear regression model can accommodate these non-numeric predictors, but these must be converted to numeric before they can go into the model. The conversion is accomplished via indicator (or dummy) variables, which always have the value either 0 or 1. Whenever a regression model contains non-numeric predictors, the SAS code must use PROC GLM instead of PROC REG, and we must include a CLASS statement to instruct SAS to create the appropriate indicator variables.

The relationship between the model equation and the SAS output is extremely important to properly interpret the results of the model. It is imperative to know how the indicator variables are defined, and how the interaction and the indicator variables are incorporated into the model. This dictates which part of the SAS output needs to be examined in order to answer specific questions.

We have examined a model that contains exactly one numeric predictor and exactly one non-numeric predictor, so the fitted model consists of one regression line for each value of the non-numeric predictor. It is possible to have more than one numeric predictor, and then each regression “line” becomes a regression “surface”. It is also possible to have more than one non-numeric predictor, and we will consider this in Chapter 5.

Section 2.6. Influence and Outliers

So far, we have concentrated on developing a model that will fit the data. In this section, we focus on the data, both by itself and in conjunction with a model. In addition, we will consider only numeric variables, so we will not be working with indicator variables. To illustrate the concepts of influence and outliers, we will use the body fat data that we first explored in Section 2.2.

We examine methods for detecting certain anomalies in the data. In some cases, the anomalies may be mistakes in the data (e.g., a misplaced decimal or a typographical error). These should be corrected to prevent them from invalidating the statistical analysis. In other cases, the anomalies are not mistakes in the data, and these “unusual” observations often lead to important insights about the data or the process that generated the data.

One observation is one complete row in the dataset. It contains a value for the response variable and a value for each of the predictors that are in the model. There can be other variables in the dataset that are not in the model, but these variables are ignored. We often call one observation a ‘point’ or a ‘data point’ because we can imagine it plotted in hyperdimensional space. There are several ways in which an observation can be unusual or ‘extreme’, including

- The combination of values for the predictors may be unusual, without any regard to the value for the response.
- The value of the response may be unusual for the specified values of the predictors.
- The data point may have a lot of influence on the model, so that if this point was removed from data, the results of the analysis could change dramatically.

2.6.1. Leverage

A data point has high leverage if its combination of values for the predictors is unusual in relation to all the other rows in the data. If there is only one predictor variable, points with high leverage appear separated (to the right or the left) of the other points. This is illustrated in Figure 2.17. The point plotted as a red star has high leverage because its X value is much larger than the other X’s in the data. The value for the response is not used to calculate leverage, so that a point with high leverage may or may not seem to follow the least squares line. In other words, the red star point in Figure 2.17 could move straight up or straight down and it would still have high leverage.

When there are multiple predictor variables, it is not possible to “see” leverage in a graph. This is because leverage is based on the combination of values for all the predictors, not just on the value of one predictor. The formula for calculating the leverage is beyond the scope of our course, and we will let SAS generate it for us. There will be one value for leverage for each observation in the dataset. It is considered “high” leverage when Leverage $> \frac{2}{n} \cdot (\# \text{parameters})$, where n is the number of observations in the data.

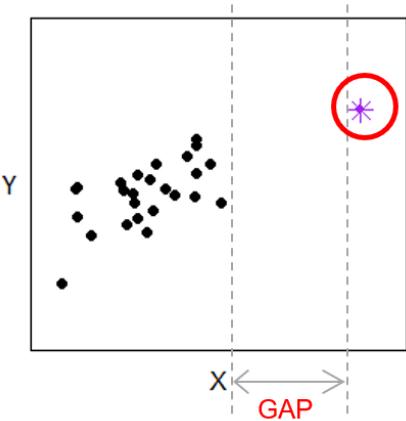


Figure 2.17. A point with high leverage

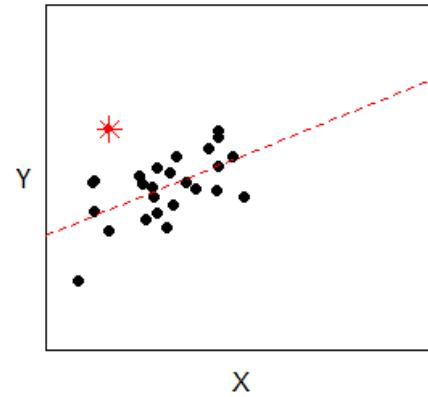


Figure 2.18. A potential outlier

2.6.2. Outliers

Outliers are observations that have unusually large or small Y values, in relation to the values of the X's that are recorded for the observation. Identification of outliers is based on the residuals, so we need to fit the model in order to find outliers in the data. Outliers have very ‘large’ residuals (either positive or negative), which means the point is ‘far’ away from the regression surface. A potential outlier is plotted as a red star in Figure 2.18. This point is farther away from the regression line than all the other points, but it is not clear if this point is actually an outlier.

We need a mechanism to decide how large a residual needs to be in order for us to declare it an outlier. We know that all the observations will have some variation around the regression surface and the amount of this variation is measured by error variance σ^2 . We do not know that value for σ^2 , but it is estimated by the MSE. Therefore, ‘large’ residuals are large relative to the MSE.

We standardize the residual by subtracting the mean and dividing by the square root of the MSE. If the residuals are independent, then the standardized residuals would follow an approximate normal

distribution. But the standardized residuals all use the same value for the MSE, which is calculated from the entire dataset. So the standardized residuals are not independent, and we do not know what probability distribution they follow.

Theoretical statisticians have shown that, if we make a small modification to the standardized residuals, then they will follow an approximate t distribution. This modification is to divide the standardized residual by the square root of (1 – leverage). These are called the studentized residuals.

- A mild, or potential, outlier has studentized residual greater than 2 or less than -2.
- An extreme outlier has studentized residual greater than 3 or less than -3.

2.6.3. Influence

The influence of a point is a measure of how much the fitted model would change if the point was removed from the data set. (Recall that a “point” is one row in the dataset.) The influence can be measured as change in the fitted values and/or a change in the estimates for the individual coefficients. A point can be influential because it has high leverage or because it is an outlier or because it is both an outlier and has high leverage.

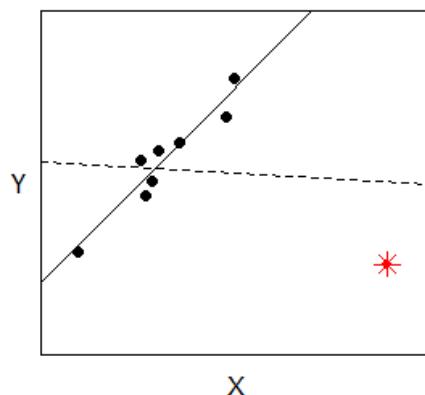


Figure 2.19. An influential point

An extremely influential point is plotted as a red star in Figure 2.19. If this point is removed from data, we obtain the estimated regression line $Y = 1.17 + 1.27X$, which is graphed as a solid line. The value for R-square is 91.3% and RMSE = 0.92. If this point is kept in the data, the estimated regression line is $Y = 14.36 - 0.07X$, which is the dashed line. The value for R-square is 0.7% and RMSE = 3.55.

To identify influential points, we employ a “leave one out” strategy. This process is repeated for each observation in the dataset. The model is fitted with the observation included, then the observation is

removed and the model is fitted again. The amount of change in the model can be measured by several different criteria:

- DFFITS - Influence on single fitted values
- Cook's distance - Influence on all fitted values
- DFBETAS – Influence on each regression coefficient

We will concentrate on Cook's distance. All of these can be generated in SAS by using the option INFLUENCE on the MODEL statement.

```
PROC REG DATA=fat;
  MODEL bodyfat = triceps midarm / INFLUENCE;
  RUN;
```

The INFLUENCE option produces the output shown in Table 2.15. Some of the rows have been removed; the full table will have one row for each observation in the data. DFFITS and DFBETAS are clearly labeled. The studentized residuals are in the column RStudent and the leverage values are in the column 'Hat Diag H'. This table does not contain Cook's distance.

Output Statistics								
Obs	Residual	RStudent	Hat Diag H	Cov Ratio	DFFITS	DFBETAS		
						Intercept	triceps	midarm
1	-1.8481	-0.8084	0.1785	1.2948	-0.3768	-0.0142	0.3087	-0.2152
2	3.4606	1.4734	0.0538	0.8654	0.3514	0.0058	-0.0755	0.0837
3	-2.8462	-1.5271	0.3988	1.3266	-1.2439	1.0563	0.0525	-1.0572
...
19	-3.0128	-1.2703	0.0648	0.9613	-0.3343	-0.1127	0.1537	-0.0321
20	0.9583	0.3839	0.0501	1.2284	0.0881	0.0140	-0.0006	-0.0024

Table 2.15. SAS output for INFLUENCE option

Outliers: To identify extreme outliers, scan the column RStudent for values that are greater than 3 or less than -3. Mild, or potential, outliers will have RStudent values greater than 2 or less than -2.

Leverage: To identify points that have high leverage, scan the column 'Hat Diag H'. For this particular model, there are $n = 20$ observations and 3 parameters (intercept, triceps and midarm), so an

observation has high leverage if the value is greater than $\frac{2}{n} \cdot (\# \text{parameters}) = \frac{2}{20} \cdot (3) = 0.3$. From Table

2.15, we can see that observation #3 has high leverage.

Influence on regression coefficients: DFBETAS measures the influence on each regression coefficient, so there is one column for each coefficient. An observation is considered influential if the absolute value is

greater than $\frac{2}{\sqrt{n}}$. For this particular model, we have $\frac{2}{\sqrt{20}} = 0.447$. From Table 2.15, we can see that

observation #3 has influence on the intercept and the slope on midarm. .

Cook's distance: The results for Cook's distance are not shown in Table 2.15. Instead, they are shown in as a graph in Figure 2.20. In the SAS output, this graph is located in the panel of plots (where we find the residual plot and the normal QQ plot). The Cook's D graph is in the middle row on the far right. The graph contains a vertical line for each observation in the data, and the height of the line is the value for Cook's D. The graph also contains a horizontal line for reference. Influential points have line segments that are taller than the reference line. For the body fat model, Figure 2.20 indicates that observation #3 is influential.

Before analyzing these data, we should make sure there are no errors in observation #3.

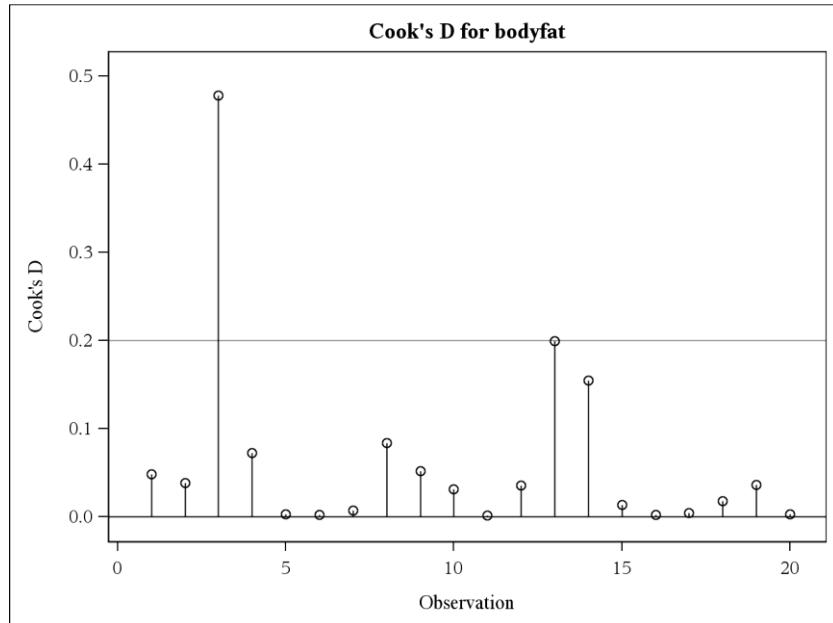


Figure 2.20. Graph for Cook's D

2.6.4. Summary

There are many different ways to identify “unusual” observations in the data. We concentrate on these three:

- Points with high leverage have an unusual combination of values for the predictor variables. They can be identified by examining the ‘Hat Diag H’ column in the SAS output.
- Outliers are points that have an unusual value for the response. They can be identified by examining the studentized residuals.
- Influential points have a large value for Cook’s distance. They can be identified by examining the graph for Cook’s D.

If we find influential points or outliers in a dataset, we do not automatically remove these observations just because they are “unusual”. First, we make sure there are no errors in the data, then we determine if there are any unique conditions under which the observation was collected (that might explain why this point is different). Ultimately, it is the experience and knowledge of the researcher that dictates whether to keep or remove a data point.

Chapter 3: Model Building

Section 3.1. Introduction

How do we decide how many and which predictors to include in a general linear regression model?

$$Y_i = \beta_0 + \sum_{j=1}^p \beta_j Z_{ji} + \varepsilon_i \quad (3.1)$$

In this chapter, we are not considering the possibility of transforming the response variable Y . We are only interested in methods for deciding which predictor variables to include in the model. Complex data sets can have dozens or more potential predictors. These can include quantitative (numeric) predictors, which can be either included or excluded from the model. It is also possible to include transformations of one or more the numeric predictor variables, but we would need to decide which predictors should be transformed and which transformation is appropriate. In addition to the numeric predictors, the data may also contain qualitative predictors. For these, we must decide whether to include or exclude all the indicator variables for each qualitative predictor. We do not transform indicator variables, and we do not remove only some of the indicators for a qualitative predictor. (All the indicators for one predictor are either in the model or not in the model.) Finally, we need to consider interactions between the predictors.

The ultimate goal is to find a set of predictors that fits the data well and generates a model that does not violate the assumptions. We are NOT looking for the ‘best’ model, because there is no such thing as a “best” model. “Best” is a relative term, and it depends on the purpose of the analysis and external constraints. If you had to choose a vehicle from those in Figure 3., which one would you choose as “best”? With linear models, different strategies can yield different different subsets of “best” predictors, and we need a systematic method for deciding which subset is preferred.



Figure 3.1. Which car is “best”?

3.1.1. Considerations

What is the nature of the study? If the data are derived from a controlled experiment, then we must consider the factors of the experiment. These include blocking factors and treatment factors, but can also include random factors which we will discuss in Chapter 6. In general, the factors involved in the controlled experiment dictate which predictors are included in the model, so the variable selection methods presented in this Chapter are usually not used for experimental data. In contrast, data derived from observational studies are often used to explore the relationships between the response and potential predictors. Variable selection methods are most often used on observational data. It is also possible for data to be derived partly from an experiment and partly observational. For these types of datasets, we would force the inclusion of the experimental variables and use the techniques of this chapter to select the observational variables.

Prior knowledge and subject-matter expertise. It is very rare for an analysis to consist entirely of variables that have never been studied. In most cases, previous research links predictor variables to the research objective, and therefore links certain predictor variables to the response. Incorporating prior knowledge may indicate that particular predictor variables be included in the model, even if the current dataset does not indicate this is warranted. In addition, we want to forcibly include predictors that we know will affect the response, even if these predictors are not of interest in our current analysis. Including such predictors allows us to control for the effects of these influences and permits more precise inference for the variables we are interested in.

Complexity of model. We want as few predictors as possible, but we also need a model that fits the data well. Models that have too few predictors can generate biased point estimates that consistently over- or under-estimate the magnitude of the relationship between Y and the X's. If there are too many predictors, the estimated parameters and predicted values may have inflated standard errors. This results in poor precision and reduces the ability to find important differences in the data.

Sample size. The complexity of the model (as measured by number of parameters) is limited by the information available in the dataset. As a general rule of thumb, we need 6 to 10 observations for each predictor in the model. Smaller datasets require smaller models, but larger datasets can accommodate more complex models.

Always check the model assumptions. Predictors may be needed, even if they do not contribute directly to the interpretation.

3.1.2. Variable Selection

It is not practical, or even possible, to look at every possible regression model. If there are 8 potential predictors, there are 256 possible models. With 10 predictors, there are over 1,000 possible models. If we consider interactions or transformations (such as $\log(X)$ or X^2), the possibilities are endless. We need a systematic approach that will generate a few candidate models, then we take a closer look at the candidate models and use personal judgement to make a final choice. There are several different strategies for generating the candidate models, and each one could produce a different candidate. At the end of the process, we have no guarantee that we have found the “best” model. Always remember that we are simply looking a model that fits the data well and does not violate the assumptions.

The basic steps in variable selection are

- Fit a series of competing models to the dataset
- Compare the models using model selection criteria
- Decide upon good candidate models
- Among the candidates, check model assumptions
- Make a final decision

As we go through this process, do not ignore common sense. Pay attention to multicollinearity issues, outliers and influential points. Do not overlook checking the assumptions of the candidate models. If a model seriously violates any of the assumptions, then the model must be discarded.

3.1.3. Criteria for Model Selection

There are several different ways to measure how well a model fits the data, and these are all based on the residuals from the fitted model. Models with small residuals are deemed a good fit to the data, and large residuals indicate a poor fit. If residuals were the only consideration, then this would be an easy task. But there are other things that can affect the goodness of fit. For example, the existence of outliers or influential points can affect the goodness of fit. In addition, more complex models (with a large number of predictors) will automatically fit the data better than a simpler model. Since we want the model to be as simple as possible, we must balance the goodness of fit with the complexity of the model. Another consideration is the purpose of the analysis. If we want to use the model for predicting new observations, measuring the goodness of fit must somehow incorporate future observations that

are not yet recorded in the current dataset. This is not an issue if we want to use the model strictly for estimating model parameters or values for the response variable.

If the goal of the analysis is to estimate or explain the data, there are several measures of goodness of fit. We will consider these:

- Coefficient of determination (R-square)
- Adjusted R-square
- Residual mean square (MSE)
- Mallow's Cp
- Akaike's Information Criterion (AIC)
- Schwarz' Bayesian Criterion (SBC)

If the goal of the analysis is to predict the response for new observations, then we will consider only the PRESS statistic to measure goodness of fit.

All of these criteria are used simply to compare candidate models. We must still use our knowledge of the subject matter to make the final selection.

Coefficient of determination (R-square)

R-square measures the proportion of total variability in the response that is explained by the fitted regression model. It is calculated according to equation (3.2), where SSReg is the sum of squares for the model, SSE is the sum of squared due to error, and SSTot is the total variability in the response.

$$0 \leq R^2 = \frac{SSReg}{SSTot} = 1 - \frac{SSE}{SSTot} \leq 1 \quad (3.2)$$

Larger values of R-square indicate a better-fitting model. There is, however, one serious drawback to using R-square to compare the goodness of fit for two different models. Every time another predictor is added to the model, the value of SSReg will increase, but the value of SSTot will remain unchanged. This implies that the value of R-square will increase as the number of predictors increases. This is a direct contradiction to our desire to have as few predictors as feasible (i.e., a parsimonious model). As a consequence, this makes R-square unsuitable for comparing the fit of two models that have different numbers of predictors.

If two models have different numbers of predictors, we cannot use R-square to compare these models. R-square will always indicate that the larger model is better. There are several ways to adjust R-square to account for the difference in the number of parameters. Some of the options are discussed below.

Adjusted R-square

Adjusted R-square modifies the value of R-square to account for the number of predictors in the model. It is calculated according to equation (3.3), where n is the sample size and p is the number of predictors in the model.

$$R_{adj}^2 = 1 - \left(\frac{n-1}{n-p-1} \right) (1 - R^2) \quad (3.3)$$

Every time a new predictor is added to the model, the value for SSE will decrease (and this is a good thing), but degrees of freedom for error will also decrease (and this is a bad thing). Adjusted R-square attempts to answer the question: Does the decrease in SSE offset the loss in error degrees of freedom?

Adjusted R-square can either increase or decrease as new predictors enter the model. While the value for (unadjusted) R-square is required to be between 0 and 1, the value for adjusted R-square may be negative (for a really bad model).

Residual mean square

This is another name for MSE, and it is sometimes abbreviated RMS. It is equivalent to adjusted R-square, except that smaller values for MSE are preferred. Adjusted R-square and MSE will always generate the same subset of predictors, so there is no reason to consider both of these criteria.

Mallow's C_p

Mallow's C_p attempts to balance two competing considerations: the drawback of including too many predictors in the model versus the mistake of excluding important predictors. Conceptually, Mallow's C_p depends on a “full” model that contains all the predictors in the dataset . The full model is used as a basis for evaluating another model, which is a “reduced” model. In other words, every predictor in the reduced model is also contained in the full model. It is important that the “full” model be a good fit to the data, then Mallow's C_p is calculated according to equation (3.4).

$$C_p = \frac{SSE(\text{reduced})}{MSE(\text{full})} - n + 2 \cdot (\# \text{parameters in reduced model}) \quad (3.4)$$

It is desirable to have a model for which C_p is close to or less than number of parameters in the model. If the value for C_p is much larger than the number of parameters, then this is considered a poor model because it is biased. Biased models consistently over-estimate or under-estimate the model parameters, and therefore produce poor estimates for the response.

Akaike's Information Criterion (AIC)

AIC is calculated according to equation (3.5), where n is the sample size, p is the number of predictors in the model, and SSE is the sum of squares due to error.

$$AIC = n \cdot \log(SSE) - n \cdot \log(n) + 2(p + 1) \quad (3.5)$$

For many models, the value of AIC will be negative. Smaller (i.e., larger negative) values of AIC are preferred. The term ' $2(p+1)$ ' is a penalty associated with the number of predictors in the model. Larger values for p generate larger values for AIC, but smaller values for AIC are preferred. As a general rule of thumb, a decrease of 2 or more AIC points usually indicates a substantial improvement in model fit.

Schwarz' Bayesian Criterion (SBC)

$$SBC = n \cdot \log(SSE) - n \cdot \log(n) + (p + 1) \cdot \log(n) \quad (3.6)$$

SBC is very similar to AIC, but the penalty for additional terms is changed from $2(p+1)$ to $(p+1) \cdot \log(n)$. If the sample size (n) is 8 or more, the penalty applied by SBC is greater than the penalty applied by AIC. This implies that SBC encourages models with fewer predictors. As with AIC, smaller values of SBC are preferred and a decrease of 2 or more SBC points usually indicates a substantial improvement in model fit.

PRESS

PRESS is an abbreviation for Prediction Sum of Squares. It is the preferred criterion if we want to use the fitted model to predict new observations. The concept behind the PRESS statistic is a “leave-one-out” strategy. The basic steps are as follows:

- Delete the i^{th} observation.

- Estimate the regression equation with the remaining ($n - 1$) observations.
- Predict the value of the i^{th} response.
- The deleted residual is the difference between observed and predicted response for the deleted observation.
- PRESS is the sum of all the squared deleted residuals.

Small values for PRESS values are desirable because they indicate small prediction errors.

3.1.4. Summary

When developing a linear regression model, there are two competing goals. We want to have as simple a model as possible, but we also want the model to fit the data well. Since complex models will generally fit the data better than simpler ones, we need to balance these two goals. We presented several different criteria to measure of goodness of fit that incorporate penalties for more complex models. We do not need to use all of these criteria to build a model. Different researchers have different preferences, and no one criterion is “best” for all situations. If the model is going to be used to predict new observations, then the choices for model selection criteria become more limited. In particular, we will use the PRESS statistic for predictive models.

Both the criteria presented in this section and the variable selection methods that will be described in the next section should be applied ONLY to data that are observational. They should NOT be applied to data derived from designed experiments. For experiments, the modeling approach is governed by the scientific question, not the data. At all costs, one should avoid data “snooping”. This is when the data are summarized and the summaries are used to develop a model. Snooping can generate spurious results (“flukes”) that are not reproducible.

Section 3.2. Procedures for Model Selection

When multiple predictor variables are available in a dataset, it is not required that all of them be used in the regression model. Decisions regarding which predictors to include (or exclude) can quickly become a very large and time-consuming problem. If there are p possible predictors, then there are 2^p models that could be generated using subsets of these predictors. For 3 predictors, there are 8 possible models and this would be manageable. But when there are 8 potential predictors, the number of possible models increases to 256, and 10 potential predictors would generate 1,024 possible models. In general, there are entirely too many models to consider them all, so we need systematic methods to identify “good” models from among the provided predictors.

There are several different procedures that we can employ to assist in identifying the “important” predictors, and these procedures may produce different results. There is no such thing as a “best” model, so there is no such thing as a “best” procedure to generate a model. These procedures are too intensive to be conducted by hand; we will use SAS to perform the calculations. These procedures are sometimes called model selection procedures, but it is more accurate to call them variable selection procedures. We will identify the model selection criteria we want to use and provide a list of potential predictors (possibly including transformations and interactions). The procedure will select a subset of the potential predictors that will generate the “best” model according to the specified criteria.

3.2.1. Automated Search Methods

There are four commonly used automated model selection procedures. These are

- **Forward Selection** starts with an intercept-only model (no predictors), and adds predictors one at a time
- **Backward Elimination** starts with all predictors in the model, and removes them one at a time
- **Stepwise** is a combination of forward selection and backward elimination
- **Best Subsets** generates all possible models using subsets of the predictors

Forward Selection

The forward selection method starts with a model that contains only the intercept (no predictor variables at all). Each of the potential predictors are examined with regard to how much it improves the fit of the model. The predictor that contributes the most is added to the model. Now the model

contains the intercept and one predictor. Each of the remaining predictors are examined with regard to how much it improves the fit of the model, and the predictor that contributes the most is added to the model. The process continues until none of the remaining predictors contribute significantly to the fit of the model.

To decide whether or not a potential predictor will contribute “significantly” to the model, the automated procedure uses a nested model F test, comparing the reduced model (without the potential predictor) to the full model (with the potential predictor). Recall that the nested model F test is testing these hypotheses:

$$\begin{aligned} H_0: & \text{reduced model is adequate} \\ H_a: & \text{full model is needed} \end{aligned}$$

If this test is decided in favor of H_0 , then the predictor should not be added to the model, but if the test is decided in favor of H_a , then the predictor should be added.

Each of the potential predictors will have its own F test, and the one with the lowest p-value (i.e., highest F statistic) is judged to be the one that contributes most to the model, and hence is the one that is added to the model. In SAS, the default level of significance for this test is 0.5, but this can be changed via the SLENTRY option (which is an abbreviation for “significance level for entry”).

In the forward selection method, potential predictors are added to the model, one at a time, until none of the remaining predictors produce a significant F test. Once a predictor enters the model, it never leaves.

Backward Elimination

The backward elimination method starts with all predictors in the model, and removes predictors one at a time. A nested model F test is used to decide which predictor to remove, and the hypotheses for this test are the same as in forward selection. If the test is decided in favor of H_0 , then the predictor should be removed from the model, but if the test is decided in favor of H_a , then the predictor should be kept in the model.

Each of the predictors in the model will have its own F test, and the one with the lowest p-value is judged to be the one that contributes the least to the model (and hence is the one that can be deleted).

from the model). In SAS, the level of significance for this test is 0.10, but this can be changed via the SLSTAY option (which is an abbreviation for “significance level to stay”).

In the backward elimination method, predictors in the model are removed one at a time until all of the remaining predictors are deemed significant to the model. Once a predictor is removed from the model, it never returns.

Stepwise

The stepwise method is a combination of forward selection and backward elimination. It starts with no predictors in the model. Then it performs one step of forward selection, followed immediately by one step of backward elimination. These two steps (one addition and one removal) are repeated until no further changes are warranted. Each decision (to either add or remove a predictor) is based on a nested model F test. In SAS, the default level of significance for a variable to be added to the model is SLENTRY = 0.15 and the default level of significance for a variable to stay in the model is SLSTAY = 0.15. These two steps are repeated until none of predictors outside the model has a significant F test (i.e., all p-values are greater than SLENTRY) and every variable in the model has a non-significant F test (i.e., all p-values are less than SLSTAY). The process may also end when the variable that is added to the model is the same variable as the one that was just removed from it.

Best Subsets

The best subsets method is not a step-by-step procedure. Instead, it generates all possible regression models using all possible subsets of the data. Then it ranks these models according to the specified criteria (e.g., adjusted R-square).

In the computer implementation of the best subset method, certain mathematical shortcuts are taken so that every possible model does not need to be examined. These shortcuts greatly reduce the amount of time required to perform this method. The shortcuts will only eliminate “poor” models; they will never eliminate a “good” one.

3.2.2. Comparison of Methods

Each method generates a subset of predictors that are considered “important” for explaining the behavior of Y. The various methods can produce the same subset of predictors, but it is also possible that each method will produce a different subset. To initiate any of these methods, we first need to

identify a collection of potential predictors (including transformations, interactions, etc.), and we will need to specify the criterion that will be used to compare the models (e.g., adjusted R-square, AIC, SBC).

All of these methods are automatic and they do not require any additional input from us. The only criterion that the computer uses to compare the models is the one that we specify. There is no guarantee that the models that are generated will satisfy the assumptions. This requirement, along with potential outliers and influential points, will still need to be examined before the model can be considered valid. Always remember to exercise common sense when choosing a final model.

When the number of potential predictors is very large, it is common to use a hybrid approach to model selection. The step-by-step methods (backward, forward and stepwise) are used to screen the potential predictors. We manually examine the list of predictors and dismiss those that have negligible effects in all the “good” models. Then apply the best subsets methods using only the remaining predictors.

To implement model selection methods in SAS, we use the SELECTION option on the MODEL statement in PROC REG. For example, if the response variable is Y and there are five potential predictors Z1, Z2, Z3, Z4 and Z5, then we would use the following MODEL statements

- For forward selection:

```
MODEL Y = Z1 Z2 Z3 Z4 Z5 / SELECTION = FORWARD;
```

- For backward elimination:

```
MODEL Y = Z1 Z2 Z3 Z4 Z5 / SELECTION = BACKWARD;
```

- For stepwise:

```
MODEL Y = Z1 Z2 Z3 Z4 Z5 / SELECTION = STEPWISE;
```

- For best subsets using the adjusted R-square criterion:

```
MODEL Y = Z1 Z2 Z3 Z4 Z5 / SELECTION = ADJRSQ;
```

- For best subsets using the Mallow's C_p criterion:

```
MODEL Y = Z1 Z2 Z3 Z4 Z5 / SELECTION = CP;
```

Additional options that can be placed on these statements are given in the following example.

Section 3.3. Example using SENIC data

This example illustrates the process of model building. It uses the SENIC dataset, which was obtained as part of the Study on the Efficacy of Nosocomial Infection Control (SENIC) to determine whether infection surveillance and control programs have reduced the rates of nosocomial (hospital-acquired) infection in U.S. hospitals. This dataset consists of a random sample of $n = 113$ hospitals selected from the original 338 hospitals surveyed. Each hospital is given an ID number, and is measured on 11 other variables. (Note that each observation in the dataset is a hospital, not a patient.) Definitions for the variables are given in Table 3.1.

We will use model selection procedures to construct a model that estimates Infection Risk. We will use only the variables that are in the dataset; we will not consider any transformations of the predictor variables.

Variable	Description
idno	Identification number for hospital (1 to 113)
Stay	Average length of stay of all patients in the hospital (measured in days)
Age	Average age of patients (in years)
InfRisk	Infection Risk: Average estimated probability of acquiring infection in hospital, in percent. (This is the response variable.)
CulRatio	Ratio of number of cultures performed to number of patients without signs or symptoms of hospital-acquired infection, times 100
XRay	Ratio of number of X-rays performed to number of patients without signs or symptoms of pneumonia, times 100
NumBeds	Average number of beds in hospital during study period
MedSch	Medical School Affiliation (1=Yes, 2=No)
Region	Geographic Region (1=Northeast, 2=North Central, 3=South, 4=West)
Census	Average number of patients in hospital per day during study period
Nurses	Average number of full-time equivalent registered and licensed practical nurses during study period (number of full-time + $\frac{1}{2}$ number of part-time)
Services	Percent of 35 potential facilities and services that are provided by the hospital

Table 3.1. Definitions for SENIC variables

An initial inspection of the variable descriptions reveals two qualitative variables: MedSchool and Region. Although their values are numeric, the numbers are simply codes for the categories. We should not treat these as numeric variables.

In SAS, the model selection procedures are implemented in PROC REG, and this accepts only numeric variables. Therefore, MedSchool and Region cannot be part of the automated model selection process unless we treat them as numeric. We don't want to ignore these variables, so we convert them to indicator variables and use the indicators in the model selection procedure. This approach is straightforward for MedSchool since there is only one indicator variable, but we must be careful with Region. There are three indicator variables for Region, and they must all be included or excluded in the model. The automated procedure in SAS will not know to enforce this restriction unless we tell it to. This is done by placing curly brackets around these three indicator variables in the MODEL statement. This is described in more detail below.

To begin the analysis, we examine the correlation matrix. We are interested in the correlation between each potential predictor variable and the response variable (InfRisk). We are also interested in the correlation between potential predictors. For this part of the analysis, we omit the indicator variables.

Pearson Correlation Coefficients, N = 113									
	Stay	Age	InfRisk	CulRatio	XRay	NumBeds	Census	Nurses	Services
Stay	1.00000	0.18891	0.53344	0.32668	0.38248	0.40927	0.47389	0.34037	0.35554
Age	0.18891	1.00000	0.00109	-0.22585	-0.01885	-0.05882	-0.05477	-0.08294	-0.04045
InfRisk	0.53344	0.00109	1.00000	0.55916	0.45339	0.35977	0.38141	0.39398	0.41260
CulRatio	0.32668	-0.22585	0.55916	1.00000	0.42496	0.13972	0.14295	0.19890	0.18513
XRay	0.38248	-0.01885	0.45339	0.42496	1.00000	0.04582	0.06291	0.07738	0.11193
NumBeds	0.40927	-0.05882	0.35977	0.13972	0.04582	1.00000	0.98100	0.91550	0.79452
Census	0.47389	-0.05477	0.38141	0.14295	0.06291	0.98100	1.00000	0.90790	0.77806
Nurses	0.34037	-0.08294	0.39398	0.19890	0.07738	0.91550	0.90790	1.00000	0.78351
Services	0.35554	-0.04045	0.41260	0.18513	0.11193	0.79452	0.77806	0.78351	1.00000

Table 3.2. Correlation matrix for the SENIC data

Correlations between predictors and the response (InfRisk) are shown in the third column of Table 3.2. Except for Age, all of the correlations are between 0.3 and 0.6, so any of these variables could be effective predictors.

As an initial assessment of possible multicollinearity issues, we also examine the correlation between predictors. In general, correlations higher than 0.7 are reason for concern. We see several high correlations, but the largest is 0.981 (between Census and NumBeds). Such a high correlation indicates that whatever information that is contained in Census is (almost) duplicated in NumBeds. We should remain alert to potential multicollinearity issues as we proceed with the analysis.

3.3.1. Fullest possible model

We are now ready to begin the model selection process. Our response is Infection Risk (InfRisk), and we consider all 10 of the potential predictors in the dataset. The diagnostic plots for this model are not shown here, but there is no evidence that the assumptions have been violated. The Parameter Estimates table, shown in Table 3.3, shows that the variance inflation factors for both NumBeds and Census are greater than 10, and this gives us cause for concern regarding multicollinearity in this model. However, the standard errors for the these two parameters are not overly large (as compared to the other standard errors in Table 3.3), so we will proceed with the analysis.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Variance Inflation
Intercept	1	-0.70331	1.21748	-0.58	0.5648	0
Stay	1	0.24240	0.07018	3.45	0.0008	2.41397
Age	1	0.01323	0.02177	0.61	0.5447	1.26553
CulRatio	1	0.05449	0.01055	5.16	<.0001	1.56380
XRay	1	0.01155	0.00526	2.19	0.0305	1.39377
NumBeds	1	-0.00349	0.00268	-1.30	0.1954	35.73915
Census	1	0.00386	0.00345	1.12	0.2655	37.76771
Nurses	1	0.00179	0.00169	1.05	0.2942	7.47018
Services	1	0.02057	0.01006	2.04	0.0435	3.13602
MedSch	1	-0.66082	0.32139	-2.06	0.0424	1.78658
Reg1	1	-1.14954	0.33917	-3.39	0.0010	2.90170
Reg2	1	-0.72403	0.29782	-2.43	0.0168	2.43661
Reg3	1	-0.78277	0.28895	-2.71	0.0079	2.48833

Table 3.3. Parameter Estimates table for the full model

This model has R-square 0.5854 and adjusted R-square 0.5356. While this does not seem extraordinarily high, it is definitely reasonable given the fact that we are trying to model something as complicated as risk of infection. *At this stage of the analysis, we are not concerned with the actual parameter estimates or the results of their t-tests.* The Parameter Estimates table is shown here solely for the variance inflation factors.

The next step is to generate alternatives to this model in an effort to get a simpler model that also fits the data well.

3.3.2. Forward selection

To implement the method of forward selection in SAS, we include additional options on the MODEL statement in PROC REG. These are given after the slash.

```
forward:  
MODEL InfRisk = Stay Age CulRatio XRay NumBeds Census Nurses Services  
          Services MedSch {Reg1 Reg2 Reg3}  
          / VIF SELECTION=FORWARD DETAILS=SUMMARY;
```

Note that all of these lines comprise a single SAS statement (that ends with the semicolon), and the entire statement goes inside PROC REG. The label ‘forward:’ is optional, but recommended. This label

will be printed on top of every page in the SAS output that is related to this model. The equation for this model uses InfRisk as the response and all the other variables as predictors. To instruct SAS to either include or exclude all three indicator variables for Region (Reg1, Reg2 and Reg3), we enclose these variables in curly brackets. After the slash, we include the familiar option VIF for the variance inflation factors and then we have two new options:

SELECTION=FORWARD	instructs SAS to select predictors using the method of forward selection.
DETAILS=SUMMARY	instructs SAS to print only the summary information for this process. Without this option, SAS would generate several pages of output detailing the results of every step in the forward selection process.

The summary table for forward selection is shown in Table 3.4Table 3.. Because we have grouped the three indicator variables for Region, SAS considers each predictor as a “group”. They are numbered in the order that they appear in the MODEL statement, so that

GROUP1 is Stay	GROUP2 is Age	GROUP3 is CulRatio
GROUP4 is XRay	GROUP5 is NumBeds	GROUP6 is Census
GROUP7 is Nurses	GROUP8 is Services	GROUP9 is MedSch
GROUP10 is Region (which consists of the three indicators Reg1, Reg2, Reg3)		

The first predictor added to the model is CulRatio (Group3), followed by Stay (Group1) and Services (Group8). After these three iterations of the forward selection process, there are 3 variables in the model and the value for R-square is 0.4934. On the 4th iteration, the three indicator variables for Region (Group10) were added to the model, so the model now contains 6 variables. The process of forward selection continues until there are 9 predictors in the model. Since Region is one of the predictors, the total number of variables in the model is 11. The final model chosen by forward selection has 11 variables and R-square is 0.5838. The value for adjusted R-square is 0.5385. The summary table also provides the value for Mallow’s C_p, which is 11.393. Note that this is close to the number of variables, and this is a good thing.

The complete list of predictors chosen by the method of forward selection is shown in Table 3.5. This is simply the Parameter Estimates table routinely generated by PROC REG. At this point in the analysis, we are not concerned with the specific parameters estimates, nor the test statistics or the p-values. We are examining this table now only to identify the variables that are included in the model and to evaluate the variance inflation factors. Since both NumBeds and Census are included in this model, there are two VIF’s that are greater than 10. This is a cause for concern, but (as mentioned earlier), the standard

errors for these two predictors are not overly large. In addition, the diagnostic plots (not shown here) do not indicate that any assumptions have been violated. We will consider this an acceptable model.

The forward selection process chose every predictor in the dataset with the exception of Age. We will explore the other methods of variable selection to see if we can find a simpler model.

Summary of Forward Selection								
Step	Group Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F	
1	GROUP3	1	0.3127	0.3127	56.7680	50.49	<.0001	
2	GROUP1	2	0.1377	0.4504	25.5479	27.57	<.0001	
3	GROUP8	3	0.0430	0.4934	17.1781	9.25	0.0029	
4	GROUP10	6	0.0436	0.5370	12.6655	3.33	0.0225	
5	GROUP4	7	0.0220	0.5590	9.3610	5.24	0.0241	
6	GROUP9	8	0.0122	0.5712	8.4122	2.97	0.0880	
7	GROUP7	9	0.0061	0.5773	8.9500	1.48	0.2270	
8	GROUP5	10	0.0020	0.5793	10.4707	0.48	0.4892	
9	GROUP6	11	0.0046	0.5838	11.3693	1.11	0.2950	

Table 3.4. Summary of forward selection

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Variance Inflation
Intercept	1	-0.07927	0.65208	-0.12	0.9035	0
Stay	1	0.25560	0.06653	3.84	0.0002	2.18293
CulRatio	1	0.05259	0.01005	5.23	<.0001	1.42710
XRay	1	0.01149	0.00525	2.19	0.0308	1.39327
NumBeds	1	-0.00334	0.00266	-1.26	0.2113	35.45328
Census	1	0.00359	0.00341	1.05	0.2950	37.12418
Nurses	1	0.00181	0.00169	1.07	0.2856	7.46552
Services	1	0.02080	0.01002	2.08	0.0405	3.13162
MedSch	1	-0.67377	0.31968	-2.11	0.0375	1.77873
Reg1	1	-1.14971	0.33811	-3.40	0.0010	2.90170
Reg2	1	-0.75089	0.29360	-2.56	0.0120	2.38295
Reg3	1	-0.78635	0.28799	-2.73	0.0075	2.48730

Table 3.5. Parameter Estimates table for forward selection

3.3.3. Backward elimination

The SAS code for backward elimination is nearly identical to that for forward selection. Simply replace the word FORWARD with the word BACKWARD.

```
backward:  
MODEL InfRisk = Stay Age CulRatio XRay NumBeds Census Nurses Services  
Services MedSch {Reg1 Reg2 Reg3}  
/ VIF SELECTION=BACKWARD DETAILS=SUMMARY;
```

Recall that the method of backward elimination starts with the full model, and removes variables one at a time until no further removals are warranted. The summary table for backward elimination, shown in Table 3.6Table 3., gives the variables that are removed from the model at each step of the process. The first variable removed is Age (Group2), followed by Census (Group6), NumBeds (Group5) and Nurses (Group7). Four variables were removed, which leaves six predictors in the model. One of these predictors is Region, so there are 8 variables in the final model. These variables, along with their variance inflation factors, are shown in the Parameter Estimates table (Table 3.7). Since neither NumBeds nor Census is included in this model, there is no evidence of multicollinearity. For this model, the value of R-square is 0.5712, the value for adjusted R-square is 0.5382, and the value for Mallow's C_p is 8.4122, which is very close to the number of variables (8).

As with the earlier models, the diagnostic plots for this model are not shown here, but they are part of the SAS output and they should be examined before making a decision about the adequacy of this model. The plots for this model show no indication that the assumptions have been violated.

Summary of Backward Elimination							
Step	Group Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	GROUP2	11	0.0015	0.5838	11.3693	0.37	0.5447
2	GROUP6	10	0.0046	0.5793	10.4707	1.11	0.2950
3	GROUP5	9	0.0020	0.5773	8.9500	0.48	0.4892
4	GROUP7	8	0.0061	0.5712	8.4122	1.48	0.2270

Table 3.6. Summary table for backward elimination

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Variance Inflation
Intercept	1	-0.32966	0.57977	-0.57	0.5708	0
Stay	1	0.27469	0.05823	4.72	<.0001	1.67090
CulRatio	1	0.05234	0.00980	5.34	<.0001	1.35780
XRay	1	0.01133	0.00521	2.17	0.0320	1.37387
Services	1	0.02546	0.00690	3.69	0.0004	1.48411
MedSch	1	-0.50274	0.29194	-1.72	0.0880	1.48249
Reg1	1	-1.10697	0.33124	-3.34	0.0012	2.78329
Reg2	1	-0.76674	0.29138	-2.63	0.0098	2.34553
Reg3	1	-0.75937	0.28139	-2.70	0.0081	2.37322

Table 3.7. Parameter Estimates table for backward elimination

3.3.4. Stepwise

The SAS code for the stepwise method is nearly identical to that for forward selection and backward elimination. Simply replace the word FORWARD (or BACKWARD) with the word STEPWISE.

```
stepwise:
MODEL InfRisk = Stay Age CulRatio XRay NumBeds Census Nurses Services
                  Services MedSch {Reg1 Reg2 Reg3}
                  / VIF SELECTION=STEPWISE DETAILS=SUMMARY;
```

Recall that the stepwise procedure starts with a model that contains only the intercept (no predictors) and makes two decisions at each step in the process:

- (1) Should any variable be added to the model? If so, which one?
- (2) Should any variable be removed from the model? If so, which one?

The summary of the stepwise selection method for the SENIC data is shown in Table 3.8Table 3.. Note that this table has a column for “Group Entered” and another column for “Group Removed”. In the first step, the predictor CulRatio (Group3) entered the model and no variable was removed. In the second step, the predictor Stay (Group1) entered the model and no variable was removed. Since the column for “Group Removed” contains no entries at all, no variables were removed from the model at any step in the process. This is perfectly acceptable. At each step, the selection process checked to see if a variable should be removed, but in every case the answer was ‘no’.

The final model chosen by the stepwise method contains 8 variables, with R-square equal to 0.5712, adjusted R-square equal to 0.5385, and Mallow’s C_p equal to 8.4122. Note that these values coincide

with the values from the backward elimination model. A closer examination of these two models reveal that *they are exactly the same model*. For a different dataset, it is entirely possible that these two methods would generate different models, but the fact that they have generated the same model gives us some assurance in the validity of this model.

Summary of Stepwise Selection									
Step	Group Entered	Group Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F	
1	GROUP3		1	0.3127	0.3127	56.7680	50.49	<.0001	
2	GROUP1		2	0.1377	0.4504	25.5479	27.57	<.0001	
3	GROUP8		3	0.0430	0.4934	17.1781	9.25	0.0029	
4	GROUP10		6	0.0436	0.5370	12.6655	3.33	0.0225	
5	GROUP4		7	0.0220	0.5590	9.3610	5.24	0.0241	
6	GROUP9		8	0.0122	0.5712	8.4122	2.97	0.0880	

Table 3.8. Summary of stepwise selection

3.3.5. Other model selection criteria

To instruct SAS to perform model selection based on a specific criterion, we include additional options on the MODEL statement. The current version of SAS (in February 2019) implements model selection for adjusted R-square and Mallow's C_p . We can tell SAS to report the values for AIC and SBC, but it will not generate models based on the these criteria.

For example, to select variables based on the models' values for adjusted R-square, use the SAS keyword ADJRSQ in the MODEL statement.

```
AdjR2:
MODEL InfRisk = Stay Age CulRatio XRay NumBeds Census Nurses Services
               MedSch {Reg1 Reg2 Reg3}
               / VIF SELECTION=ADJRSQ DETAILS=SUMMARY BEST=6 AIC SBC;
```

The additional options for AIC and SBC will report those two values for each model, but the models will be sorted according to adjusted R-square since that is the keyword in the SELECTION option. The option "BEST=6" tells SAS to print the six "best" models. Because we have specified adjusted R-square as the selection criteria, the six "best" models are those that have the six highest values for adjusted R-square and the models will be presented in order of decreasing adjusted R-square (so that the "best" model is at the top).

If we wanted to use Mallow's C_p as the selection criterion, we would replace SELECTION=ADJRSQ with SELECTION=CP in the MODEL statement. Then the six "best" models according the C_p criterion would be presented in a table, with the best model in the top row.

The output for the adjusted R-square criterion is shown in Table 3.9. The largest (best) adjusted R-square is 0.5403 and this comes from a model that uses 9 variables. The specific variables that are included in this model are given in the last column. Note that this model includes 7 of the original 10 predictors, but there are 9 variables in the model because the categorical variable Region uses three indicator variables. The second row in Table 3.9 provides the second-best model (according to adjusted R-square), but this model is the same as the one generated by forward selection. The third row gives the third-best model according to adjusted R-square, but this is the same as the one generated by both backward elimination and stepwise.

The best model according to adjusted R-square was not identified by any of the earlier methods. This is because the best subsets procedure is used to generate these models, and the best subsets method is substantially different than the step-by-step approach of the earlier methods.

Number in Model	Adjusted R-Square	R-Square	AIC	SBC	Variables in Model
9	0.5403	0.5773	-12.0045	15.26934	Stay CulRatio XRay Nurses Services MedSch Reg1 Reg2 Reg3
11	0.5385	0.5838	-9.7702	22.95843	Stay CulRatio XRay NumBeds Census Nurses Services MedSch Reg1 Reg2 Reg3
8	0.5382	0.5712	-12.3954	12.15110	Stay CulRatio XRay Services MedSch Reg1 Reg2 Reg3
10	0.5380	0.5793	-10.5370	19.46425	Stay CulRatio XRay NumBeds Nurses Services MedSch Reg1 Reg2 Reg3
10	0.5378	0.5791	-10.4883	19.51293	Stay CulRatio XRay NumBeds Census Services MedSch Reg1 Reg2 Reg3
9	0.5371	0.5743	-11.2130	16.06085	Stay CulRatio XRay Census Services MedSch Reg1 Reg2 Reg3

Table 3.9. Six best models according to adjusted R-square

Number in Model	C(p)	R-Square	Adjusted R-Square	AIC	SBC	Variables in Model
8	8.4122	0.5712	0.5382	-12.3954	12.15110	Stay CulRatio XRay Services MedSch Reg1 Reg2 Reg3
9	8.9500	0.5773	0.5403	-12.0045	15.26934	Stay CulRatio XRay Nurses Services MedSch Reg1 Reg2 Reg3
7	9.3610	0.5590	0.5296	-11.2182	10.60091	Stay CulRatio XRay Services Reg1 Reg2 Reg3
9	9.6666	0.5743	0.5371	-11.2130	16.06085	Stay CulRatio XRay Census Services MedSch Reg1 Reg2 Reg3
9	10.1929	0.5721	0.5347	-10.6352	16.63867	Stay CulRatio XRay NumBeds Services MedSch Reg1 Reg2 Reg3
9	10.2322	0.5720	0.5346	-10.5922	16.68163	Stay Age CulRatio XRay Services MedSch Reg1 Reg2 Reg3

Table 3.10. Six best models according to Mallow's Cp criterion

The models generated by the Mallow's C_p criterion (SELECTION=CP) are shown in Table 3.10. The best model according to the C_p criterion is the third best model according to the adjusted R-square criterion. The second best Mallow's C_p model is the first best according to adjusted R-square. It is not uncommon to have duplicate models generated by the various methods. In fact, it is reassuring to have duplicates because it gives these models more credibility.

Note that all the models in Table 3.10 include the predictors Stay, CulRatio, XRay, Services, and the three indicator variables for Region. The model in the first row also contains MedSch. The model in the second row includes both MedSch and Nurses, but this is the only model that contains Nurses.

3.3.6. Choose the final model

We have generated several candidate models, with some duplications, and now we must choose one model that would be used for subsequent analysis. It is not mandatory that we consider every model that has been generated by every method, but, at the same time, we do not want to be overly restrictive in the final comparison. All of the models that have been generated have reasonable values for adjusted R-square, so none of them can be discarded on the basis on this criterion. We have not seen the diagnostic plots or the other criteria for all of these models, since SAS provides details only for the first best model generated by each method.

The decisions that must be made now are subjective ones, and other reasonable statisticians may decide differently than the author. I recommend that, due to multicollinearity concerns, we discard the models that include both NumBeds and Census. This will eliminate the full model, the model generated by forward selection and the 2nd best and 5th best models generated by the adjusted R-square criterion. Among the remaining models, there are some duplicates. We can eliminate the stepwise model, the 3rd model from adjusted R-square and the 1st model from Mallow's Cp because these are identical to the backward elimination model. We can also eliminate the 2nd Cp model because it is the same as the 1st adjusted R-square model. Finally, we can eliminate the 4th Cp model because it is the same as the 6th adjusted R-square model.

This leaves us with 7 candidate models. While this may seem like a lot of models to consider, it is far less than the 1,024 models that are possible with 10 predictors. The predictors included in the candidate models are shown in Table 3.11. Note that all of the candidate models include the predictors Stay, CulRatio, XRay, Services and Region, but they differ in the remaining predictors.

Based on the information in Table 3.11, none of the candidate models exhibit serious deficiencies. Their values for adjusted R-square are all very close, and each model has a value for Mallow's Cp that is close to the number of variables. (The value of Mallow's Cp for candidate model #3 has not been included in any of the SAS output thus far, hence the question mark in Table 3.11.) If we use adjusted R-square to select a model, then candidate model #2 would be the best. But AIC and SBC both indicate that candidate model #1 is best. Since all of these models are so close, the final decision may be made on the basis of the research objective or subject matter expertise. For example, if previous research has indicated that age is definitely related to infection risk, then candidate model #7 would be the likely choice since it is the only one that includes the predictor age. On the other hand, if the analysis is being conducted for a worker's union of nurses, then candidate models #2 or #3 would be a good choice because they are the only ones that include the predictor nurses.

Predictors	Candidate Models						
	#1	#2	#3	#4	#5	#6	#7
	backward	AdjRsq #1	AdjRsq #4	AdjRsq #6	Cp #3	Cp #5	Cp #6
Stay	X	X	X	X	X	X	X
CulRatio	X	X	X	X	X	X	X
XRay	X	X	X	X	X	X	X
Services	X	X	X	X	X	X	X
Region	X	X	X	X	X	X	X
MedSch	X	X	X	X	-	X	X
Nurses	-	X	X	-	-	-	-
NumBeds	-	-	X	-	-	X	-
Census	-	-	-	X	-	-	-
Age	-	-	-	-	-	-	X
# variables	8	9	10	9	7	9	9
AdjRsq	0.5382	0.5403	0.5380	0.5371	0.5296	0.5347	0.5346
AIC	-12.40	-12.00	-10.54	-11.21	-11.22	-10.64	-10.59
SBC	12.15	15.27	19.46	16.06	10.60	16.64	16.68
Cp	8.41	8.95	?	9.67	9.36	10.19	10.23

Table 3.11. Candidate models for the SENIC data

Before making the final decision, it is imperative to examine the diagnostic plots for each model. Any model that violates the assumptions must be discarded. The two main diagnostic plots (the residual plot and the QQ plot) have already been generated for the “best” model for each criterion, but we have not yet examined the plots for influence and outliers. It is sometimes advantageous to generate these graphs by themselves, without all the other output from PROC REG. To accomplish this in SAS, we will need to use the output delivery system (ODS). This is a more advanced technique in SAS, and it is provided here only for your knowledge and convenience. It is not a core part of the content of this course.

Here is the SAS code that will generate only the panel of diagnostic plots for each model.

```

PROC REG DATA=senic;
  ODS SELECT DiagnosticsPanel;
  MODEL InfRisk = Stay CulRatio XRay Services Reg1 Reg2 Reg3 MedSch;
  MODEL InfRisk = Stay CulRatio XRay Services Reg1 Reg2 Reg3 MedSch Nurses;
  MODEL InfRisk = Stay CulRatio XRay Services Reg1 Reg2 Reg3 MedSch Nurses NumBeds;
  MODEL InfRisk = Stay CulRatio XRay Services Reg1 Reg2 Reg3 MedSch Census;
  MODEL InfRisk = Stay CulRatio XRay Services Reg1 Reg2 Reg3;
  MODEL InfRisk = Stay CulRatio XRay Services Reg1 Reg2 Reg3 MedSch NumBeds;
  MODEL InfRisk = Stay CulRatio XRay Services Reg1 Reg2 Reg3 MedSch Age;
RUN;
```

There is no need to provide a label for each model because SAS will number them sequentially and these numbers will match the candidate model numbers in Table 3.11. The ODS SELECT statement is quite useful for limiting the amount of output generated by SAS procedures. “DiagnosticsPanel” is a key word in SAS that refers to the page of graphs generated by PROC REG. Every graph and every table in the customary SAS output has an official ODS name, and the ODS SELECT statement is used to print only the specified graphs and/or tables.

The output from this SAS code is not shown here, but the diagnostic plots are a very important part of model selection and they should not be ignored. All of these candidate models have very similar diagnostic plots.

- All the residual plots look excellent. They have no discernible pattern.
- All the QQ plots look excellent. The points are covering the line.
- Every model exhibits a few mild outliers (studentized residuals bigger than 2), but there are no extreme outliers (bigger than 3).
- Every model has a handful of points with mildly high leverage, but nothing extreme.
- The Cook’s D plot for every model shows at least one point that is influential. While this is mildly disconcerting, every model exhibits the same pattern, so it is not a basis to distinguish these models.

There is no obvious conclusion to this example. The final model should definitely include the predictors Stay, CulRatio, XRay, Services and Region, but the fate of the other predictors is not clear. Whether we choose to include or exclude the other predictors, the model will fit the data reasonably well and the model assumptions will not be violated.

Section 3.4. Prediction Models

The model selection methods and criteria presented earlier in this chapter apply to models that are going to be used for estimating model parameters and values for the response variable. There is another main purpose of regression models, and that is to predict new observations. There is no assurance that a model that is a good fit to the existing data will also be successful for future predictions. This could due to influential factors that were unknown during model building, or perhaps there is a different correlation structure among the predictors. The key idea is that we want to test the

model in the environment in which it is going to perform. To assess the predictive ability of a model, we need to use different methods and criteria.

3.4.1. Deleted Residuals

The predictive ability of a regression model is based on the concept of deleted residuals. If we “pretend” that we did not have the i^{th} observation in a dataset, how accurately would the model estimate its Y value? To make this assessment, we temporarily delete the i^{th} observation and then fit the model using the remaining $n - 1$ observations. Using the fitted model, we estimate \hat{Y}_i , the Y value for the i^{th} observation. If the estimated Y had been generated using all the data, we would use the notation \hat{Y}_i . To indicate that this estimate has been obtained using only the other $n - 1$ observations, we put parentheses in the subscript: $\hat{Y}_{(i)}$. The difference between this estimate for \hat{Y}_i and the observed value for Y_i is called the deleted residual, denoted by $r_{(i)}$. The i^{th} deleted residual is defined in equation (3.7).

$$r_{(i)} = Y_i - \hat{Y}_{(i)} \quad (3.7)$$

3.4.2. PRESS Statistic

PRESS is an abbreviation for prediction sum of squares. It is calculated as the sum of the squared deleted residuals.

$$\text{PRESS} = \sum_{i=1}^n (Y_i - \hat{Y}_{(i)})^2 \quad (3.8)$$

Small values for PRESS are desirable, since that indicates small prediction error.

To provide an example of how the PRESS statistic can be used, consider the candidate models for the SENIC data that we developed in the previous section, and look more closely at candidate models #4 and #5. Candidate model #5 had the fewest predictors (Stay, CulRatio, XRay, Services and Region) and it also had the smallest value for adjusted R-square (although all the values for adjusted R-square were very close). Candidate model #4 contained the same predictors as model #5, plus MedSchool and Census, and its value for adjusted R-square was slightly higher. Based solely on adjusted R-square, it would seem that model #4 is slightly better than model #5.

To gauge the predictive abilities of these models, we use SAS to calculate their respective PRESS statistics. This is done via the PRESS option on the MODEL statement. This option only instructs SAS to calculate the statistic; SAS does not automatically print the statistic. In order to print it, we need some extra SAS code.

```
PROC REG DATA=senic NOPRINT OUTEST=PressInfo;
'#4': MODEL InfRisk = Stay CulRatio XRay Services Reg1 Reg2 Reg3 MedSch Census / PRESS;
'#5': MODEL InfRisk = Stay CulRatio XRay Services Reg1 Reg2 Reg3 / PRESS;
OUTPUT PRESS=press;
RUN;
PROC PRINT DATA=PressInfo;
VAR _MODEL_ _PRESS_;
RUN;
```

The NOPRINT option on the PROC REG statement suppresses printing all of the output. The OUTEST option tells SAS to create a new dataset that will contain estimates generated by the model. We are allowed to choose any name for this new dataset, and we are calling it PressInfo. The PRESS option on each MODEL statement tells SAS to calculate the PRESS statistic, and the OUTPUT statement will write the PRESS statistic to this new dataset. Then the PROC PRINT statement prints the contents of the new dataset. There will be a lot of columns in the new dataset, but we are only interested in the column labeled _PRESS_. The VAR statement will print only the label for the model (_MODEL_) and the PRESS statistic (_PRESS_). The output is shown in Table 3.12

Obs	_MODEL_	_PRESS_
1	'#4'	106.484
2	'#5'	105.355

Table 3.12. PRESS statistics

The PRESS statistics are 105.355 for candidate model #5 and 106.484 for candidate model #4. Since smaller values for PRESS are preferred, we would conclude that candidate model #5 has better predictive ability than #4 (but not by much). Note that candidate model #5 has fewer predictors, but it has better predictive ability. This gives rise to a counter-intuitive fact:

Adding more predictors is sometimes detrimental to the model.

3.4.3. K-fold Cross Validation

PRESS is only one method for assessing the predictive ability of the model. It is sometimes called ‘leave one out’ cross validation, abbreviated LOOCV. K-fold cross validation is similar to PRESS, but operates on groups of observations instead of individual observations. To perform K-fold cross validation, we split the observations in the existing dataset into K groups of approximately equal size.

For each group

- Temporarily remove this group from the data
- Fit the model using the observations in the other groups
- Predict the response value of the observations that were removed
- Calculate the residuals

When these steps have been completed for all the groups, there will be one residual for each observation in the original data set. The sum of these squared residuals measures the predictive ability of the model.

There are no built-in SAS functions or options that will automatically perform K-fold cross validation. Some knowledge of programming is required to implement this approach. The choice of the groups (both the number of groups and the observations that go into each group) can affect the value of the residuals. It is customary to perform K-fold cross validation more than once on the same dataset (using different group assignments) to get a reliable measure of predictive ability.

3.4.4. Mean Square for Prediction

Another method to assess predictive ability of the model is to split the data into two groups: a training set and a testing set. The training set should contain approximately 2/3 of the observations, and the remaining 1/3 are in the testing set. Observations in the training set are used to estimate the regression equation, then the estimated equation is used to predict the response values for the observations in the testing set. Residuals are calculated only for the testing set.

The mean square for prediction (MSPR) is the sum of the squared residuals. It is analogous to MSE, but it is specifically designed to assess the predictive ability of the model. The formula for calculating MSPR is given in equation (3.9).

$$MSPR = \frac{1}{n^*} \sum_{k=1}^{n^*} (Y_k - \hat{Y}_k)^2 \quad (3.9)$$

where

n^* is the number of observations in the testing set

Y_k is the observed response value for the k^{th} observation in the testing set, and

\hat{Y}_k is the predicted response value for the k^{th} observation in the testing set

MSPR is a measure of the error, so smaller is better.

As with K-fold cross validation, there are no built-in SAS functions or options to calculate MSPR, so some knowledge of programming is required to implement this approach. In addition, the value for MSPR depends on which observations are in the training set and which are in the testing set. It is customary to perform this calculation more than once to get a reliable measure of predictive ability.

Chapter 4: One-Way ANOVA

In this module, we begin analyzing data that are collected as part of controlled, randomized experiments. This type of data is fundamentally different from data that are collected in an observational study, although the datasets may look similar and similar statistical techniques can be applied to both. In the first half of this book, we investigated regression analysis, which is typically used for observational data. Observational data are recorded without interfering with the course of events, that is, the researcher is passive observer. There is no attempt to manipulate the conditions under which the data are recorded. For example, in Chapter 1 we performed regression analysis on the lead contamination in trees as function of the traffic volume in a nearby highway. Neither the traffic volume nor the lead contamination was controlled or manipulated in any way. The existing values were simply recorded, and regression analysis was used to quantify the relationship between the variables. In Chapter 3, we used regression analysis to select predictors (i.e., build a model) that could be used to estimate infection risks in hospitals based on existing characteristics of the hospitals. No attempt was made to manipulate these characteristics.

Experimental data is essentially different in two ways.

- For experimental data, the design of the experiment dictates the model. We do not try to build a model from a list of possible predictors. Predictors are chosen in advance and we do not exclude them from the model.
- For experimental data, the predictor variables will primarily be indicator variables (see Section 2.5.). We do not transform indicator variables, but it is sometimes necessary to transform the response variable.

There are other important differences between the analysis of experimental vs. observational data, but first we need to understand the basic framework for experimental designs.

Section 4.1. Principles of Experimental Design

To obtain experimental data, researchers actively intervene to control the study conditions and record the responses. There are several key components of any experiment. Some of these are

- **Factors.** The factors are the predictor variables, which are chosen in advance by the researcher. The values for the factors are also chosen in advance by the researcher.
- **Levels.** The levels are the pre-defined values of the predictor variables.
- **Treatments.** Each treatment is a specific combination of the levels of the factors.
- **Experimental unit (EU).** An EU is the entity to which the treatments are applied. There will be one row in the dataset for each EU.
- **Replication.** The number of times each treatment is applied to an EU.
- **Response variable.** This is the variable that the researchers are most interested in.

To illustrate these components, consider the following example.

The potential for long-term effects from concussions suffered by NFL football players is currently a great concern. Medical professionals and scientists want to develop a better helmet to protect the players from injuries, but it is not clear what type of helmet is best. To gain insight into this problem, researchers designed an experiment to compare three brands of helmets (A, B, and C). Ten helmets of each brand were acquired. Five helmets of each brand were subjected to a sudden impact on the front of the helmet, and the other five were subjected to a sudden impact on the rear of the helmet. The response variable is the amount of impact on the interior of the helmet.

- This experiment has two factors: Brand and Location
- The factor Brand has three levels: A, B, and C
- The factor Location has two levels: Front and Rear
- There are 6 treatments:
 1. A-Front
 2. B-Front
 3. C-Front
 4. A-Rear
 5. B-Rear
 6. C-Rear
- The experimental units (EUs) are the helmets. Since there are 3 brands and 10 helmets of each brand, there are 30 EUs. There will be 30 rows in the dataset.
- The number of replications is 5. There are 30 experimental units and 6 treatments, so each treatment is applied to 5 experimental units.
- The response variable is the amount of impact on the interior of the helmet

There are many types of experimental designs. The most basic is a one-way design, in which there is only one factor with k levels. In one-way designs, each level is treatment. When there are multiple factors, we can have multi-way factorial designs (e.g., two-way, three-way, etc.). For these designs, the treatments consist of every possible combination of the levels of the factors. Another common type of experimental design is a block design. A block is simply a factor created by the researcher to help explain the variation in the response. A blocking factor is not considered part of the treatments, but it is incorporated into the experimental design because it is known (or suspected) to have an effect on the response. There are many other experimental designs (e.g., split plot, strip plot) but we will not consider those in this course.

4.1.1. Experimental error

If we completely ignore the treatments and examine only the values for the response variable, we can calculate the total variability in the response. (This is the total sum of squares that we encountered in regression analysis.) When analyzing experimental data, it is believed that the treatments should help explain some of the total variability in the response. This is analogous to regression analysis, in which we believed that the predictor variables should help explain some of the total variability in the response. The excess variability (i.e., the variability that is not accounted for by the treatments) is called experimental error. Reducing the experimental error is vital for statistical precision, but some amount of experimental error will always be present. Possible sources of experimental error include

- Natural variation among EUs
- Variability in measuring the response
- Failure to include factors that affect the response
- Inability to exactly reproduce experimental conditions from one EU to another

In the football helmet example, natural variation among the EUs could be due to slight inconsistencies in the manufacture of the helmets, which might cause one helmet to be stronger than another even if they are of the same brand and appear to be identical. Variability in measuring the response could occur as a result of re-calibration of the equipment that measures the interior impact. If two helmets of the same brand are painted different colors, and if the chemical formulation of the paint affects the strength of the helmet, then this would be an example of failing to include a factor that affects the response.

Finally, it is never possible to exactly reproduce the experimental conditions from one EU to another. Slight changes in conditions such as temperature and/or humidity might have a slight effect on the

response. In theory, experimental designs should incorporate all factors that have a non-negligible effect on the response. The negligible effects of the other, unmeasured, factors are incorporated into experimental error.

4.1.2. Randomization

Randomization between the treatments and the experimental units is a key component of all experimental designs. The decision regarding which EU gets which treatment is based on a random assignment. This is an *absolutely critical* part of every experimental design, because it is the only way that we have reasonable assurance that any differences we may see in the response variable are due to the treatments and not due to differences among the EUs.

In practice, randomization can be implemented in two different ways:

- 1) If the researcher is creating the conditions that define the treatments, then the EUs are randomly assigned to a treatment.
- 2) If the treatments are naturally occurring characteristics (e.g., gender, species), then the EUs are randomly selected from the population that has those characteristics.

In the football helmet example, researchers employed both of these methods. The 10 helmets of each brand were randomly selected from the population of helmets from that brand (randomization method 2). The factor location (with levels front and back) was specifically chosen by the researcher, with ‘front’ randomly assigned to 5 helmets of each brand so that ‘back’ was assigned to the remaining 5 helmets of each brand (randomization method 1).

For small studies, randomization can be implemented by creating several pieces of paper and labeling them, one for each EU. Put all the papers in a container and shuffle them. Without looking, repeatedly select a piece of paper and record the EU, then discard that piece of paper. The first EU that is selected is assigned to the first treatment, the second EU that is selected is assigned to the second treatment, etc. Cycle through all the treatments until all the EUs are assigned. Note that this process is identical to that employed by most state lotteries (and bingo games!)

For larger studies, this process can be accomplished using computer software to generate the random assignment. One way to do this is to sequentially number all the experimental units (from 1 to n), then randomly rearrange the numbers. Partition the rearranged numbers into k groups, where k is the

number of treatments. All EUs in the first group are assigned to the first treatment, those in the second group are assigned the second treatment, etc.

4.1.3. Replication

Replications are the number of independent repetitions of a treatment, that is, it is the number of experimental units that are assigned to each treatment. If there is no replication, we cannot estimate the experimental error. Without the experimental error, we cannot estimate variability of the treatment effects, so we have no way of knowing how “typical” our results are.

In ideal situations, the experimental design will be *balanced*, that is, there are the same number of EUs assigned to each treatment. Balanced data provide for more precise inference, but datasets are typically not balanced even when the original experimental design is balanced. This is because some data values may be missing. Missing data can occur for a variety of reasons, including laboratory equipment failure, omissions in recording the data, personnel issues, weather-related issues, etc. In general, we start with an experimental design that is balanced, and try to avoid missing data.

4.1.4. Examples

Oil Containers

Plastic containers for holding cooking oil are molded in a machine that has two feeders, each feeding into three molding stations (for a total of six stations). Plastic is extruded through the feeders into continuous cylinders between two halves of each mold. As the molds close, the cylinders are pinched off and air is blown into them to form the container shapes. The production manager is concerned that the molded containers at different stations are not the same weight. To investigate whether or not this is true, a random sample of 8 containers is taken from each station.

- Experimental design: This is a one-way design.
- Factor(s): There is only one factor, Station.
- Levels: The factor Station has 6 levels.
- Treatments: There are 6 treatments, corresponding to the 6 levels of the factor.
- Experimental unit: The EUs are the motor oil containers.
- Number of replications: There are 8 replications. These are the 8 containers at each station.

- Number of EUs: Since there are 6 stations and 8 containers at each station, there are a total of 48 EUs (so 48 observations in the dataset).
- The response variable is the weight of each container.

Cloth Dyeing Experiment

The quality control department of a fabric finishing plant is studying the effect of several factors on dyeing for a cotton cloth used to manufacture shirts. Two operators (1 or 2), three cycle times (40, 50 or 60 minutes) and two temperatures (300° or 350°) were selected, and three small specimens of cloth were dyed under each set of conditions. The finished cloth was compared to a standard, and a numerical score was assigned.

- Experimental design: This is a three-way factorial design.
- Factor(s): The three factors are Operator, Cycle Time, and Temperature.
- Levels: Operator has 2 levels (1 and 2), Cycle Time has 3 levels (40, 50 and 60 minutes), and Temperature has 2 levels (300° and 350°)
- Treatments: There are $2 \times 3 \times 2 = 12$ treatments, which are shown in Table 4.1Table 4..
- Experimental unit: The EUs are the small specimens of cloth.
- Number of replications: There are 3 replications, since 3 EUs were dyed under each set of conditions.
- Number of EUs: Since there are 12 treatments and 3 replications, there are $12 \times 3 = 36$ EUs (so 36 observations in the dataset).
- The response variable is the numeric score assigned to each finished cloth specimen.

Treatment	Operator	Cycle Time	Temp.
1	1	40	300
2	1	40	350
3	1	50	300
4	1	50	350
5	1	60	300
6	1	60	350

Treatment	Operator	Cycle Time	Temp.
7	2	40	300
8	2	40	350
9	2	50	300
10	2	50	350
11	2	60	300
12	2	60	350

Table 4.1. The twelve treatments for the cloth dyeing experiment.

4.1.5. Causation vs. association

When we are dealing with observational data, we can determine whether or not a predictor variable is associated with the response variable, but we cannot conclude that a change in the predictor variable will cause a change in the response. This is because there can be other variables, not recorded in the

dataset or included in the analysis, that affect both the predictors and the response. In order to conclude that a change in a predictor *causes* a change in the response, we must have data from a properly designed experiment, so that all variables that are known (or suspected) to affect the response are incorporated as factors in the experimental design.

4.1.6. Summary

There is a fundamental difference between observational data and experimental data, although the datasets may look similar. With experimental data, the predictor variables are chosen in advance. The values for the predictors are also chosen in advance. We do not choose which predictors we want to include in the analysis – if the predictor is part of the experimental design, then it will be included in the analysis.

With experimental data, we use different terminology than we do with observational data. The predictor variables are now called factors, and the pre-defined values for the factors are called levels. Combinations of the levels of the factors are called the treatments, and the entities to which the treatments are applied are called the experimental units. The number of times each treatment is applied to an experimental unit is called the number of replications.

In order to determine if the predictors cause a change in the response variable, we must have experimental data. Causation cannot be determined from observational data.

Section 4.2. Single Factor Studies

The most basic experimental design contains only one factor. This factor may have many levels, and each level is considered a treatment. This type of experimental design is called a one-way completely randomized design (CRD) because the experimental units are randomly assigned to the treatments without any restrictions on the randomization.

We are usually interested in comparing the mean response for the treatments, and the main question we want to answer is this: Do all treatments have (statistically) the same mean response, or do some of the treatments have a different mean? We answer this question with an ANOVA F test. In this section, we present the details of how this test is constructed.

The ANOVA F test is an extension of the ordinary two-sample t test. Recall that the t test is comparing the means of two groups (populations). The response variable Y is some measured characteristic of each population, and we assume that the values for each population follow a normal distribution. The t test is testing the null hypothesis $H_0: \mu_1 = \mu_2$ vs. the alternative $H_a: \mu_1 \neq \mu_2$. This is illustrated in Figure 4..

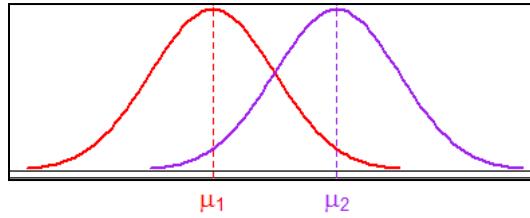


Figure 4.1. Two populations for a t test

The ANOVA F test for a one-way design extends the ordinary two-sample t test to include more than two populations, as illustrated in Figure 4.2. The hypotheses are

$$\begin{aligned} H_0 &: \text{all population means are equal to each other} \\ H_a &: \text{at least one population has a different mean} \end{aligned} \tag{4.1}$$

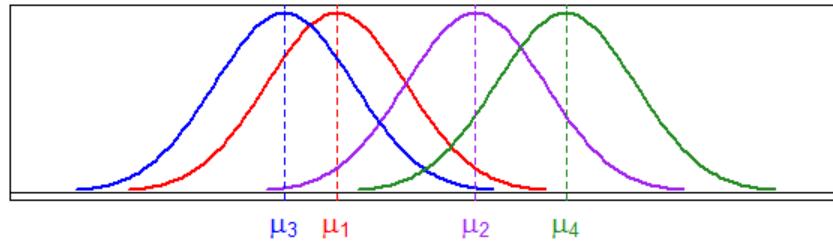


Figure 4.2. Multiple populations for an ANOVA F test

We conduct this test by comparing two variances. This may seem odd at first, since the hypotheses state that we are comparing means, not variances. We illustrate this concept with an example.

4.2.1. Example: Effect of caffeine

In an effort to determine the effect of caffeine on muscle and nervous systems, the following experiment was conducted. Thirty male college students were recruited and randomly assigned to consume one of three dosages of caffeine: 0 mg, 100 mg or 200 mg. Two hours later, the student repeatedly taps his finger on the table, and the researcher records the number of taps per minute. We want to use these data to answer the question: Does caffeine affect the mean number of finger taps?

The recorded data are shown in Table 4.2 and a scatterplot of the data is shown in Figure 4.3. Note that the scatterplot appears “stacked”. This is because there are 10 replications for each level of Caffeine, so the dataset contains 10 Y values for each X. There should be 10 points in each vertical column of points, but some of the points are hidden because there are duplicate values for Y.

0 mg	100 mg	200 mg
242	248	246
245	246	248
244	245	250
248	247	252
247	248	248
248	250	250
242	247	246
244	246	248
246	243	245
242	244	250

Table 4.2. Caffeine data

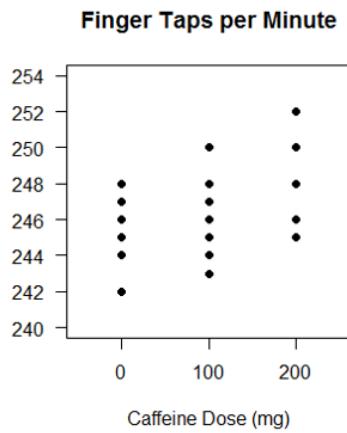


Figure 4.3. Scatterplot of caffeine data

We will use these data to determine if caffeine dosage affects the mean number of finger taps. This test involves comparing two types of variability: the within group variability and the between group variability, where each group corresponds to one treatment (0, 100 or 200 mg).

4.2.2. Within- and between- group variability

Within-group variability is based on the variance of the values within each treatment. The variance of the number of finger taps for dosage 0 mg is estimated using only the recorded values for 0 mg. The same is true for 100 mg and 200 mg – the variance for each of these two groups are estimated from the observations that were measured in that treatment. This is illustrated in Figure 4.4. Note that each treatment group has a different mean, denoted by the horizontal bar within each group, but this is not enough information for us to declare that the population means are different. Some amount of variability will always be present, so we need a way to determine if the sample means are different *enough* that we could conclude the population means are different.

The variability within each treatment group is characterized by the amount of vertical spread in the points for each group. We measure this numerically via the variance. If we combine the variability for all the groups, we have a measure of the within-group variability.

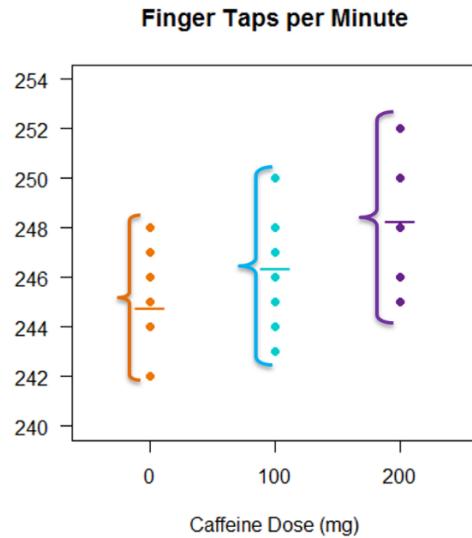


Figure 4.4. Within-group variability for caffeine example

We now consider the other type of variability -- the between-group variability. This measures the variability among the treatment means. Using the sample mean for each group, we calculate the variance of these means. This is illustrated in Figure 4.5. To perform the hypotheses test, we need to decide if the variability among the treatment means is “large” relative to the variability within treatments. We will answer this question with an F test, similar to what we have done with regression analysis.

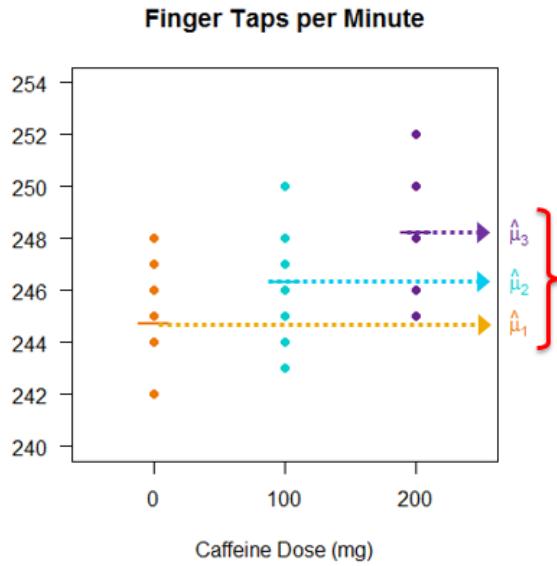


Figure 4.5. Between-group variability for caffeine example

We do not expect the sample means to all be identically equal, even if the population means are equal. There is always sampling variability associated with values calculated from a sample. If the variability among the means (i.e., between the groups) is much larger than the variability within the groups, then the sample means are relatively widespread and it is believable that the true population means are different. On the other hand, if the variability between groups is about the same as (or less than) the variability within groups, then the sample means are NOT more widespread, and it is believable that the true population means are all the same.

Figure 4.6 illustrates the results of two different fictitious experiments. Both experiments contain 3 treatments (labeled A, B and C in the graphs) and both experiments produce approximately the same sample means for each treatment. In the graph on the left, the within-group variability is small, and this makes it easier to detect differences among the treatment means. In the graph on the right, the within-group variability is much larger, and this makes it more difficult to detect differences among the means.

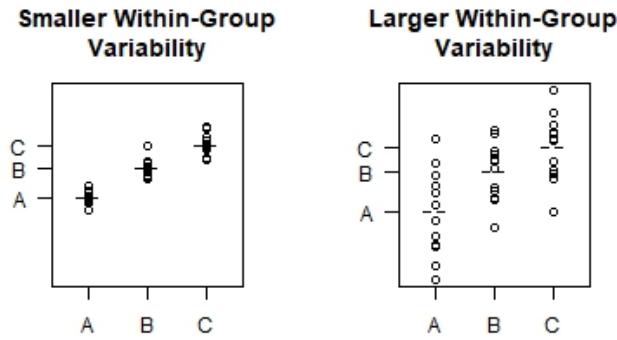


Figure 4.6. Comparing large and small within-group variability

To be precise about how the within- and between-group variability is calculated, we need some notation.

n_i = sample size for the i^{th} treatment

Y_{ij} = observed value for the j^{th} EU the i^{th} treatment

$\bar{Y}_{i\cdot}$ = sample mean for the i^{th} treatment

$\bar{Y}_{\cdot\cdot}$ = the “grand mean”, a.k.a. “overall mean”, i.e., the mean of all the observations

s_i^2 = sample variance for the i^{th} treatment

N = total number of observations

Treatments	Observed Values				Sample means	Sample variances
Dose 0 mg	Y_{11}	Y_{12}	...	$Y_{1,10}$	$\bar{Y}_{1\cdot}$	s_1^2
Dose 100 mg	Y_{21}	Y_{22}	...	$Y_{2,10}$	$\bar{Y}_{2\cdot}$	s_2^2
Dose 200 mg	Y_{31}	Y_{32}	...	$Y_{3,10}$	$\bar{Y}_{3\cdot}$	s_3^2

Table 4.3. Notation for a one-way experimental design

Treatments	Sample size	Sample means	Sample variances
Dose 0 mg	$n_1 = 10$	$\bar{Y}_{1\cdot} = 244.8$	$s_1^2 = 5.73$
Dose 100 mg	$n_2 = 10$	$\bar{Y}_{2\cdot} = 246.4$	$s_2^2 = 4.27$
Dose 200 mg	$n_3 = 10$	$\bar{Y}_{3\cdot} = 248.3$	$s_3^2 = 4.90$

Table 4.4. Summary statistics for the caffeine example

The summary statistics for the caffeine example are shown in Table 4.4. The total within-group variability is calculated as the sum of the total squared deviations within each group. These are directly related to the sample variance for each group. We continue to use the same notation that we used with regression. The total squared deviation is called the sum of squares, abbreviated SS.

The total squared deviations within each group is the sum of squares for the group. These are calculated as follows:

$$\text{For dose 0: } SS_1 = (n_1 - 1)s_1^2 = (9)5.73 = 51.57$$

$$\text{For dose 100: } SS_2 = (n_2 - 1)s_2^2 = (9)4.27 = 38.43$$

$$\text{For dose 200: } SS_3 = (n_3 - 1)s_3^2 = (9)4.90 = 44.10$$

We sum these values to get the within-group variability.

$$SS_{Within} = \sum_{\text{groups}} SS_{group} = \sum_{\text{groups}} (n_i - 1)s_i^2 \quad (4.2)$$

For the caffeine example, $SS_{Within} = SS_1 + SS_2 + SS_3 = 51.57 + 38.43 + 44.10 = 134.1$. This is the **sum of squares due to error (SSE)**. For the caffeine example, SSE = 134.1.

We now turn our attention to the between-group variability. This is the variability among the treatment means and it is also called the sum of squares due to treatments. From the summaries in Table 4.4, we first calculate the grand mean.

$$\bar{Y}_{..} = \frac{1}{N} \sum_i \sum_j Y_{ij} = \frac{1}{N} \sum_{\text{groups}} n_i \bar{Y}_i. \quad (4.3)$$

For the caffeine data, the grand mean is $\bar{Y}_{..} = \frac{1}{30} \{(10)244.8 + (10)246.4 + (10)248.3\} = 246.5$. The total variation in the treatment means is the sum of squares *between* treatments.

$$SS_{Between} = SS_{Treatments} = \sum n_i (\bar{Y}_{i..} - \bar{Y}_{..})^2 \quad (4.4)$$

For the caffeine example,

$$\begin{aligned} SS_{Between} &= SS_{Treatments} = n_1(\bar{Y}_{1..} - \bar{Y}_{..})^2 + n_2(\bar{Y}_{2..} - \bar{Y}_{..})^2 + n_3(\bar{Y}_{3..} - \bar{Y}_{..})^2 \\ &= (10)(244.8 - 246.5)^2 + (10)(246.4 - 246.5)^2 + (10)(248.3 - 246.5)^2 = 61.4 \end{aligned}$$

4.2.3. Conduct the hypothesis test

Now that we have the sums of squares, we next determine the degrees of freedom and calculate the mean squares. This is exactly what we did in regression analysis, except that now we are using the word “Treatment” instead of the word “Model”. In a one-way analysis of variance with t treatments, the degrees of freedom for treatments is $t - 1$ and the degrees of freedom for error is $N - t$. The mean square is the sum of squares divided by the degrees of freedom.

For the caffeine example, there are $t = 3$ treatments and $N = 30$ observations, so

$$\text{For Treatments: } df \text{ Treatments} = t - 1 = 3 - 1 = 2 \text{ and } MS \text{ Treatments} = 61.4 / 2 = 30.7$$

$$\text{For Error: } df \text{ Error} = N - t = 30 - 3 = 27 \text{ and } MS \text{ Error} = 134.1 / 27 = 4.967$$

To test the hypotheses in equation (4.1), the test statistic is

$$F = \frac{MS \text{ Between}}{MS \text{ Within}} = \frac{MS \text{ Treatments}}{MS \text{ Error}} = \frac{MSTrt}{MSE} \quad (4.5)$$

This follows an F distribution with numerator degree of freedom equal to $df \text{ Treatment}$ and denominator degrees of freedom equal to $df \text{ Error}$. We will reject H_0 when the test statistic is large.

For the caffeine example, the test statistic is $F = 30.7 / 4.967 = 6.18$, and the degrees of freedom are 2 and 27, respectively. From the probability table for the F distribution, the critical value is 3.35 (assuming $\alpha = 0.05$). Since the test statistic is greater than the critical value, we reject H_0 . The conclusion can be stated this way: At significance level 0.05, the sample provides convincing evidence that the mean number of finger taps per minute is different for different caffeine doses.

The calculations necessary to conduct this hypothesis test are summarized in an ANOVA table. This is the same ANOVA table that we used with regression analysis, except that we are now using the term “Treatment” instead of “Model” and we are using the letter t (for number of treatments) instead of p (for number of predictors). This is shown in Table 4.5.

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F	p-value
Treatment	$t - 1$	SS Treatment	MS Treatment	MS Trt / MSE	
Error	$N - t$	SS Error	MS Error		
Corrected Total	$N - 1$	SS Total			

Table 4.5. ANOVA table for one-way analysis of variance

In the next section, we will introduce the SAS code necessary to generate the ANOVA table. All of the results generated by SAS will match our hand calculations, with the exception of the p-value. We will rely on SAS to perform these calculations and generate the p-value.

Section 4.3. Linear Models

There are two basic methods to write a model for an analysis of variance problem. We first present the cell means model. The equation for a cell means model explicitly includes a parameter for each treatment mean. Each parameter represents the population mean for one of the treatments, and the estimate for each parameter is the corresponding sample mean. This method for expressing an analysis of variance model is fairly straightforward because it allows us to compare treatment means by directly comparing two (or more) model parameters. The cell means model is useful for understanding how a linear model can be applied to an ANOVA analysis, but it is not implemented by SAS. The model implemented by SAS is called an effects model. In an effects model, one treatment is assigned to be a reference treatment and all the other treatments are compared to the reference treatment. The difference between the mean of the reference treatment and the mean of another treatment is called the effect of the “other” treatment. The model contains one parameter for the reference treatment plus one parameter for each effect.

4.3.1. Cell means model

The equation for a cell means model can be written

$$Y_{ij} = \mu_i + \varepsilon_{ij} \quad (4.6)$$

where

$$i = 1, 2, \dots, t \quad (t \text{ is the number of treatments})$$

$$j = 1, 2, \dots, n_i \quad (\text{treatment } i \text{ has sample size } n_i)$$

Y_{ij} is the observed response for the j^{th} EU in the i^{th} treatment

μ_i is the true mean response for treatment i (i.e., the population mean)

ε_{ij} is the error for the the j^{th} EU in the i^{th} treatment

We assume that the ε_{ij} are independent, normally distributed, with mean 0 and constant variance σ^2 .

For the caffeine data, the cell means model (equation (4.6)) generates these least squares estimates

$$\hat{\mu}_1 = \bar{Y}_{1\cdot}, \quad \hat{\mu}_2 = \bar{Y}_{2\cdot}, \quad \text{and} \quad \hat{\mu}_3 = \bar{Y}_{3\cdot}$$

Note that these estimates are the sample means for each treatment.

One frequently asked question is “How can this be a linear model? Where are the X’s?” The model in equation (4.6) can be written another way, and this requires indicator variables. When there are t treatments, there will be t indicator variables, defined as

$$X_1 = \begin{cases} 1 & \text{for treatment 1} \\ 0 & \text{otherwise} \end{cases} \quad X_2 = \begin{cases} 1 & \text{for treatment 2} \\ 0 & \text{otherwise} \end{cases} \quad \dots \quad X_t = \begin{cases} 1 & \text{for treatment } t \\ 0 & \text{otherwise} \end{cases}$$

Then the cell means model can be written

$$Y_{ij} = \beta_1 X_{1ij} + \beta_2 X_{2ij} + \dots + \beta_t X_{tij} + \varepsilon_{ij} \quad (4.7)$$

Note that there is no intercept in the cell means model.

To illustrate how the indicator variables are incorporated into equation (4.7) for the caffeine data, we specify the subscripts (i and j) and assign values to the three indicator variables. These are shown in Table 4.6. The cell means model for these data is

$$Y_{ij} = \beta_1 X_{1ij} + \beta_2 X_{2ij} + \beta_3 X_{3ij} + \varepsilon_{ij}$$

The least squares estimates for the parameters of this model are

$$\hat{\beta}_1 = \bar{Y}_{1..}, \quad \hat{\beta}_2 = \bar{Y}_{2..}, \quad \text{and} \quad \hat{\beta}_3 = \bar{Y}_{3..}$$

Note that the estimates for the β 's are exactly the same as the estimates for the μ 's in equation (4.6). These estimates are simply the sample means for each treatment.

Taps Y_{ij}	Treatment	i	j	X_1	X_2	X_3
242	0mg	1	1	1	0	0
245	0mg	1	2	1	0	0
...
242	0mg	1	10	1	0	0
248	100 mg	2	1	0	1	0
246	100 mg	2	2	0	1	0
...
244	100 mg	2	10	0	1	0
246	200mg	3	1	0	0	1
248	200mg	3	2	0	0	1
...
250	200 mg	3	10	0	0	1

Table 4.6. Subscripts and indicators for cell means model

Regardless of which version of the cell means model is used, the parameter estimates are generated via the least squares criterion. Because they are least squares estimates, they are unbiased and have minimum variance of all unbiased unbiased estimators. (See the Gauss-Markov Theorem in Section 1.2.)

Before we can perform inference (hypothesis tests and/or confidence intervals), we need to know the standard errors of these estimates. From statistical theory, we know that

$$\text{var}(\hat{\mu}_i) = \text{var}(\hat{\beta}_i) = \frac{\text{var}(Y_{ij})}{n_i} = \frac{\sigma^2}{n_i}.$$

Since the value of σ^2 is not known, we use its estimate $\hat{\sigma}^2 = \text{MSE}$, and this produces an estimated variance

$$\hat{\text{var}}(\hat{\mu}_i) = \hat{\text{var}}(\hat{\beta}_i) = \frac{\hat{\sigma}^2}{n_i} = \frac{\text{MSE}}{n_i}.$$

The standard error is the square root of the estimated variance

$$SE(\hat{\mu}_i) = SE(\hat{\beta}_i) = \sqrt{\frac{\text{MSE}}{n_i}}.$$

For each estimated treatment mean, the standard error depends on the sample size for the treatment. If all the treatments have the same sample size, then all the standard errors will be the same. This is what we call “balanced” data. If the treatments have different sample sizes, then the standard errors will be different. This is what we call “unbalanced” data.

For the caffeine example, each treatment has a sample size of 10. We have previously calculated $\text{MSE} = 4.967$ and the degrees of freedom for error, $\text{dfE} = 27$. Since this is balanced data, the standard errors are all the same.

$$SE(\hat{\mu}_1) = SE(\hat{\mu}_2) = SE(\hat{\mu}_3) = \sqrt{\frac{\text{MSE}}{n_i}} = \sqrt{\frac{4.967}{10}} = 0.705$$

To construct confidence intervals and conduct hypothesis tests, use critical values from a t distribution with degrees of freedom equal to dfE . For the caffeine data, to construct 95% confidence intervals for the means, we use the critical value $t_{\alpha/2, \text{dfE}} = t_{0.025, 27} = 2.052$. (Recall that we always use $\alpha/2$ for confidence intervals and for two-sided tests.)

The generic form of a confidence interval for population mean is (point estimate) \pm (margin of error), where (margin of error) = (critical value) \times SE. Since the caffeine data is balanced, the margin of error will be the same for all treatments. (In other words, we have the same level of precision for all the treatments.) The margin of error is $2.052 \times 0.705 = 1.45$. The 95% confidence intervals for the caffeine treatment means are

- For dose 0 mg: 244.8 ± 1.45 , or (243.35, 246.25) finger taps
- For dose 100 mg: 246.4 ± 1.45 , or (244.95, 247.85) finger taps
- For dose 200 mg: 248.3 ± 1.45 , or (246.85, 249.75) finger taps

Later in this section, we will see these standard errors and confidence intervals in the SAS output.

We can also perform hypothesis tests for the population means. For some constant C , the hypotheses for the i^{th} treatment mean are $H_0 : \mu_i = C$ vs. $H_a : \mu_i \neq C$. The test statistic is

$$t = \frac{\hat{\mu}_i - C}{SE(\hat{\mu}_i)} = \frac{\hat{\mu}_i - C}{\sqrt{\frac{MSE}{n_i}}} \quad (4.8)$$

The critical value is from the t distribution, with degrees of freedom dfE. We reject H_0 if $|t|$ is greater than the critical value. We can get SAS to perform this test, but only when $C = 0$. For most datasets, this particular test is not very informative. Instead, we are usually interested in comparing the means of two different treatments. This type of test will be discussed later.

4.3.2. Effects model

A second way to write an ANOVA model is the effects model. One treatment (i.e., one level of the factor) is selected to be the reference level and all other treatments are compared to the mean of the reference level. The equation for this model is

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad (4.9)$$

where

$i = 1, 2, \dots, t - 1$ (the t treatment groups, excluding the reference level)

$j = 1, 2, \dots, n_i$ (the EU's within each treatment)

Y_{ij} is the observed response for the j^{th} EU in the i^{th} treatment

μ is the true (population) mean for the reference level

τ_i is the effect of the i^{th} treatment

ε_{ij} is the error for the j^{th} EU in the i^{th} treatment

We assume that the ε_{ij} are independent and normal, with mean 0 and constant variance σ^2 .

The effects model and the cell means model produce identical estimates for the treatment means, as well as hypothesis tests and confidence intervals regarding these means. The only difference between the cell means and effects models is in how they are parameterized. For example, in the cell means

model the mean for the i^{th} treatment is simply μ_i , but in the effect model this mean is $\mu + \tau_i$. This implies that $\tau_i = \mu_i - \mu$, which is simply the difference between the i^{th} treatment mean and the reference level mean. This is why the τ 's are called the 'effects'. If τ_i is positive, then the effect of treatment i is that it produces a larger mean than the reference level. If τ_i is negative, then the effect of treatment i is that it produces a smaller mean than the reference level.

Any of the treatments could be the reference level treatment, but in the following discussion we will use the subscript R to represent the reference level. The effects model (in equation (4.9)) has a total of t parameters that must be estimated: there are $t - 1$ τ 's (the effects) and one μ (the reference level mean). From the cell means model, we know that the least squares estimate for the i^{th} treatment mean is $\bar{Y}_{i\cdot}$ (which is just the sample mean for the i^{th} treatment), so in the effects model the least squares estimate for μ is $\hat{\mu} = \bar{Y}_R$. (Note that this has the subscript R for the reference level.) Since the τ 's are the differences in means, the least squares estimate for τ_i is $\hat{\tau}_i = \bar{Y}_{i\cdot} - \bar{Y}_R$.

4.3.3. SAS code for the effects model

While it may be easier to understand the cell means model, most software (including SAS) implements analysis of variance via the effects model. The predictor variable is the factor that defines the treatments. In most analysis of variance problems, the predictor will be a categorical (non-numeric) variable. Other names for this type of variable are categorical or classification variables. Since the predictor is not numeric, we cannot use PROC REG to perform analysis of variance. Instead, we must use PROC GLM. (GLM is an abbreviation for General Linear Model.) We first encountered PROC GLM in Section 2.5, where we modeled the time to pain relief (the response) as a function of headache severity and gender. In that example, headache severity was a numeric variable and gender was a classification variable. In most analysis of variance problems, all of the predictors will be classification variables and they will be incorporated into the model via indicator variables.

The basic code for performing analysis of variance on the caffeine example is shown below.

```
PROC GLM DATA=caffeine;
  CLASS Dose;
  MODEL Taps = Dose / SOLUTION ;
  LSMEANS dose / STDERR CL;
  run;
```

The CLASS statement defines the predictor Dose as a classification variable. This statement instructs SAS to create the necessary indicator variables for this predictor. We do not construct the indicator variables ourselves, but we need to understand how SAS constructs these variables so that we can understand how to interpret the SAS output. By default, SAS will choose the last level of the factor as the reference level. This is the last level alphabetically; it is not the last value that is in the dataset. Since the values for Dose are 0, 100 and 200, SAS considers these alphabetic (not numeric) values, and will take the last one (200) as the reference level. It is possible to change the reference level, but this should be done only if there is a good reason to do so. In the caffeine example, it makes more sense to make Dose 0 the reference level, and this is accomplished with a slight modification to the CLASS statement.

```
CLASS Dose (REF='0') ;
```

Note that there are quotes around 0, because SAS considers all classification variables as alphabetic.

The MODEL statement defines the response variable (on the left side of the equals sign) and the predictor(s) on the right side of the equals sign. The SOLUTION option on the MODEL statement tells SAS to print a table containing the least squares estimates (and other quantities) for each parameter that is in the model. The SOLUTION option can be removed if this table is not desired.

Both the CLASS and the MODEL statements are required for PROC GLM. The LSMEANS statement is optional, but recommended. Since SAS will be using the effects model (not the cell means model), none of the explicit parameter estimates will provide the estimated treatment means. The LSMEANS statement will generate a table of these means. and the options STDERR and CL will produce the standard errors and confidence limits, respectively, for the treatment means. As with all SAS options, these should be included only when they are needed.

The SAS code and output for analyzing the caffeine data are shown in the next section.

4.3.4. SAS code and output for the caffeine data

```
DATA caffeine;
INPUT Dose Taps @@;
DATALINES;
  0 242   0 245   0 244   0 248   0 247
  0 248   0 242   0 244   0 246   0 242
100 248 100 246 100 245 100 247 100 248
100 250 100 247 100 246 100 243 100 244
200 246 200 248 200 250 200 252 200 248
200 250 200 246 200 248 200 245 200 250
;
PROC GLM DATA=caffeine PLOTS=DIAGNOSTICS;
  CLASS Dose (REF='0');
  MODEL Taps = Dose / SOLUTION;
  LSMEANS Dose / STDERR CL;
run;
```

Note: The @@ at the end of the INPUT statement tells SAS to read multiple (Dose, Taps) pairs of data on the same line.

The first page of the GLM output contains two tables. The Class Level Information table provides every classification variable that was specified in the CLASS statement, and lists the levels for each classification variable. The order of the levels is important to understand the other output for this model. The last value in this list is the reference level. Since we have used the option REF='0' on the CLASS statement, the last value is 0 and this is the reference level. If we did not include this option on the CLASS statement, then the values would be ordered 0 100 200, and the reference level would be 200. The second table on this page of the output provides the total number of observations that SAS read from the provided data. Always make sure SAS is using all of the provided data.

The GLM Procedure

Class Level Information		
Class	Levels	Values
Dose	3	100 200 0

Number of Observations Read	30
Number of Observations Used	30

The second page of the SAS output contains 5 tables. The first table is the ANOVA table, and this is followed by a table that contains basic information about the model. The third table is the Type I sum of squares (SS) table. IGNORE THE TYPE I SS TABLE. The 4th table is the Type III sum of squares table, and we will use this in our interpretation of the results when the model contains more than one factor. The last table is the Parameter Estimates table, and it was generate only because we included the SOLUTION option on the MODEL statement. The message at the bottom of this page can be ignored. This message will appear every time there is a classification variable.

The GLM Procedure
Dependent Variable: Taps

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	61.4000000	30.7000000	6.18	0.0062
Error	27	134.1000000	4.9666667		
Corrected Total	29	195.5000000			

R-Square	Coeff Var	Root MSE	Taps Mean
0.314066	0.904098	2.228602	246.5000

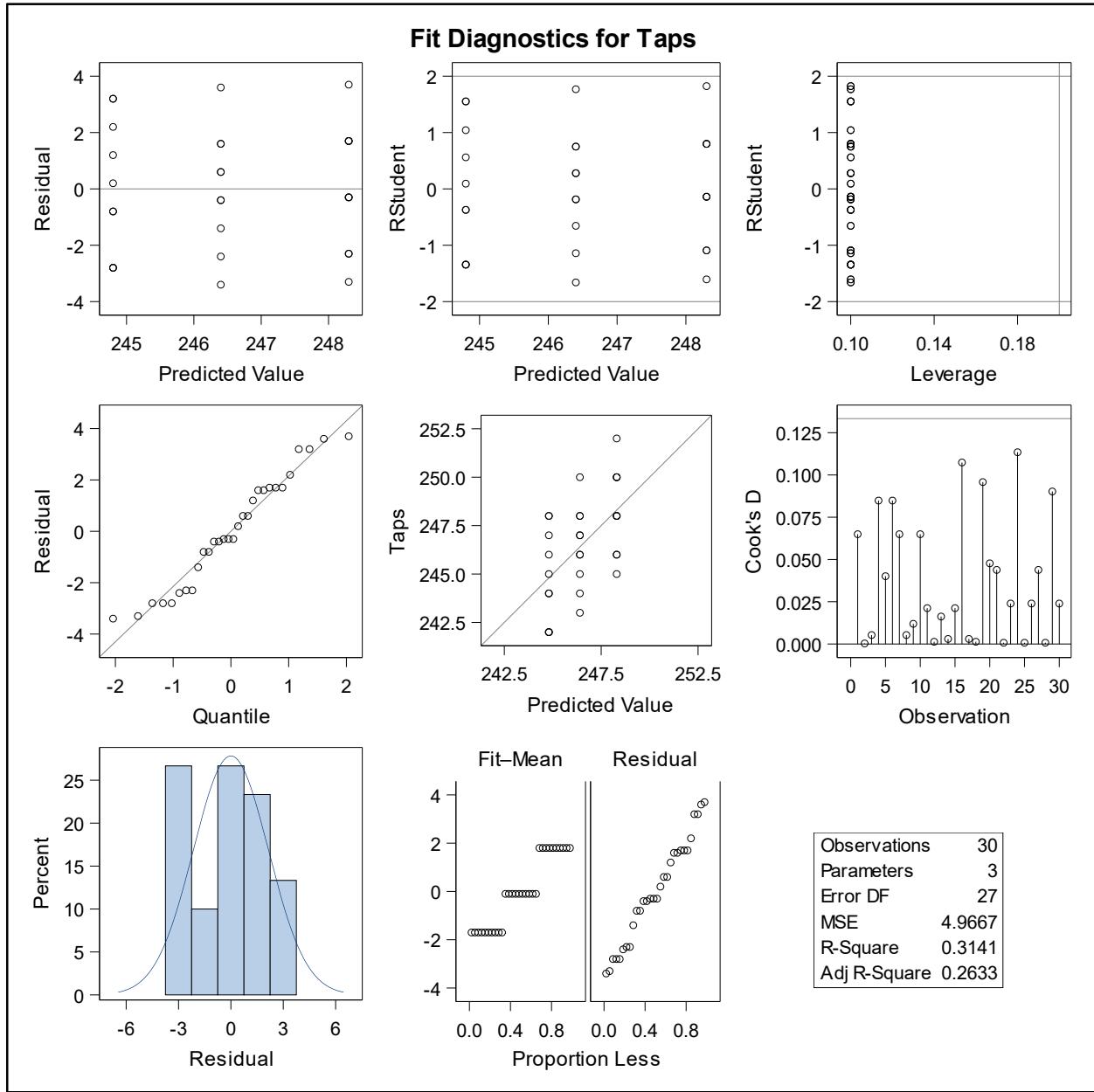
Source	DF	Type I SS	Mean Square	F Value	Pr > F
Dose	2	61.4000000	30.7000000	6.18	0.0062

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Dose	2	61.4000000	30.7000000	6.18	0.0062

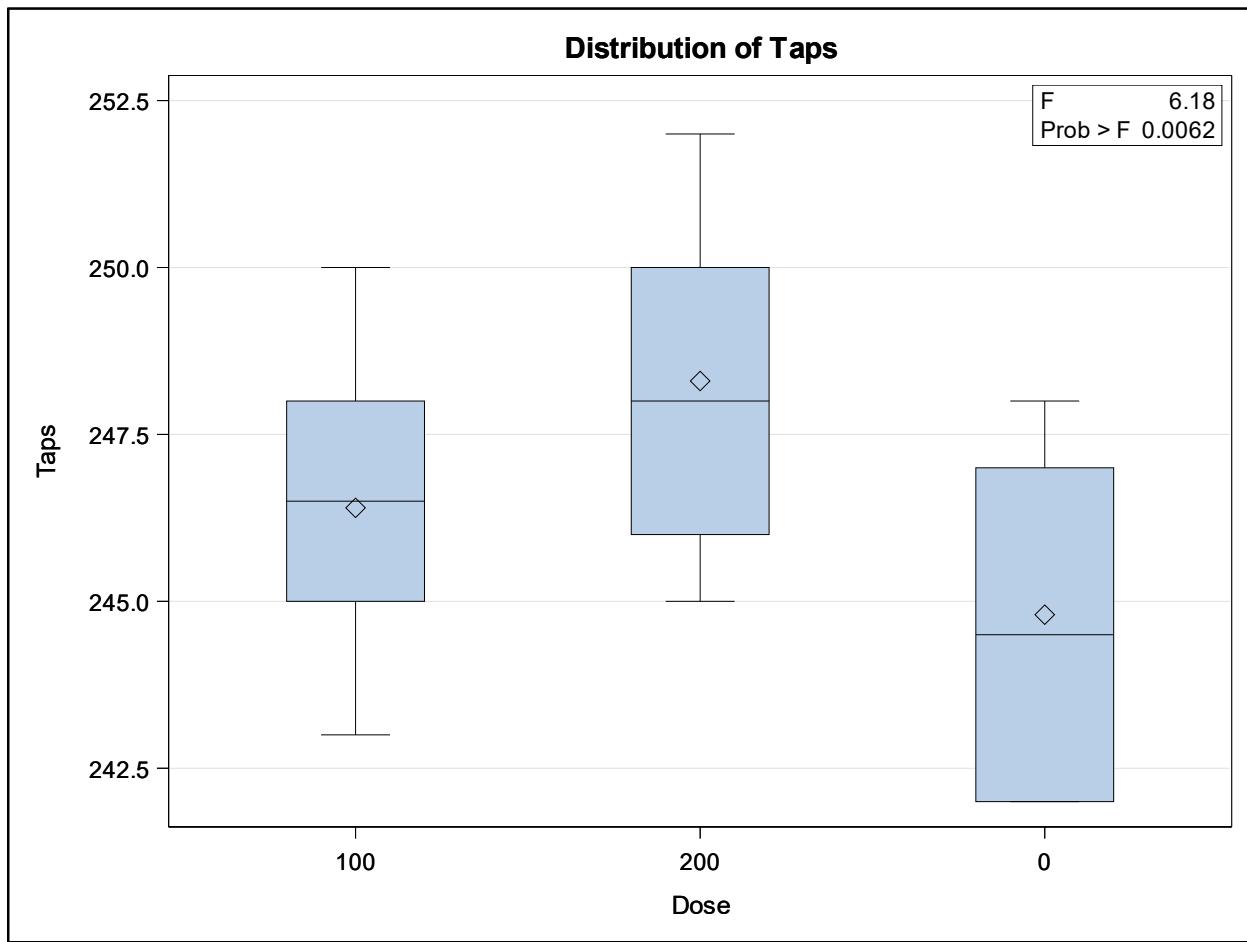
Parameter	Estimate		Standard Error	t Value	Pr > t
Intercept	244.8000000	B	0.70474582	347.36	<.0001
Dose 100	1.6000000	B	0.99666109	1.61	0.1200
Dose 200	3.5000000	B	0.99666109	3.51	0.0016
Dose 0	0.0000000	B	.	.	.

Note: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

The next page in the SAS output contains the diagnostic plots for this model. We will use these plots exactly the way we used them in regression analysis. The graph in the upper left is the residual plot, and the one below that is the normal QQ plot. We use these to decide whether or not the model assumptions appear to be satisfied. In several of these plots, the points appear to be “stacked” instead of randomly scattered. This is because there are only 3 possible values for X (Dose) in the data, so there are multiple Y values for each X.



The next page of the SAS output contains side-by-side boxplots of the observed response values for each treatment. The estimated means are shown as small diamonds inside each box. This graph is useful for getting an overall “feel” of the data, but it is not particularly useful for any specific part of the analysis. If the shaded portions of the boxes appear to have extremely different sizes, then this could be an indication that the assumption of equal variances has been violated, and this concern would need to be investigated further. Note that the levels for Dose are in the order that was defined in the Class Level Information table (on the first page of the output). Dose 0 is last because it is the reference level. The inset (in the upper right corner) provide the test statistic and p-value for testing the ANOVA hypotheses. These values are the same as those in the ANOVA table.



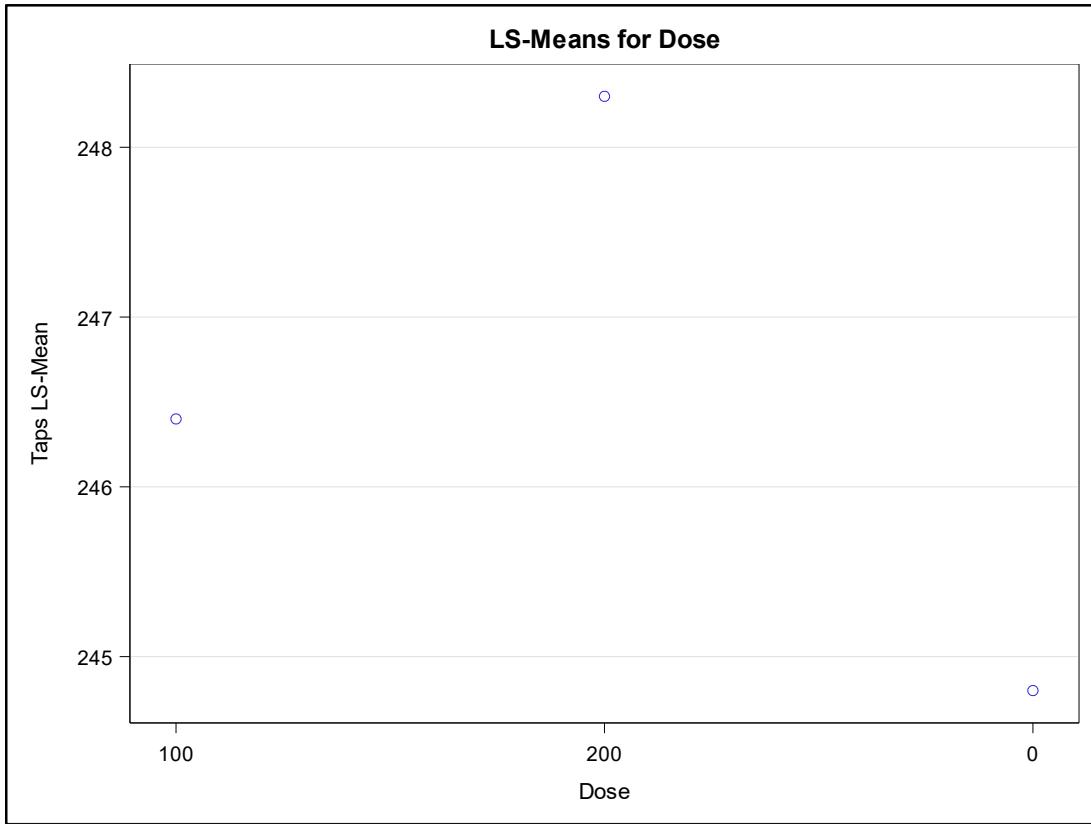
The last page of the SAS output is the result of the LSMEANS statement. If this statement is not included in the SAS code, then this page of output will not be generated. The table contains, for each level of Dose, the estimated treatment mean and the p-value for testing

$$H_0 : (\text{treatment mean}) = 0 \quad \text{vs.} \quad H_a : (\text{treatment mean}) \neq 0$$

These tests are not usually very informative, and they are not part of a typical analysis of variance. This table also contains the standard errors and the confidence limits for each treatment mean, but this is only because we included these options on the LSMEANS statement. Notice that the estimates, the standard errors and the confidence limits match the ones we calculated “by hand”.

The GLM Procedure
Least Squares Means

Dose	Taps LSMEAN	Standard Error	Pr > t	95% Confidence Limits	
100	246.400000	0.704746	<.0001	244.953981	247.846019
200	248.300000	0.704746	<.0001	246.853981	249.746019
0	244.800000	0.704746	<.0001	243.353981	246.246019



4.3.5. Steps to interpret the SAS output

1. Look at the two tables on the first page of the output. Make sure that all levels of the factor are listed, that there are no unexpected levels, and that all observations are being used.
2. Look at the diagnostic plots on the third page of the output to verify that the model assumptions appear to be satisfied.
3. Look at the ANOVA table on the second page of the output, which is reproduced here.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	61.4000000	30.7000000	6.18	0.0062
Error	27	134.1000000	4.9666667		
Corrected Total	29	195.5000000			

The total degrees of freedom is 29, and this is always one less than the number of observations ($N = 30$). Since there are $t = 3$ treatments, the degrees of freedom for Model is $t - 1 = 3 - 1 = 2$. The degrees of freedom for error is $N - t = 30 - 3 = 27$. The sum of squares for Model and for Error were calculated “by hand” earlier, as were the two mean squares and the test statistic. Both the test statistic (“F Value”) and the p-value (“Pr > F”) are for testing the null hypothesis that all treatments have the same mean versus the alternative that at least one treatment has a different mean. This test is often referred to as the “overall ANOVA F test”. Since the p-value is less than 0.05, we reject the null hypothesis and conclude that at least one mean is different. (The analysis will not be complete until we try to identify which mean is different, but that will be discussed later.)

Note: If the p-value for the overall ANOVA F test is greater than 0.05, then we would conclude that all treatments have the same mean. In this event, there should be no further investigations into comparing the treatment means. It would appropriate to summarize the treatment means (for example, by reporting their estimates and/or confidence intervals), but it would not be appropriate to conduct additional hypothesis tests to compare specific means. When the overall ANOVA F test indicates that all treatments have the same mean, it is pointless (and incorrect) to try to find any differences.

4. Since the overall ANOVA F test is significant for the caffeine data, the next step in our analysis should be to identify which treatment means are different. There are several different things

we could do, but since this is our first example, we will look only at the Parameter Estimates table. This is the last table on the second page of the output, and is reproduced here.

Parameter	Estimate		Standard Error	t Value	Pr > t
Intercept	244.8000000	B	0.70474582	347.36	<.0001
Dose 100	1.6000000	B	0.99666109	1.61	0.1200
Dose 200	3.5000000	B	0.99666109	3.51	0.0016
Dose 0	0.0000000	B	.	.	.

To properly interpret the Parameter Estimates table, we must compare it to the equation for the effects model (equation (4.9)). The caffeine data has 3 treatments and we are using Dose=0 as the reference treatment. So the model is

$$Y_{ij} = \mu + \tau_1 + \tau_2 + \varepsilon_{ij}$$

where μ is the mean for the reference treatment, τ_1 is the effect for Dose 100 and τ_2 is the effect for Dose 200. (We know that Dose 100 comes before Dose 200 because that is the order in the Class Level Information table.) In the Parameter Estimates table, the line labeled “Intercept” corresponds to the reference treatment (μ) , the line labeled “Dose 100” is for τ_1 and the line labeled “Dose 200” is for τ_2 .

For the line labeled “Intercept”: The estimate for μ is $\hat{\mu} = 244.8$. This is the estimated mean for Dose 0 (the reference level). Note that the standard error for this estimate (0.7047) matches what we calculated “by hand” (0.705). The test statistic (“t Value”) and p-value (“Pr>|t|”) are for testing whether or not this treatment mean is equal to 0. We are usually not interested in this test.

For the line labeled “Dose 100”: The estimate for τ_1 is $\hat{\tau}_1 = 1.6$. This is the estimated EFFECT for Dose 100. To get the estimated MEAN for Dose 100, we need to calculate it.

$$\hat{\mu}_1 = \hat{\mu} + \hat{\tau}_1 = 244.8 + 1.6 = 246.4.$$

The standard error is not what we calculated earlier, because this is the standard error for the estimated effect and we had previously calculated the standard error for the estimate mean. The test statistic and p-value are for testing whether or not the effect is equal to 0, i.e., whether or not $\tau_1 = 0$. Since the effect is the difference between the mean for this treatment and the

mean for the reference treatment, the null hypothesis for this test can be written $\mu_1 - \mu = 0$. In other words, the test reported on the “Dose 100” line is testing whether or not the mean for Dose 100 is equal to the mean for the reference level. The p-value is 0.12, which is greater than 0.05, so we conclude that these two means are equal.

For the line labeled “Dose 200”: The estimate for τ_2 is $\hat{\tau}_2 = 3.5$. This is the estimated EFFECT for Dose 200. To get the estimated MEAN for Dose 200, we calculate it as

$$\hat{\mu}_2 = \hat{\mu} + \hat{\tau}_2 = 244.8 + 3.5 = 248.3.$$

The standard error matches the standard error on the Dose 100 line, since both of these are for an estimated effect (not a mean) and we are working with balanced data. The test statistic and p-value are for testing whether or not the effect is equal to 0, i.e., whether or not $\tau_2 = 0$, which is the same as whether or not $\mu_2 - \mu = 0$. In other words, the test reported on the “Dose 200” line is testing whether or not the mean for Dose 200 is equal to the mean for the reference level. The p-value is 0.0016, which is less than 0.05, so we conclude that these two means are different.

Summary of the Parameter Estimates table. From the Parameter Estimates table, we conclude that the mean for Dose 0 is not 0 (which is not surprising), the means for Dose 0 and Dose 100 are the statistically the same, and the means for Dose 0 and Dose 200 are significantly different. Note that there is nothing in the Parameter Estimates table that provides any information regarding whether or not the mean for Dose 100 is equal to the mean for Dose 200. To make this determination, we need to include additional statements in the SAS code (e.g., LSMEANS).

5. At some point during the process of performing an analysis of variance, summary information regarding the treatment means needs to be examined. This information will be included in the SAS output as a result of an LSMEANS statement. In the caffeine example, this is on the last page of the SAS output and is reproduced here.

Dose	Taps LSMEAN	Standard Error	Pr > t	95% Confidence Limits	
100	246.400000	0.704746	<.0001	244.953981	247.846019
200	248.300000	0.704746	<.0001	246.853981	249.746019
0	244.800000	0.704746	<.0001	243.353981	246.246019

The values in this table correspond to the treatment means (as opposed to the results in the Parameter Estimates, which are primarily effects). Note that Dose 0 is listed last because this is the reference level. Also note that the standard errors round to 0.705, which is what we calculated “by hand”. The standard errors and the confidence limits are included in this table only because we included the options STDERR and CL on the LSMEANS statement. The p-values in this table (in the column “Pr > |t|”), are for testing whether or not each treatment mean is equal to 0. As mentioned earlier, these tests are usually not very interesting, since we typically want to compare the means of two treatments.

Section 4.4. Model Diagnostics

Every statistical procedure has underlying assumptions. If the assumptions are violated, then the statistical procedure is not valid. When the assumptions are violated, the p-values for all hypothesis tests could be incorrect and levels of confidence for all confidence intervals could also be incorrect. We can never “prove” that the assumptions are satisfied. Instead, we look for evidence that they are violated. If we do not find evidence that they are violated, then we assume they are satisfied.

In the previous section, we saw that analysis of variance models can be expressed either in terms of cell means or in terms of effects. For both of these two types of models, the assumptions are the same as for regression models. Specifically, we assume that the errors (1) are independent, (2) follow a normal distribution, (3) have mean 0, and (4) have constant variance, which we denote by σ^2 . To check these assumptions, we follow the same steps that we did with regression models. First, fit the model and get the diagnostic plots. Use the QQ plot to check the assumption of normality. If the points on this plot drastically depart from the diagonal line, then this is evidence that the assumption of normality has been violated. Use the residual plot to check for equal variances. If there is a drastic difference in the vertical spread of points, then this is evidence of nonconstant variance. When we are performing analysis of variance, the replication (i.e., many values for Y for each value of X) permits additional ways to check the assumptions. We can still use the graphs to make visual assessments, but it is more reliable to use official hypothesis tests.

4.4.1. Violation of independence

Violations of the assumption of independence can **severely** affect the conclusions of an analysis. To check this assumption, we must consider how the data were collected; there is nothing in the SAS output that can be used to evaluate independence. Intuitively, observations (and therefore the errors) are statistically independent if the information in one observation does not affect the other observations. There are several typical ways in which this assumption can be violated. For example, when multiple observations are taken from the same object over time, then the data are referred to as repeated measures. It is not appropriate to use analysis of variance techniques on repeated measures data. Another example is when the same object is measured more than once. This is usually done to improve the accuracy of the measurement, but it produces multiple observations (rows) in the data for the same object. This is called subsampling, because there should be only one row of data for each object in the sample. Subsampling also occurs when a group of objects receive the same treatment, but measurements are taken on individuals in the group. For example, a classroom of students all receive the same “new” lesson plan, but test scores are measured on individual students.

Example violation #1. Suppose that a medical researcher wishes to compare two medicines for reducing cholesterol. Each patient in the study has cholesterol measured every week. The observations over time on each patient would not be independent. For instance, a patient with higher than average cholesterol one week would likely be followed by a higher than average cholesterol level the next week. However, observations associated with one patient would be independent of observations associated with another patient because they are treated individually.

Example violation #2. Suppose a researcher randomly selects specimens of soil from a contaminated area. Each soil specimen is subdivided into three parts and a measurement of a pesticide in the soil is made on each of the three parts. Because the parts are taken from the same soil specimen, the 3 parts would not be independent. These parts are called subsamples. For instance, a higher than normal reading on one subsample would likely be associated with a higher than normal reading on another. In this case the original soil specimens are the experimental units, and there needs to be one response value in the data for each experimental unit. In a typical scenario, we would average the three subsample readings to obtain a single reading for each soil specimen.

Example violation #3. Suppose we are interested in how the oven temperature affects the quality of baked bread. We make 5 loaves and put them in the oven at the same time. The measurements on

these loaves might not be independent. For instance, if the loaves are baked a bit longer than called for, this would affect all of the loaves in the same way, so the observations on the loaves that were baked at the same time (in the same oven) would not be independent of one another.

Analyzing dependent data as if observations are independent can have serious consequences. In particular, p-values for hypothesis tests and the confidence levels for confidence intervals can be dramatically affected. It is beyond the scope of this text to discuss methods that may be applied when the independence assumption is violated. However, you should be aware of time-dependent data and subsampling so that you do not mistakenly apply the wrong methods to such data.

4.4.2. Violation of normality

Modest deviations from normality will have little effect on the p-values and confidence levels associated with analysis of variance. We say that ANOVA is robust to a minor violation of this assumption. If the distribution of the errors (as estimated by the residuals) is somewhat symmetric and mound-shaped, then we will assume the assumption of normality has not been violated. This can sometimes be difficult to determine via graphs. Since ANOVA datasets have replication (i.e., multiple observed Y values for each value of the predictor) there are numerous hypothesis tests that can be used to assess normality. These tests cannot be applied to regression datasets, since regression datasets typically do not have replication.

Tests for normality include Shapiro-Wilk, Kolmogorov-Smirnov, Cramèr-von Mises and Anderson-Darling. For all of these tests, the null hypothesis is that the residuals follow a normal distribution. All of these tests can be generated in SAS via PROC UNIVARIATE with the NORMAL option. The exact SAS code will be given later. For the caffeine data, the results of these tests are shown in Table 4.7. Since the p-values are all greater than 0.05, we do not reject the null hypothesis and conclude that the assumption of normal errors appears to be satisfied.

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.948843	Pr < W	0.1574
Kolmogorov-Smirnov	D	0.124262	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.059538	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.444496	Pr > A-Sq	>0.2500

Table 4.7. Normal tests for the caffeine example

4.4.3. Violation of equal variance

As with the assumption of normal errors, there are several formal hypothesis tests that can be used to assess equality of variance. These tests cannot be used in regression analysis because they require that the dataset has replication. Many of the formal tests for equality of variance are sensitive to violations of the normality assumption. One recommended test is the Brown-Forsythe test, which is a modification of a test first proposed by Levene. The modification makes the test more robust to violations of normality. The Brown-Forsythe test can be applied only when the ANOVA model contains one factor. In the next chapter, we will consider ANOVA models that contain two factors and the Brown-Forsythe test cannot be applied to those models. The null hypothesis for the Brown-Forsythe test is that all treatments have the same variance, so having a significant p-value (less than 0.05) indicates that the assumption of equal variances has been violated.

To get SAS to perform the Brown-Forsythe test, we need to include an additional statement in PROC GLM. The syntax for this statement is

```
MEANS treatment / HOVTEST = BF;
```

In this statement, HOVTEST is an abbreviation for homogeneity of variance test and BF is an abbreviation for Brown-Forsythe. For the caffeine example, we replace the word treatment with the actual treatment variable (Dose), and SAS will generate Table 4.8. Since the p-value (0.7565) is greater than 0.05, we do not reject the null hypothesis. We conclude that the variances are equal for all the treatments.

Brown and Forsythe's Test for Homogeneity of taps Variance ANOVA of Absolute Deviations from Group Medians					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
dose	2	0.8667	0.4333	0.28	0.7565
Error	27	41.5000	1.5370		

Table 4.8. Brown-Forysthe test for the caffeine example

4.4.4. SAS code for the caffeine example

Here is the complete code for performing analysis of variance on the caffeine data.

```
DATA caffeine;
INPUT Dose Taps @@;
DATALINES;
  0 242  0 245  0 244  0 248  0 247
  0 248  0 242  0 244  0 246  0 242
  100 248 100 246 100 245 100 247 100 248
  100 250 100 247 100 246 100 243 100 244
  200 246 200 248 200 250 200 252 200 248
  200 250 200 246 200 248 200 245 200 250
;
PROC GLM DATA=caffeine;
  CLASS Dose (REF='0');
  MODEL Taps = Dose / SOLUTION ;
  LSMEANS Dose / STDERR CL;
  MEANS Dose / HOVTEST=BF;           /* Brown-Forsythe test */
  OUTPUT OUT=myresults RESIDUAL=resids;
  RUN;
PROC UNIVARIATE DATA=myresults NORMAL;
  VAR resids;
  RUN;
```

Most of this code has already been discussed. Recall that the @@ symbol at the end of the INPUT statement tells SAS to read multiple (Dose, Taps) pairs of values on each line of data. The CLASS statement in PROC GLM creates the indicator variables and the REF='0' option forces SAS to use Dose 0 as the reference level. The LSMEANS statement generates the treatment means, and the options STDERR and CL will produce the standard errors and the confidence limits. The MEANS statement is new, and it is required solely to generate the Brown-Forsythe test. The rest of the output from the MEANS statement is also generated by the LSMEANS statement (so there will be some duplication in the SAS output). The OUTPUT statement is also new. This statement tells SAS to create a new dataset that will contain specific results from fitting the model. The name of this new dataset is myresults, which is defined by OUT= option. This dataset will contain the residuals from the fitted model, but in the new dataset the variable name for these residuals is resids. The only reason we create this new dataset is so that we can generate the tests for normality. These tests are generated via PROC UNIVARIATE with the NORMAL option. Note that PROC UNIVARIATE is using the new dataset (myresults) and we are testing the variable that contains the residuals (resids).

The relevant parts of the SAS output from this code have already been discussed. Next, we consider another example.

4.4.5. Example: Check assumptions for a different dataset

Consider an example dataset that contains observed values for 10 replications in each of five treatment groups (A, B, C, D and E). The data are shown in Figure 4.7.

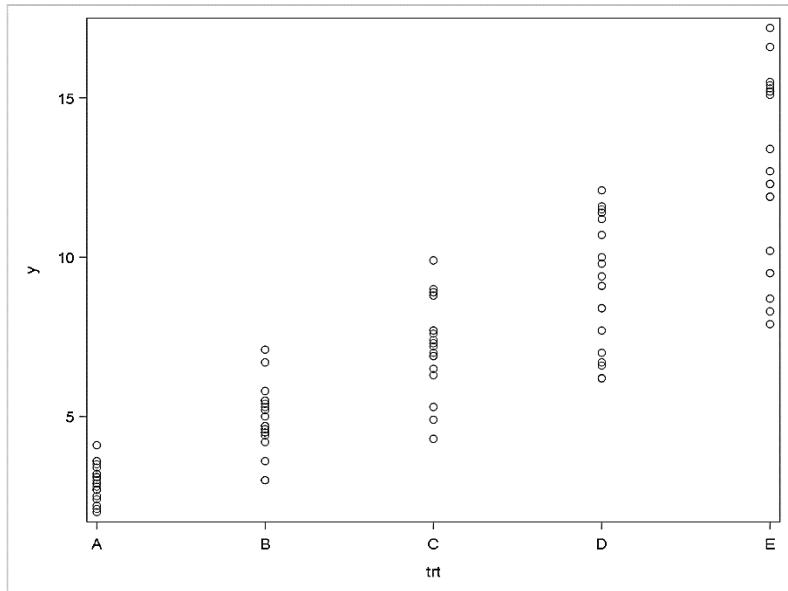


Figure 4.7. Scatterplot of simulated data

An ANOVA model was fit to the data, and the diagnostic plots are shown in Figure 4.8. The QQ plot (in the middle) looks acceptable even though the histogram of the residuals (on the right) shows a higher than expected peak for residuals near 0. The residual plot (on the left) shows a distinct wedge shape, and this provides convincing evidence that these five treatments do not have the same variance. The Brown-Forsythe test, shown in Table 4.9, further supports this conclusion. The p-value for this test is less than 0.0001, so we reject the hypothesis that the variances are equal.

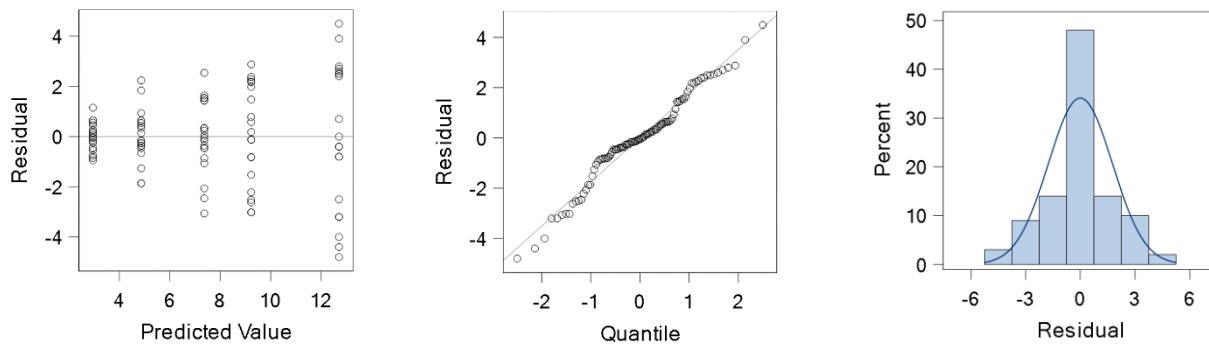


Figure 4.8. Diagnostic plots for simulated data

Brown and Forsythe's Test for Homogeneity of γ Variance ANOVA of Absolute Deviations from Group Medians					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
trt	4	51.4644	12.8661	13.63	<.0001
Error	95	89.6460	0.9436		

Table 4.9. Brown-Forsythe test for simulated data

To alleviate the difficulties with unequal variances, we need to transform the response variable. In most ANOVA applications, it is not appropriate to transform a predictor variable because these variables go into the model as indicator (0/1) variables. In general, performing transformations on indicator variables has no effect on the model. There are several transformations that we could apply to the response variable in an effort to stabilize the variance. It is customary to use $\log(Y)$ instead of Y as the response variable, and this often alleviates the problem. When we apply the logarithmic transformation to Y in our current example, we obtain the diagnostic plots in Figure 4.9. Note that the wedge pattern in the residual plot has completely disappeared, and there is no longer any evidence of nonconstant variance. This is confirmed by the Brown-Forsythe test (Table 4.10), which has p-value 0.6888. The QQ plot looks about the same as before, but the strong central peak in the histogram is no longer present. When taken together, the residual plots for the transformed data show no evidence that the assumptions have been violated.

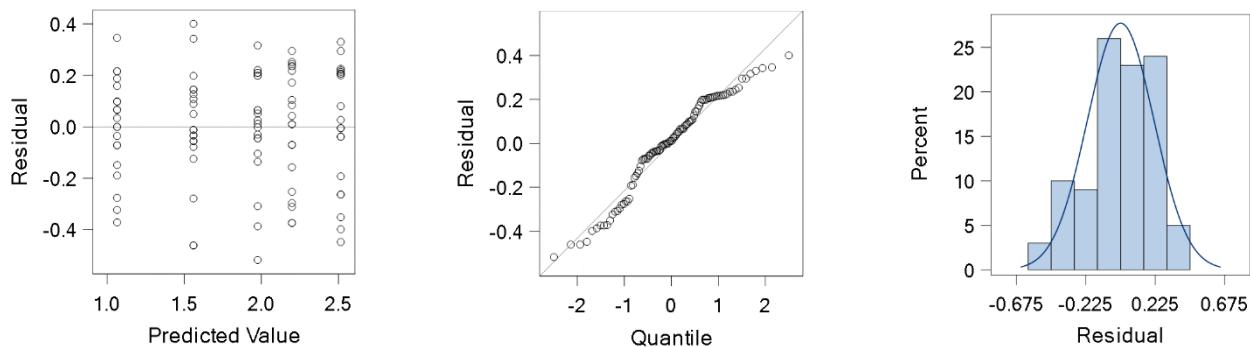


Figure 4.9. Diagnostic plots for the transformed data

Brown and Forsythe's Test for Homogeneity of logY Variance					
ANOVA of Absolute Deviations from Group Medians					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
trt	4	0.0402	0.0101	0.56	0.6888
Error	95	1.6914	0.0178		

Table 4.10. Brown-Forsythe test for the transformed data

4.4.6. Other transformations and back-transformations

In some datasets, it is entirely possible that using a logarithmic transformation of Y will not alleviate difficulties with the assumptions. Using $\log(Y)$ works well in many cases, and it is the first transformation that should be tried to eliminate nonconstant variances. This is why $\log(Y)$ is sometimes called the variance-stabilizing transformation. If using $\log(Y)$ does not work, then a different transformation could be tried. One possibility is to raise Y to some exponent. This approach offers great flexibility, since the exponent can be any real number. For example, using the exponent $\frac{1}{2}$ is the same as the square root transformation and using the exponent -1 is the same as the reciprocal transformation. (These transformations were discussed in more detail in Section 1.7.) Since there are an infinite number of possible exponents, it can be tricky to decide which exponent is most appropriate. The Box-Cox procedure can be used to choose an appropriate power. Details of the Box-Cox procedure are beyond the scope of this textbook, but additional information can be found online.

Whenever the values for Y are transformed, SAS will use the transformed values in all of its calculations. This forces the results that are generated by SAS to also be transformed. SAS does not recognize that a transformation has taken place, so it is incumbent on the statistician to properly interpret the output. This will involve back-transformations for any estimates that involve Y.

For example, suppose an experiment is conducted to compare the effectiveness of three different antibiotics used to treat ear infections. The response variable could be the number of weeks until the infection is cleared. Now suppose a log transformation is applied to the response variable. This transformation affects the values for Y, but it also affects the units that are associated with these values. The original Y values were measured in weeks, but $\log(Y)$ is measured in log-weeks. In this situation, it is inappropriate to present the results of the transformed model exactly as they appear in the SAS output. Any estimate involving Y (including both point estimates and interval estimates) need to be back-transformed so that the values will be in weeks instead of log-weeks. The exponential function e^y is the

inverse of the logarithmic function $\log(y)$, so a confidence interval such as (1.15, 2.13) that is measured in log-weeks needs to undergo a back-transformation to report the values in weeks. The appropriate confidence interval would be $(e^{1.15}, e^{2.13})$, or (3.16, 8.41) weeks.

A cautionary note: Please pay attention to the issues involved in transforming the Y values. If you report the confidence interval (1.15, 2.13), you are saying that you are 95% confident that the mean time to eliminate the infection is somewhere between 1.15 and 2.13. Everyone will assume these values are in weeks, because the original values were measured in weeks. The actual interval is between 3.16 and 8.41 weeks. There is a huge difference between these two intervals. It is a simple mistake, but it can have dire consequences. If you report the wrong interval, you could lose credibility with your co-workers and your supervisor.

4.4.7. SAS code for the simulated example

```

DATA simulated;
INPUT trt $ y @@;
DATALINES;
A 3.1 A 3.2 A 3.2 A 3.0 A 2.5 A 3.1 A 2.7 A 3.6 A 2.9 A 2.0
A 2.7 A 2.2 A 2.9 A 2.1 A 2.4 A 3.4 A 3.6 A 4.1 A 3.5 A 2.8
B 5.5 B 4.6 B 5.4 B 4.4 B 4.6 B 3.6 B 3.0 B 4.7 B 7.1 B 5.8
B 5.2 B 4.7 B 5.3 B 5.0 B 3.0 B 5.5 B 4.2 B 4.5 B 6.7 B 4.5
C 7.0 C 7.7 C 6.3 C 5.3 C 6.9 C 6.9 C 9.9 C 8.8 C 8.9 C 4.9
C 7.7 C 6.5 C 7.3 C 7.4 C 9.0 C 8.9 C 7.2 C 4.3 C 7.6 C 8.8
D 6.2 D 7.0 D 10.7 D 9.1 D 9.4 D 6.6 D 11.4 D 9.1 D 7.7 D 11.2
D 12.1 D 11.6 D 11.5 D 11.4 D 6.2 D 8.4 D 9.8 D 8.4 D 6.7 D 10.0
E 15.5 E 10.2 E 7.9 E 12.3 E 15.4 E 9.5 E 15.3 E 9.5 E 17.2 E 11.9
E 15.1 E 8.7 E 13.4 E 16.6 E 15.2 E 15.2 E 11.9 E 8.3 E 12.7 E 12.3
;
PROC SGPLOT data=simulated;
SCATTER X=trt Y=y;
RUN;
PROC GLM DATA=simulated PLOTS=DIAGNOSTICS;
CLASS trt;
MODEL y = trt;
MEANS trt / HOVTEST=BF;
RUN;
DATA two;
SET simulated;
logY = log(y);
RUN;
PROC GLM DATA=two PLOTS=DIAGNOSTICS;
CLASS trt;
MODEL logY = trt;
MEANS trt / HOVTEST=BF;
RUN;

```

Section 4.5. Multiple Comparisons

Before we jump into the topic of multiple comparisons, we need to understand how we arrive at this point in the analysis. If the overall ANOVA F test is not significant (i.e., if its p-value is greater than 0.05), then we would declare that there are no significant differences among the means. We could then report the estimated values for these means (either as point estimates or confidence intervals, back-transformed if necessary), but then the analysis would stop. We would never attempt to compare specific means, because the overall ANOVA F test tells us that the differences among these means is not significant.

If the overall ANOVA F test is significant, then we conclude that there is at least one significant difference among the treatment means. However, this test provides no information about which mean (or means) might be different. To explore this, we perform "post-hoc" tests to compare specific treatments that we are interested in. For some experiments, the research objective dictates the comparisons of interest, but for other experiments we might want to compare every treatment to every other treatment (i.e., all possible pairs). Regardless of the specific comparisons, any time we perform multiple tests with the same data we increase the likelihood that we will make a Type I error (reject H_0 when H_0 is true). To keep the Type I error rate at α , we must make some adjustments. There are numerous methods of adjustment, but we will consider these five:

- Fisher's Least Significant Difference (LSD)
- Tukey's Honest Significant Difference (HSD)
- Bonferroni's adjustment
- Scheffe's method
- Dunnett's method

These five methods are not hypothesis tests – they are methods for adjusting the p-values of tests. These methods are applicable only when are performing multiple hypothesis tests with the same data.

Everything that is in this section is based on two events: (1) the assumptions are not violated and (2) the overall ANOVA F test is significant. If either of these two things is not true, then we will never perform the procedures described in this section.

4.5.1. Which means are different?

We are assuming that the model assumptions are not violated and that the overall ANOVA F test is significant. The overall F test tells us that at least one treatment has a different mean, but it does not tell us which treatment(s) are different. To make this determination, we need to perform additional tests.

If we want to know whether or not the mean for treatment i is equal to the mean for treatment j , we would need to test the hypotheses

$$H_0 : \mu_i = \mu_j \text{ vs. } H_a : \mu_i \neq \mu_j \quad (4.10)$$

which can be written

$$H_0 : \mu_i - \mu_j = 0 \text{ vs. } H_a : \mu_i - \mu_j \neq 0 \quad (4.11)$$

If these were the only two treatments in the dataset, then we could perform an ordinary two-sample t test. The test statistic is

$$t = \frac{\bar{Y}_{i\bullet} - \bar{Y}_{j\bullet}}{s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \quad (4.12)$$

where s_p is the pooled standard deviation, defined as

$$s_p = \sqrt{\frac{(n_i - 1)s_i^2 + (n_j - 1)s_j^2}{n_i + n_j - 2}} \quad (4.13)$$

We would reject H_0 if the absolute value of the test statistic is greater than the critical value from the t distribution with degrees of freedom $n_i + n_j - 2$.

If there are $t = 3$ treatments, we would need to test three hypotheses

- trt 1 vs trt 2. $H_0 : \mu_1 = \mu_2$ vs. $H_a : \mu_1 \neq \mu_2$
- trt 1 vs trt 3. $H_0 : \mu_1 = \mu_3$ vs. $H_a : \mu_1 \neq \mu_3$
- trt 2 vs trt 3. $H_0 : \mu_2 = \mu_3$ vs. $H_a : \mu_2 \neq \mu_3$

If there are $t = 4$ treatments, we would need to test six hypotheses

- trt 1 vs trt 2. $H_0 : \mu_1 = \mu_2$ vs. $H_a : \mu_1 \neq \mu_2$
- trt 1 vs trt 3. $H_0 : \mu_1 = \mu_3$ vs. $H_a : \mu_1 \neq \mu_3$
- trt 1 vs trt 4. $H_0 : \mu_1 = \mu_4$ vs. $H_a : \mu_1 \neq \mu_4$
- trt 2 vs trt 3. $H_0 : \mu_2 = \mu_3$ vs. $H_a : \mu_2 \neq \mu_3$
- trt 2 vs trt 4. $H_0 : \mu_2 = \mu_4$ vs. $H_a : \mu_2 \neq \mu_4$
- trt 3 vs trt 4. $H_0 : \mu_3 = \mu_4$ vs. $H_a : \mu_3 \neq \mu_4$

Every time we perform a hypothesis test, the probability we incorrectly reject H_0 is α . This is the Type I error rate, which is usually set to $\alpha = 0.05$. If we perform 3 tests, the probability we incorrectly reject at least one null hypothesis is $1 - (0.95)^3 = 0.142$, so we would make at least one Type I error about 14% of the time. If we perform 6 tests, the probability we incorrectly reject at least one null hypothesis is $1 - (0.95)^6 = 0.265$ so we would make at least one Type I error about 27% of the time. This inflation of the Type I error rate is guaranteed to occur, and this is why we do not use ordinary t tests to compare multiple treatment means in an analysis of variance.

4.5.2. Example: Inflation of the Type I error rate

The simulated data shown in Table 4.11 were randomly generated from normally distributed populations that all have mean 100 and standard deviation 2. We should find no significant differences among the means.

Treatment	A	B	C	D	E
Responses	97	97	100	98	103
	99	101	101	100	106
	97	101	100	99	100
	100	103	99	101	100
	101	98	99	101	102
	101	103	98	99	100
	100	96	100	102	101
	102	102	98	101	100
Means	99.625	100.125	99.375	100.125	101.500
Variances	3.411	7.554	1.125	1.839	4.571

Table 4.11. Simulated data with 5 treatments

We used SAS to fit an ANOVA effects model to these data. The diagnostic plots looked fine, and the overall ANOVA F test produced the p-value 0.2354. As we expected, this indicates that all of the treatments have the same mean. If we erroneously ignore this test, and continue the analysis by comparing the treatment means via ordinary t tests, an interesting thing happens. This is summarized in Table 4.12.

Compare Treatments	p-value	Reject at $\alpha=.05?$	Correct Decision?
A to B	0.3379	No	✓
A to C	0.3724	No	✓
A to D	0.2735	No	✓
A to E	0.0408	Yes	✗
B to C	0.2417	No	✓
B to D	0.5000	No	✓
B to E	0.1414	No	✓
C to D	0.1191	No	✓
C to E	0.0123	Yes	✗
D to E	0.0734	No	✓

Table 4.12. Results of multiple t tests

According to the ordinary t tests, we would reject two of these ten tests. Specifically, we would decide that treatments A and E have different means, and that treatments C and E have different means. But this is simulated data, so we know that all of these treatments have the same population mean and we should not reject any of these hypotheses. Our error rate would be 2/10, or 0.20, and this is quadruple the desired error rate of 0.05.

4.5.3. Controlling the Type I error rate

To maintain the desired Type I error rate, we must make adjustments to our criteria for rejecting H_0 in the individual tests. Many methods for adjustment have been proposed, but we will consider these five:

- Fisher's Least Significant Difference (LSD)
- Tukey's Honest Significant Difference (HSD)
- Bonferroni's adjustment
- Scheffe's method
- Dunnett's method

Fisher's Least Significant Difference (LSD)

Sir Ronald A. Fisher was a British mathematician, statistician and geneticist, and made extraordinary contributions to numerous scientific fields in the first half of the 20th century. Fisher's LSD makes two changes to the ordinary t test. One change affects the degrees of freedom and the other change affects the pooled standard deviation. The ordinary t test uses degrees of freedom based on the sample sizes of the two groups that are being compared ($n_i + n_j - 2$), but Fisher's LSD uses the degrees of freedom for error (dfE) from the overall ANOVA, which incorporates the sample sizes for all the groups in the dataset, not just the ones currently being tested.. This changes affects the critical value from the t distribution. The second change involves the pooled standard deviation. The ordinary t test combines the variances of the two groups being tested, but Fisher's LSD use the root MSE from ANOVA, which uses the variances from all the treatment groups. These two changes make it a little bit harder to reject the individual tests, and this reduces the overall Type I error rate.

To define how Fisher's method works, let t^* represent the critical value for the t distribution with degrees of freedom dfE. The test statistic uses MSE instead of the pooled standard deviation, and we reject H_0 if

$$\frac{|\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}|}{\sqrt{MSE} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} > t^* \quad (4.14)$$

In other words, we declare that the mean for treatment i is significantly different than the mean for treatment j whenever

$$|\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}| > t^* \sqrt{MSE} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \quad (4.15)$$

The expression on the right side of the inequality is the smallest difference between means that is statistically significant, so it is called the least significant difference, or LSD. If the data are balanced (so that $n_i = n_j$), then Fisher's LSD becomes a constant that can be used to compare any pair of treatments in the dataset.

$$LSD = t^* \sqrt{MSE} \sqrt{\frac{2}{n}} \quad (4.16)$$

When we apply this technique to the simulated data, we have $dfE = 35$, $MSE = 3.7$ and $n_i = 8$ for every group. The critical value is $t^* = 2.030$, so Fisher's LSD is

$$LSD = (2.030) \sqrt{3.7} \sqrt{\frac{2}{8}} = 1.952 \quad (4.17)$$

To compare two treatments, we need only to look at the difference between their estimated means and compare it to 1.952, as shown in Table 4.13. The test comparing treatments A & E is no longer significant, and this reduces the Type I error rate to $1/10 = 0.10$. This is still higher than our desired rate of 0.05, but it is better than the 0.20 error rate that we had before making Fisher's adjustment.

Compare Treatments	Difference	Significant?	Correct Decision?
A to B	0.5	No	✓
A to C	0.25	No	✓
A to D	0.5	No	✓
A to E	1.875	No	✓
B to C	0.75	No	✓
B to D	0	No	✓
B to E	1.375	No	✓
C to D	0.75	No	✓
C to E	2.125	Yes	✗
D to E	1.375	No	✓

Table 4.13. Fisher's LSD for simulated data

Tukey's Honest Significant Difference (HSD)

The adjustment for multiple comparisons that has been proposed by American mathematician John Tukey does not involve the t distribution. Instead, it considers the distribution of the largest sample mean minus the smallest sample mean. This difference follows a Studentized Range distribution. The distribution has two parameters: the numerator degrees of freedom is the number of treatment groups and the denominator degrees of freedom is dfE.

Tukey's HSD is defined by

$$HSD = (q^*) \sqrt{\frac{MSE}{n}} \quad (4.18)$$

where q^* is the critical value from the Studentized Range distribution and n is the sample size for each group.

For the simulated data, $MSE = 3.7$ and $n = 8$. There are 5 treatment groups and $dfE = 35$, so $q^* = 4.066$ and $HSD = (4.066) \sqrt{\frac{3.7}{8}} = 2.765$. From Table 4.13, the largest difference in the sample means is 2.125 (between treatments C and E). This difference is not significant, so none of the differences are significant. Using Tukey's method, we would have decided correctly (we would not reject) for every test.

Note: It is possible to use Tukey's adjustment when the sample sizes are not equal, but it requires another modification and the resulting procedure is called Tukey-Kramer. We will let SAS perform these calculations, and SAS will automatically make the necessary adjustments in the event the sample sizes are not equal.

Bonferroni's adjustment

The adjustment proposed by Italian mathematician Carlo Bonferroni does not adjust the test statistic or the reference distribution for these tests. Instead, it adjusts either the p-values or the significance level. In order to use Bonferroni's adjustment, we must know in advance the number of tests we are going to perform. If the number of tests is k , then the significance level for each test is reduced to α/k . Alternatively, the p-value for each test can be multiplied by k and then compared to α . Either of these approaches ensure that the overall significance level (for all the tests) is equal to α . For the simulated data, we are performing 10 tests, so the significance level for each test would become $0.05/10 = 0.005$.

Using the p-values in Table 4.12, we see none of them are smaller than 0.005, so we would not reject any of these tests. As with Tukey's adjustment, Bonferroni's adjustment would produce the correct decision on every test.

When Bonferroni's method is performed in SAS, the p-values are adjusted so that we will compare each reported p-value to our customary significance level ($\alpha = 0.05$).

Scheffè's method

The adjustment proposed by American statistician Henry Scheffè does not involve the t distribution at all. Instead, it is a modification of the overall ANOVA F test. The critical value for this method uses the original significance level $\alpha = 0.05$. The critical value is denoted by F^* and it from the F distribution with numerator degrees of freedom $t - 1$ and denominator degrees of freedom dfE .

Scheffè's least significant difference is defined by

$$\sqrt{(t-1)(F^*)(MSE)\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \quad (4.19)$$

If the difference of two treatment means is larger than this value, then the difference is declared to be statistically significant.

For the simulated data, Scheffè's least significant difference is

$$\sqrt{(5-1)(2.64)(3.7)\left(\frac{1}{8} + \frac{1}{8}\right)} = 3.13$$

The largest difference in treatment means is 2.125 (between treatments C and E). The largest difference is not significant, so none of the differences are significant. With Scheffè's method, we would have decided correctly for every test.

Dunnett's method

The method developed by Canadian statistician Charles Dunnett can only be used when one of the treatments is a control treatment and we want to compare each of the other treatments to the control. Dunnett's method is not appropriate if we want to compare all possible pairs of treatments. Dunnett's method requires a special distribution called (appropriately enough) Dunnett's distribution. There are

two parameters for this distribution: the number of treatments (including the control) and the degrees of freedom for error (in the ANOVA table). Two treatments are significantly different if the difference between their sample means is greater than

$$(d^*) \sqrt{\frac{2 \cdot MSE}{(\# \text{replications})}} \quad (4.20)$$

where d^* is the critical value from Dunnett's distribution.

This method is not appropriate for the simulated data because it does not have a control treatment.

4.5.4. Comparison of methods

Bonferroni's method is the simplest to apply, but it can be overly strict. In other words, this method can fail to find important differences among the treatment means. This is particularly true if the number of tests is more than a few. Suppose, for example, that there are 20 tests. Then Bonferroni's method would require that a p-value be less than $0.05/20 = 0.0025$ in order to be significant. Such a low threshold may be difficult to reach, even if there is a significance difference between two means.

Sir R. A. Fisher was one of the first statisticians to recognize the problem of inflated Type I error rates, and his method was one of the first potential remedies. Since then, many other methods have been proposed that improve his method. To date, there is no general consensus regarding which method is "best", so Fisher's method remains viable. This method is automatically incorporated into PROC GLM, unless another method is specified.

Tukey's method is preferred when we are testing all pairs of means.

Dunnett's method is only used when we are comparing each treatment to a control treatment.

Scheffè's method is primarily for a collection of tests that are suggested by the data. This occurs when initial results from an analysis are used to uncover previously unknown patterns in the data, and formal hypothesis tests are used to determine if the patterns are significant. This is sometimes called 'data snooping', and it is generally not a good thing to do because it is likely to lead to spurious results (results that cannot be duplicated in another experiment). The likelihood of obtaining spurious results can be greatly reduced by using Scheffè's method.

4.5.5. SAS code for multiple comparisons

To compare treatment means in SAS, we use PROC GLM with either a MEANS statement or an LSMEANS statement. The example SAS code, shown below, contains numerous MEANS and LSMEANS statements to illustrate all the adjustment methods described in the this section. **It is not appropriate to use all of these methods in a single data analysis.** The method of adjustment should be chosen based on the merits of the method and type of analysis that is being performed.

```
DATA comps;
  INPUT trt $ value @@;
  DATALINES;
A 97 A 99 A 97 A 100 A 101 A 101 A 100 A 102
B 97 B 101 B 101 B 103 B 98 B 103 B 96 B 102
C 100 C 101 C 100 C 99 C 99 C 98 C 100 C 98
D 98 D 100 D 99 D 101 D 101 D 99 D 102 D 101
E 103 E 106 E 100 E 100 E 102 E 100 E 101 E 100
;

PROC GLM DATA=comps PLOTS=DIAGNOSTICS;
  CLASS trt;
  MODEL value = trt ;
  LSMEANS trt / PDIFF;           * equivalent to Fisher's LSD;
  LSMEANS trt / PDIFF ADJUST=TUKEY;
  LSMEANS trt / PDIFF ADJUST=BON;
  LSMEANS trt / PDIFF ADJUST=SCHEFFE;
  LSMEANS trt / PDIFF=CONTROL('A') ADJUST=DUNNETT;
  MEANS trt / HOVTEST=BF LSD LINES;
  MEANS trt / TUKEY LINES;
  MEANS trt / TUKEY LINESTABLE;
  RUN;
```

Many of these statements contain the option PDIFF. This is an abbreviation for pairwise differences, and it instructs SAS to not only compute the treatment means but also conduct the hypothesis tests to compare every possible pair of means. The ADJUST= option defines the adjustment method for multiple comparisons.

The initial part of the SAS output contains the ANOVA table additional tables and the diagnostic plots. The interpretation of this part of the output has already been discussed, and will not be repeated here. It is important to note that the assumptions do not appear to be violated, so we can interpret the overall F test. The p-value for this test is 0.2354, so we conclude there are no significant differences among the treatment means. If we were doing this analysis on a “real” dataset, the analysis would stop here. Since we are using this example to explore the various methods of adjustment for multiple comparisons, we will continue interpreting the output.

We now consider the output from the first LSMEANS statement : LSMEANS trt / PDIFF;

The first table is automatically generated by the LSMEANS statement. The second table is the result of the PDIFF option and it contains the p-values for comparing all pairs of means. This LSMEANS statement also generates two graphs that are not shown here.

The first table provides the estimated mean for each treatment, and it also has an LSMEAN Number. This number is needed to interpret the second table. For example, to compare the means for treatments C and E, find the LSMEAN numbers for these two treatments in the top table. These numbers are 3 and 5. The p-value for testing these two means is in the second table, at the intersection of i = 3 and j = 5, which is 0.0338. The p-values in the second table have been adjusted via Fisher's method (because this is the adjustment method unless another method has been specified). There is nothing in the SAS output that tells you Fisher's method of adjustment has been applied for these p-values.

The GLM Procedure
Least Squares Means

trt	value LSMEAN	LSMEAN Number
A	99.625000	1
B	100.125000	2
C	99.375000	3
D	100.125000	4
E	101.500000	5

Least Squares Means for effect trt Pr > t for H0: LSMean(i)=LSMean(j)					
Dependent Variable: value					
i/j	1	2	3	4	5
1		0.6064	0.7964	0.6064	0.0593
2	0.6064		0.4407	1.0000	0.1617
3	0.7964	0.4407		0.4407	0.0338
4	0.6064	1.0000	0.4407		0.1617
5	0.0593	0.1617	0.0338	0.1617	

This is the output from the statement: LSMEANS trt / PDIFF ADJUST=TUKEY;

The first table is the same as before, but the p-values in the second table have been adjusted according to Tukey's method. The fact that Tukey's method has been applied is reported in the header of this page in the output, but it is not in the tables themselves. This LSMEANS statement also produces two graphs that are not shown here.

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey

trt	value LSMEAN	LSMEAN Number
A	99.625000	1
B	100.125000	2
C	99.375000	3
D	100.125000	4
E	101.500000	5

Least Squares Means for effect trt Pr > t for H0: LSMean(i)=LSMean(j)					
Dependent Variable: value					
i/j	1	2	3	4	5
1		0.9848	0.9989	0.9848	0.3113
2	0.9848		0.9348	1.0000	0.6133
3	0.9989	0.9348		0.9348	0.2000
4	0.9848	1.0000	0.9348		0.6133
5	0.3113	0.6133	0.2000	0.6133	

At the risk of being redundant, the statement LSMEANS trt / PDIFF ADJUST=BON; also generates two tables and two graphs. The first table has not changed, and the second table contains p-values that have been adjusted via Bonferroni's method. This statement also produces two graphs that are not shown here. A test is significant if the p-value in the table is less than 0.05.

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Bonferroni

trt	value LSMEAN	LSMEAN Number
A	99.625000	1
B	100.125000	2
C	99.375000	3
D	100.125000	4
E	101.500000	5

Least Squares Means for effect trt Pr > t for H0: LSMean(i)=LSMean(j)					
Dependent Variable: value					
i/j	1	2	3	4	5
1		1.0000	1.0000	1.0000	0.5928
2	1.0000		1.0000	1.0000	1.0000
3	1.0000	1.0000		1.0000	0.3378
4	1.0000	1.0000	1.0000		1.0000
5	0.5928	1.0000	0.3378	1.0000	

Here are the two tables generated by the statement `LSMEANS trt / PDIFF ADJUST=SCHEFFE;`
The first table has not changed, and the second table contains p-values that have been adjusted via Scheffe's method. This statement also produces two graphs that are not shown here.

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Scheffe

trt	value LSMEAN	LSMEAN Number
A	99.625000	1
B	100.125000	2
C	99.375000	3
D	100.125000	4
E	101.500000	5

Least Squares Means for effect trt Pr > t for H0: LSMean(i)=LSMean(j)					
Dependent Variable: value					
i/j	1	2	3	4	5
1		0.9913	0.9994	0.9913	0.4468
2	0.9913		0.9608	1.0000	0.7280
3	0.9994	0.9608		0.9608	0.3198
4	0.9913	1.0000	0.9608		0.7280
5	0.4468	0.7280	0.3198	0.7280	

Below is the output for the statement `LSMEANS trt / PDIFF=CONTROL ('A') ADJUST=DUNNETT;`
This looks different than the others because this is comparing treatment A to each of the other treatments. The p-values should be compared to 0.05. The output for this statement also includes two graphs that are not shown here.

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Dunnett

trt	value LSMEAN	H0:LSMean=Control
		Pr > t
A	99.625000	
B	100.125000	0.9587
C	99.375000	0.9968
D	100.125000	0.9587
E	101.500000	0.1781

The next section of output is a result of the statement `MEANS trt / HOVTEST=BF LSD LINES;`
The first table provides the results of the Brown-Forsythe test (option HOVTEST=BF), but this is not shown here. There is also a graph containing side-by-side boxplots, but it is not shown here. This

statement also generates the table and graph shown below. The table is the result of the LSD option and the graph is generated via the LINES option.

For our dataset, the value for Fisher's LSD is 1.9525, which matches 1.952 that we calculate by hand (equation (4.17)). In the graph, the treatments are listed in order of their estimated means and the brightly bold lines show which treatments have statistically similar means. This is a visual way to interpret the table of p-values that were generated via the PDIFF option without specifying the method of adjustment. The graph indicates that treatments A, B, C and D have similar means and that treatments A, B, D and E have similar means. This is another way of saying the A and E have different means, but the other treatments have similar means.

The GLM Procedure

t Tests (LSD) for value

Alpha	0.05
Error Degrees of Freedom	35
Error Mean Square	3.7
Critical Value of t	2.03011
Least Significant Difference	1.9525

Fisher's LSD

value t Grouping for Means of trt (Alpha = 0.05)

Means covered by the same bar are not significantly different.

trt Estimate

E	101.50
B	100.13
D	100.13
A	99.6250
C	99.3750

This graph is part of the output generated by the statement MEANS trt / TUKEY LINES;

This output has the same format as the options LSD LINES, except that Tukey's adjustment is used (instead of Fisher's), and this has changed the result. When Tukey's adjustment is applied, all of the treatments have similar means, as indicated by the single vertical line in the graph.

The GLM Procedure

Tukey's Studentized Range (HSD) Test for value

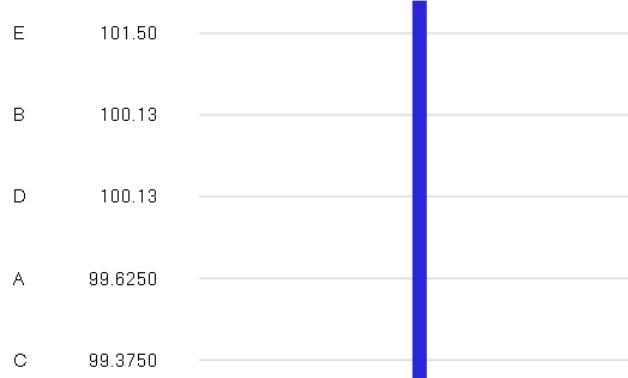
Alpha	0.05
Error Degrees of Freedom	35
Error Mean Square	3.7
Critical Value of Studentized Range	4.06586
Minimum Significant Difference	2.7651

Tukey's HSD

value Tukey Grouping for Means of trt (Alpha = 0.05)

Means covered by the same bar are not significantly different.

trt Estimate



The last statement in the SAS code is MEANS trt / TUKEY LINESTABLE;

This statement generates the same output as the previous MEANS statement, but the LINESTABLE option generates one more table, which is shown below. The information in this table is exactly the same that is presented in the graph on the previous page. The column labeled “Tukey Grouping” contains all A's, and this indicates that all the treatments have statistically similar means. If any of the means had been different, there would be some B's (or even C's) in this column. Any treatment means that would have the letter B is significantly different from any mean that does not have the letter B.

Means with the same letter are not significantly different.			
Tukey Grouping	Mean	N	trt
A	101.5000	8	E
A			
A	100.1250	8	B
A			
A	100.1250	8	D
A			
A	99.6250	8	A
A			
A	99.3750	8	C

This concludes the fairly exhaustive examples involving adjustments for multiple comparisons. It should be emphasized that all of these techniques should not be applied to every dataset.

- If you want to compare all pairs of means, use Tukey's adjustment.
- If you want to compare each treatment to a control treatment, use Dunnett's adjustment.
- If you are not concerned about potentially missing significant differences, you can use Bonferroni's adjustment.
- If you look at preliminary results from the data and you use this information to decide which tests to perform next, then use Scheffe's adjustment.
- If none of these conditions apply to your current situation, you can use Fisher's adjustment.

Section 4.6. Contrasts

In the previous section, we looked at hypothesis tests that compare one treatment mean to another treatment mean. While this is a customary and important part of post-hoc testing in analysis of variance, it is often the case that more complicated comparisons need to be made. These additional tests might involve three or more treatment means, and they are structured to be what statisticians call contrasts. A contrast is simply a linear combination of the treatment means, which has the form

$$\text{contrast} = c_1\mu_1 + c_2\mu_2 + \dots + c_t\mu_t \quad (4.21)$$

The coefficients (the c 's) must sum to 0, and some of the coefficients may equal 0.

For example, suppose there are 3 treatments (with population means μ_1 , μ_2 and μ_3) and we want to compare the mean of treatment 1 to the mean of treatment 2. The null hypothesis is $\mu_1 = \mu_2$, which can be written $\mu_1 - \mu_2 = 0$. Since this test involves exactly two means, it could be performed via the methods in the previous section. But this is also a contrast because the equation can be written $(1)\mu_1 + (-1)\mu_2 + (0)\mu_3 = 0$, so that $c_1 = 1$, $c_2 = -1$, and $c_3 = 0$. Since the coefficients sum to 0, this is a contrast.

As another example, suppose there are 3 treatments and we want to compare the mean of treatment 1 to the average of treatments 2 and 3. The equation for the null hypothesis would be

$\mu_1 - \frac{1}{2}(\mu_2 + \mu_3) = 0$ which could be written $(1)\mu_1 + (-0.5)\mu_2 + (-0.5)\mu_3 = 0$. This is a contrast because

the coefficients sum to 0. A test involving this contrast might be appropriate if treatment 1 was a control treatment and treatments 2 and 3 were new treatments. Contrasts are often used when the treatments are comprised of more than one factor. This is illustrated in the following example.

4.6.1. Contrast coefficients and tests

Consider the following scenario. An experiment was conducted to determine the effects of two types of preservatives (A, B) in different amounts (100, 400) on prevention of bacteria growth. A control treatment was also included. The five treatments are

1. (A, 100)
2. (A, 400)
3. (B, 100)
4. (B, 400)
5. Control (no preservative)

Let $\mu_1, \mu_2, \mu_3, \mu_4$, and μ_5 be the population means for these five treatments. Note that these treatments (and their means) are listed in alphabetic/numeric order, since A100 comes before A200, which comes before B100, etc. All the contrast coefficients must follow this order.

Here are some contrasts that may be of interest, along with their coefficients

- **(A, 100) vs. Control**

$$H_0 : \mu_1 = \mu_5 \text{ which can be written } H_0 : (1)\mu_1 + (0)\mu_2 + (0)\mu_3 + (0)\mu_4 + (-1)\mu_5 = 0.$$

The coefficients are 1, 0, 0, 0 and -1.

- **Average of A's vs. Control**

$$H_0 : \frac{\mu_1 + \mu_2}{2} = \mu_5, \text{ which can be written } H_0 : (0.5)\mu_1 + (0.5)\mu_2 + (0)\mu_3 + (0)\mu_4 + (-1)\mu_5 = 0$$

The coefficients are 0.5, 0.5, 0, 0 and -1.

It would be easier to clear the fractions (multiply by 2) and use the coefficients 1, 1, 0, 0 and -2.

- **Average of 100's vs. average of 400's**

$$H_0 : \frac{\mu_1 + \mu_3}{2} = \frac{\mu_2 + \mu_4}{2}, \text{ which can be written } H_0 : \left(\frac{1}{2}\right)\mu_1 + \left(-\frac{1}{2}\right)\mu_2 + \left(\frac{1}{2}\right)\mu_3 + \left(-\frac{1}{2}\right)\mu_4 + (0)\mu_5 = 0$$

The coefficients are 0.5, -0.5, 0.5, -0.5 and 0.

It would be easier to clear the fractions and use the coefficients 1, -1, 1, -1 and 0.

- **Average of treatments vs. Control**

$$H_0 : \frac{\mu_1 + \mu_2 + \mu_3 + \mu_4}{4} = \mu_5, \text{ which can be written } H_0 : \left(\frac{1}{4}\right)\mu_1 + \left(\frac{1}{4}\right)\mu_2 + \left(\frac{1}{4}\right)\mu_3 + \left(\frac{1}{4}\right)\mu_4 + (-1)\mu_5$$

The coefficients are 0.25, 0.25, 0.25, 0.25 and -1.

It would be easier to clear the fractions and use the coefficients 1, 1, 1, 1 and -4.

These coefficients are summarized in Table 4.14.

Contrast	Coefficients				
	c_1	c_2	c_3	c_4	c_5
(A, 100) vs. Control	1	0	0	0	-1
Average of A's vs. Control	0.5	0.5	0	0	-1
	or	1	1	0	-2
Average of 100's vs. average of 400's	0.5	-0.5	0.5	-0.5	0
	or	1	-1	1	-1
Average of treatments vs Control	0.25	0.25	0.25	0.25	-1
	or	1	1	1	-4

Table 4.14. Examples of contrast coefficients

The contrast coefficients cannot be rounded. They must sum to 0 exactly. If (for some unknown reason) you wanted to compare the average of the first three treatments to the control treatment, the coefficients would be $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, -1)$. The fractions cannot be rounded to 0.33. Instead, the fractions should be cleared. This is done by multiplying every coefficient by 3, so we would use the coefficients $(1, 1, 1, 0, -3)$. For any contrast, the coefficients can be multiplied by -1 , and the results of the hypothesis test will remain unchanged.

Hypothesis tests involving contrasts always take the form

$$H_0 : \text{population contrast} = 0$$

$$H_a : \text{population contrast} \neq 0$$

The test statistic is $t = \frac{\text{estimated contrast}}{\text{standard error of estimated contrast}}$.

The test statistic follows a t distribution with degrees of freedom dfE.

For the population contrast $c_1\mu_1 + c_2\mu_2 + \dots + c_t\mu_t$,

- the estimated contrast is $c_1\bar{Y}_1 + c_2\bar{Y}_2 + \dots + c_t\bar{Y}_t$.
- the standard error is $SE = \sqrt{MSE} \sqrt{\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \dots + \frac{c_t^2}{n_t}}$

We reject H_0 if the absolute value of the test statistic is greater than the critical value.

Example 4.6.1

Suppose there are 3 treatments with observed values as shown in the table. (Note that the last value for treatment 2 is missing.) The ANOVA table is also shown.

Treatment	1	2	3
Data	4	1	2
	5	2	3
	5	3	4
	8		5
Means	5.5	2.0	3.5

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	21.64	10.82	5.41	0.0327
Error	8	16.00	2.00		
Corrected Total	10	37.64			

We want to test whether the mean of treatment 1 is significantly different than the average of treatments 2 and 3, so the population contrast is $(1)\mu_1 + (-0.5)\mu_2 + (-0.5)\mu_3$.

The estimated contrast is $(1)5.5 + (-0.5)2.0 + (-0.5)3.5 = 2.75$

$$\text{Standard error} = \sqrt{2.00} \sqrt{\frac{1^2}{4} + \frac{(-0.5)^2}{3} + \frac{(-0.5)^2}{4}} = 0.89$$

$$\text{Test statistic} = t = \frac{2.75}{0.89} = 3.09$$

The critical value is 2.306 (from the t table with $df = dfE = 8$ and $\alpha/2 = 0.025$).

Since the test statistic is greater than the critical value ($3.09 > 2.306$), we reject H_0 and conclude that the mean for treatment 1 is significantly different from the average of the means for treatments 2 and 3.

4.6.2. Linear and quadratic trends

There is a special group of contrasts that can be used only when the treatments are numeric. These contrasts test for linear and/or quadratic trends in the means. It is also possible to test for higher-order trends (e.g., cubic or quartic), but we will consider only linear and quadratic trends. The concept is similar to regression analysis, but the tests are conducted within the framework of analysis of variance.

Examples in which trend analysis would be appropriate include

- Wooden beams are tested at different amount of pressure (100 psi, 200 psi, 300 psi, 400 psi) to see how strong they are.
- Different amounts of nitrogen (0%, 5%, 10%) are applied to plots of wheat to see how this affects yields.
- Frozen food products are stored at different temperatures (-5°F, -10°F, -15°F) to see how long they last before spoiling.

To explore the concept of trends, suppose there are five treatments: 0, 5, 10, 15 and 20. Since the treatments are numeric, we can plot them against their means. Then we look for (and test for) patterns in the means. The graphs in Figure 4.10 illustrate some possible trends.

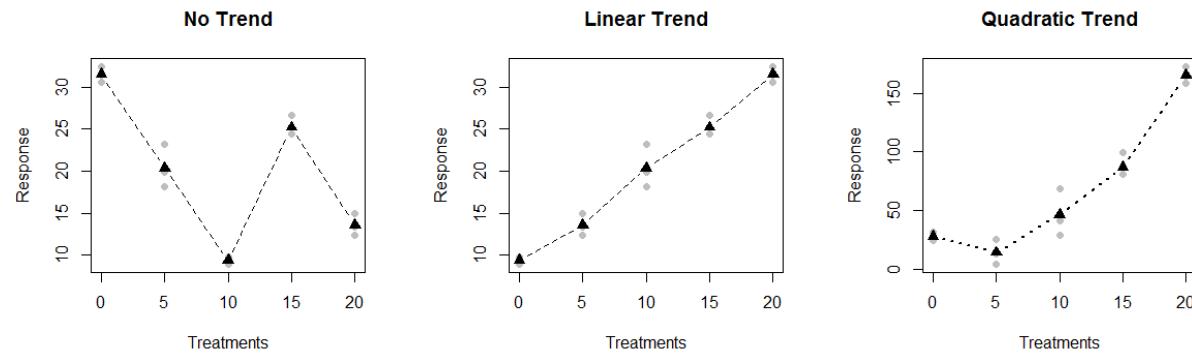


Figure 4.10. Linear and quadratic trends

Notice that the middle graph in Figure 4.10 shows a linear trend in the means. This is similar to regression analysis, but it is a little more flexible than regression. In regression, we would model the means as a perfectly straight line, but a linear contrast in an ANOVA setting allows for a little bit of variation.

We can use contrasts to test for trends in the treatment means, but only when the treatments are numeric. The treatments do not have to be evenly spaced, but the computations are simpler if they are evenly spaced. We will consider trends only when the treatments are evenly spaced. If there are three treatments, then we can test for a linear trend and we can test for a quadratic trend. If there are four treatments, we can test for linear, quadratic and cubic trends. Each additional treatment permits the inclusion of a higher-order trend. We will focus on testing linear and quadratic trends, regardless of the number of treatments.

When there are three treatments, the contrast coefficients are

- for linear trends: $-1, 0, 1$
- for quadratic trends: $1, -2, 1$

When there are four treatments, the contrast coefficients are

- for linear trends: $-3, -1, 1, 3$
- for quadratic trends: $1, -1, -1, 1$

Example 4.6.2

Beans were planted in 3 different densities: 10 plants per plot, 20 plants per plot, and 30 plants per plot. Yield was measured on each plot. The data and graph below show that the yield increases linearly with density. We will test to see if this trend is significant.

Density	10	20	30
	11.9	16.1	20.3
Yield	11.1	15.0	18.6
	13.9	16.9	17.5
Means	12.3	16.0	18.8

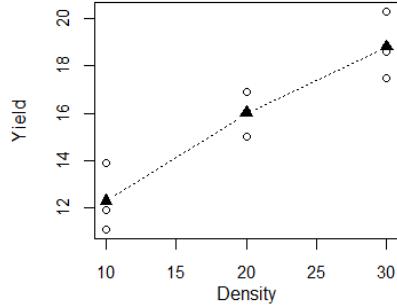


Figure 4.11. Data and graph for bean yields

We use SAS to get the ANOVA table. In particular, we need the MSE and dfE. These are $MSE = 1.66$ and $dfE = 6$.

The treatment means are 12.3, 16.0, and 18.8. For three treatments, the coefficients for linear contrast are $-1, 0$ and 1 , so the estimated contrast is $(-1)(12.3) + (0)(16.0) + (1)(18.8) = 6.5$.

$$\text{The standard error of the estimate is } SE = \sqrt{\text{MSE}} \sqrt{\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \frac{c_3^2}{n_3}} = \sqrt{1.66} \sqrt{\frac{(-1)^2}{3} + \frac{0}{3} + \frac{1^2}{3}} = 1.052 .$$

The test statistic is $t = 6.5 / 1.052 = 6.18$.

To get the critical value, we use the t table with $df = dfE = 6$ and $\alpha/2 = 0.025$. The critical value is 2.447.

Since the test statistic is greater than the critical value ($6.18 > 2.447$), we reject H_0 and conclude that there is a significant linear trend.

A closer inspection of the graph in Figure 4.11 reveals a slight curvature in the pattern for the means. This indicates that there might be a quadratic (X^2) pattern in the means. To test this, we will use a quadratic trend. The treatment means, MSE and dfE remain the same as with a linear trend, but the contrast coefficients are now 1, -2 and 1.

The estimated contrast is $(1)(12.3) + (-2)(16.0) + (1)(18.8) = -0.9$.

$$\text{The standard error is } SE = \sqrt{1.66} \sqrt{\frac{1^2}{3} + \frac{(-2)^2}{3} + \frac{1^2}{3}} = 1.822$$

The test statistic is $t = -0.9 / 1.822 = -0.49$.

The critical value does not change (it is still 2.477).

Since the absolute value of the test statistic is not greater than the critical value (0.49 is not greater than 2.477), we do not reject H_0 . At $\alpha = 0.05$, we conclude that there is not a significant quadratic trend.

4.6.3. Other polynomial contrasts

Contrasts can be used to compare treatment means, regardless of whether the treatments are quantitative or qualitative. The specific contrasts that are used for trends can be applied ONLY when the treatments are numeric. We have examined only linear and quadratic trends, with treatment levels that are evenly spaced. The numeric treatments are not required to be evenly spaced, but if they are not evenly spaced the contrast coefficients will be different. If there are more than 3 treatments, then higher-degree contrasts (e.g., cubic, 4th degree, etc.) can be tested. Table 4.15 provides the contrast coefficients for linear and quadratic trends for up to 8 treatments, assuming the treatments are evenly spaced.

Number of Treatments	Contrast	Coefficients						
		-1	0	1				
3	linear	-1	0	1				
	quadratic	1	-2	1				
4	linear	-3	-1	1	3			
	quadratic	1	-1	-1	1			
5	linear	-2	-1	0	1	2		
	quadratic	2	-1	-2	-1	2		
6	linear	-5	-3	-1	1	3	5	
	quadratic	5	-1	-4	-4	-1	5	
7	linear	-3	-2	-1	0	1	2	3
	quadratic	5	0	-3	-4	-3	0	5
8	linear	-7	-5	-3	-1	1	3	5
	quadratic	7	1	-3	-5	-5	-3	1

Table 4.15 Coefficients for linear and quadratic trends

When the treatments are not evenly spaced, the coefficients for linear and quadratic trends depend on the levels of the treatments and their sample sizes. These are not calculated by hand -- a SAS program is given below. In the program, the variable 'a' contains the levels of the treatments and the variable 'b' contains the sample size for each treatment. ORPOL is a built-in SAS function to calculate orthogonal polynomials. For example, the following code generates the coefficients for linear and quadratic trends for data in which the quantitative treatments are 10, 20, 40 and 80 (note that these are not evenly spaced) and the sample size is 5 for each treatment.

```
PROC IML;
a = {10 20 40 80};
b = {5 5 5 5};
coeff = ORPOL(a,2,b);
print a;
print b;
print coeff;
run;
quit;
```

a			
10	20	40	80

b			
5	5	5	5

coeff		
0.2236068	-0.229366	0.2368565
0.2236068	-0.14596	-0.047371
0.2236068	0.0208514	-0.343442
0.2236068	0.3544745	0.1539567

The first column of the coefficients are constants that can be ignored. The second column has the coefficients for the linear trend (-0.229366, -0.14596, ...) . The third column has the coefficients for the quadratic trend (0.2368565, -0.047371, ...). These coefficients need to be rounded so they sum to 0.

4.6.4. SAS code for contrasts

To compute contrasts and test for significance of a contrast, we add a CONTRAST statement to PROC GLM. We can provide a descriptive name for each contrast, to make it easier to find the results in the output. In the CONTRAST statement, the order of the coefficients must match the order that SAS is using. Refer to the Class Level Information table at the beginning of the PROC GLM output to verify the order.

Example 4.6.3

Using the bean data as given in Example 2, we can use CONTRAST statements to test for linear and quadratic trends. The code is shown below. Most of this code should be fairly routine by now, but the new statements (the CONTRAST statements) are in bold.

Consider the components of the linear CONTRAST statement: `CONTRAST 'Linear' Density -1 0 1;`

- The keyword CONTRAST asks that a contrast be computed.
- The name ‘Linear’ (in quotes) gives a name to the contrast. The name can be anything, but it is best to provide a descriptive name. This name will be printed in the SAS output, so the results on the contrast will be easy to find.
- The next word (Density) is the name of quantitative predictor variable (the one that defines the treatments). It must be the same as one of the terms on the right side of the MODEL statement.
- The numbers (-1 0 1) are the coefficients of the the contrast. In this case, these are the coefficients for computing the linear contrast involving 3 equally spaced treatments. Different coefficients will produce different contrasts (so remember to change the name of the contrast whenever the coefficients change).
- Don’t forget the semicolon at the end of the statement.

The SAS statement for the quadratic contrast is constructed in a similar way.

```
DATA beans;
INPUT Density Yield @@;
DATALINES;
10 11.9 10 11.1 10 13.9
20 16.1 20 15.0 20 16.9
30 20.3 30 18.6 30 17.5
;
PROC GLM DATA=beans PLOTS=DIAGNOSTICS;
CLASS Density;
MODEL Yield = Density;
CONTRAST 'Linear' Density -1 0 1;
CONTRAST 'Quadratic' Density 1 -2 1;
RUN;
```

Before we consider the results of the two CONTRAST statements, we need to perform a preliminary analysis of variance. There is no apparent violation of the assumptions, and the overall ANOVA F test is significant ($p = 0.0025$), so it is valid to consider the post-hoc tests (i.e., the contrasts). From the Class Level Information table, the order of the treatments is 10, 20, 30.

Class Level Information		
Class	Levels	Values
Density	3	10 20 30

Note: If the “REF=” option had been used on the CLASS statement, then the order of the treatments would be altered. The coefficients in the CONTRAST statement must match the order of the treatments as given in the Class Level Information table.

The results for all the CONTRAST statements are given in a separate table.

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
Linear	1	63.37500000	63.37500000	38.18	0.0008
Quadratic	1	0.40500000	0.40500000	0.24	0.6389

The name for each contrast was specified in the code. (If we had given the name George to the linear contrast, then this table would have George instead of Linear.) These are the same two contrasts that we examined “by hand” earlier in this section. We calculated test statistics based on the t distribution and SAS is calculating test statistics based on the F distribution. They are equivalent because, when the numerator degrees of freedom is equal to 1, then $F = t^2$.

For the linear contrast, we calculated the test statistic $t = 6.18$, so $F = 6.18^2 = 38.19$, and SAS has $F = 38.18$. The difference is due to roundoff error in our hand calculations. From our hand calculations, we concluded that there is a significant linear trend. SAS generated the p-value 0.0008, so we would arrive at the same conclusion.

For the quadratic contrast, we calculated the test statistic $t = -0.49$, so $F = (-0.49)^2 = 0.24$, and this is what SAS has. From our hand calculations, we concluded that there is not a significant quadratic trend. SAS generated the p-value 0.6389, so we would arrive at the same conclusion.

It is absolutely imperative that the numeric treatments be in order from smallest to largest. This will automatically be true if SAS chooses the reference level, but it may not be true if the reference level is specified by the ‘REF=’ option in the SAS code. If we included the option ‘REF=10’ in Example 2, then the order of the treatments would put 10 last. This would be clearly indicated in the Class Level Information table.

Class Level Information		
Class	Levels	Values
Density	3	20 30 10

The coefficients for a linear contrast with three treatments are -1, 0 and 1, but this assumes the treatments are in numeric order. The coefficient -1 goes with treatment 10, coefficient 0 goes with treatment 20 and coefficient 1 goes with treatment 30. The null hypothesis for the linear contrast is

$$H_0 : (-1)\mu_{10} + (0)\mu_{20} + (1)\mu_{30} = 0 \quad (4.22)$$

If the option ‘REF=10’ had been included in the SAS code, then 10 would be the reference level. The reference level is always last, so to test for a linear trend we would need to rearrange the coefficients

$$H_0 : (0)\mu_{20} + (1)\mu_{30} + (-1)\mu_{10} = 0 \quad (4.23)$$

so the correct CONTRAST statement would be

```
CONTRAST 'Linear' Density 0 1 -1;
```

4.6.5. Examples of contrasts

Example 4.6.4

Continuing with the bean data from Example 2, suppose we wish to compare the mean yield of density 10 to the average of the yields of densities 20 and 30. The contrast is $\mu_1 - (\mu_2 + \mu_3)/2$ when expressed in terms of population means. Thus the contrast coefficients are 1, -0.5, and -0.5. The following contrast statement would test for significance of the contrast

```
contrast '1 vs avg of 2 and 3' Density 1 -0.5 -0.5;
```

We could also multiply the coefficients by 2 to convert to whole numbers. Thus we could test the same contrast with this statement

```
contrast '1 vs avg of 2 and 3' Density 2 -1 -1;
```

Example 4.6.5

To use the CONTRAST statement in SAS, the sum of the coefficients must be exactly 0. There is no allowable error for rounding. Suppose we have 4 treatments and wish to compare treatment 1 to the average of treatments 2, 3, and 4. This would give us the following contrast for population means:

$\mu_1 - (\mu_2 + \mu_3 + \mu_4)/3$. However, if we were to set the coefficients as 1, -.333, -.333, and -.333 we would get an error because they do not add exactly to 0. We could use 1, -.333, -.333, and -.334 (changing the last -.333 to -.334), but a better way is to multiply the coefficients by 3 and use 3, -1, -1 and -1.

Example 4.6.6

Suppose we have three pain medications labeled Control, Aspirin and Tylenol. Suppose we wish to compare control response to the average response of Aspirin and Tylenol. SAS will put the treatments in alphabetic order: Aspirin, Control, Tylenol. Thus the coefficients for Control vs. the average of Aspirin and Tylenol should be -.5, 1, -.5 (or -1, 2, -1). If we were to use 1, -.5, -.5 this would compare Aspirin to the average of Control and Tylenol. If in doubt as to the order SAS is using, look at the Class Level Information table at the beginning of the PROC GLM output. This is the order of the treatments that should be used in the CONTRAST statement. If the code contains a MEANS or an LSMEANS statement, then the order of the treatments in those tables will also match the order that SAS is using. (Using the MEANS or LSMEANS tables will be easier when there is more than one factor.)

4.6.6. Summary

All of the contrasts presented in this section are considered part of “post-hoc” analysis. Other post-hoc tests include the pairwise differences that were discussed in Section 4.5. The only valid reason to consider post-hoc tests is if (1) the model assumptions are not violated and (2) the overall ANOVA F test is significant. If either of these two conditions is not satisfied, then the results of post-hoc tests should not be interpreted.

Much of the discussion in this section involved linear, quadratic and higher-order trends. These are simply special cases of contrasts, and they are applicable only when the treatment levels are numeric.

Section 4.7. Power and Sample Size

If there truly are difference between the treatment means, we want our statistical analysis to detect these differences. Our ability to do this depends primarily on two things:

- (1) A properly designed and implemented study
- (2) Sample sizes that are large enough to overcome the background variation (“noise”) in the measurements, so that we can accurately estimate the true (population) treatment means.

Details regarding item (1) are discussed in an Experimental Design course. We will focus on item (2).

4.7.1. Criteria for determining the sample size

The appropriate sample size is directly linked to the amount of background variation that is inherent in the data. If the background variation is large, then it will be difficult to detect a difference in treatment means, so the sample size needs to be larger. If the background variation is small, then it will be easier to detect a difference in the treatment means, so the sample size can be smaller. In general, larger sample sizes give rise to more precise inference, and makes it easier to detect differences among the treatment means. The background variation can be measured several different ways, giving rise to several different criteria that can be used to determine the appropriate sample size.

In some situations, we want to select a sample size so that our estimates for the treatment means are within some pre-specified tolerance. For example, we might want to estimate the yield of a agricultural crop within ± 100 bushels, or estimate weight gain of an animal within ± 10 pounds. This tolerance is measured by the margin of error when constructing confidence intervals for the treatment means. For this situation, we would specify the desired margin of error and determine the sample size necessary to achieve the margin of error. The specified tolerance could also be expressed in terms of Fisher’s LSD, so that we would determine the sample size necessary to achieve a specified LSD.

It is also possible to select a sample size so that the overall ANOVA F test has a pre-specified power. Recall that power is the probability that the null hypothesis is reject when it is not true, that is, when there are differences among the means. In other words, power is the probability that we correctly reject the null hypothesis. To calculate the power, we need to know the true values of the treatment means (or at least the differences between them), so there will be different values for power depending on the true values for the means. The relationship between power and the treatment means is

generally displayed in a graph called the power curve, which is described in more detail later in this section.

Regardless of which criteria we use to determine the sample size, we must know how much variability to expect in the data. In other words, we need to know the standard deviation of the population. Since this value is not known, we must either make an intelligent guess as to its value, or we can use the value of the error standard deviation ($\hat{\sigma} = \sqrt{MSE} = RMSE$) from a previous study.

Margin of error criterion

If we are using the margin of error criterion to determine the sample size, we assume that all treatments have the same number of observations. We denote this value by n , so that $n_1 = n_2 = \dots = n_t$, where t is the number of treatments. We wish to select a sample size so that the margin of error (MOE) of the sample mean will be a desired value. Let dMOE be the desired MOE. To Find the sample, we solve for n in this equation

$$dMOE = t^* \times \frac{\hat{\sigma}}{\sqrt{n}} \quad (4.24)$$

where t^* is the critical value. For planning purposes, we can approximate $t^* = 2$, because the critical value for a 95% confidence is about 2 (except for very small sample sizes). The solution is

$$n = \left(t^* \times \frac{\hat{\sigma}}{dMOE} \right)^2 \approx \left(2 \times \frac{\hat{\sigma}}{dMOE} \right)^2 = \frac{4 \times MSE}{(dMOE)^2} \quad (4.25)$$

For example, suppose an analysis of variance was done on the grams of fat in bags of four brands of potato chips. The analysis of variance table is shown below.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	164.09	54.70	11.32	0.0002
Error	18	87.00	4.83		
Corrected Total	21	251.09			

Table 4.16. ANOVA table for potato chip example

In a future study, we would like to estimate the mean fat content with a desired margin of error 0.5 grams of fat. The value of the MSE is 4.83, so we use it in our sample size formula. We have

$$n = \frac{4 \times \text{MSE}}{(\text{dMOE})^2} = \frac{4 \times 4.83}{(0.5)^2} \approx 77.28 \Rightarrow n = 78 \quad (4.26)$$

For all sample size calculations, we always go UP to the next larger integer, because rounding down would decrease the precision below the desired level. To estimate the mean fat of the potato chips within ± 0.5 grams, we would need a sample size of 78 bags for each brand of chips. If this sample size seems too large, it simply tells us that our expectations for the future study are unrealistic. The amount of natural variability in the fat content is too great to be able to estimate the mean fat this precisely, unless the sample size is very large.

LSD criterion

When we use the LSD criterion, we are specifying the difference between the treatment means that we want to be able to detect. This is sometimes called a “meaningful” difference. For example, if the true mean fat in one brand of potato chips is 3.5 grams and the true mean fat in a different brand is 3.6 grams, it may not be important to be able to detect a difference of 0.1 grams. To calculate the sample size necessary to achieve a desired LSD (denoted dLSD), we will again assume that all treatments have the same sample size (so that $n_i = n_j = n$) and that the critical value is approximately equal to 2. We begin with the definition of the LSD

$$\text{dLSD} = t^* \times \hat{\sigma} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} = t^* \times \hat{\sigma} \sqrt{\frac{2}{n}} \quad (4.27)$$

Solving for n , we have

$$n = \left(\frac{t^* \hat{\sigma} \sqrt{2}}{\text{dLSD}} \right)^2 = \left(\frac{2\sqrt{2} \hat{\sigma}}{\text{dLSD}} \right)^2 = \frac{8 \times \text{MSE}}{(\text{dLSD})^2} \quad (4.28)$$

Continuing with the potato chip example, suppose we would like to have an LSD value of 2.0 for comparing the means. We have MSE = 4.83, so the sample size for each group would be

$$n = \frac{8 \times 4.83}{2^2} = 9.66 \Rightarrow n = 10 \quad (4.29)$$

As before, we round this value UP to the next integer. We would need 10 bags of each brand of chips to achieve a desired LSD of 2.0 grams.

4.7.2. Power of the ANOVA F test

The power of a statistical test is the probability that the null hypothesis is rejected when it should be (i.e., when the alternative hypothesis is true). Since this is a correct decision, we want the power to be as large as possible. There is always uncertainty in any statistical test, so it is not feasible to achieve power = 1 (i.e., 100% accuracy), but we usually want the power to be in the range of 0.5 to 0.9. In general, larger samples produce greater power, but the sample size usually becomes unrealistic when the power exceeds 0.9. The power also depends on the difference between the true population means, that is, larger differences generate greater power.

When using the power to make sample size determinations, there are four quantities of interest. These are

- The population standard deviation. As before, we can use an estimate from a previous study.
- The population means for which we would like to have a significant F test.
- The level of significance of the test, usually 5%.
- The desired power of the test, typically between 0.5 and 0.9.

We can provide any three of these four quantities to SAS, and SAS will calculate the fourth.

For example, suppose we want to do a new study involving four brands of potato chips. From a previous study, we have $\hat{\sigma} = \text{MSE} = 2.20$, which we assume to be the population standard deviation in the new study. We would like to perform the ANOVA F test at the 5% level of significance and have probability 0.8 of detecting a difference between means if the true population means for the four groups are in fact 15, 16, 17 and 18. The SAS code to calculate the required sample size is shown below.

```
PROC POWER;
  ONEWAYANOVA
    GROUPMEANS = 15 | 16 | 17 | 18
    STDDEV = 2.2
    ALPHA = 0.05
    NPERGROUP =
    POWER = 0.5 0.6 0.7 0.8 0.9
;
RUN;
```

Note that the placement of the semicolons in this code is a bit unusual. The PROC POWER statement ends with a semicolon, but the next semicolon does not appear until just before the RUN statement. All of the code in between these semicolons is one SAS statement, but they are put on separate lines to make them easier to read. The GROUPMEANS line identifies the assumed values for the population

treatment means, separated by vertical bars. The standard deviation (STDDEV) is the value of MSE from the previous study, and ALPHA is the significance level of the test. The value for NPERGROUP is not specified (a dot indicates a missing value), so this is what we want SAS to calculate. The line for POWER is specifying several values for the desired power, and SAS will calculate the sample size necessary to achieve each value of power. The output from this code is shown below.

Fixed Scenario Elements	
Method	Exact
Alpha	0.05
Group Means	15 16 17 18
Standard Deviation	2.2

Computed N Per Group			
Index	Nominal Power	Actual Power	N Per Group
1	0.5	0.529	7
2	0.6	0.603	8
3	0.7	0.725	10
4	0.8	0.817	12
5	0.9	0.906	15

Table 4.17. PROC POWER output

The first table in the output contains the specifications we provided and the results of the calculations are in the second table. For example, to achieve power 0.5 we would need 7 observations in each group. For this sample size, the actual power is slightly greater at 0.529. The actual power is the power we get when the sample size is rounded up to an integer. The nominal power is the target power we specified. It is usually not possible to achieve the nominal power exactly, since that would require that the sample size be a fraction.

This scenario is based on having group means of 15, 16, 17 and 18, so that the difference between groups means is at least 1. If we want our ANOVA F test to correctly detect these differences 50% of the time, we would need a sample of size 7 for each group. If we want to detect these differences 90% of the time, we would need a sample of size 15 for each group. Note that these are the sample sizes for each group because we used the option NPERGROUP in our SAS code.

These calculations relied on the value for the population standard deviation, which we presumed was 2.2 from an earlier study. What if this value is not correct? It is possible to get SAS to use a variety of values for the standard deviation, but the results are presented in graph rather than a table. This type of graph is called a power curve.

The SAS code for generating power curves for the potato chip example is shown below.

```
PROC POWER;
  ONEWAYANOVA
    TEST=OVERALL
    GROUPMEANS = 15 | 16 | 17 | 18
    STDDEV = 1.8 to 2.6 by 0.4
    ALPHA = 0.05
    NTOTAL = 8 to 200 by 4
    POWER = .
;
PLOT X=N MIN=8 MAX=200;
RUN;
```

The value for the standard deviation is no longer forced to be equal to 2.2, instead it can be any number between 1.8 to 2.6, in increments of 0.4. SAS will generate separate power curves for $\sigma = 1.8, 2.2$ and 2.6 . This code does not include an option for NPERGROUP; instead it uses NTOTAL which is the total sample size for all groups combined. This value is allowed to range between a minimum of 8 (2 per group) to a maximum of 200 (50 per group), in increments of 4. We have not specified any values for power, since that is what we want SAS to calculate. The PLOT statement generates the graph in Figure 4.12.

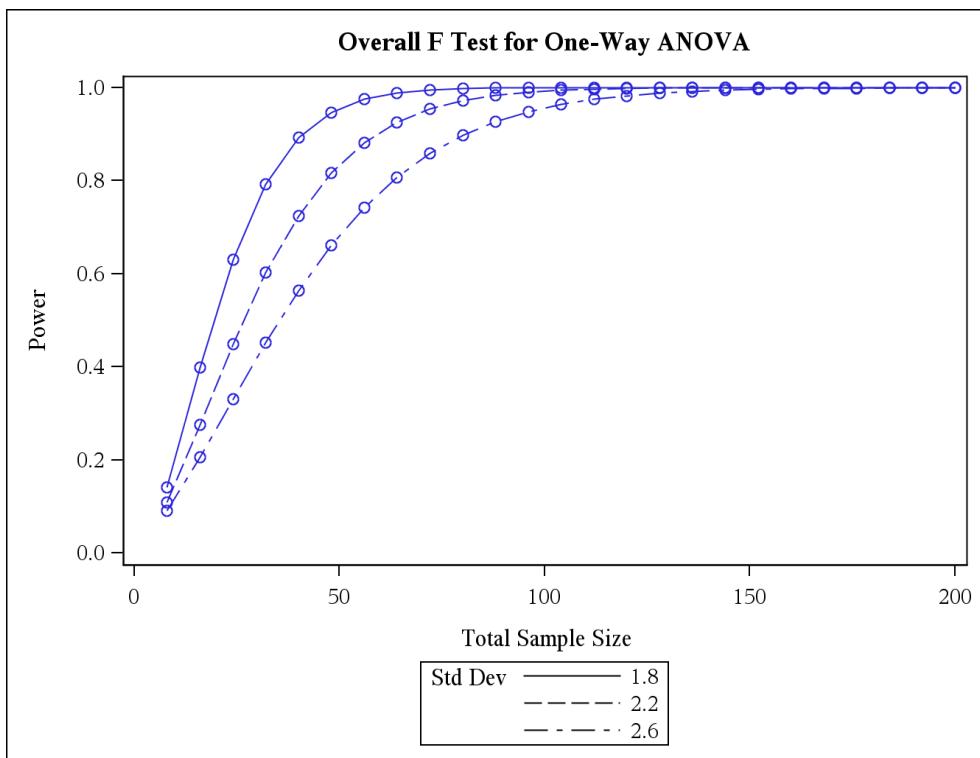


Figure 4.12. Power curves for potato chip example

There are three power curves in Figure 4.12, corresponding to the three values we specified for the standard deviation. To interpret this graph, suppose we want power 0.8 to detect a difference of 1 in the sample means. (This is the difference between the sample means 15, 16, 17 and 18 as specified in PROC POWER.) Draw a horizontal line from Power=0.8 (on the y-axis) and note where the horizontal line intersects each curve. The line intersects the first curve at approximately $N = 40$, the second curve at approximately $N = 48$, and the third curve at approximately $N = 60$. These values are the total sample size, so they need to be divided by the number of groups (4) to get the sample size for each group. To achieve 80% power, the required sample sizes are

- 10 per group, if the population standard deviation is 1.8 (first curve)
- 12 per group, if the population standard deviation is 2.2 (second curve)
- 15 per group, if the population standard deviation is 2.6 (third curve)

Sample size determination and power curves are part of the planning process before collecting the data. They are not part of the analysis after the data has been collected.

Chapter 5: Two-Way ANOVA

Section 5.1. Definitions and models

In the previous chapter, we examined ANOVA models that contain exactly one factor. These are called one-way designs. We now consider ANOVA models that contain exactly two factors, and these are called two-way designs. Every two-way ANOVA model can be analyzed as one-way design, but there are definite advantages to explicitly including both factors in the analysis.

Both of the factors in a two-way ANOVA model are classification variables, each with multiple levels. We will consider only factorial treatment structures, so that each level of the first factor is combined with each level of the second factor. The treatments are all the possible combinations of the levels of the two factors. For example, suppose that cookies are made from three different recipes (R1, R2 and R3), and each recipe is baked at two different oven temperatures (T1 and T2). This would produce a total of six treatments: (R1, T1), (R2, T1), (R3, T1), (R1, T2), (R2, T2) and (R3, T2).

The first part of the analysis of two-way data is to provide a descriptive statistical summary of the two-way means. This includes the following steps:

1. Put the data in an appropriate (e.g., spreadsheet) format.
2. Obtain the mean of each treatment.
3. Arrange the means in a two-way table according to the two factors.
4. Obtain a two-way plot of the means
5. Interpret the results.

These steps simply provide a statistical summary of the data. It is not a complete statistical analysis because it does not provide any information regarding whether or not the treatment means are equal, or which treatments (if any) have a different mean. The treatment means that are estimated from the data are only point estimates for the population treatment means. The statistical summary provides only the point estimates; it does not provide confidence intervals for the means. The descriptive analysis is usually carried out using statistical software, as the next example shows.

5.1.1. Fabric data example

An experiment was conducted to study the effects of treating fabric with inorganic salts on the flammability of fabric. Two concentrations and three salts were used, and a vertical burn test was used on three specimens of cloth for each concentration and salt combination. The response variable is the temperature at which the fabric specimen ignites.¹ The data are shown in Table 5.1Table 5..

Concentration	Salt		
	Untreated	CaCO ₃	CaCl ₂
1	812, 827, 876	733, 728, 720	725, 727, 719
2	945, 881, 919	786, 771, 779	756, 781, 814

Table 5.1. Data for fabric example

The estimated means are usually presented in a two-way table, similar to the format of the original data. This is shown in Table 5.2. These are the means for each treatment, but they are often called the estimated cell means or two-way means. A graph of these means is called a mean profile plot. This is shown in Figure 5..

From the graph, we can see that fabrics treated with either CaCO₃ or CaCl₂ ignite at lower temperatures than fabric that was untreated, and that fabrics that received Concentration 1 ignite at lower temperatures than fabric that received Concentration 2. It is not yet known if these differences are statistically significant. Later in the analysis, we will perform official hypothesis tests to determine this.

The format of the data in Table 5.1 is not suitable for statistical software. In order to read this data into SAS, we need one column for Concentration, one columns for Salt, and one column for the Temperature. This is shown in Table 5.3.

Estimated Mean Temperature to Ignite			
Concentration	Salt		
	Untreated	CaCO ₃	CaCl ₂
1	838.33	727.00	723.67
2	915.00	778.67	783.67

Table 5.2. Estimated treatment means

¹ Hsieh and Hardin, "Effects of Selected Inorganic Salts on Cotton Flammability", **Textile Research Journal**, Vol. 54, No. 3, 1984, pp. 171-179.

Conc.	Salt	Temp.	
1	Untreated	812	\
1	Untreated	827	Treatment 1. estimated mean = $(812 + 827 + 876) / 3 = 838.33$
1	Untreated	876	/
2	Untreated	945	\
2	Untreated	881	Treatment 2. estimated mean = $(645 + 881 + 919) / 3 = 915.00$
2	Untreated	919	/
1	CaCO ₃	733	\
1	CaCO ₃	728	Treatment 3. estimated mean = $(733 + 728 + 720)/3 = 727.00$
1	CaCO ₃	720	/
2	CaCO ₃	786	\
2	CaCO ₃	771	Treatment 4. estimated mean = $(786 + 771 + 779)/3 = 778.67$
2	CaCO ₃	779	/
1	CaCl ₂	725	\
1	CaCl ₂	727	Treatment 5. estimated mean = $(725 + 727 + 729)/3 = 723.67$
1	CaCl ₂	719	/
2	CaCl ₂	756	\
2	CaCl ₂	781	Treatment 6. estimated mean = $(726 + 781 + 814)/3 = 783.67$
2	CaCl ₂	814	/

Table 5.3. Spreadsheet format for fabric data

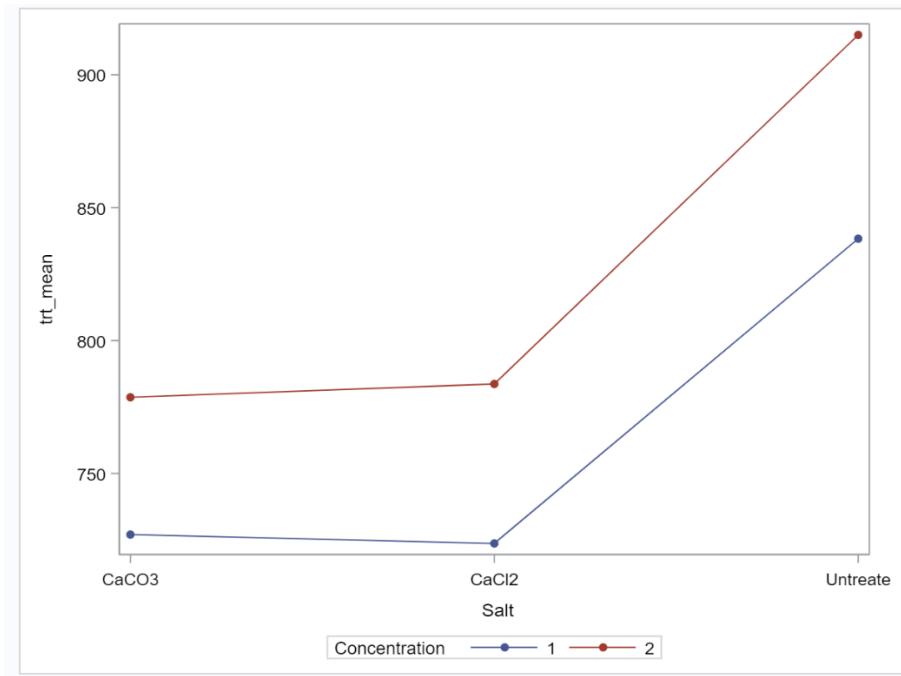


Figure 5.1. Mean profile plot for fabric data

5.1.2. Notation for two-way ANOVA

Let A denote one factor, with levels $i = 1, 2, \dots, a$.

Let B denote the other factor, with levels $j = 1, 2, \dots, b$.

μ_{ij} is the population mean for the treatment defined by the i^{th} level of A and j^{th} level of B.

Y_{ijk} is the observed response for the k^{th} subject in the i^{th} level of A and j^{th} level of B.

ε_{ijk} is the random “noise” for the k^{th} subject in the i^{th} level of A and j^{th} level of B.

The population means are often called cell means because they can be represented in one cell of a two-way table, as shown in Table 5.4. For this table, we have two levels for factor A (so $a = 2$) and three levels for B (so $b = 3$). This produces a total of 6 treatments.

Population Treatment Means			
Factor levels	B1	B2	B3
A1	μ_{11}	μ_{12}	μ_{13}
A2	μ_{21}	μ_{22}	μ_{23}

Table 5.4. Two-way layout for treatment means

When we average all of the population means for a given level of a factor, we obtain a marginal mean.

For example, the marginal mean for B1 is $\mu_{B1} = \frac{1}{2}(\mu_{11} + \mu_{21})$ and the marginal mean for A2 is

$\mu_{A2} = \frac{1}{3}(\mu_{21} + \mu_{22} + \mu_{23})$. We call these marginal means because they are often displayed in the margins of the two-way table, as shown in Table 5.5. The overall mean (a.k.a. the grand mean) is the average of all the treatment means, and we denote it as μ (with no subscripts). This is in the lower right cell of the two-way table.

Population Means				
Factor levels	B1	B2	B3	Marginal means for A
A1	μ_{11}	μ_{12}	μ_{13}	μ_{A1}
A2	μ_{21}	μ_{22}	μ_{23}	μ_{A2}
Marginal means for B	μ_{B1}	μ_{B2}	μ_{B3}	μ

Table 5.5. Population means and marginal means for a two-way table

The cell means, marginal means and the overall mean for the fabric data are shown in Table 5.6. These are all sample means, and they are estimates for the corresponding population means. (For this situation, the population would be all the fabric treated by one of these combinations, not just the fabric in the experiment.)

Estimated Mean Temperature to Ignite				
Concentration	Untreated	Salt		Marginal Means for Concentration
		CaCO ₃	CaCl ₂	
1	838.33	727.00	723.67	763.00
2	915.00	778.67	783.67	825.78
Marginal Means for Salt	876.67	752.84	753.67	794.39

Table 5.6. Estimated means and marginal means for the fabric data

5.1.3. Population main effects and interactions

The means calculated from the sample data are estimates for the population means. We now turn our attention to the main effects and interactions *for the population*. We first consider the main effect for each factor.

The main effects for the levels of factor A are the differences between the marginal means for factor A and the overall mean. When A has two levels,

- the main effect of level A1 is $\mu_{A1} - \mu$, and
- the main effect of level A2 is $\mu_{A2} - \mu$

The main effects for the levels of factor B are the differences between the marginal means for factor B and the overall mean. When B has three levels,

- the main effect of level B1 is $\mu_{B1} - \mu$, and
- the main effect of level B2 is $\mu_{B2} - \mu$, and
- the main effect of level B3 is $\mu_{B3} - \mu$

For the fabric data

- the estimated effect for Concentration 1 (level A1) is $763.00 - 794.39 = -31.39$
- the estimated effect for Concentration 2 (level A2) is $825.78 - 794.39 = 31.39$
- the estimated effect for Untreated (level B1) is $876.67 - 794.39 = 82.28$
- the estimated effect for Salt CaCO₃ (level B2) is $752.84 - 753.67 = -41.55$
- the estimated effect for Salt CaCl₂ (level B3) is $753.67 - 794.39 = -40.72$

When the effect is negative the marginal mean is below the overall mean, and when the effect is positive the marginal mean is above the overall mean. From the estimated effects, we expect fabric treated with Concentration 1 to ignite at a lower temperature than Concentration 2, and that fabric treated with either of the two salts to ignite at a lower temperature than the untreated fabric. *It needs to be emphasized that these values are for sample data, so they are only estimates of the true effects in the population. We have not yet conducted statistical tests to determine if these effects are significant.*

Now that we have defined the effects, the next step is to examine whether or not the effects are additive. The effects of factors A and B are additive if, for each treatment combination (A_i, B_j), the population treatment mean can be expressed as

$$\mu_{ij} = \mu + (\mu_{Ai} - \mu) + (\mu_{Bj} - \mu) \quad (5.1)$$

In other words, the effects are additive if

$$(\text{mean for } A_i \text{ and } B_j) = (\text{grand mean}) + (\text{effect of } A_i) + (\text{effect of } B_j) \quad (5.2)$$

This is usually written

$$\mu_{ij} = \mu + \alpha_i + \beta_j \quad (5.3)$$

where α_i is the main effect of level i of A and β_j is the main effect of level j of B.

If equation (5.2) is not true, then we say there is an *interaction* and we must include an interaction effect in equation (5.3). The interaction effect is denoted by $(\alpha\beta)_{ij}$. It is defined as the difference between the treatment mean and the additive effect, that is,

$$(\alpha\beta)_{ij} = \mu_{ij} - (\mu + \alpha_i + \beta_j) \quad (5.4)$$

so that the treatment mean can be expressed as

$$\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} \quad (5.5)$$

For the fabric data, we can use the estimated means in Table 5.6 to determine if the means are additive. Consider the treatment combination Concentration 1 and Salt CaCl₂. The estimated treatment mean is $\hat{\mu}_{13} = 723.67$. The estimated effect of Concentration 1 is $\hat{\alpha}_1 = \hat{\mu}_{A1} - \hat{\mu} = 763.00 - 794.39 = -31.39$ and the estimated effect for Salt CaCl₂ is $\hat{\beta}_3 = \hat{\mu}_{B3} - \hat{\mu} = 753.67 - 794.39 = -40.72$. We substitute these estimates into equation (5.3). Does $723.67 = 794.39 + -31.39 + -40.72$? No, 723.67 is not equal to 722.28, so the estimated means are not additive. This does not imply that the population means are not additive. The difference between 723.67 and 722.28 seems small, so the interaction we see in the sample may not be statistically significant.

5.1.4. Models for two-way ANOVA

Every two-way ANOVA model can be written as

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk} \quad (5.6)$$

where ε_{ijk} are independent, normally distributed, with mean 0 and constant variance σ^2 .

There are two types of models that can be used for a two-way ANOVA.

- An **additive model** uses the equation

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \quad (5.7)$$

- An **interaction model** uses the equation

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \quad (5.8)$$

When performing a two-way analysis of variance, it is customary to first fit an interaction model. After verifying that the assumptions are not violated, we perform a hypothesis test to determine if the interaction terms are needed in the model. If they are not needed, we can consider using an additive model.

5.1.5. Interactions in plain English

The concept of interactions can be quite confusing at first, but statisticians use this term exactly the way it is used in the medical community. For example, consider a clinical trial (i.e., an experiment), in which people who routinely consume regular (non-diet) soda are randomly assigned to one of four treatment groups. The purpose of the experiment is to identify which treatment is more effective at enhancing weight loss. One treatment group takes diet pills and switches from regular to diet soda. Another treatment group also takes diet pills, but continues to drink regular soda. A third group does not take diet pills, but switches to diet soda. A fourth group does not take diet pills and does not switch to diet soda. After one week, the weight loss is recorded.

This experiment consists of two factors (Pills and Soda), and each factor has 2 levels. The response variable is the weight loss (in pounds) after one week. The treatment means are shown in Table 5.7.

		Diet soda	
		Yes	No
Diet pills	Yes	5	2
	No	3	1

Table 5.7. Treatment means for weight loss example

It is not surprising that the people who took the diet pills lost more weight, on average, than those who did not take the pills (marginal means 3.5 and 2 pounds, respectively). It is also not surprising that the people who switched to diet soda lost more weight, on average, than those who continued to drink regular soda (marginal means 4 and 1.5 pounds, respectively).

There are two ways to view the interaction between Soda and Pills:

View #1. For each level of Soda, what is the effect of Pills?

- When Soda = Yes (they switched to diet soda), the folks who took the pills lost 2 more pounds, on average, than those who did not take the pills. (this is 5 vs 3)
- When Soda = No (they did not switch to diet soda), the folks who took the pills lost 1 more pound, on average, than those who did not take the pills. (this is 2 vs 1)

These comparisons illustrate that the effect of the pills is greater (more weight loss) for those who switched to diet soda.

View #2. For each level of Pills, what is the effect of Soda?

- When Pills = Yes, the folks who switched to diet soda lost 3 more pounds than those who did not switch. (this is 5 vs 2)
- When Pills = No, the folks who switched to diet soda lost 2 more pounds than those who did not switch. (this is 3 vs 1)

These comparisons illustrate that the effect of switching to diet soda is greater (more weight loss) for those who were taking the diet pill.

To summarize, the effect of switching to diet soda depends on whether or not the person was taking the diet pill. Also, the effect of the diet pill depends on whether or not the person switched to diet soda.

This is precisely what an interaction is – the effect of one factor is not the same for all levels of the other factor.

There are many other examples of interactions in the medical community. For example, blood pressure medication can interact with cold medication. In general, most cold medications will increase blood pressure, even for people who are not taking blood pressure medication. But the effect of the cold medication may be greater for people who are taking blood pressure medication. Most of the common drug interactions are included in the warning labels. Have you ever taken a prescription medication that had the warning “Do not take with alcohol”? This is because alcohol interacts with the medication. For example, mixing alcohol with some diabetes medications can cause abnormally low blood sugar levels, nausea, headaches, and rapid heartbeat. Alcohol by itself does not cause these conditions, and neither does the medication by itself. But the *combination* of both medications can be problematic. This is an interaction.

Section 5.2. Hypotheses for two-way ANOVA

In the last section, we considered both additive models and interaction (i.e., non-additive) models for two-way ANOVA. In this section, we define important hypotheses that are tested in a two-way analysis. These hypotheses involve the main effects and the interaction effects. In later sections, we will conduct statistical tests of these hypotheses to determine which effects are significant which are not. In this section, we are interested in defining the hypotheses and understanding what they mean.

5.2.1. Hypotheses for main effects

The hypotheses for main effects are defined in terms of the interaction model, defined in equation (5.8):

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

The hypotheses for the main effects of factor A are

$$\begin{aligned} H_0 : \alpha_i &= 0 \text{ for all levels of A} \\ \text{vs. } H_a : \text{not all } \alpha_i \text{'s are equal to 0} \end{aligned} \tag{5.9}$$

These hypotheses have been defined in terms of the effects for the levels of A, but they can be also defined in terms of the marginal means for the factor. If the null hypothesis is true (so that none of the levels of A have an effect on the response), then all of the levels of A will have the same marginal mean. To see why this is true, suppose that $\alpha_i = 0$ for all levels of A (so that the null hypothesis is true). By definition, $\alpha_i = \mu_{Ai} - \mu$. So if this difference is always 0, then $\mu_{Ai} = \mu$ for all levels of A. This tells us that all the marginal means are equal (and they are equal to the grand mean). With this understanding, the hypotheses for the main effects of factor A can be expressed as

$$\begin{aligned} H_0 : \text{the marginal means for all levels of A are equal} \\ \text{vs. } H_a : \text{at least one of these marginal means is different} \end{aligned} \tag{5.10}$$

The hypotheses defined in equations (5.9) and (5.10) *are the same set of hypotheses*. They are simply two different ways of thinking about the main effects of factor A. If H_0 is true, we say that there are no A main effects. If H_a is true, we say there are A main effects.

The hypotheses for the main effects of factor B are defined similarly. In terms of effects, the hypotheses are

$$\begin{aligned} H_0 &: \beta_j = 0 \text{ for all levels of B} \\ \text{vs. } H_a &: \text{not all } \beta_j \text{'s are equal to 0} \end{aligned} \quad (5.11)$$

In terms of marginal means, the hypotheses are

$$\begin{aligned} H_0 &: \text{the marginal means for all levels of B are equal} \\ \text{vs. } H_a &: \text{at least one of these marginal means is different} \end{aligned} \quad (5.12)$$

If H_0 is true, we say that there are no B main effects. If H_a is true, we say there are B main effects.

5.2.2. Hypotheses for interaction effects

Continuing with the interaction model in equation (5.8), the interaction hypotheses are

$$\begin{aligned} H_0 &: (\alpha\beta)_{ij} = 0 \text{ for all treatments} \\ \text{vs. } H_a &: \text{at least one of the } (\alpha\beta)_{ij} \text{'s is not 0} \end{aligned} \quad (5.13)$$

If H_0 is true, we say that there is no interaction. If H_a is true, we say there is an interaction. If there is no interaction, then the interaction model in equation (5.8) reduces to the additive model in equation (5.7).

Interactions can be written as contrasts of treatment means. Consider the population means formed by combining levels A1 and A2 with levels B1 and B2, as shown in Table 5.8. If there is no interaction between these four treatments, then the difference between B1 and B2 when A is at level A1 will be equal to the difference between B1 and B2 when A is at the level A2. This would be the contrast $(\mu_{11} - \mu_{12}) - (\mu_{21} - \mu_{22})$, and is illustrated in Table 5.8 (a). Another way to view the same contrast is to compare the difference between A1 and A2 when B is at level B1 to the difference between A1 and A2 when B is at level B2. This would be the contrast $(\mu_{11} - \mu_{21}) - (\mu_{12} - \mu_{22})$, and is shown in Table 5.8 (b). Regardless of which way you view this contrast, if the contrast is equal to 0 then there is not an interaction between these two levels of the two factors. If the contrast is not 0, then there is an interaction.

Factor levels	B1	B2	B3
A1	μ_{11}	μ_{12}	μ_{13}
A2	μ_{21}	μ_{22}	μ_{23}

(a)

Factor levels	B1	B2	B3
A1	μ_{11}	μ_{12}	μ_{13}
A2	μ_{21}	μ_{22}	μ_{23}

(b)

Table 5.8. Two ways to view an Interaction contrast

Example

Consider the fictitious population means in Table 5.9. The marginal means for factor A are both equal to

12, so there are no main effects for factor A. The marginal means for factor B are 10, 16 and 10.

Because they are not all the same, there are main effects for factor B.

Factor Levels	B1	B2	B3	Marginal Means
A1	9	18	9	12
A2	11	14	11	12
Marginal Means	10	16	10	12

Table 5.9. Fictitious population means

The interaction contrasts are

- Between A1 & A2 and B1 & B2: $(9 - 11) - (18 - 14) = -6$
- Between A1 & A2 and B1 & B3: $(9 - 11) - (9 - 11) = 0$
- Between A1 & A2 and B2 & B3: $(18 - 9) - (14 - 11) = 6$

Since some of the contrasts are not 0, there is an interaction between A and B.

5.2.3. An Agronomy Example

For a hypothetical agronomy study, we will consider two cases of population means to illustrate how to use marginal means and contrasts to determine whether or not there are main effect and interaction.

For this example, we will assume that the means are *population* means (not sample means).

An agronomist studied how an insecticide and a herbicide (bug and weed killers) affected the growth of plants. The agronomist applied combinations of two levels of insecticide (0 and 4) and three levels of herbicide (0, 1 and 2) to containers of plants and measured their dry weights after the growing period.

Case 1

For this case, we will use the population means shown in Table 5.10. We will show that there are main effects for both insecticide and herbicide, but there is no interaction.

Insecticide	Herbicide			Marginal Means
	H = 0	H = 1	H = 2	
I = 0	100	70	70	80
I = 4	90	60	60	70
Marginal Means	95	65	65	75

Table 5.10. Agronomy means for case 1

There are main effects for insecticide because the marginal means (80 and 70) are not the same.

Similarly, there are main effects for herbicide because the marginal means (95, 65 and 65) are not the same. Since we are assuming these means are population means, these effects are significant. (If they were sample means, we would need to perform hypothesis tests to determine significance.)

There is no interaction because every interaction contrast involving these means is equal to 0. A simple way to show this is to note that difference between I = 0 and I = 4 is the same for all levels of H. In other words,

- For H = 0: $100 - 90 = 10$
- For H = 1: $70 - 60 = 10$
- For H = 2: $70 - 60 = 10$

Since all of these differences are the same, the effect of insecticide is the same for all levels of herbicide.

Case 2

For this case, we will use the population means shown in Table 5.11. We will show that there are main effects for both insecticide and herbicide, and there is also an interaction.

Insecticide	Herbicide			Marginal Means
	H = 0	H = 1	H = 2	
I = 0	100	90	80	90
I = 4	90	70	50	70
Marginal Means	95	80	65	80

Table 5.11. Agronomy means for case 2

There are main effects for insecticide because the marginal means (90 and 70) are not the same.

Similarly, there are main effects for herbicide because the marginal means (95, 80 and 65) are not the same. As with Case 1, we are assuming these means are population means, so these effects are significant. If they were sample means, we would need to perform hypothesis tests to determine significance.

To show that there is an interaction, we have to find just one interaction contrast that is not equal to 0.

Consider the contrast involving $I = 0$ and $I = 4$ with $H = 0$ and $H = 1$. The interaction contrast is

$(100 - 90) - (90 - 70) = -20$. We can also use the simpler technique used in Case 1 to determine if there is an interaction. For the means in Case 2, the differences between $I = 0$ and $I = 4$ are

- For $H = 0$: $100 - 90 = 10$
- For $H = 1$: $90 - 70 = 20$
- For $H = 2$: $80 - 50 = 30$

Since all of these differences are not the same, the effect of insecticide is the not same for all levels of herbicide. Thus there is an interaction between herbicide and insecticide.

To understand the nature of the interaction, we should always try to interpret the interaction in the context of the study being done. Note that when $I = 0$, the means are 100, 90 and 80. For each level increase in herbicide, the mean dry weight decreases 10 units. A different pattern is present when $I = 4$. The means are 90, 70 and 50. For each level increase in herbicide, the mean dry weight decreases 20 units. This is the nature of interaction – the effect of herbicide depends on the level of insecticide. This can also be seen by comparing the means for $I = 0$ and $I = 4$ at each level of H , which were summarized above. When H is 0, 1 and 2, the differences are 10, 20 and 30, respectively. The interaction occurs because the insecticide “accelerates” the effect of herbicide.

5.2.4. Interaction plots

In an interaction plot, the treatment means are plotted so that levels of one factor are on the x axis and the levels of the other factor are joined by line segments. This plot is sometimes called a mean profile plot. If the line segments are parallel, there is no interaction. If the line segments are not parallel, then there is an interaction.

Interaction plots for both cases of the agronomy study are shown in Figure 5.2. For Case 1, there is no interaction so the interaction plot shows parallel line segments. For Case 2, there is an interaction so the lines are not parallel.

It is important to remember that in the fictitious agronomy example, we are assuming that the treatment mean are *population* means, so we do not need to perform a hypothesis test to determine if the interaction is significant. Whenever we are working with sample data, we may have interaction plots in which the line segments are not strictly parallel, but we will need to conduct formal hypothesis tests to determine if interaction is significant.

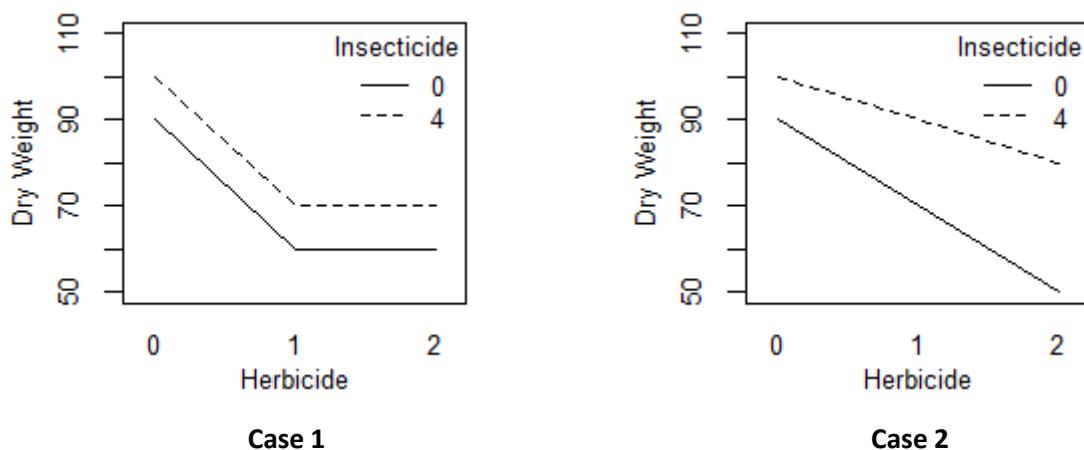


Figure 5.2. Interaction plots for agronomy example

Section 5.3. ANOVA Table and F-tests

We are still working with two-way data, focusing on main effects and interaction effects. So far, we have used *population* means to define these effects, but now we consider sample data. Definitions and calculations for the main effects and interaction effects remain the same, but we will need to consider formal hypothesis tests to determine if the effects are statistically significant. In other words, “small” differences we may see in the sample data might be due to sampling variability, and they may not indicate that there is a true difference in the population.

We begin with the interaction model: $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$

There are three main hypothesis tests of interest. These are:

- the main effects for factor A:
 $H_0 : \alpha_i = 0$ for all levels of A
vs. $H_a : \text{not all } \alpha_i\text{'s are 0}$
- the main effects for factor B:
 $H_0 : \beta_j = 0$ for all levels of B
vs. $H_a : \text{not all } \beta_j\text{'s are 0}$
- the interaction effects:
 $H_0 : (\alpha\beta)_{ij} = 0$ for all treatments
vs. $H_a : \text{not all } (\alpha\beta)_{ij}\text{'s are 0}$

Each test will have a test statistic that depends on the point estimates for the parameters and the variability in the response. The total variability in the response is separated into three components that are attributed to, i.e., explained by (1) the A main effects, (2) the B main effects, and (3) the interaction effects.

5.3.1. Point estimates

For all of the model parameters (α 's, β 's and $(\alpha\beta)$'s) and all of the means (treatment means, marginal means and grand mean), the symbols that are used to denote the population values are also used to denote the point estimate. To distinguish between the parameter and the estimate, we include a caret (a “hat”) for the point estimate. We use the least squares criterion to obtain the estimates, which is analogous to how we obtained the estimates for regression analysis.

The least squares estimates for the the means can be described as

- the grand mean
 $\hat{\mu}$ = the average of all the observations
- the treatment (i.e., cell) means
 $\hat{\mu}_{ij}$ = the average of all observations in treatment (A_i, B_j)
- the marginal mean for the i^{th} level of factor A
 $\hat{\mu}_{Ai}$ = average of all the cell means involving factor A_i
- the marginal mean for the j^{th} level of factor B
 $\hat{\mu}_{Bj}$ = average of all the cell means involving factor B_j

Note: These definitions are based on Type 3 sums of squares, which lead to Type 3 estimates and hypothesis tests. There are also Type 1, Type 2 and Type 4 definitions, but we will not consider those in this course. Differences between the four types arise only when the data are not balanced.

To obtain the estimates for the main effects and the interaction effects, we use the definitions for these effects and the least squares estimates for the means. These estimates are

- effect of the i^{th} level of Factor A: $\hat{\alpha}_i = \hat{\mu}_{Ai} - \hat{\mu}$
- effect of the j^{th} level of Factor B: $\hat{\beta}_j = \hat{\mu}_{Bj} - \hat{\mu}$
- interaction effect: $(\alpha\beta)_{ij} = \hat{\mu}_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j$

5.3.2. Partitioning the total variation

Test statistics for the hypotheses are based on the points estimates defined above, but they also depend on separating (partitioning) the the total variation in the response. This is same approach that we used in regression. The total variation in the response is split into two basic parts: the variation that is explained by the model (SSModel) and the remaining unexplained variation which is called the sum of squares due to error (SSE).

$$\begin{aligned}
SST_{\text{tot}} &= \sum(Y_{ijk} - \bar{Y})^2 = \sum(Y_{ijk} - \hat{\mu})^2 \\
&= \sum(Y_{ijk} - \hat{Y}_{ijk} + \hat{Y}_{ijk} - \hat{\mu})^2 \\
&= \sum(Y_{ijk} - \hat{Y}_{ijk})^2 + \sum(\hat{Y}_{ijk} - \hat{\mu})^2 \\
&= \text{SSE} + \text{SSModel}
\end{aligned}$$

We continue partitioning by splitting the model sum of squares (SSModel) into components for the effects.

$$\begin{aligned}
\hat{Y}_{ijk} &= \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + (\hat{\alpha}\hat{\beta})_{ij} \\
\hat{Y}_{ijk} - \hat{\mu} &= \hat{\alpha}_i + \hat{\beta}_j + (\hat{\alpha}\hat{\beta})_{ij} \\
\hat{Y}_{ijk} - \hat{\mu} &= (\hat{\mu}_{Ai} - \hat{\mu}) + (\hat{\mu}_{Bj} - \hat{\mu}) + (\hat{\mu}_{ij} - \hat{\mu}_{Ai} - \hat{\mu}_{Bj} + \hat{\mu})
\end{aligned}$$

Square each term and sum over all observations.

a is the number of levels for A

b is the number of levels for B

r is the number of observations in each treatment (the number of replications)

$$\begin{aligned}
\sum(\hat{Y}_{ijk} - \hat{\mu})^2 &= rb \sum(\hat{\mu}_{Ai} - \hat{\mu})^2 + ra \sum(\hat{\mu}_{Bj} - \hat{\mu})^2 + r \sum(\hat{\mu}_{ij} - \hat{\mu}_{Ai} - \hat{\mu}_{Bj} + \hat{\mu})^2 \\
\text{SSModel} &= \text{SSA} + \text{SSB} + \text{SSAB}
\end{aligned}$$

These sums of squares, in addition to the sum of squares for error (SSE) are used to construct the mean squares, which are used to construct the test statistics.

For the error

- degrees of freedom = dfE = $abr - ab$
- mean square = MSE = SSE / dfE

For the A main effect (with a levels)

- degrees of freedom = dfA = $a - 1$
- mean square = MSA = SSA / dfA
- test statistic = F = MSA / MSE

For the B main effect (with b levels)

- degrees of freedom = dfB = $b - 1$
- mean square = MSB = SSB / dfB
- test statistic = F = MSB / MSE

For the interaction

- degrees of freedom = $df_{AB} = (a - 1)(b - 1)$
- mean square = $MS_{AB} = SS_{AB} / df_{AB}$
- test statistic = $F = MS_{AB} / MSE$

Each of the test statistics follows an F distribution. The numerator and denominator degrees of freedom for the F distribution match the degrees of freedom in the fraction used to calculate the test statistic. As with regression, the degrees of freedom, sums of squares, mean squares and test statistics can be summarized in an ANOVA table, as shown in Table 5.12. We will rely on software to perform most of these calculations (especially the sums of squares and the p-values).

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
A main effect	$a - 1$	SSA	SSA/dfA	MSA / MSE	
B main effect	$b - 1$	SSB	SSB / dfB	MSB / MSE	
AB interaction	$(a - 1)(b - 1)$	SSAB	SSAB / dfAB	MSAB / MSE	
Error	$abr - ab$	SSE	SSE / dfE		
Total	$abr - 1$	SSTot			

Table 5.12. Generic ANOVA table for a two-way interaction model

5.3.3. Fabric data, re-visited

In Section 5.1.1, we examined data in which fabric was treated with one of three different inorganic salts, at one of two levels of concentration, for the purpose of measure their effect on the flammability of the fabric. The response variable is the temperature at which the fabric ignites. The data contains 3 replications (see Table 5.3). The ANOVA table for the two-way interaction model is shown in Table 5.13.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Concentration	1	17734.72	17734.72	34.11	<.0001
Salt	2	60928.78	30464.39	58.60	<.0001
Concentration*Salt	2	486.11	243.06	0.47	0.6375
Error	12	6238.67	519.89		
Corrected Total	17	85388.28			

Table 5.13. ANOVA table for the fabric data

We can easily verify the degrees of freedom:

- Factor A is Concentration: number of levels = $a = 2$; $dfA = a - 1 = 1$
- Factor B is Salt: number of levels = $b = 3$; $dfB = b - 1 = 2$
- Interaction: $dfAB = (a - 1)(b - 1) = (1)(2) = 2$
- Error: $dfE = abr - ab = (2)(3)(3) - (2)(3) = 12$
- The total degrees of freedom is always one less than the sample size:
 $dfTot = abr - 1 = (2)(3)(3) - 1 = 17$

We can also easily verify the values for the mean squares ($MS = SS/df$) and the test statistics

($F = MS/MSE$):

- $MS(\text{Concentration}) = 17734.72 / 1 = 17734.72$
- $MS(\text{Salt}) = 60928.78 / 2 = 304.64.39$
- $MS(\text{Interaction}) = 486.11 / 2 = 243.06$
- $MSE = 6268367 / 12 = 519.89$
- Test statistic for Concentration: $F = 17734.72 / 519.89 = 34.11$
- Test statistic for Salt: $F = 30464.39 / 519.89 = 58.60$
- Test statistic for interaction: $F = 243.06 / 519.89 = 0.47$

It is strongly recommended that the basic elements of the ANOVA table be checked at the beginning of every two-way analysis. Very minor mistakes in the data, such as extra (or missing) spaces or mis-matched capitalization, can create unexpected levels for a factor. These can be easily identified by examining the degrees of freedom in the ANOVA table. In addition, the total degrees of freedom should always be one less than the total number of observations (in all treatment groups). If the ANOVA table contains an incorrect value for the total degrees of freedom, the most likely mistake is that some of the data has been overlooked.

To interpret the ANOVA table, examine the interaction first. The interaction is not significant ($p = 0.6375$). This indicates two things:

- If the type of salt impacts temperature, the effect is statistically the same for both levels of concentration.
- If the concentration impacts temperature, the effect is statistically the same for all three types of salts.

The result of the interaction test will dictate which tests needs to be considered next. If the interaction is significant, the analysis becomes more complicated. This will be discussed later. For the fabric data, the interaction is not significant. This tells us that we can interpret the other tests in the ANOVA table, specifically the tests for main effects.

- Salt does affect temperature. ($F = 58.60$, $p < 0.0001$)
- Concentration does affect temperature ($F = 34.11$, $p < 0.0001$)

These tests do not tell us *how* the temperature is affected, i.e., whether the temperature is higher or lower, or by what quantity. To make these determinations, we will need to perform post-hoc tests.

5.3.4. Nested model F test, re-visited

Each of the hypothesis tests for main effects and interactions can be viewed as a nested models F test. This test was first discussed in Section 2.4.5. It requires two models. One model is called a full model and the second model called the reduced model. The full model contains every term that is in the reduced model, plus some additional terms. The nested model F test is used to decide which model is more appropriate for the data, that is, if the additional terms are needed to adequately describe the data.

To apply this test to a two-way analysis, the full model is the interaction model. We will consider the test for interaction effects, which has the null hypothesis $H_0 : (\alpha\beta)_{ij} = 0$ for all treatments . If this hypothesis is true, then all the interaction terms can be removed from the interaction model. This results in the additive model, which is the reduced model. Thus the test for interaction is simply a nested model F test that compares the interaction model (the full model) to the additive model (the reduced model). The “additional terms” are the interaction terms. If this test is significant (the p-value is less than the significance level), then the interaction model should be used. If this test is not significant, then the additive model can be used.

Section 5.4. t tests and contrasts

In the previous section, we examined how to use F tests to determine whether or the *effects* are significant. In this section, we turn our attention to the *means*. We can get tests and confidence intervals for

- one mean (a treatment mean or a marginal mean)
- the difference of two means
- contrasts of means

All tests and confidence intervals require three things:

1. a point estimate
2. a standard error for the point estimate
3. the reference distribution (to get the critical value and/or the p-value)

For the treatment means and the marginal means, we have already calculated point estimates. We now get the standard errors for these estimates. The standard error measures of any sample mean (either a treatment mean or marginal mean) measures how accurately the sample mean approximates the

corresponding population mean. The standard of the estimate is $SE = \sqrt{\frac{MSE}{n}}$, where n is the number of observations that go into computing the estimated mean. We illustrate the calculations using the fabric data. The sample means and the number of observations are shown in Table 5.14.

		Salt			
		Untreated mean (n)	CaCO ₃ mean (n)	CaCl ₂ mean (n)	Marginal mean (n)
Concentration	1	838.33 (3)	727.00 (3)	723.67 (3)	763.00 (9)
	2	915.00 (3)	778.67 (3)	783.67 (3)	825.78 (9)
	Marg. mean (n)	876.67 (6)	752.84 (6)	753.67 (6)	

Table 5.14. Sample means for the fabric data

From the ANOVA table (Table 5.13), $MSE = 519.89$. This produces the following standard errors:

- For each of the cell means, $n = 3$ so $SE = \sqrt{\frac{519.89}{3}} \approx 13.16$
- For the marginal means for Concentration, $n = 9$ so $SE = \sqrt{\frac{519.89}{9}} \approx 7.60$
- For the marginal means for Salt, $n = 6$, so $SE = \sqrt{\frac{519.89}{6}} \approx 9.31$

We can perform hypothesis tests and construct confidence intervals for any of the sample means. For individual treatment means and marginal means, we are usually interested in confidence intervals. Formal hypothesis tests are often less informative than confidence intervals for these means. For combinations of means, such as the difference of two means or a contrast involving several means, we are usually interested in testing whether or not the means are equal (i.e., if the difference in means is equal to 0). We can also construct confidence intervals for the difference of means and for contrasts. When multiple means are involved, the standard errors are different than those described about.

5.4.1. Confidence intervals for means

The basic form of a confidence interval is: (point estimate) \pm (critical value) \times (SE). The critical value comes from the t distribution, with degrees of freedom equal to dfE . For the fabric data, $dfE = 12$, and the two-sided critical value for significance level 0.05 is 2.179. With this information, we can construct the following confidence intervals for any of the means shown in Table 5.14. For example,

- Treatment mean for CaCl_2 and Concentration 2:
 $783.67 \pm (2.179)(13.16) = 783.67 \pm 28.68$, or $(754.99, 812.35)$
- Marginal mean for Concentration 1:
 $763.00 \pm (2.179)(7.60) = 763.00 \pm 16.56$, or $(746.44, 779.56)$
- Marginal mean for Untreated:
 $876.67 \pm (2.179)(9.31) = 876.67 \pm 20.3$, or $(856.37, 896.97)$

Note that the confidence intervals become more narrow when a greater number of observations are used to calculate the point estimate and the standard error. This is because more data increases the precision of our estimates, and this results in narrower confidence intervals.

When we construct confidence intervals or perform hypothesis tests for the *difference* of two means, we need to make sure that the two means do not have any observations in common. This would be the case for

- any two of the treatment means
- any two of the marginal means for factor A (Concentration)
- any two of the marginal means for factor B (Salt)

We will not consider the difference between, for example, the marginal mean for concentration 1 and the marginal mean for salt CaCl_2 , because these two means use some of the same observations.

When the two means do not have any observations in common, we obtain the point estimate of difference simply by subtracting the two individual point estimates. The standard error of the difference is $\text{SE}_{\text{diff}} = \sqrt{\text{SE}_1^2 + \text{SE}_2^2}$, where SE_1 and SE_2 are the standard errors of the two individual point estimates.

The critical value is the same as with a single mean.

For example, the point estimate for the difference between the marginal means of CaCO_3 and CaCl_2 is $752.84 - 753.67 = -0.83$. The standard error of the difference is $\sqrt{9.31^2 + 9.31^2} \approx 13.17$. Using the same critical value (2.179), we obtain a 95% confidence interval for the difference:

$$-0.83 \pm (2.179)(13.17) = -0.83 \pm 28.70, \text{ or } (-29.53, 27.87)$$

We are 95% confident that the mean temperature to ignite for CaCO_3 is between 29.53 degrees lower than and 27.87 degrees higher than the mean temperature to ignite for CaCl_2 . Since this confidence interval contains 0, we would conclude (at significance level 0.05) that there is not a significant difference in the mean temperature for these two salts.

5.4.2. Hypothesis tests for the difference of two means

Although confidence intervals can be used to determine if the difference between two means is statistically significant, it is often the case that we want to perform an official hypothesis test. We should always obtain the same conclusion from the hypothesis test as we do from the confidence interval. To compare two means (μ_1 and μ_2), the hypotheses are

$$H_0: \mu_1 = \mu_2 \text{ vs. } H_a: \mu_1 \neq \mu_2$$

which can also be written

$$H_0: \mu_1 - \mu_2 = 0 \text{ vs. } H_a: \mu_1 - \mu_2 \neq 0$$

The test statistic is $t = \frac{|\hat{\mu}_1 - \hat{\mu}_2|}{\text{SE diff}}$, and the critical value is from the t distribution with degrees of freedom equal to dfE. (This is exactly the same critical value that we have been using). If the test statistic is greater than the critical value, we reject H_0 and conclude the means are significantly different.

For example, we will perform a hypothesis test to decide if the marginal mean for CaCO₃ is equal to the marginal mean for CaCl₂. To simplify the notation, we will use subscript 1 for CaCO₃ and subscript 2 for CaCl₂. The sample (marginal) means are $\hat{\mu}_1 = 752.84$ and $\hat{\mu}_2 = 753.67$. The standard errors of the estimated means are SE₁ = 9.31 and SE₂ = 9.31, so the standard error of the difference is

$$\sqrt{9.31^2 + 9.31^2} \approx 13.17. \text{ The test statistic is } t = \frac{|752.84 - 753.67|}{13.17} \approx 0.063. \text{ Using the critical value 2.179,}$$

we do not reject H_0 (because the test statistic is not greater than the critical value). At significance level 0.05, there is not enough evidence to conclude that the mean temperature at which fabric ignites is different for fabric treated with CaCO₃ versus CaCl₂.

5.4.3. Contrasts

Contrasts for two-way ANOVA are constructed in a manner similar to what we have done in one-way ANOVA. (Refer to Section 4.6. for a discussion of contrasts in one-way ANOVA.) A contrast can be used to compare any linear combination of means, provided the coefficients sum to 0. To illustrate, we will use the fabric data to test if the average temperature for salts CaCO₃ and CaCl₂ at concentration 1 is equal to the average temperature for the same two salts at concentration 2.

Concentration	Treatment Means		
	Untreated	CaCO_3	CaCl_2
1	$\hat{\mu}_{11} = 838.33$	$\hat{\mu}_{12} = 727.00$	$\hat{\mu}_{13} = 723.67$
2	$\hat{\mu}_{21} = 915.00$	$\hat{\mu}_{22} = 778.67$	$\hat{\mu}_{23} = 783.67$

Table 5.15. Treatment means for the fabric data

The treatment means are summarized in Table 5.15.

The average of CaCO_3 and CaCl_2 at concentration 1 is $\frac{1}{2}(\mu_{12} + \mu_{13})$.

The average of CaCO_3 and CaCl_2 at concentration 2 is $\frac{1}{2}(\mu_{22} + \mu_{23})$.

The contrast we want to test is $\frac{1}{2}(\mu_{12} + \mu_{13}) - \frac{1}{2}(\mu_{22} + \mu_{23})$. We need to eliminate the parentheses, so the contrast is written $\frac{1}{2}\mu_{12} + \frac{1}{2}\mu_{13} - \frac{1}{2}\mu_{22} - \frac{1}{2}\mu_{23}$. We will be testing whether or not the contrast is equal to 0. Since the right hand side of the equation is 0, it is easier to clear the fractions (multiply by 2) and write the contrast as $\mu_{12} + \mu_{13} - \mu_{22} - \mu_{23}$.

The point estimate for the contrast is

$$(1)\hat{\mu}_{12} + (1)\hat{\mu}_{13} + (-1)\hat{\mu}_{22} + (-1)\hat{\mu}_{23} = 727.00 + 723.67 - 778.67 - 783.67 = -111.67$$

The standard error of the contrast requires the MSE. From the ANOVA table (Table 5.13), the MSE is 519.89. The standard error is

$$\sqrt{\text{MSE}} \sqrt{\frac{1^2}{3} + \frac{1^2}{3} + \frac{(-1)^2}{3} + \frac{(-1)^2}{3}} = \sqrt{519.89} \sqrt{\frac{4}{3}} = 26.33$$

We are now ready to test the hypotheses $H_0 : \text{contrast} = 0$ vs. $H_a : \text{contrast} \neq 0$

$$\text{The test statistic is } t = \frac{\text{point estimate}}{\text{standard error}} = \frac{-111.67}{26.33} = -4.241.$$

The critical value is 2.179 (from the t distribution with degrees of freedom $\text{dfE} = 12$ and $\alpha/2 = 0.025$).

The absolute value of the test statistic is greater than the test statistic, so we reject H_0 . At significance level 0.05, we conclude that the average temperature for salts CaCO_3 and CaCl_2 at concentration 1 is NOT equal to the average temperature for the same two salts at concentration 2.

5.4.4. Summary

We have performed two similar-sounding hypothesis tests, and reached opposite conclusions. In the first test, we were asking the question: Is there a difference between the average temperature for CaCO_3 and the average temperature for CaCl_2 ? Our conclusion is that there is no difference. In the second test, we were asking the question: Is the average temperature for CaCO_3 and CaCl_2 at concentration 1 equal to the average temperature of CaCO_3 and CaCl_2 at concentration 2? Our conclusion is that there IS a difference.

Both of these hypothesis test involve the same four treatment means: (1) CaCO_3 at concentration 1, (2) CaCO_3 at concentration 2, (3) CaCl_2 at concentration 1, and (4) CaCl_2 at concentration 2. In the first test we are comparing the average of (1) and (2) to the average of (3) and (4). In the second test, we are comparing the average of (1) and (3) to the average of (2) and (4). This apparently minor change in the hypotheses results in a completely different test and a completely different conclusion.

When performing a two-way analysis of variance, remember that the experimental design dictates the model. We do NOT use results derived from the data to construct a model. Always include the interaction term in the model, unless there is a specific reason to exclude it. The most common reason to exclude the interaction term is that there is not enough observations in the data set to estimate the additional parameters in the model. As long as there are at least two observations for each treatment, there will be sufficient data to include the interaction term, but the precision of the estimates will be limited. In general, we like to have at least 3 observations for each parameter in the model.

When analyzing the results of a two-way analysis, examine the results of the interaction test before examining the tests for the main effects. If the interaction is not significant, then we can interpret the main effects. On the other hand, if the interaction is significant, then we must IGNORE the main effects tests. Instead, we compare means of interest individually (for example, by using contrasts). This complication will be addressed in Section 5.6.

There have been a lot of hand calculations in this section. While it is important to understand how these calculations are performed, we will be using software to generate the values we need. The main points you need to understand are

- (1) how to define what you need to answer specific questions regarding a two-way analysis
- (2) how to get software to generate the values you need
- (3) how to interpret the results

Section 5.5. SAS statements for two-way ANOVA

In this section, we define the SAS statements that are required to perform a two-way analysis of variance. We continue to use the fabric data. The basic SAS statements that we used for one-way analysis of variance (in Chapter 4) still apply to two-way analysis, but the additional factor requires additional components in the code.

The most basic code will generate the ANOVA table, but not much else. We continue to use PROC GLM (not PROC REG), because PROC REG does not allow categorical predictors.

```
PROC GLM DATA=fabric;
  CLASS Concentration Salt;
  MODEL Temperature = Concentration | Salt;
  RUN;
```

The CLASS statement identifies the two categorical predictors. SAS will automatically define the indicator variables that are needed for these predictors. Note that we did NOT change the reference level for either of these predictors, so the reference level will be the last one alphabetically (or numerically) for each predictor. Since Concentration has two levels (1 and 2), it has exactly one indicator variable:

$$X_1 = \begin{cases} 1 & \text{if Concentration=1} \\ 0 & \text{otherwise} \end{cases}$$

Salt has three levels (CaCO_3 and CaCl_2 and Untreated), so it has exactly two indicator variables. SAS will put the levels in alphabetical order, so that Untreated is the reference level. The indicator variables are

$$X_2 = \begin{cases} 1 & \text{if Salt = CaCl}_2 \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad X_3 = \begin{cases} 1 & \text{if Salt = CaCO}_3 \\ 0 & \text{otherwise} \end{cases}$$

The MODEL statement defines Temperature as the response, with Concentration and Salt as predictors. SAS will automatically include the main effects for Concentration and Salt. The vertical bar between Concentration and Salt tells SAS to also include the interaction in the model. If we did NOT want to include the interaction, we would simply remove the vertical bar and just leave a space between Concentration and Salt.

The output from this code is shown on the next page. Most of the output will be familiar because it closely resembles the output for a one-way analysis.

The GLM Procedure

Class Level Information		
Class	Levels	Values
Concentration	2	1 2
Salt	3	CaCO3 CaCl2 Untreat

Number of Observations Read	18
Number of Observations Used	18

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	79149.61111	15829.92222	30.45	<.0001
Error	12	6238.66667	519.88889		
Corrected Total	17	85388.27778			

R-Square	Coeff Var	Root MSE	Temperature Mean
0.926938	2.870266	22.80107	794.3889

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Concentration	1	17734.72222	17734.72222	34.11	<.0001
Salt	2	60928.77778	30464.38889	58.60	<.0001
Concentration*Salt	2	486.11111	243.05556	0.47	0.6375

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Concentration	1	17734.72222	17734.72222	34.11	<.0001
Salt	2	60928.77778	30464.38889	58.60	<.0001
Concentration*Salt	2	486.11111	243.05556	0.47	0.6375

(The output also contains an interaction plot, but that will be discussed later.)

The first four tables in this output are similar to what we encountered in one-way ANOVA. The last two tables are new. They provide the partitioned sums of squares, as discussed in Section 5.2. For the fabric data, the last two tables are identical, *but this is only because the data are balanced*. If the data are not balanced, the values in these two tables will be different. We will always use the Type III sums of squares, which are in the last table. To avoid confusion, you can tell SAS to omit the Type I sums of squares table by including the SS3 option on the MODEL statement in the SAS code:

```
MODEL Temperature = Concentration | Salt / SS3;
```

One peculiar aspect of the SAS output is that it does not provide a detailed ANOVA table. Instead it provides a summary ANOVA table that includes a single line for the model:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	79149.61111	15829.92222	30.45	<.0001
Error	12	6238.66667	519.88889		
Corrected Total	17	85388.27778			

The details of the model (i.e., the partitioned sums of squares) are provided in the Type III sums of squares table.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Concentration	1	17734.72222	17734.72222	34.11	<.0001
Salt	2	60928.77778	30464.38889	58.60	<.0001
Concentration*Salt	2	486.11111	243.05556	0.47	0.6375

Note that the detailed table has a total of $1 + 2 + 2 = 5$ degrees of freedom, which matches the degrees of freedom for Model in the summary table. Also note that the total Type III sums of squares in the detailed table ($17734.72222 + 60928.77778 + 486.11111 = 79149.61111$), which matches the sum of squares for Model in the summary table. For simplicity of explanation, we will temporarily round these values to 1 decimal. The total variability in the response (the temperature at which fabric ignites) is 85,388.3. The model explains 79,149.6 of this variability, which is 92.69% (so R-Square = 0.9269). The remaining (unexplained) variability is SSE = 6,238.7. Within the model, the predictor Salt explains most of the variability (60,928.8), but Concentration explains 17,737.7 and the interaction explains only 486.1.

The degrees of freedom for each component in the summary table corresponds to the number of indicator variables for that component, which is always one less than the number of levels for the predictor. Since Concentration has 2 levels, it has one indicator variable (which was defined earlier), and its degrees of freedom is 1. Salt has 3 levels, so it has two indicator variables (also defined earlier), and its degrees of freedom is 2.

The degrees of freedom for the interaction is always the product of the degrees of freedom for the two original factors, which in this dataset is $1 \times 2 = 2$. To see why this is true, we need to understand how the interaction terms are included in the model.

For the main effect of Concentration, the model contains one variable. It is the indicator variable X_1 as defined earlier. For the main effect of Salt, the model contains two variables. These are the indicator variables X_2 and X_3 as defined earlier. To get the interaction terms for the model, each indicator variable for the first factor is multiplied by each indicator variable for the second factor. For the fabric data, this generates two interaction terms: $X_1 \cdot X_2$ and $X_1 \cdot X_3$, so the degrees of freedom for interaction is 2.

In addition to the intercept, every term in the model (including both main effects terms and interaction terms) has a multiplier. The values for these multipliers are usually non-informative, so they are not regularly part of a two-way analysis. This table can be very useful when constructing CONTRAST statements, which is described later in this section. If you want to see this table, include the SOLUTION option on the MODEL statement.

```
MODEL Temperature = Concentration | Salt / SS3 SOLUTION;
```

This will generate the SOLUTION table, as shown below.

Parameter	Estimate		Standard Error	t Value	Pr > t
Intercept	915.0000000	B	13.16420511	69.51	<.0001
Concentration 1	-76.6666667	B	18.61699741	-4.12	0.0014
Concentration 2	0.0000000	B	.	.	.
Salt CaCO3	-136.3333333	B	18.61699741	-7.32	<.0001
Salt CaCl2	-131.3333333	B	18.61699741	-7.05	<.0001
Salt Untreate	0.0000000	B	.	.	.
Concentration*Salt 1 CaCO3	25.0000000	B	26.32841023	0.95	0.3611
Concentration*Salt 1 CaCl2	16.6666667	B	26.32841023	0.63	0.5386
Concentration*Salt 1 Untreate	0.0000000	B	.	.	.
Concentration*Salt 2 CaCO3	0.0000000	B	.	.	.
Concentration*Salt 2 CaCl2	0.0000000	B	.	.	.
Concentration*Salt 2 Untreate	0.0000000	B	.	.	.

The model does not contain a term for any reference level or for any interaction that involves a reference level. There are lines in the SOLUTION table for the reference levels, but their estimated coefficients are all equal to 0 and the other columns are undefined (denoted by a period). It is possible

to construct the treatment means and the marginal means from this table, but this is definitely NOT recommended. Instead, you can use an LSMEANS statement.

```
LSMEANS Concentration Salt Concentration*Salt;
```

This statement will generate the marginal means for Concentration, the marginal means for Salt, and the two-way treatment means for all Concentration*Salt combinations. You can include additional options on the LSMEANS statement to generate standard errors (STDERR) and/or confidence intervals (CL) for the estimated means, and you can also generate pairwise tests of means (PDIFF).

```
LSMEANS Concentration Salt Concentration*Salt / STDERR CL PDIFF;
```

The LSMEANS statement will generate several pages of tables and graphs. Only the tables will be discussed here. The first set of tables are for Concentration.

Concentration	Temperature LSMEAN	Standard Error	H0:LSMEAN=0	H0:LSMean1=LSMean2
			Pr > t	Pr > t
1	763.000000	7.600357	<.0001	<.0001
2	825.777778	7.600357	<.0001	

Concentration	Temperature LSMEAN	95% Confidence Limits	
1	763.000000	746.440244	779.559756
2	825.777778	809.218022	842.337534

Least Squares Means for Effect Concentration					
		Difference Between Means	95% Confidence Limits for LSMean(i)-LSMean(j)		
i	j		-86.196810	-39.358746	
1	2	-62.777778			

The first table provides the estimated means and the standard errors. These values are for the marginal means for Concentration, and they match what we have calculated by hand. The first table also provides the p-values for testing whether or not each marginal mean is equal to 0 (**H0:LSMEAN=0**) and the last column (**H0:LSMean1=LSMean2**) contains the p-value for testing if the marginal means are equal to each other. This last column is included in the output only because we included the PDIFF option in the code.

The second table repeats the values for estimated marginal means, and it provides the confidence limits for the means. The confidence limits are included in the output only because we included the CL option in the code.

The third table provides the point estimate for the difference between the two marginal means, and the confidence limits for the difference.

The next set of tables correspond to the predictor Salt. The means reported in these tables are the marginal means for the three levels of Salt.

Salt	Temperature LSMEAN	Standard Error	Pr > t	LSMEAN Number
CaCO3	752.833333	9.308499	<.0001	1
CaCl2	753.666667	9.308499	<.0001	2
Untreat	876.666667	9.308499	<.0001	3

Least Squares Means for effect Salt Pr > t for H0: LSMean(i)=LSMean(j)			
Dependent Variable: Temperature			
i/j	1	2	3
1		0.9506	<.0001
2	0.9506		<.0001
3	<.0001	<.0001	

Salt	Temperature LSMEAN	95% Confidence Limits	
CaCO3	752.833333	732.551857	773.114810
CaCl2	753.666667	733.385190	773.948143
Untreat	876.666667	856.385190	896.948143

Least Squares Means for Effect Salt				
i	j	Difference Between Means	95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-0.833333	-29.515672	27.849006
1	3	-123.833333	-152.515672	-95.150994
2	3	-123.000000	-151.682339	-94.317661

The first two tables are similar to what we have encountered with one-way analysis. The first table provides point estimates and standard errors for the marginal means for Salt, and it provides the p-values for testing whether each of these means is equal to 0. The LSMEAN Number, in the right-most

column, is used in the second and fourth tables. The second table is the result of the PDIFF option. We interpret this table exactly the way we did in one-way ANOVA. The third table repeats the values for the estimated marginal means, and also provides the confidence limits for these means (option CL in the code). The last table provides the point estimates and confidence limits for the difference of marginal means. The values in the columns marked “i” and “j” correspond to the LSMEAN Numbers in the first table.

The next set of tables correspond to the treatment means. These are included in the output because we included `Concentration*Salt` in the LSMEANS statement. Since there are 6 treatments in this dataset, all of these tables will not fit on one page, so they will be discussed one at a time.

The first table, shown below, provides the point estimates and the standard errors for the treatment means. Note that all the standard errors are equal because this is balanced data. The p-values (in the column `Pr > |t|`) are testing whether or not each mean is equal to 0. These tests are usually not part of the analysis. The last column provides the LSMEAN Number for each treatment, which are used in the tables below.

Concentration	Salt	Temperature LSMEAN	Standard Error	Pr > t 	LSMEAN Number
1	CaCO3	727.000000	13.164205	<.0001	1
1	CaCl2	723.666667	13.164205	<.0001	2
1	Untreat	838.333333	13.164205	<.0001	3
2	CaCO3	778.666667	13.164205	<.0001	4
2	CaCl2	783.666667	13.164205	<.0001	5
2	Untreat	915.000000	13.164205	<.0001	6

The second table, shown below, is the result of the PDIFF option in the code. It provides the p-values for comparing every pair of treatment means. These p-values have not been adjusted for multiple comparisons because we did not include an “ADJUST=” option in the code. We interpret this table exactly the way we did in one-way ANOVA.

Least Squares Means for effect Concentration*Salt Pr > t for H0: LSMean(i)=LSMean(j)						
Dependent Variable: Temperature						
i/j	1	2	3	4	5	6
1		0.8609	<.0001	0.0168	0.0102	<.0001
2	0.8609		<.0001	0.0120	0.0073	<.0001
3	<.0001	<.0001		0.0076	0.0125	0.0014
4	0.0168	0.0120	0.0076		0.7928	<.0001
5	0.0102	0.0073	0.0125	0.7928		<.0001
6	<.0001	<.0001	0.0014	<.0001	<.0001	

The third table, shown below, repeats the point estimates for the treatment means (that were originally reported in the first table), but this table also include the confidence limits for each treatment mean. This table is produced because we included the option CL on the LSMEANS statement.

Concentration	Salt	Temperature LSMEAN	95% Confidence Limits	
1	CaCO ₃	727.000000	698.317661	755.682339
1	CaCl ₂	723.666667	694.984328	752.349006
1	Untreate	838.333333	809.650994	867.015672
2	CaCO ₃	778.666667	749.984328	807.349006
2	CaCl ₂	783.666667	754.984328	812.349006
2	Untreate	915.000000	886.317661	943.682339

The last table provides point estimates and confidence limits for the difference between each pair of treatment means. The values in columns “i” and “j” correspond to the LSMEAN Numbers in the first table. For example, if we compare Salt CaCO₃ to Salt CaCl₂, when both are at concentration 2, we would use LSMEAN numbers 4 and 5. The estimated difference between these two means is -5.00, and the 95% confidence interval for the difference is (-45.56, 35.56). Since the confidence interval contains 0, we would conclude that there is not a significant difference between these two treatment means.

Least Squares Means for Effect Concentration*Salt				
i	j	Difference Between Means	95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	3.333333	-37.229619	43.896286
1	3	-111.333333	-151.896286	-70.770381
1	4	-51.666667	-92.229619	-11.103714
1	5	-56.666667	-97.229619	-16.103714
1	6	-188.000000	-228.562953	-147.437047
2	3	-114.666667	-155.229619	-74.103714
2	4	-55.000000	-95.562953	-14.437047
2	5	-60.000000	-100.562953	-19.437047
2	6	-191.333333	-231.896286	-150.770381
3	4	59.666667	19.103714	100.229619
3	5	54.666667	14.103714	95.229619
3	6	-76.666667	-117.229619	-36.103714
4	5	-5.000000	-45.562953	35.562953
4	6	-136.333333	-176.896286	-95.770381
5	6	-131.333333	-171.896286	-90.770381

The SAS output also contains many graphs, most of which the author finds uninformative. The exception is the interaction plot, which appears as a result of the PROC GLM statement. This plot, shown in Figure 5.3, shows the estimated treatment means. Levels for Concentration are on the x axis and levels for Salt are connected by line segments. If the line segments are not approximately parallel, this provides an indication that interaction may be significant. The graph simply serves as a visual aid. To make a final determination, we would need to interpret the hypothesis test for interaction. The plot in Figure 5.3 has Concentration on the x axis because this is the classification variable that is listed first in the CLASS statement. To put Salt on the x axis, put Salt first in the CLASS statement.

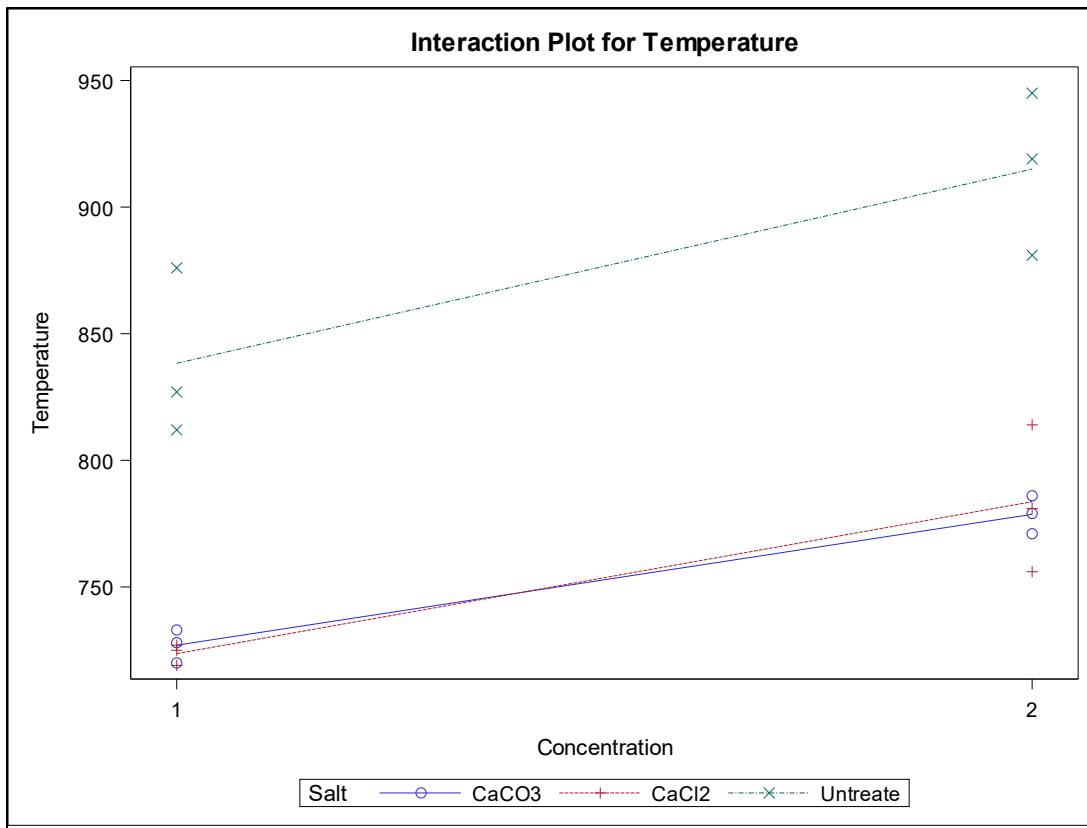


Figure 5.3. Interaction plot for the fabric data

5.5.1. Contrasts in two-way ANOVA

Writing SAS CONTRAST statements for a two-way analysis requires careful attention. Intuitively, we express contrasts in terms of the treatment means (μ_{ij} 's), but SAS requires that these be written in terms of effects (α 's and β 's). To make this adjustment, we will use the fact that

$$\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}.$$

In Section 5.3., we tested the contrast that compared the average temperature for salts CaCO_3 and CaCl_2 at concentration 1 to the average temperature for these two salts at concentration 2. Via hand calculations, we found the test statistic for testing the contrast is 4.241. We now describe how to write the statements to get SAS to perform the calculations.

First, determine the order of the levels that SAS is using. These can be found in the Class Level Information table. This is the first table of the PROC GLM output, and is reproduced in Table 5.16. From this table, we see that Concentration is the first factor, generically referred to as factor A. The subscript $i = 1$ corresponds to concentration level 1, and $i = 2$ corresponds to concentration level 2. Salt is the second factor, generically referred to as factor B. The subscript $j = 1$ is for CaCO_3 , $j = 2$ is for CaCl_2 , and $j = 3$ is for Untreated. The order of the subscripts for the interaction terms is $(i, j) = (1,1), (1,2), (1,3), (2,1), (2,2)$, and $(2,3)$.

Class Level Information		
Class	Levels	Values
Concentration	2	1 2
Salt	3	CaCO_3 CaCl_2 Untreat

Table 5.16. Class Level Information table for the fabric data

The next step is to write the desired contrast using the correct subscripts. The average of CaCO_3 and CaCl_2 at concentration 1 is the average of means μ_{11} and μ_{12} . The average of CaCO_3 and CaCl_2 at concentration 2 is the average of means μ_{21} and μ_{22} . The contrast we want to test is

$\frac{1}{2}(\mu_{11} + \mu_{12}) - \frac{1}{2}(\mu_{21} + \mu_{22}) = 0.5\mu_{11} + 0.5\mu_{12} - 0.5\mu_{21} - 0.5\mu_{22}$. Since we will be testing whether or not the contrast is equal to 0, it is recommended that the fractions be cleared from the contrast, so we will multiply by 2 and use the contrast $\mu_{11} + \mu_{12} - \mu_{21} - \mu_{22}$.

Before we can write the CONTRAST statement in SAS, we need to express the contrast in terms of effects rather than means. This requires that we make the substitution $\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$, and then collect like terms.

$$\begin{aligned}
 & \mu_{11} + \mu_{12} - \mu_{21} - \mu_{22} \\
 &= (\mu + \alpha_1 + \beta_1 + (\alpha\beta)_{11}) + (\mu + \alpha_1 + \beta_2 + (\alpha\beta)_{12}) - (\mu + \alpha_2 + \beta_1 + (\alpha\beta)_{21}) - (\mu + \alpha_2 + \beta_2 + (\alpha\beta)_{22}) \\
 &= \mu + \alpha_1 + \beta_1 + (\alpha\beta)_{11} + \mu + \alpha_1 + \beta_2 + (\alpha\beta)_{12} - \mu - \alpha_2 - \beta_1 - (\alpha\beta)_{21} - \mu - \alpha_2 - \beta_2 - (\alpha\beta)_{22} \\
 &= 2\alpha_1 - 2\alpha_2 + (\alpha\beta)_{11} + (\alpha\beta)_{12} - (\alpha\beta)_{21} - (\alpha\beta)_{22}
 \end{aligned}$$

The subscripts need to match the order identified above, so we must insert placeholders for the missing interaction terms

$$\begin{aligned}
 & \mu_{11} + \mu_{12} - \mu_{21} - \mu_{22} \\
 &= 2\alpha_1 - 2\alpha_2 + (1)(\alpha\beta)_{11} + (1)(\alpha\beta)_{12} + (0)(\alpha\beta)_{13} + (-1)(\alpha\beta)_{21} + (-1)(\alpha\beta)_{22} + (0)(\alpha\beta)_{23}
 \end{aligned}$$

The CONTRAST statement in SAS is

```
CONTRAST 'ContrastName' Concentration 2 -2 Concentration*Salt 1 1 0 -1 -1 0;
```

The result from this statement shown in the table below. The test statistic is $F = 17.99$, with p-value 0.0011. We reject the hypothesis that the contrast equals 0. We conclude that the average temperature for the two salts at concentration 1 is not the same as the average for the two salts at concentration 2.

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
ContrastName	1	9352.083333	9352.083333	17.99	0.0011

When we performed these calculations by hand, we found the test statistic $t = 4.241$. We were using a t test, but SAS performs an equivalent F test. Because the degrees of freedom for the contrast is equal to 1, these tests generate the same p-value and their test statistics are related by $F = t^2$ ($17.99 = 4.241^2$).

Section 5.6. Examples of two-way ANOVA

There are three examples in this section that describe the process of performing a two-way analysis. For two factors A and B, the basic steps of the analysis are

1. Fit the interaction model.
2. Check the model assumptions. If the model assumptions are violated, STOP. No further analysis can be done with this model.
3. Make a two-way table of the means or an interaction plot of the means to help you interpret the results of the analysis.
4. Check the p-value of the interaction to determine if it is significant. The results of this test will dictate what you do next.
 - a. If the interaction is significant, compare treatment means. DO NOT interpret the main effects for either factor. The treatment means can be compared via the PDIFF option or with CONTRAST statements.
 - b. If the interaction is not significant, check the p-values for each of the main effects. If a main effect is significant, compare the marginal means of that factor.
5. Summarize the important point of the analysis. Use the two-way table of means or the interaction plot to help interpret the results of the ANOVA.

5.6.1. Steel springs data

Steel springs are made in large batches. The percentage of good springs in a batch depends on the temperature at which the springs are made and the amount of carbon in the steel. The values of temperature are 1500 and 1600 and the values of carbon are 0.5, 0.6 and 0.7, so there are 6 treatments. The data are shown in Table 5.17.

We fit an interaction model and obtain the residual plot and the normal probability plot. This can be accomplished with the following code.

```
PROC GLM DATA=steelsprings PLOTS=DIAGNOSTICS;
CLASS Carbon Temp;
MODEL Percent = Carbon | Temp;
RUN;
```

Trt	Temp	Carbon	Percent
1	1500	0.5	75
1	1500	0.5	69
1	1500	0.5	70
2	1500	0.6	68
2	1500	0.6	68
2	1500	0.6	67
3	1500	0.7	69
3	1500	0.7	69
3	1500	0.7	64
4	1600	0.5	81
4	1600	0.5	76
4	1600	0.5	82
5	1600	0.6	79
5	1600	0.6	78
5	1600	0.6	78
6	1600	0.7	73
6	1600	0.7	76
6	1600	0.7	74

Table 5.17. Steel springs data

The two diagnostic plots are shown in Figure 5.4. Neither of these graphs show any indication that the assumptions have been violated. To verify the assumption of equal variances, we can fit a one-way model and use the Brown-Forsythe test. (This test is applicable only in one-way designs.) This step is not necessary for this example, but it could be informative if the residual plot showed any irregularities.

To perform the Brown-Forsythe test, we use the following code.

```
PROC GLM DATA = steelsprings PLOTS=NONE;
CLASS Trt;
MODEL Percent = Trt;
MEANS Trt / HOVTEST=BF;
RUN;
```

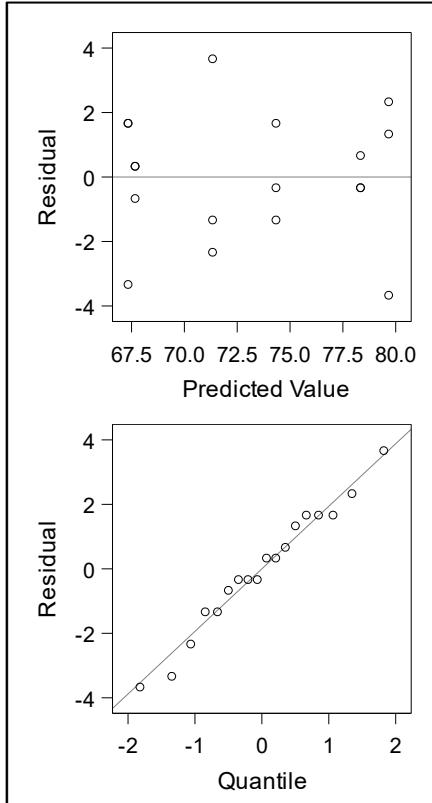


Figure 5.4. Diagnostic plots for steel springs data

The results of the Brown-Forsythe are shown in Table 5.18. As expected, the p-value for this test is 0.8017, so we are confident in assuming the variances are equal.

Brown and Forsythe's Test for Homogeneity of Percent Variance ANOVA of Absolute Deviations from Group Medians					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Trt	5	9.1111	1.8222	0.46	0.8017
Error	12	48.0000	4.0000		

Table 5.18. Brown-Forsythe test for steel springs data

Next, we will get a table of the two-way means. This is accomplished with the following code. PROC MEANS uses the CLASS statement to produce summary statistics for each treatment, as shown in Table 5.19.

```
PROC MEANS DATA=steelsprings;
CLASS Temp Carbon;
VAR Percent;
RUN;
```

Analysis Variable : Percent							
Temp	Carbon	N Obs	N	Mean	Std Dev	Minimum	Maximum
1500	0.5	3	3	71.3333333	3.2145503	69.0000000	75.0000000
	0.6	3	3	67.6666667	0.5773503	67.0000000	68.0000000
	0.7	3	3	67.3333333	2.8867513	64.0000000	69.0000000
1600	0.5	3	3	79.6666667	3.2145503	76.0000000	82.0000000
	0.6	3	3	78.3333333	0.5773503	78.0000000	79.0000000
	0.7	3	3	74.3333333	1.5275252	73.0000000	76.0000000

Table 5.19. PROC MEANS output for steel springs data

This is not the way two-way tables of means is usually presented, so we use this output to manually produce a table in the customary two-way layout, as shown in Table 5.20

		Carbon		
		0.5	0.6	0.7
Temperature	1500	71.3333	67.6667	67.3333
	1600	79.6667	78.3333	74.3333

Table 5.20. Two-way estimated means for steel springs data

These means can also be displayed in an interaction plot. This plot is automatically generated by PROC GLM using the following code. (This code produces all of the output needed to complete the analysis.)

```

PROC GLM DATA=steelsprings PLOTS=DIAGNOSTICS;
CLASS Carbon Temp;
MODEL Percent = Carbon | Temp / SS3;
LSMEANS Temp Carbon Temp*Carbon / PDIFF;
RUN;

```

The interaction plot is shown in Figure 5.5. Note that increasing the amount of Carbon tends to decrease the percentage of good springs, but the amount of decrease is different for the two temperatures. This indicates that there may be a significant interaction, so we will need to check this.

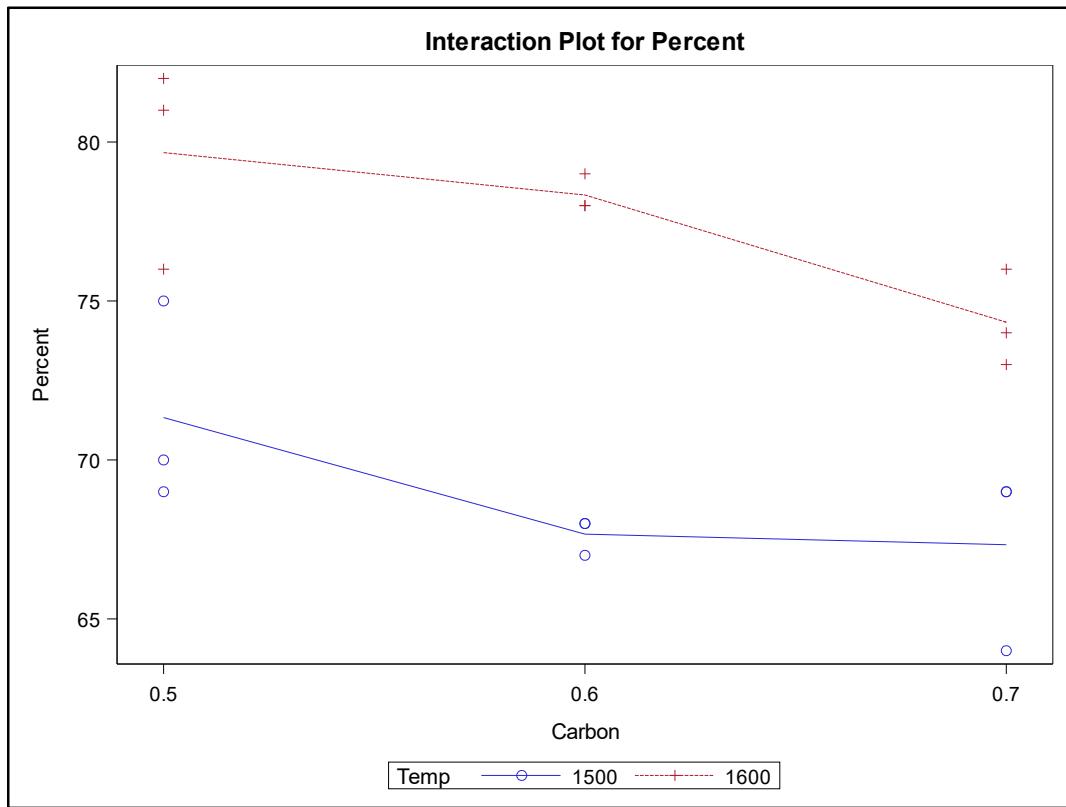


Figure 5.5. Interaction plot for steel springs data

The ANOVA table and Type III sums of squares table are shown in Table 5.21. The test for “Model” has p-value less than 0.0001, so some combination of the factors is affecting the percentage of good springs. The test for interaction has $p = 0.4074$. We conclude the interaction is NOT significant, so we can interpret the main effects.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	413.7777778	82.7555556	15.52	<.0001
Error	12	64.0000000	5.3333333		
Corrected Total	17	477.7777778			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Carbon	2	65.4444444	32.7222222	6.14	0.0146
Temp	1	338.0000000	338.0000000	63.37	<.0001
Carbon*Temp	2	10.3333333	5.1666667	0.97	0.4074

Table 5.21. ANOVA tables for steel springs data

The main effect of Carbon is significant ($F=6.14$, $p=0.0146$), and the main effect of Temp is also significant ($F=63.37$, $p<.0001$). The next step is to compare the marginal means for each of these factors. To do this, we look at the output generated by the LSMEANS statement. We will NOT consider all of the output generated by this statement. We ONLY need to look at the tests that compare the marginal means for Carbon and the tests that compare the marginal means for Temp.

We will not look at the two-way pairwise comparisons because the interaction is not significant.

The LSMEANS output for the marginal means of Temp is shown in Table 5.22. Since there are only two levels for Temp, the result of the PDIFF option is provided in the last column of this table. (There is not a separate matrix of p-values.) Note that this p-value is the same as for Temp in the Type III sums of squares table. Again, this is because there are only two levels for Temp. The estimated marginal mean for Temp 1500 (68.78 percent) is lower than the marginal mean for Temp 1600 (77.44 percent), and we conclude the difference is significant ($p < 0.0001$).

Temp	Percent LSMEAN	H0:LSMean1=LSMean2
		Pr > t
1500	68.7777778	<.0001
1600	77.4444444	

Table 5.22. Comparison of marginal means for Temp

To complete the analysis, we need to look at the marginal means for Carbon. There are three levels for Carbon, so this output in Table 5.23 looks similar to what we saw with one-way ANOVA. The only significant difference is between Carbon 0.5 and Carbon 0.7 (lsmeans 1 and 3), with p-value 0.0044.

Note that the p-values in this table have not been adjusted for multiple comparisons. If we used a Bonferroni adjustment, a test would be significant if its p-value is less than $0.05/3 = 0.0167$. This adjustment does not alter our interpretation of this table.

Carbon	Percent LSMEAN	LSMEAN Number
0.5	75.5000000	1
0.6	73.0000000	2
0.7	70.8333333	3

Least Squares Means for effect Carbon Pr > t for H0: LSMean(i)=LSMean(j)			
Dependent Variable: Percent			
i/j	1	2	3
1		0.0853	0.0044
2	0.0853		0.1301
3	0.0044	0.1301	

Table 5.23. Comparison of marginal means for Carbon

The final step is to summarize the findings. This should be tailored to suit the audience that will be receiving this information, and the description needs to be in the context of the original problem. The author would summarize the analysis this way:

We have found that temperature affects the percentage of good springs in a batch ($p < .0001$), and it appears that a higher temperature produces a higher percentage. For temperature 1500, the expected percentage is 68.8, but for temperature 1600 the expected percentage increases to 77.4. The amount of carbon also affects the percentage of good springs and lower amounts of carbon produce higher percentages. The expected percentages for Carbon 0.5, 0.6 and 0.7 are 75.5, 73.0 and 70.8, respectively. The only significant difference is between Carbon 0.5 and 0.7 ($p = 0.0044$).

SAS code for the steel springs example

```
DATA steelsprings;
  INPUT Trt Temp Carbon Percent;
  DATALINES;
  1 1500 0.5 75
  1 1500 0.5 69
  1 1500 0.5 70
  2 1500 0.6 68
  2 1500 0.6 68
  2 1500 0.6 67
  3 1500 0.7 69
  3 1500 0.7 69
  3 1500 0.7 64
  4 1600 0.5 81
  4 1600 0.5 76
  4 1600 0.5 82
  5 1600 0.6 79
  5 1600 0.6 78
  5 1600 0.6 78
  6 1600 0.7 73
  6 1600 0.7 76
  6 1600 0.7 74
;

PROC MEANS DATA=steelsprings;
  CLASS Temp Carbon;
  VAR Percent;
  RUN;

/* Do a one-way analysis SOLELY to get the Brown-Forsythe test */
TITLE 'ONE-WAY ANALYSIS';
PROC GLM DATA = steelsprings PLOTS=NONE;
  CLASS Trt;
  MODEL Percent = Trt;
  MEANS Trt / HOVTEST=BF;
  RUN;
TITLE ' '; * this clears out the title;

PROC GLM DATA=steelsprings PLOTS=DIAGNOSTICS;
  CLASS Carbon Temp;
  MODEL Percent = Carbon | Temp / SS3;
  LSMEANS Temp Carbon Temp*Carbon / PDIFF;
  RUN;
```

5.6.2. Preservative data

Two different amounts of a preservative were applied to packages of a food product. There were 32 packages in all, 16 that received 100 units of preservative and the other 16 that received 400 units of the preservative. Four of the packages with 100 units of preservative and four of the packages with 400 units of preservative were selected for analysis each week, and this was done for a total of four weeks. Bacterial counts were made on each package. The logarithms of bacterial counts were analyzed. The data are shown in Table 5.24.

		Logarithm of Bacteria Count			
		Week			
Preservative		1	2	3	4
100	2.85	3.81	4.46	5.60	
	2.21	3.45	4.47	5.33	
	3.20	3.20	4.15	5.94	
	3.29	3.60	5.14	5.50	
400	1.93	2.51	2.99	3.44	
	2.82	3.05	3.01	3.50	
	2.93	3.66	3.50	3.25	
	2.80	2.63	3.55	3.15	

Table 5.24. Preservative data

The complete SAS code is shown at the end of this example. Portions of the relevant output will be extracted as we proceed through the analysis.

We fit an interaction model to the data and examined the diagnostic plots. There are no apparent violations of the assumptions, so we continue with the analysis. The mean profile plot, shown in Figure 5.6, indicates the two amounts of preservative produce about the same bacterial counts in the first two weeks, but appear to separate after that. An important question is this: At what week is there a significant difference between the two means?

The test for interaction test has $F = 10.99$, with $p < 0.0001$ (see Table 5.25), so the interaction is significant.

Due to the significant interaction, we cannot interpret the main effects for either week or the amount of preservative. The tests involving marginal means are meaningless in the presence of interaction. Instead, we must look at comparisons of the two-way (i.e., treatment) means.

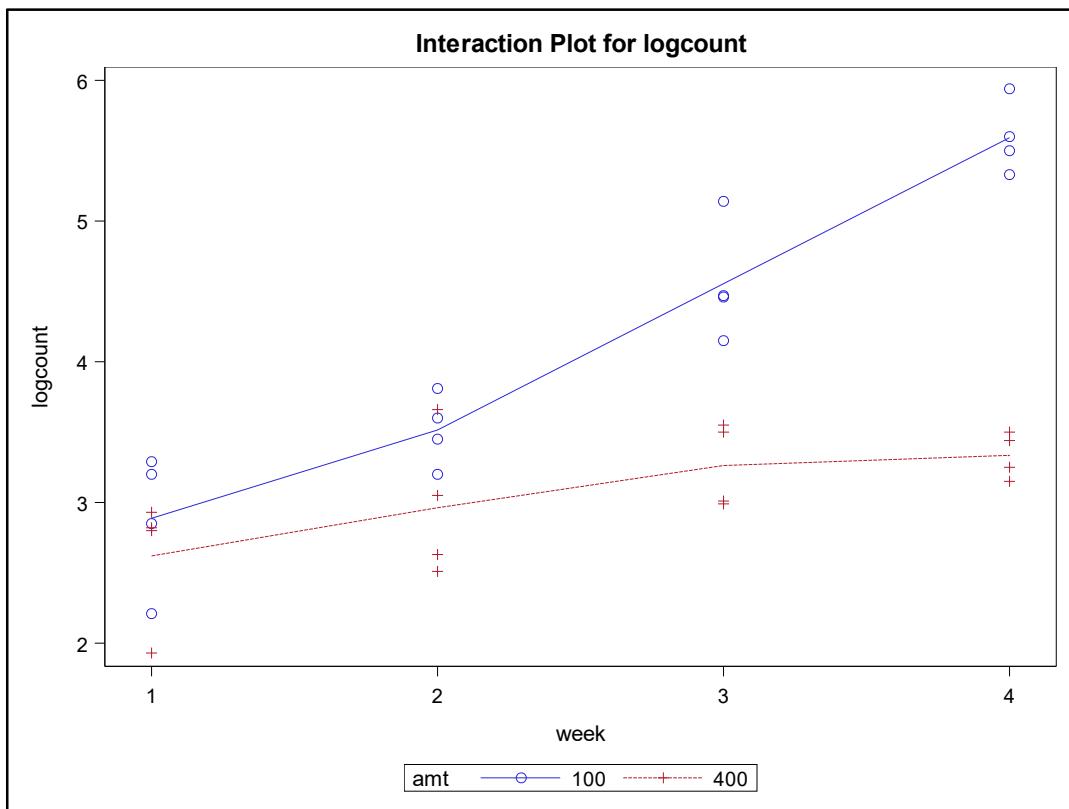


Figure 5.6. Interaction plot for preservative data

Source	DF	Type III SS	Mean Square	F Value	Pr > F
amt	1	9.54845000	9.54845000	66.44	<.0001
week	3	13.50180000	4.50060000	31.31	<.0001
week*amt	3	4.73890000	1.57963333	10.99	<.0001

Table 5.25. Detailed ANOVA table for the preservative data

There are many comparisons that could be made among the treatment means, but we need to focus our attention on tests that provide information about the two preservatives. It would not make any sense to compare preservative 100 at week 1 to preservation 400 at week 4. If a significant difference was found between these two treatment means, we would not know if the difference is due to the difference in preservative or to the difference in weeks. A more logical approach is to compare the two preservatives at each week. This will require four tests, one for each week.

Since we will be looking at 4 post-hoc tests, we need to consider an adjustment for multiple comparisons. If we apply a Bonferroni correction, the level of significance for each of these four tests should be $0.05/4 = 0.0125$. We use SAS to generate the matrix of p-values for all pairwise comparisons, but we do not tell SAS to make an adjustment for multiple comparisons. We will make that adjustment manually.

The two-way means and table of unadjusted p-values for the pairwise comparisons are shown in Table 5.27. We need to extract the results for four tests, and compare the p-values to 0.0125.

1. Is $(\text{amt } 100) = (\text{amt } 400)$ when week = 1?
These are lsmean numbers 1 and 2. $p = 0.3283 > 0.0125$. No significant difference.
2. Is $(\text{amt } 100) = (\text{amt } 400)$ when week = 2?
These are lsmean numbers 3 and 4. $p = 0.0503 > 0.0125$. No significant difference.
3. Is $(\text{amt } 100) = (\text{amt } 400)$ when week = 3?
These are lsmean numbers 5 and 6. $p < .0001 < 0.0125$. There is a significant difference.
4. Is $(\text{amt } 100) = (\text{amt } 400)$ when week = 4?
These are lsmean numbers 7 and 8. $p < .0001 < 0.0125$. There is a significant difference.

week	amt	logcount LSMEAN	LSMEAN Number
1	100	2.88750000	1
1	400	2.62000000	2
2	100	3.51500000	3
2	400	2.96250000	4
3	100	4.55500000	5
3	400	3.26250000	6
4	100	5.59250000	7
4	400	3.33500000	8

Table 5.26. Estimated treatment means

Least Squares Means for effect week*amt Pr > t for H0: LSMean(i)=LSMean(j)									
Dependent Variable: logcount									
i/j	1	2	3	4	5	6	7	8	
1		0.3283	0.0279	0.7820	<.0001	0.1746	<.0001	0.1080	
2	0.3283		0.0027	0.2136	<.0001	0.0247	<.0001	0.0135	
3	0.0279	0.0027		0.0503	0.0007	0.3556	<.0001	0.5083	
4	0.7820	0.2136	0.0503		<.0001	0.2742	<.0001	0.1774	
5	<.0001	<.0001	0.0007	<.0001		<.0001	0.0007	0.0001	
6	0.1746	0.0247	0.3556	0.2742	<.0001		<.0001	0.7891	
7	<.0001	<.0001	<.0001	<.0001	0.0007	<.0001		<.0001	
8	0.1080	0.0135	0.5083	0.1774	0.0001	0.7891	<.0001		

Table 5.27. Comparisons of two-way means for preservative data

This concludes the analysis, but now we need to summarize the results.

The data measures the effectiveness of a preservative, which presumably inhibits bacterial growth. Common sense tells us that the amount of bacteria will increase with time. From the interaction plot, it is clear that the two levels of preservatives have different trajectories over time, but the plot does not indicate if the difference is significant. The interaction test indicates that the difference in trajectories is significant ($F = 10.99$, $p < 0.0001$). If the food product is stored for 1 or 2 weeks, there is not a significant difference in mean log bacteria count between the two amounts of preservative ($p = 0.3283$ and $p = 0.0503$). Significant differences emerge at week 3 and continue to week 4 (both $p < 0.0001$).

To put this in practical terms, if the food product is stored for 2 weeks or less, either 100 or 400 units the preservative will result in (statistically) the same bacterial count. If the food product is stored for longer periods, 400 units of the preservative should be used to inhibit bacterial growth.

SAS code for the preservative example

```
DATA preservatives;
INPUT package amt week logcount @@;
DATALINES;
1 100 1 2.85 2 100 1 2.21 3 100 1 3.20 4 100 1 3.29
5 100 2 3.81 6 100 2 3.45 7 100 2 3.20 8 100 2 3.60
9 100 3 4.46 10 100 3 4.47 11 100 3 4.15 12 100 3 5.14
13 100 4 5.60 14 100 4 5.33 15 100 4 5.94 16 100 4 5.50
17 400 1 1.93 18 400 1 2.82 19 400 1 2.93 20 400 1 2.80
21 400 2 2.51 22 400 2 3.05 23 400 2 3.66 24 400 2 2.63
25 400 3 2.99 26 400 3 3.01 27 400 3 3.50 28 400 3 3.55
29 400 4 3.44 30 400 4 3.50 31 400 4 3.25 32 400 4 3.15
;
PROC GLM DATA=preservatives plots=diagnostics;
CLASS week amt;
MODEL logcount = amt week amt*week / SS3;
LSMEANS amt week amt*week / PDIFF;
RUN;
```

5.6.3. Average daily gain data

An animal scientist compared two diets to see if one is better than the other in terms of enhancing weight gain in cattle. The cattle were divided into two groups of 16 animals (bulls and steers). Half of each group were given diet 1 and the other half were given diet 2. After a fixed period of time, the average daily weight gain (ADG, in pounds) was measured for each animal. The data are shown in Table 5.28.

The purpose of the analysis is to compare the two diets, but it is possible that the diets may work differently for bulls and steers.

Diet 1	Bull	1.87	1.60	1.76	2.13	1.81	2.64	1.33	1.80
	Steer	1.12	3.20	2.72	3.95	2.03	2.29	2.83	2.03
Diet 2	Bull	1.49	2.08	2.29	1.55	1.39	1.71	1.98	1.92
	Steer	2.61	2.67	3.20	2.72	2.40	1.87	2.93	3.52

Table 5.28. Average daily gain data

We fit an interaction model to the data, and produced the diagnostic plots shown in Figure 5.7. The QQ plot looks fine, but the residual plot indicates that there might be a violation of equal variances. Before proceeding with the analysis, we create a new variable in the dataset to identify each of the four treatments: (1, Bull), (2, Bull), (1, Steer) and (2, Steer).

Using the new treatment variable, we fit a one-way model to the data for the sole purpose of generating the Brown-Forsythe test. The p-value for this test is 0.0821, so we conclude that equality of variances has not been violated and we return to the two-way interaction model.

The overall ANOVA F test (in Table 5.29) indicates that there are differences among the treatment means ($F = 5.71$, $p = 0.0035$), so the next step is to evaluate a potential interaction. In the mean profile plot (Figure 5.8), the lines are not perfectly parallel, but the test for interaction indicates the interaction is not significant ($F=0.53$, $p=0.4739$). Therefore, we can examine the main effects for Group and Diet. There is not a significant difference between the two diets ($F=0.15$, $p=0.7007$), but the difference between bulls and steers is significant ($F=16.45$, $p = 0.0004$).

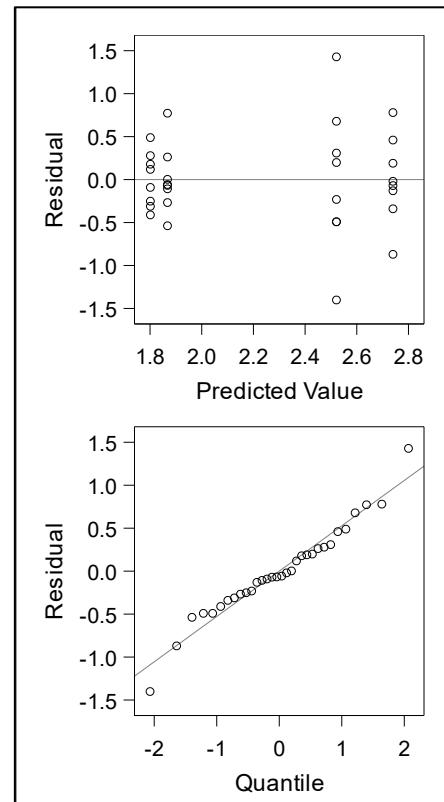


Figure 5.7. Diagnostic plots for ADG data

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	5.28107500	1.76035833	5.71	0.0035
Error	28	8.63292500	0.30831875		
Corrected Total	31	13.91400000			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Group	1	5.07211250	5.07211250	16.45	0.0004
Diet	1	0.04651250	0.04651250	0.15	0.7007
Diet*Group	1	0.16245000	0.16245000	0.53	0.4739

Table 5.29. ANOVA and Type III sums of squares tables for the ADG data

The only significant difference we have detected is between Bulls and Steers. We want to examine this difference more closely. For this, we look at the collection of LSMEANS tables for Group. We have included the options PDIFF and CL to obtain a confidence interval for the difference of these two means, as shown below. The difference is taken as Bull – Steer, and the point estimate is negative, so Bulls gain, on average, 0.796 less than Steers. The 95% confidence interval for the difference is $(-1.198, -0.394)$.

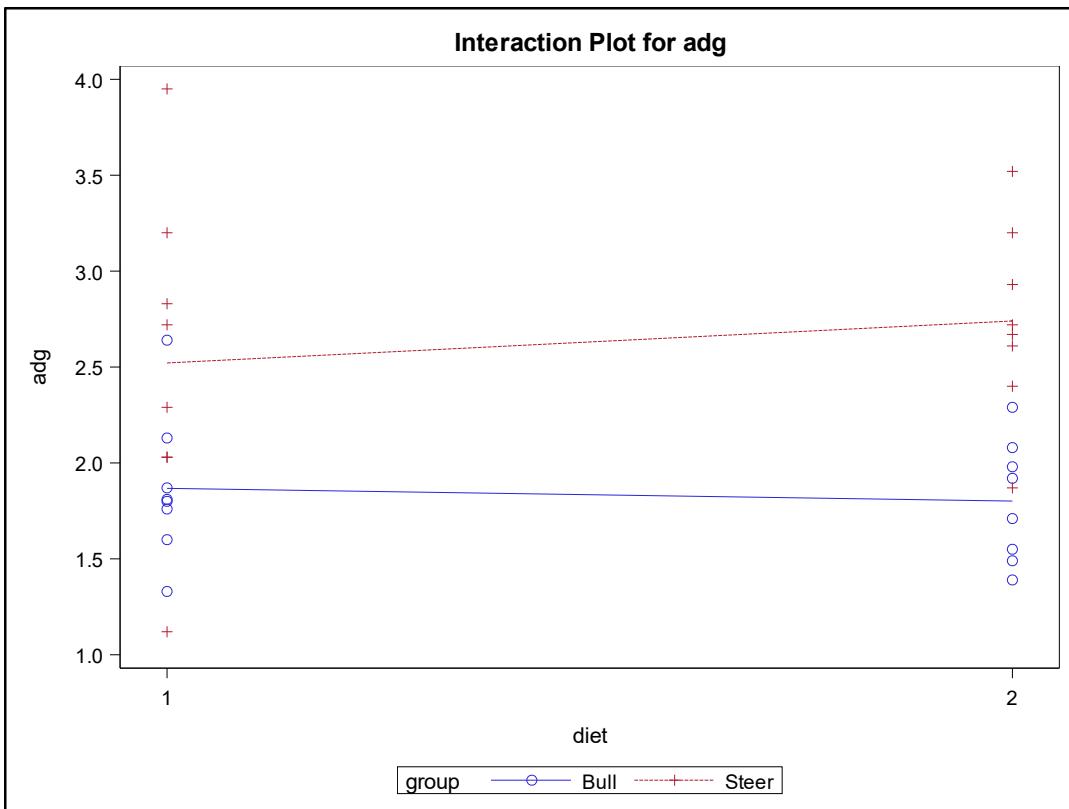


Figure 5.8. Interaction plot for the ADG data

group	adg LSMEAN	H0:LSMean1=LSMean2	
			Pr > t
Bull	1.83437500		0.0004
Steer	2.63062500		

group	adg LSMEAN	95% Confidence Limits	
Bull	1.834375	1.550023	2.118727
Steer	2.630625	2.346273	2.914977

Least Squares Means for Effect group					
i	j	Difference Between Means	95% Confidence Limits for LSMean(i)-LSMean(j)		
1	2	-0.796250	-1.198384	-0.394116	

Table 5.30. LSMEANS tables for Group in the ADG data

We could summarize the analysis this way.

There is not a significant difference in mean average daily weight gain between the two diets ($F = 0.15$, $p = 0.7007$). Regardless of the diet, steers gain, on average, 0.8 pounds more than bulls. We are 95% confident that average daily weight gain for steers is between 0.39 and 1.20 pounds more than bulls.

SAS code for Average Daily Gain data

```
DATA DietData;
INPUT group $ diet adg @@;
/*-----*/
/* The next 4 lines create a new variable (trt) to use in a one-way analysis. */
/* This is needed ONLY because we want to check the Brown-Forsythe test, and */
/* this test is available only for one-way designs. */
/*-----*/
IF group = 'Bull' AND diet = 1 THEN trt = 1;
IF group = 'Bull' AND diet = 2 THEN trt = 2;
IF group = 'Steer' AND diet = 1 THEN trt = 3;
IF group = 'Steer' AND diet = 2 THEN trt = 4;
DATALINES;
Bull 1 1.87 Bull 1 1.60 Bull 1 1.76 Bull 1 2.13
Bull 1 1.81 Bull 1 2.64 Bull 1 1.33 Bull 1 1.80
Steer 1 1.12 Steer 1 3.20 Steer 1 2.72 Steer 1 3.95
Steer 1 2.03 Steer 1 2.29 Steer 1 2.83 Steer 1 2.03
Bull 2 1.49 Bull 2 2.08 Bull 2 2.29 Bull 2 1.55
Bull 2 1.39 Bull 2 1.71 Bull 2 1.98 Bull 2 1.92
Steer 2 2.61 Steer 2 2.67 Steer 2 3.20 Steer 2 2.72
Steer 2 2.40 Steer 2 1.87 Steer 2 2.93 Steer 2 3.52
;
/*-----*/
/* The first GLM is a one-way analysis (using 'trt'). */
/* The only thing we want from this analysis is the */
/* Brown-Forsythe test. */
/*-----*/
PROC GLM DATA=DietData PLOTS=NONE;
CLASS trt;
MODEL adg = trt;
MEANS trt / HOVTEST=BF;
RUN;
/*-----*/
/* This is the interaction model. */
/*-----*/
PROC GLM DATA=DietData PLOTS=DIAGNOSTICS;
CLASS diet group;
MODEL adg = group | diet / SS3;
LSMEANS group diet group*diet / PDIFF CL;
RUN;
```

5.6.4. Summary

We have presented three examples of two-way analysis of variance. In Example 1 (steel springs) and in Example 3 (average daily gain), the interaction was not significant, so we examined the main effects tests. Each main effects test is used to determine if all the marginal means for the factor are equal to each other, or if at least one of the marginal means is different. In Example 1, the factor Temperature had only two levels, so testing if the all the marginal means are equal is the same as testing if the marginal mean for level 1 (Temp = 1500) is the same as the marginal mean for level 2 (Temp = 1600). It would not be necessary to examine LSMEANS to make this decision, because this test is included in the Type III sums of squares table. However, you would need to use LSMEANS if you want confidence intervals for the marginal means or confidence interval for the difference of the marginal means. The other factor in Example 1 was Carbon, and it had 3 levels. The main effects test for Carbon was significant, so the marginal means for the three levels of Carbon are not all the same. Since this factor has 3 levels, we need to look at the output of the LSMEANS statement with the PDIFF option to determine which means are different.

Example 3 (average daily gain) is similar to Example 1 (steel springs) in that the interaction is not significant, so we can interpret the main effects tests. Both of the factors in Example 3 had only two levels, so the LSMEANS statement is needed only to get confidence interval for the means (or the difference of the two means).

Example 2 (preservatives) was different than both Example 1 and Example 3 in that the preservative data had a significant interaction. This greatly complicates the analysis, and the analysis becomes even more complicated if the number of levels is greater. When there is a significant interaction, you must IGNORE the main effects tests. Instead, you have to examine tests that involve the two-way (treatment) means.

There is one more thing about the preservative data (Example 2) that needs to be made clear. The explanation that came with the data made it obvious that there were 32 packages of food product to begin with, and that each of these packages was measured one time to determine the bacteria count. If there had been a slightly different description of the data (and the data itself remained the same), then the analysis we performed could be completely incorrect.

For example, suppose the description of the data collection process indicated that there were 8 packages of food product to begin with, and at the end of each week a small slice was taken from each

package to determine the bacteria count. This process was repeated for four weeks. The values in the dataset for log bacteria count could be exactly the same, and the two factors (weeks and amount of preservative) are exactly the same, but the analysis we did would be entirely incorrect. This is because the bacteria measurements *on the same package* from week to week would be related to each other, simply because they are taken from the same package. This would violate the assumption of independence. All of the models we consider in this course require that the observations be independent. If this is not the case, then more sophisticated models are needed.

Chapter 6: Generalizations

Section 6.1. Three-Way ANOVA

There can be any number of factors in a designed experiment. So far, we have looked at one-factor and two-factor experiments. We now turn our attention to three-factor experiments. As with earlier ANOVA analyses, we restrict our attention to factorial experimental designs in which the treatments are defined by the combinations of each level of each factor with each level of every other factor.

Sometimes, these types of experimental data are referred to by the number of levels of each factor. For example, a $2 \times 2 \times 2$ (or 2^3) design is an experiment with 3 factors and each factor has 2 levels. This produces a total of $2 \times 2 \times 2 = 8$ treatments. A $2 \times 3 \times 4$ design also contains 3 factors, but one factor has 2 levels, another factor has 3 levels and the remaining factor has 4 levels. This produces a total of $2 \times 3 \times 4 = 24$ treatments.

A three-way analysis of variance will be similar to a two-way analysis, but the existence of a third factor increases the complexity of the analysis. The results will still involve multiple F tests and t tests, for both main effects and interactions. After verifying the assumptions, it is extremely important that the interaction tests be considered first. The presence (or absence) of significant interactions will dictate which hypotheses should be considered and which should be ignored.

Studies involving several factors are often focused on identifying how the factors work together to affect the outcome, as opposed to simply comparing the means of the combinations. Questions that are typically asked with ANOVA are:

- Which factors affect the response?
- Can any factor be ignored?
- Do any of the factors interact?
- Which combination of the levels of the factors produce the largest (or smallest) mean?

6.1.1. Water heater example

A study was done to determine factors that may affect the efficiency of a solar water heater. The factors are

- the capacity of the water heater
- the flow rate of the water through the system
- the length of exposure of the solar collector to direct sunlight

Each factor has two levels. Capacity is either 80 or 120 gallons, flow rate is either high or low, and exposure is either 4 or 6 hours. This produces a total of 8 treatments. The response variable is the efficiency of the water heater, where higher values indicate greater efficiency. The experimental design incorporates two replicates for each treatment, for a total of 16 observations.

To get the treatment means, we average the two replications for each treatment. These are called the three-way means, as illustrated in Table 6.1Table 6..

Capacity	Flow	Exposure	Efficiency		Capacity	Flow	Exposure	Three-way means
120	high	6	41.6	⇒	120	high	6	41.45
120	high	6	41.3	⇒	120	high	4	39.80
120	high	4	39.9	⇒	120	low	6	52.15
120	high	4	39.7	⇒	120	low	4	43.95
120	low	6	51.9	⇒	80	high	6	38.80
120	low	6	52.4	⇒	80	high	4	36.25
120	low	4	43.0	⇒	80	low	6	50.75
120	low	4	44.9	⇒	80	low	4	42.40
80	high	6	39.2					
80	high	6	38.4					
80	high	4	37.5					
80	high	4	35.0					
80	low	6	50.2					
80	low	6	51.3					
80	low	4	41.3					
80	low	4	43.5					

Table 6.1. Three-way means for water heater data

In a three-way analysis of variance, there are also two-way means and one-way means. While these are (technically) marginal means as defined in the previous chapter, we rarely call them marginal means when there are more than two factors. This is because there are multiple two-way means and multiple one-way means, so the designation “marginal” mean is not specific enough.

Two-way means are the means for each of the combinations of the levels of two factors when averaged over the levels of the third factor. For a three-factor analysis of variance, this produces three sets of two-way means.

- The two-way means for Capacity*Flow are found by averaging over the levels of Exposure.
- The two-way means for Capacity*Exposure are found by averaging over the levels of Flow.
- The two-way means for Flow*Exposure are found by averaging over the levels of Capacity.

Calculations for the two-way means in the water heater data are shown in Table 6.2.

Capacity*Flow Two-Way Means

Capacity	Flow	Average of		Mean
120	high	41.45	39.80	40.625
120	low	52.15	43.95	48.050
80	high	38.80	36.25	37.525
80	low	50.75	42.40	46.575

Capacity*Exposure Two-Way Means

Capacity	Exposure	Average of		Mean
120	6	41.45	52.15	46.800
120	4	39.80	43.95	41.875
80	6	38.80	50.75	44.775
80	4	36.25	42.40	39.325

Flow*Exposure Two-Way Means

Flow	Exposure	Average of		Mean
high	6	41.45	38.80	40.125
high	4	39.80	36.25	38.025
low	6	52.15	50.75	51.450
low	4	43.95	42.40	43.175

Table 6.2. Two-way means for the water heater data

One-way means are the means for the levels of one factor averaged over the levels of the other two factors.

- One-way means for Capacity are found by averaging over all the levels of Exposure and Flow.
- One-way means for Flow are found by averaging over all the levels of Capacity and Exposure.
- One-way means for Exposure are found by averaging over all the levels of Capacity and Flow.

Calculations of the one-way means for the water heater data shown in Table 6.3.

Factor	Levels	Average of				Mean
Capacity	120	41.45	39.80	52.15	43.95	44.3375
	80	38.80	36.25	50.75	42.40	42.0500

Factor	Levels	Average of				Mean
Flow	high	41.45	39.80	38.80	36.25	39.0750
	low	52.15	43.95	50.75	42.40	47.3125

Factor	Levels	Average of				Mean
Exposure	6	41.45	52.15	38.80	50.75	45.7875
	4	39.80	43.95	36.25	42.40	40.6000

Table 6.3. One-way means for the water heater data

6.1.2. Generic 3-way ANOVA table

Analysis of a three-factor experiment relies on an ANOVA table. The ANOVA table will contain the customary columns for the degrees of freedom, the sums of squares, the mean squares, as well as the test statistics (for the F tests) and the p-values for these tests. Using the generic labels A, B and C to represent the three factors, the rows in the ANOVA table correspond to these effects:

- one row for each of the three main effects: A, B and C
- one row for each of the three two-way interactions: A*B, A*C, and B*C
- one row for the three-way interaction: A*B*C

In addition, there is one row for the error and one row for the total.

For each of the effects (both main effects and interactions), the test statistics and p-values in the ANOVA table are testing specific hypotheses, but there are different ways to interpret them. These are described in Table 6.4.

When interpreting a three-way ANOVA table, we use the significant terms to direct our attention to the means that need to be compared. The general guideline is as follows. If any factor is involved in a significant interaction with another factor, then you must consider the means that involve both factors. If there are no significant interactions among any of the factors, then you should look at the one-way means for each factor. If factor A has a significant interaction with factor B, but A does not interact with C and B does not interact with C, then you should look at the two-way means for A*B and the one-

way means for C. If A interacts with B and B interact with C, then you must consider the three-way means for A*B*C.

Source	Null hypothesis
A	All the one-way means for A are equal. i.e., The effect of A is not significant.
B	All the one-way means for B are equal. i.e., The effect of B is not significant.
A*B	All the two-way means for A*B are equal. i.e., There is not a significant interaction between A and B.
C	All the one-way means for C are equal. i.e., The effect of C is not significant.
A*C	All the two-way means for A*C are equal. i.e., There is not a significant interaction between A and C.
B*C	All the two-way means for B*C are equal. i.e., There is not a significant interaction between B and C.
A*B*C	There is not a significant three-way interaction.

Table 6.4. Null hypotheses for tests reported in a three-way ANOVA table

6.1.3. SAS code for water heater data

The SAS code needed to analyze the water heater data is shown below.

```

DATA heaters;
INPUT Eff Cap Flo $ Exp @@;
DATALINES;
41.6 120 high 6    41.3 120 high 6    39.9 120 high 4    39.7 120 high 4
51.9 120 low 6     52.4 120 low 6     43.0 120 low 4     44.9 120 low 4
39.2  80 high 6    38.4  80 high 6    37.5  80 high 4    35.0  80 high 4
50.2  80 low 6     51.3  80 low 6     41.3  80 low 4     43.5  80 low 4
;
PROC GLM DATA=heaters PLOTS=DIAGNOSTICS;
CLASS Cap Flo Exp;
MODEL Eff = Cap | Flo | Exp / SS3;
LSMEANS Cap | Flo | Exp / PDIFF;
RUN;

```

Most of this code should be familiar. Recall that the dollar sign after Flo in the INPUT statement tells SAS that Flo is a character variable. Including the option PLOTS=DIAGNOSTICS in the PROC GLM

statement ensures that the output will contain the diagnostic plots needed to verify the model assumptions. The vertical bars in the MODEL and LSMEANS statements have not yet been used in any of our examples. These vertical bars instruct SAS to include all the main effects and all the interactions.

Since there are three factors, there are

- three main effects: Cap, Flo and Exp
- three two-way interactions: Cap*Flo, Cap*Exp, and Flo*Exp
- one three-way interaction: Cap*Flo*Exp

To exclude some of the interactions, remove the vertical bars (from both the MODEL and LSMEANS statements) and add the interactions you want to include. For example, to fit a model that contains all the main effects and two-way interactions (but excludes the three-way interaction), these two statements would be

```
MODEL Eff = Cap Flo Exp Cap*Flo Cap*Exp Exp*Cap / SS3;  
LSMEANS Cap Flo Exp Cap*Flo Cap*Exp Exp*Cap / PDIFF;
```

In general, we include all the interactions in the MODEL statement, provided the dataset has sufficient replication. However, the LSMEANS statement with the PDIFF option will generate a lot of output because pairwise comparisons will be done for all 3-way, 2-way and 1-way means. Instead of including all of these in the output, it is recommended to first fit the model (including all interactions), but do not include the LSMEANS statement in the code. Interpret the ANOVA table to determine which means are important, then add the appropriate LSMEANS statement to the code and re-run it.

To interpret the output of the water heater code, the first step is to check the assumptions. For a three-way ANOVA analysis, this is accomplished exactly the same way as in a one-way or two-way analysis. For the water heater data, the residual plot shows no apparent pattern and the QQ plot shows no major departures from normality, so there is no evidence that the assumptions have been violated. (These graphs are not shown here.)

The next step is to examine the Class Level Information table, shown in Table 6.5. As with one-way and two-way data, the last value listed for each classification variable is the reference level for that variable. As always, this part of the output should not be overlooked because mistakes in the data are often revealed here.

The next part of the output is the summary ANOVA table, shown in Table 6.6. This part of the output is equivalent to doing a one-way analysis of variance on the eight treatments. Note that the degrees of

freedom for Model is 7, which is one less than the number of treatments. The significant F test ($F = 59.56$, $p < .0001$) indicates that at least one main effect and/or interaction is significant. If this test had been not significant (i.e., if $p > 0.05$), then the analysis would stop with the conclusion that none of the factors have any effect on the response.

Class Level Information		
Class	Levels	Values
Cap	2	120 80
Flo	2	high low
Exp	2	4 6

Number of Observations Read	16
Number of Observations Used	16

Table 6.5. Class Level Information table for water heater data

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	441.1843750	63.0263393	59.56	<.0001
Error	8	8.4650000	1.0581250		
Corrected Total	15	449.6493750			

Table 6.6. Summary ANOVA table for water heater data

The detailed ANOVA table (Table 6.7) shows that the three-way interaction is not significant ($p=0.7249$), so we can consider the tests for the various two-way interactions.

- The Cap*Flo interaction is not significant ($p=0.1528$).
- The Cap*Exp interaction is not significant ($p=0.6236$).
- The Flo*Exp interaction IS significant ($p=0.0003$)

Since the Flo*Exp interaction is significant, we cannot consider the one-way means for Flo or the one-way means for Exp. In other words, we must ignore the information on the line labeled “Flo” and ignore the information on the line labeled “Exp”. Instead, we must compare the corresponding two-way means, which will be done later in the analysis. Note that Cap is not involved in any significant (Cap*Flo $p=0.1528$, Cap*Exp $p=0.6236$, Cap*Flo*Exp $p=0.7249$), so we need to examine the test for the main effect of Cap. The p-value for this test is 0.0021, which tells us the one-way means for Cap are not all the same. We need to conduct further analysis to determine which levels of Cap have different means.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Cap	1	20.9306250	20.9306250	19.78	0.0021
Flo	1	271.4256250	271.4256250	256.52	<.0001
Cap*Flo	1	2.6406250	2.6406250	2.50	0.1528
Exp	1	107.6406250	107.6406250	101.73	<.0001
Cap*Exp	1	0.2756250	0.2756250	0.26	0.6236
Flo*Exp	1	38.1306250	38.1306250	36.04	0.0003
Cap*Flo*Exp	1	0.1406250	0.1406250	0.13	0.7249

Table 6.7. Detailed ANOVA table for the water heater data

Thus far in the analysis we have interpreted tests for interactions and main effects, as reported in the ANOVA table. These tests tell us which means we need to examine. The output we need in order to continue with the analysis is generated by the LSMEANS statement. We will *not* look at all the output generated by the LSMEANS statement. From our interpretation of the interaction tests, we need to consider only the one-way means for Cap, and the two-way means for Flo*Exp.

The LSMEANS statement as presented in the original code

```
LSMEANS Cap | Flo | Exp / PDIFF;
```

will generate tests that compare the levels for all the one-way means, for all the two-way means and for all the three-way means. Since we are going to use only a small part of all these tests, we can limit the amount of SAS output by modifying this line of code to generate only the tests we are interested in.

```
LSMEANS Cap Flo*Exp / PDIFF;
```

The LSMEANS output for Capacity is shown in Table 6.8. The only reason we are looking at this table is because (1) Cap is not involved in any significant interaction, and (2) the main effect for Cap is significant. If either of these two conditions were different, then we would never look at this table. Since there are only two levels for Cap, the PDIFF option in the code does not produce a table of p-values. Instead, it produces an extra column in this table. The main effect means are 44.3375 for capacity 120 and 42.0500 for capacity 80. Since we know the means are significantly different, we can conclude that capacity 120 water heaters have greater efficiency. Because there is no significant interaction involving Capacity, we can say that capacity 120 water heaters have greater efficiency regardless of the setting for flow rate or length of exposure to direct sunlight.

Cap	Eff LSMEAN	H0:LSMean1=LSMean2	
			Pr > t
120	44.3375000		0.0021
80	42.0500000		

Table 6.8. LSMEANS for one-way means of Capacity

Next, we examine the two-way means for Flo*Exp. The relevant output is shown in Table 6.9.

Flo	Exp	Eff LSMEAN	LSMEAN Number
high	4	38.0250000	1
high	6	40.1250000	2
low	4	43.1750000	3
low	6	51.4500000	4

Least Squares Means for effect Flo*Exp Pr > t for H0: LSMean(i)=LSMean(j)				
Dependent Variable: Eff				
i/j	1	2	3	4
1		0.0203	0.0001	<.0001
2	0.0203		0.0030	<.0001
3	0.0001	0.0030		<.0001
4	<.0001	<.0001	<.0001	

Table 6.9. Two-way means for Flo*Exp in water heater data

We will not consider all the pairwise tests presented in Table 6.9. Instead, we are interested in these specific comparisons:

- (1) Is there a difference between high and low flow rate when exposure is 4?
- (2) Is there a difference between high and low flow rate when exposure is 6?
- (3) Is there a difference between exposure 4 and 6 when the flow rate is low?
- (4) Is there a difference between exposure 4 and 6 when the flow rate is high?

We are excluding tests that compare, for example, high flow and exposure 4 to low flow and exposure 6. (This would be lsmean numbers 1 and 4.) We are not interested in this particular test because its results would not be informative. We can see from Table 6.9 that this test is significant, but it does not tell us if the significant difference in mean efficiency is due to the difference in flow rate or to the difference in exposure (or perhaps both). For a typical three-way analysis, we are trying to isolate the effect of each factor. When dealing with interactions, it is customary to fix the level of one factor (e.g., set the flow rate to low) and compare the levels of the other factor.

Since we are looking at exactly four tests, we can apply a Bonferroni adjustment for multiple comparisons and declare a test is significant if its p-value is less than $0.05/4 = 0.0125$

Note that there is not a significant difference between “high 4” and “high 6” (lsmeans 1 and 2, $p = 0.0203$). We can conclude that, for water heaters with a high flow rate, the exposure to sunlight does not have an effect on mean efficiency. All of the other pairwise tests are significant. We conclude that for water heaters with a low flow rate, the exposure to sunlight does have an effect on mean efficiency (lsmeans 3 and 4, $p < 0.0001$), and that longer exposure results in greater mean efficiency (51.450 to 43.175). For water heaters with exposure 4, the flow rate does have an effect on the mean efficiency (lsmeans1 and 3, $p=0.0001$), and that low flow rate produces greater mean efficiency (43.175 to 38.025). For water heaters with exposure 6, the flow rate does have an effect on the mean efficiency (lsmeans 2 and 4, $p < 0.0001$), and that low flow rate produces greater mean efficiency (51.450 to 40.125).

In summary, we conclude that a large capacity heater has greater efficiency and that a low flow rate is advantageous especially when combined with a solar collector that is exposed for a longer time to direct sunlight.

6.1.4. Alternate analysis for water heater data

The previous analysis of the water heater data presumed that we were interested in determining which factors or combination of factors affect the efficiency. While this is usually the goal of analysis of variance, it is also possible that the main objective of the analysis could simply be to determine which combination of factors produce the greatest mean efficiency. If this is the primary objective, we can employ a simpler approach. Instead of including the main effects and all the interactions in the model, we include only the three-way interaction and then perform pairwise tests on the three-way means to determine what combination produces the greatest mean efficiency. The SAS code this analysis is shown below. (The DATA step is omitted because it has not changed.)

```
PROC GLM DATA=heaters PLOTS=DIAGNOSTICS;
  CLASS Cap Flo Exp;
  MODEL Eff = Cap*Flo*Exp;
  LSMEANS Cap*Flo*Exp / PDIFF ADJUST=TUKEY LINES;
  RUN;
```

The result of the LINES option is shown in Figure 6.1Figure 6.. The mean efficiency for low flow rate and exposure 6 is the statistically the same (52.15 and 50.75), regardless of the capacity of the water heater.

From this graph, we can conclude that exposure 6 and a low flow rate will produce the greatest mean efficiency regardless of the capacity. (*Note: We cannot claim that exposure 6, low flow rate and capacity 120 produces the greatest mean efficiency. There is not a significant difference between the top two treatments, so it is not statistically valid to prefer one of these two treatments over the other.*)

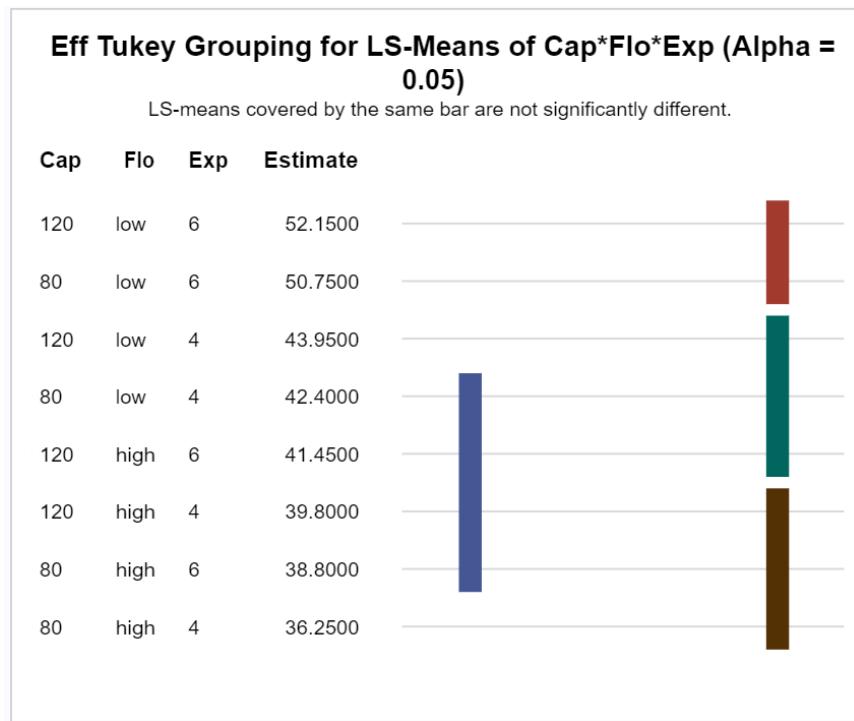


Figure 6.1. Result of LINES option for water heater data

The LINES graph provides a straightforward way to evaluate the relative differences between the various treatments, but it does not provide p-values of the tests used to generate this graph. For the p-values, we need to examine the tabular results of the PDIFF option. These are shown in Table 6.10.

To compare Capacity 120 to 80 for low flow rate and exposure 6, we need to use lsmean numbers 4 and 8. The p-value for this comparison is 0.8521, which is consistent with the “no significant difference” indicated in the LINES chart. The first break in the LINES chart is between “80 low 6” and “120 low 4”. These are lsmean numbers 4 and 7. The p-value for comparing these means is 0.0024, which is consistent with the significant difference indicated in the LINES chart.

Cap	Flo	Exp	Eff LSMEAN	LSMEAN Number
80	high	4	36.2500000	1
80	high	6	38.8000000	2
80	low	4	42.4000000	3
80	low	6	50.7500000	4
120	high	4	39.8000000	5
120	high	6	41.4500000	6
120	low	4	43.9500000	7
120	low	6	52.1500000	8

Least Squares Means for effect Cap*Flo*Exp Pr > t for H0: LSMean(i)=LSMean(j)								
Dependent Variable: Eff								
i/j	1	2	3	4	5	6	7	8
1		0.3194	0.0046	<.0001	0.0954	0.0131	0.0010	<.0001
2	0.3194		0.0896	<.0001	0.9667	0.2850	0.0138	<.0001
3	0.0046	0.0896		0.0006	0.3018	0.9744	0.7868	0.0002
4	<.0001	<.0001	0.0006		<.0001	0.0003	0.0024	0.8521
5	0.0954	0.9667	0.3018	<.0001		0.7389	0.0453	<.0001
6	0.0131	0.2850	0.9744	0.0003	0.7389		0.3377	<.0001
7	0.0010	0.0138	0.7868	0.0024	0.0453	0.3377		0.0007
8	<.0001	<.0001	0.0002	0.8521	<.0001	<.0001	0.0007	

Table 6.10. Pairwise differences of three-way means for the water heater data

6.1.5. Summary

Analysis of three-way ANOVA data is a process. There is not a “one size fit all” approach to performing this kind of analysis. There are many hypothesis tests that must be examined, and there is not a single path through these that can be applied to every data set. Using A, B and C to represent the three factors, the basic steps are:

1. Check the assumptions. If any of the assumptions are clearly violated, stop the analysis.
Transform the response variable and re-fit the model.
2. Examine the overall ANOVA F test. This is the test on the “Model” line in the summary ANOVA table. If this test is not significant (i.e., if the p-value is greater than 0.05), then stop the analysis.
The conclusion is that none of the factors have a significant effect on the response.

3. Examine the 3-way interaction test.
 - a. If this test is significant, then examine the three-way means. Do not consider tests for the two-way interactions or the main effects, and do not consider any tests involving the one-way or the two-way means.
 - b. If this test is not significant, continue with step 4.
4. Examine the tests for the two-way interactions. There will be three of these: one for A*B, one for A*C and one for B*C.
 - a. If a factor is not involved in any significant interaction, then look at the one-way means for this factor.
 - b. If a factor is involved in exactly one significant interaction, then look at the two-way means for the factors that have the significant interaction.
 - c. If a factor is involved in two significant interactions, then look at the three-way means.

General guidelines for “looking at the means”

For one-way means, we want to compare each level of the factor to every other level of the factor. In other words, we want to look at all pairwise differences. This will follow the same procedure that we used with one-way ANOVA, and we should adjust the p-values via Tukey’s method to account for multiple comparisons. For example, if factor A has three levels (a1, a2, a3), and A is not involved in any significant interaction, then we would need to consider three tests:

- Test #1: Is a1 = a2?
- Test #2: Is a1 = a3?
- Test #3: Is a2 = a3?

Because A is not involved in a significant interaction, each test would use the one-way means for A, and the result of each test would apply to all levels of all the other factors.

When we “look at” two-way means, we want to compare two levels of one factor while keeping the level of the other factor fixed. For example, suppose A has 3 levels (a1, a2, a3) and B has 3 levels (b1, b2, b3), and there is a significant A*B interaction. To compare the levels of A, each of the three tests identified above would need to be conducted for each level of B.

- Test #1: Is a1 = a2? becomes three tests:
 - (a) Is a1 = a2 when B=b1? i.e., Is (a1, b1) = (a2, b1)?
 - (b) Is a1 = a2 when B=b2? i.e., Is (a1, b2) = (a2, b2)?
 - (c) Is a1 = a2 when B=b3? i.e., Is (a1, b3) = (a2, b3)?

- Test #2: Is $a_1 = a_3$? becomes three tests:
 - (a) Is $a_1 = a_3$ when $B=b_1$? . . . i.e., Is $(a_1, b_1) = (a_3, b_1)$?
 - (b) Is $a_1 = a_3$ when $B=b_2$? . . . i.e., Is $(a_1, b_2) = (a_3, b_2)$?
 - (c) Is $a_1 = a_3$ when $B=b_3$? . . . i.e., Is $(a_1, b_3) = (a_3, b_3)$?
- Test #3: Is $a_2 = a_3$? becomes three tests:
 - (a) Is $a_2 = a_3$ when $B=b_1$? . . . i.e., Is $(a_2, b_1) = (a_3, b_1)$?
 - (b) Is $a_2 = a_3$ when $B=b_2$? . . . i.e., Is $(a_2, b_2) = (a_3, b_2)$?
 - (c) Is $a_2 = a_3$ when $B=b_3$? . . . i.e., Is $(a_2, b_3) = (a_3, b_3)$?

A similar approach would be used to compare levels of B.

Note that there are many two-way tests that we do not consider. For example, we would not compare the (a_1, b_1) mean to the (a_3, b_3) mean. In special circumstances, there may be a perfectly logical reason to examine this test, but for the purposes of a general three-way analysis of variance this is not a test we typically consider. This is because this test is not very informative. If this test is significant, we will not know if the difference is due to the difference between levels of A, or due to the difference between levels of B, or perhaps both.

Since we are performing multiple tests using the same data, some method of controlling the type I error rate should be employed. Tukey's method is preferred when we are comparing all pairs of means, but in the A*B example, we are looking at 9 tests that compare the levels of A and another 9 tests to compare the levels of B, for a total of 18 tests. The total number of pairwise tests is 36, so we are considering only half of them. Bonferroni's method could be used, so that we would declare a significance difference only if the p-value is less than $0.05/18 \approx 0.0028$. This may be overly strict, so another method (e.g., Scheffé or Tukey) may be preferred.

It is entirely possible that a three-way ANOVA could include additional tests to compare specific means. These additional tests may require the use of a CONTRAST statement. Inclusion of these additional tests is not a standard part of an analysis of variance, because they would be directly tied to a specific research objective.

Section 6.2. Randomization and Blocking

There are two properties that we would like all experiments to have, regardless of the area of study.

1. We would like bias to be small so that the experiment does not unfairly favor one treatment over another.
2. We would like the random variation to be small so that the effects of the treatments can be more clearly seen.

Randomization and blocking are two statistical tools that can be used to reduce bias and reduce the adverse effects of excessive random variability in experimentation.

Experimentation is the process of applying treatments to experimental units for the purpose of measuring the responses. The statistical design of an experiment is, in its most basic form, the plan that determines which experimental units go with which treatments. Randomization and blocking are essential elements in devising the plan for data collection.

Consider an example from agronomy in which we wish to compare four types of fertilizer (treatments) in order to compare their effect on the yields of wheat. Suppose we have a field with 12 plots (numbered 1 through 12) to use as our experimental units, as shown in Figure 6.2.

1	3	5	7	9	11
2	4	6	8	10	12

Figure 6.2. Diagram of agricultural field divided into 12 plots

A simple and straightforward way of administering the fertilizer treatments to the plots would be to assign fertilizer #1 to plots 1,3 and 5, fertilizer #2 to plots 7, 9 and 11, fertilizer #3 to plots 2, 4 and 6, and fertilizer #4 for plots 8, 10, and 12. This would be easy to implement, since each type of fertilizer would be applied to adjacent plots, as shown in Figure 6.3. This is an example of a systematic assignment. It is not random and it has the potential to create insurmountable problems with the statistical analysis. Note that there are 3 plots that receive each treatment, so the design is balanced. This is a good thing, but the treatments were not assigned randomly and this is a bad thing.

1	3	5	7	9	11
----- Treatment = 1 -----					
2	4	6	8	10	12
----- Treatment = 3 -----					

Figure 6.3. A non-random assignment

Suppose, for example, that there was a “bad” spot in the field. For instance, if a fungus were dormant in the soil of several adjacent plots but his was not known at the start of the study, a systematic application of the treatments to the plots could result in most of one treatment being applied to the “bad” plots. As illustrated in Figure 6.4, the “bad” spot (represented by the oval) affects treatment 3 more than treatment 1, and it does not affect treatments 2 or 4 at all. This type of assignment could make treatment 3 look different than the other treatments, when the difference is actually attributable to the differences among the experimental units (the plots) rather than the treatments. One method for combatting this difficulty is to randomly assign the treatments to the experimental units.

1	3	5	7	9	11
----- Treatment = 1 -----					
2	4	6	8	10	12
----- Treatment = 3 -----					

Figure 6.4. A “bad” spot in the field affects some treatments more than others

6.2.1. Completely randomized design (CRD)

In a completely randomized experimental design (abbreviated CRD), the experimental units are randomly assigned to the treatments. For this example, think of putting 12 numbers in a hat and drawing them out one by one. The first three numbers that are drawn would be assigned to treatment #1, the next three to treatment #2, etc. Here is one possible random assignment: Treatment #1 is assigned to plots 4, 6 and 11; Treatment #2 is assigned to plots 2, 5 and 9; Treatment #3 is assigned to

plots 3, 8 and 12; and Treatment #4 is assigned to plots 1, 7 and 10. This assignment is depicted in Figure 6.5.

1	3	5	7	9	11
Trt = 4	Trt = 3	Trt = 2	Trt = 4	Trt = 2	Trt = 1
2	4	6	8	10	12
Trt = 2	Trt = 1	Trt = 1	Trt = 3	Trt = 4	Trt = 3

Figure 6.5. One possible random assignment

Note that this design is also balanced (three plots receive each treatment), but now the “bad” spot that dominates plot 4 and affects plots 2, 3 and 6, is now affecting all three treatments.

The random assignment of treatments to experimental units guards against biases that could occur because of unknown or uncontrollable differences among the experimental units

6.2.2. Randomized complete block (RCB) design

The completely randomized design works best if the experimental units are homogeneous, that is, if they are as alike as possible. If this is not the case, differences among the experimental units can cause the experimental error to be large, which makes it difficult to obtain statistically significant differences among the treatments. Unfortunately, in many situations such as field trials or experiments dealing with people or animals, variability among experimental units is unavoidable. If the differences among the experimental units is known in advance of the study, then a technique such as blocking can be employed to reduce the experimental error.

To explore the concept of blocking, consider an agricultural field that has a slope, so that as we move from west to east (i.e., from left to right in the diagram), the plots tend to have increased moisture due to a gradient (slope) of the land. Since the moisture content of the plot could affect the crop yield (the response), we should incorporate this knowledge into the experimental design. It is possible to ignore the differences in moisture and employ a completely randomized design. This would provide valid results from an analysis of variance, but we can get more precise results if we incorporate knowledge regarding moisture by using a method known as blocking.

To create a block design, we group the experimental units according to similar characteristics. This ensures that experimental units within a block are as alike as possible, but experimental units in different blocks may be different. To apply this procedure to agronomy example, we would group the plots according to their moisture content. This would result in the four western-most plots to be in one block, the four middle plots in another block, and the four eastern-most plots in a third block. Since we are assuming moisture varies from west to east, plots within a block should have similar moisture content. This is illustrated Figure 6.6.

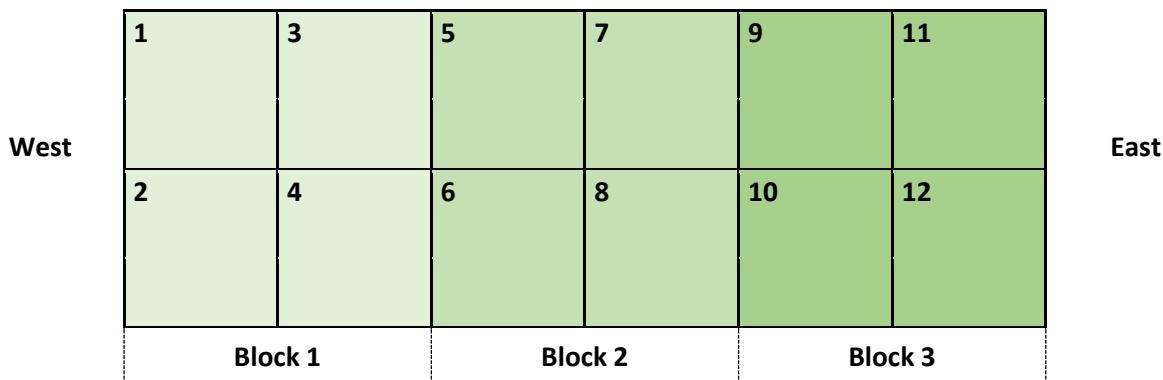


Figure 6.6. Diagram of blocking for agronomy study

When an experimental design incorporates blocks in this manner, we call it a randomized complete block design, abbreviated RCB. Each block has the same number of experimental units as there are treatments, and each treatment must appear exactly once in each block. Within each block, experimental units are randomly assigned to treatments.

One way to achieve a randomized block assignment in the agronomy example is to put four numbers (1 to 4) in a hat and draw them out one by one. These numbers represent the four treatments. For random assignments within Block 1, select numbers from the hat one by one. The first number is assigned to Plot 1, the second number to Plot 2, the third number to Plot 3 and the last number to Plot 4. Now put all the number back into the hat and repeat the process for Block 2. The first number is assigned to Plot 5, the second number to Plot 6, the third number to Plot 7 and the last to Plot 8. Then put the numbers back into the hat and repeat this process for Block 3. One such random assignment is shown in Figure 6.7.

	1	3	5	7	9	11	
West	Trt = 2	Trt = 1	Trt = 3	Trt = 4	Trt = 4	Trt = 1	East
2	4	6	8	10	12		
	Trt = 3	Trt = 4	Trt = 1	Trt = 2	Trt = 3	Trt = 2	
	Block 1		Block 2		Block 3		

Figure 6.7. Randomization for a block design

When blocks are incorporated into the experimental design, the experimental units within blocks are homogeneous. This gives us small within-block random variability. By using ANOVA in the right way, only the variability with blocks will contribute to the MSE, and this allows us to get more precise comparisons among the treatments even though the experimental units as a whole may have a lot of variability among them.

Blocking applies to more than just the physical characteristics of experimental units. It also applies experimental conditions that change over time or place. We block in order to create experimental conditions with each block as consistent as possible. For example, if we are baking bread using three recipes and we have two ovens for doing the baking, we may block on ovens. We would do all three recipes with one oven (block 1) and all three recipes with the other oven (block 2). If the experiment must be done over several days, we can block on days, doing all three recipes each day.

Both the completely randomized design and the randomized block design have factorial treatment structures. This means that the treatments are formed from the combinations of the factors, so that each level of each factor is combined with each level of every other factor. Suppose the fertilizer treatments in our agronomy experiment consist of combinations of nitrogen (at levels 0 or 10) and phosphorus (at levels 0 or 5). We form the four treatments as

- Treatment 1 = (0 nitrogen, 0 phosphorus)
- Treatment 2 = (0 nitrogen, 5 phosphorus)
- Treatment 3 = (10 nitrogen, 0 phosphorus)
- Treatment 4 = (10 nitrogen, 5 phosphorus)

To make the random assignment of treatments to the experiment units (in either a completely randomized design or a randomized block design), we assign the treatments to the experimental units. We do not randomly assign nitrogen levels and phosphorus levels individually.

There are many other types of experimental designs, including Latin squares, incomplete block designs, split-plot designs, strip-plot designs and repeated measures designs. Each type of design requires an adjustment to the basic analysis of variance techniques we discuss in this book. More information about these designs can be found in any textbook or course on experimental design. The completely randomized design is analyzed using the ANOVA tools we have already developed, that is, one-way, two-way and three-way ANOVA, with contrasts and multiple comparisons as appropriate. In the next section, we present an analysis for the completely randomized design.

6.2.3. Analysis of the RCB design

To analyze a randomized complete block design, we use the ANOVA tools we have developed and simply consider Block as one of the factors. For most RCB designs, we will assume that the effect of the blocking factor is additive. In other words, we assume that there is no interaction between the blocks and the treatments. This implies that, in going from one block to another, the average response will either increase the same amount or decrease the same amount as a result of the changing block conditions, regardless of the effects of the treatments. For example, consider the agronomy example we examined in the previous section. Suppose that, as we move from west to east, the soil moisture of the field becomes more favorable for growing wheat. If this causes the yields to increase the same amount (on average) regardless of the effects of the treatments, then the blocking factor is additive.

The mathematical model for an RCB design can be expressed as

The response is equal to . . .
an overall effect . . .
plus an effect due to treatment . . .
plus an effect due to block . . .
plus some random error.

In symbols, the RCB model is written $Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$, where

- Y_{ij} is the response for the i^{th} treatment in the j^{th} block
- μ is the overall mean response (over all EU's and all blocks)
- α_i is the effect of the i^{th} treatment
- β_j is the effect of the j^{th} block

ε_{ij} is the random error for the i^{th} treatment in the j^{th} block

This is simply an additive ANOVA model in which one of the factors is Treatment and the other factor is Block. To implement this model in SAS, you can use this generic code. The treatment factor is denoted ‘trt’ and the blocking factor is ‘blk’.

```
PROC GLM;
  CLASS trt blk;
  MODEL response = trt blk / SS3;
  LSMEANS trt blk / PDIFF;
  RUN;
```

Note that this is the standard code for a two-way analysis of variance, except that the interaction term ($\text{trt} * \text{blk}$) is not included.

It is always recommended that the SAS output be examined for any obvious errors before jumping into a detailed analysis. One of the easy things to check is the degrees of freedom (df). Suppose there are t treatments and b blocks. Because each treatment appears exactly once in each block, the total number of observations is $t * b$. The degrees of freedom (df) are calculated as

- df Total = $t * b - 1$
- df Treatment = $t - 1$
- df Block = $b - 1$
- df Error = (df Total) – (df Treatment) – (df Block) = $(t * b - 1) - (t - 1) - (b - 1) = (t - 1)(b - 1)$

For every ANOVA, we must have a measure of experimental error, that is, we must have an MSE. Because we assume that the RCB has only one observation for each block by treatment combination, we cannot use the observations within blocks to compute MSE. However, because there is no interaction between blocks and treatments, the MS for block by treatment interaction is affected only by random error. Thus we may use the block by treatment mean square as the MSE in an RCB design.

6.2.4. Analysis of the agronomy data

Suppose that the observations in our agronomy example are as depicted in Figure 6.8. The same data is shown in Table 6.11. Notice how much the responses vary in going from west to east for each treatment. The differences (East – West) for each treatment are

- for Treatment 1: $48.3 - 40.5 = 7.8$
- for Treatment 2: $47.0 - 39.4 = 7.6$
- for Treatment 3: $46.2 - 38.3 = 7.9$
- for Treatment 4: $46.1 - 38.1 = 8.0$

The difference is about the same for each treatment. This indicates that an additive model (that excludes a block by treatment interaction) seems appropriate.

		1	3	5	7	9	11	
		Trt = 2	Trt = 1	Trt = 3	Trt = 4	Trt = 4	Trt = 1	
West		39.4	40.5	43.0	42.0	46.1	48.3	East
	2	4	6	8	10	12		
	Trt = 3	Trt = 4	Trt = 1	Trt = 2	Trt = 3	Trt = 2		
	38.3	38.1	45.4	44.1	46.2	47.0		
		Block 1		Block 2		Block 3		

Figure 6.8. Hypothetical data for RCB agronomy example

	Block 1 (West)	Block 2 (Middle)	Block 3 (East)
Trt 1	40.5	45.4	48.3
Trt 2	39.4	44.1	47.0
Trt 3	38.3	43.0	46.2
Trt 4	38.1	42.0	46.1

Table 6.11. Hypothetical data for RCB agronomy example

If we ignore the blocking factor and analyze the data as a one-way completely randomized design, the SAS code would be

```
PROC GLM;
CLASS Treatment;
MODEL Response = Treatment;
RUN;
```

This code produces the ANOVA table shown in Table 6.12. Because the variability from one block to the next has been ignored, it appears that there is not a significant difference between the treatments ($p=0.8442$).

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	12.65	4.22	0.27	0.8442
Error	8	124.09	15.51		
Corrected Total	11	136.74			

Table 6.12. ANOVA table for CRD

When we account for the blocking factor and do an analysis of a block design, the basic SAS code produces the output in Table 6.13.

```
PROC GLM;
CLASS Treatment Block;
MODEL Response = Treatment Block / SS3;
RUN;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Treatment	3	12.65	4.22	46.13	0.0002
Block	2	123.55	61.77	675.93	<.0001
Error	6	0.55	0.09		
Corrected Total	11	136.74			

Table 6.13. ANOVA table for RCB

Note that the degrees of freedom, sum of squares and mean square for Treatment in the randomized block design (3, 12.65 and 4.22, respectively) match those values for Model in the the completely randomized design. The value for Error in the completely randomized design is now split into Block and Error. The total degrees of freedom for Error in the original design is 8, but in the block design this is split into 2 degrees of freedom for Block and the remaining 6 is still in Error. The same is true for the sum of squares for Error. The original sum of squares for Error (124.09) has been split into Block (123.55) and the rest remains in Error (0.55). (Note: 123.55 + 0.55 should equal 124.09. The difference is due to roundoff error.) The additive relationship does NOT continue to the mean square, since the mean square is defined as the sum of squares divided by the degrees of freedom.

Incorporating the blocking factor into the analysis has impacted the analysis in several ways. A large portion of the sum of squares for error has been moved into the sum of squares for block. Since the sum of squares for error is smaller, the mean square for error ($MSE = SSE/dfE$) is also smaller. The MSE is the denominator of the F statistics, so a smaller denominator makes the F statistics larger. Finally, larger F statistics are more likely to produce significant results. To summarize this chain of events, incorporating the blocking factor has increased the value of the F statistics and this makes it more likely to find significant differences. This series of changes in the agronomy data are most evident in the p-value for the test for Treatment. In the original analysis (that ignored blocks) there is no evidence of a treatment effect ($p=0.8442$). When the blocks are incorporated into the analysis, the treatment effect is clearly evident ($p=0.0002$).

The RCB analysis of the agronomy data used the labels 1 through 4 to represent the treatments. In reality, the treatments are comprised of two factors defined by the combinations of Nitrogen (with

levels 0 and 10) and Phosphorus (with levels 0 and 5). To incorporate both factors into the analysis, we replace the variable Treatment with the factors that comprise the treatments, and we need to explicitly include the interaction between the two treatment factors. This is accomplished with the following code.

```
PROC GLM;
CLASS Nitrogen Phosphorus Block;
MODEL Response = Nitrogen Phosphorus Nitrogen*Phosphorus Block / SS3;
RUN;
```

Note that we are NOT including any interaction with Block, since we are assuming there is no interaction between the treatments and the blocks. However, there can be interaction between the two treatment factors (Nitrogen and Phosphorus), but neither of these will interact with Block. The ANOVA table for this model is shown in Table 6.14. Note that “Treatment” in the previous analysis has been split into its component parts: Nitrogen, Phosphorus and the interaction. Each component has 1 degree of freedom, so the degrees of freedom for “Treatment” is 3. The sums of squares are also additive. The sum of squares for “Treatment” in the previous analysis is equal to the sum for Nitrogen, Phosphorus and the interaction ($12.65 = 0.08 + 2.08 + 0.48$; the difference is due to roundoff error). These will always be additive when the design is balanced (i.e., when there is an equal number of observations for each treatment).

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Nitrogen	1	10.08	10.08	110.33	<.0001
Phosphorus	1	2.08	2.08	22.80	0.0031
Nitrogen*Phosphorus	1	0.48	0.48	5.25	0.0618
Block	2	123.55	61.77	675.93	<.0001
Error	6	0.55	0.09		
Corrected Total	11	136.74			

Table 6.14. ANOVA table for two-factor block design for agronomy data

6.2.5. Other cases of blocking

In the previous discussion, we assumed that each treatment was observed exactly once in each block. This is an example of a complete block design. It is sometimes necessary to have fewer experimental units per block than there are treatments. For instance, there may be four treatments but only three experimental units per block. This is an example of an incomplete block design. The analysis of an incomplete block design is very similar to that of a complete block design, but the manner in which the treatments are randomly assigned to the experimental units within each block must be carefully

monitored. The author strongly recommends that you consult your friendly statistician if you plan to use incomplete blocks in a study.

It is also possible to have more experimental units in a block than there are treatments. For instance, there may be eight experimental units per block, but there are only four treatments. In this case, the treatments would need to be randomly assigned to the experimental units within each block so that each treatment occurs twice in each block. This will keep the design balanced, which will allow for more precise inference. The analysis of this type of block design is also very similar to a complete block design.

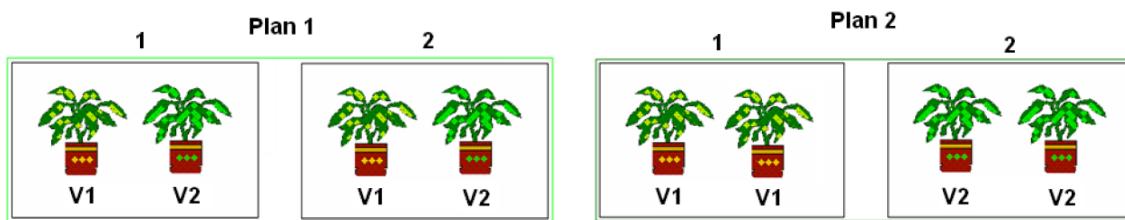
Any time blocking is employed in any experimental design, it is important to distinguish the blocking factor(s) from the treatment factor(s). Blocking factors are created specifically to create groups of experimental units that will share similar characteristics. The levels of a blocking factor define these groups. *The levels of a blocking factor are not randomly assigned.* In general, we are not interested in comparing the levels of a blocking factor. In fact, we expect that the levels of a blocking factor will have different mean responses. This is why we created the blocks in the first place. In Table 6.14, the p-value for blocks is less than 0.0001. This tells us that the mean response for the levels of the blocking factor are not all the same. In other words, *blocking was effective at identifying a source of variation in the response.* If the F test for blocks is not significant, then blocking was not effective.

6.2.6. Example questions

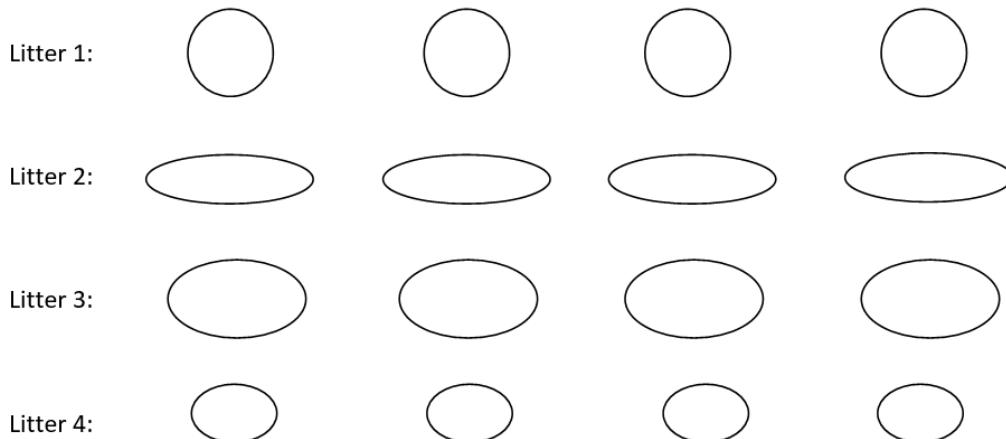
The following questions are provided to help solidify the concepts of experimental design and blocking. The answers to these questions are at the end of this section.

1. A researcher wishes to study the effects of two types of preservatives (BHA, BHT) at two amounts (100, 400) on the growth of bacteria in meats. There are 16 identical samples of meat to which the treatments may be applied.
 - a. How many treatments are there?
 - b. What are the experimental units?
 - c. Show an example of how you would assign the experimental units to the treatments in a completely random design.
 - d. Suppose the samples of meat are hamburger and the samples come from 4 different batches (with 4 samples per batch). How would this affect the randomization you would use?

2. A horticultural scientist wishes to compare the growth of two varieties of tropical plants, labeled V1 and V2. The plants are to be placed in growth chambers where the temperature can be controlled. Two growth chambers are available for the study. There are two plans being considered. In plan 1, one plant of each variety is placed in each growth chamber. In plan 2, two plants of one variety are placed in one growth chamber and two plants of the other variety are placed in the other growth chamber. The growth chambers are made by the same company and are identical. A picture is shown below. Which plan is preferable and why?



3. An engineer is interested in changing the air flow of a commercial freezer to see how this affects the amount of time it takes to freeze a batch of pizzas. The air flow is set to "high" in the morning and 10 batches of pizza are run through the freezer. In the afternoon, the air flow is set at "low" and another 10 batches of pizzas are run through the freezer.
- What are the possible biasing factors that could affect the validity of this experiment?
 - How would you re-design the experiment to mitigate the effect of the biasing factors?
4. An animal scientist wishes to see the effect of four different diets (1, 2, 3 and 4) on the average daily weight gain (ADG) of baby pigs. The experimenter has 16 piglets available for the study. They come from 4 litters with 4 piglets chosen from each litter. It is expected that litter will affect weight gain. A diagram with ovals depicting piglets is shown below.



- a. Show one possible randomization for a randomized complete block design where litter is the block. Write the treatments in the ovals.
- b. Give the degrees of freedom for diet, litter and error.
- c. Using the data provided below, write the SAS code necessary to analyze the data and interpret the results.
- d. Analyze the same data as a one-way ANOVA with diet as the only factor. Compare the results to those in part c. Would you say that including litter as a blocking factor is a good idea for this experiment?

Litter	Diet	ADG
1	1	0.79
1	2	0.75
1	3	0.75
1	4	0.83
2	1	0.74
2	2	0.75
2	3	0.71
2	4	0.80
3	1	0.74
3	2	0.73
3	3	0.79
3	4	0.82
4	1	0.85
4	2	0.81
4	3	0.83
4	4	0.87

6.2.7. Answers to example questions

Answer 1

- a. The 4 treatments are: 1=(BHA, 100), 2=(BHA, 400), 3=(BHT, 100), 4=(BHT, 400)
- b. The experimental units are the samples of meat. (There are 16 EUs in this experiment.)
- c. Put 16 numbers in a hat corresponding to the 16 samples of meat. Draw out 4 numbers to identify the samples that will be assigned to treatment 1, another 4 numbers to identify the samples assigned to treatment 2, etc. Here is one example:

Treatment	Samples			
BHA, 100	4	3	14	12
BHA, 400	10	15	5	16
BHT, 100	2	8	1	11
BHT, 400	13	9	6	7

- d. We would use batch as a blocking factor. We would assign the four samples from each batch randomly to the four treatments. Here is one way this could be done.

Treatment	Sample number			
	Batch 1	Batch 2	Batch 3	Batch 4
BHA, 100	2	1	3	1
BHA, 400	3	4	1	3
BHT, 100	4	3	4	2
BHT, 400	1	2	2	4

Answer 2

Plan 1 is better. If there are any differences between the growth chambers in terms of condition, or where it is placed in the lab, or anything else that could affect how it performs, then having both varieties in each chamber will allow the unique characteristics of each chamber to affect the varieties the same way. For instance, if the temperature knob on growth chamber 1 sets the temperature slightly higher than indicated by the dial, this greater temperature could affect the growth of the varieties within the chamber the same way so that the relative comparison between varieties would be unchanged.

With Plan 2, any differences between the chambers will be completely confounded with the treatments, so if we see a difference in growth between the two varieties, we will not be sure whether to attribute the difference to the varieties themselves or the differences between chambers, or a combination of the two.

Answer 3

Differences in the ambient environment from morning to afternoon could affect the result. For instance, if it is a lot warmer in the afternoon than in the morning, the higher outside temperature might affect how the freezer performs. This effect would completely confound with the effect of air flow, making it impossible to know whether difference in freezing time would be due to the flow, the change in ambient temperature, or both.

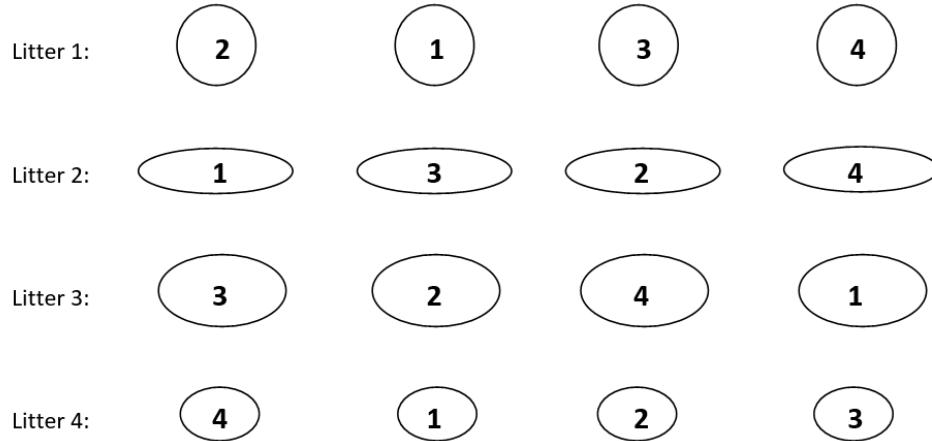
One remedy would be to do 5 batches at high flow and 5 batches at low flow in the morning, and then do the same in the afternoon. Suppose it takes 15 minutes to complete a run. We could form blocks of 30 minutes in which to complete runs of high and low flow as shown below. This would be repeated in the afternoon giving us 10 blocks of two treatments each.

	8:00 AM		8:30 AM		9:00 AM		9:30 AM		10:00 AM	
order	1	2	3	4	5	6	7	8	9	10
flow	low	high	high	low	high	low	low	high	low	high

An engineer might complain that this is too hard to do or too expensive and take a shortcut such as running all low flow first in the morning and all high flow next in the morning, then repeating this pattern in the afternoon. While this would work better than the original plan, it still has the problem that bias could occur since ambient condition change throughout the day.

Answer 4

- a. Here is one possible randomization. The restriction is that each treatment (1 to 4) appears exactly once in each litter.



- b. There are 16 observations, so the total degrees of freedom is 15. There are 4 diets, so df Diet = 3. There are 4 litters, so df Litter = 3. The df Error is $15 - 3 - 3 = 9$. Note that we could also get df Error by realizing that the error mean square is the diet by litter mean square, so df Error = $3 \times 3 = 9$.

c. Here is the SAS code to analyze this data.

```

DATA piglets;
INPUT Litter Diet ADG @@;
DATALINES;
  1 1 0.79    1 2 0.75    1 3 0.75    1 4 0.83
  2 1 0.74    2 2 0.75    2 3 0.71    2 4 0.80
  3 1 0.74    3 2 0.73    3 3 0.79    3 4 0.82
  4 1 0.85    4 2 0.81    4 3 0.83    4 4 0.87
;
PROC GLM DATA=piglets;
  CLASS Litter Diet;
  MODEL ADG = Litter Diet / SS3;
  LSMEANS Diet / PDIFF ADJUST=TUKEY;
  RUN;
PROC GLM DATA=piglets;
  CLASS Diet;
  MODEL ADG = Diet / SS3;
  RUN;

```

The type III ANOVA table is shown below. It shows a significant diet effect ($p=0.0068$) and a significant litter effect ($p=0.0016$). The significant litter effect simply indicates that blocking effective at explaining the excess variation in these data.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Litter	3	0.01800000	0.00600000	12.27	0.0016
Diet	3	0.01160000	0.00386667	7.91	0.0068

Here are the marginal means for diet. Since we are looking at all pairwise tests, the p-values have been adjusted via Tukey's method. We see that diet 4 has significantly bigger mean than the other three diets, but diets 1, 2 and 3 do not differ from each other.

Diet	ADG LSMEAN	LSMEAN Number
1	0.78000000	1
2	0.76000000	2
3	0.77000000	3
4	0.83000000	4

Least Squares Means for effect Diet Pr > t for H0: LSMean(i)=LSMean(j)				
Dependent Variable: ADG				
i/j	1	2	3	4
1		0.5970	0.9165	0.0446
2	0.5970		0.9165	0.0069
3	0.9165	0.9165		0.0172
4	0.0446	0.0069	0.0172	

- d. Here is the analysis of variance for diet alone. We see that this analysis indicates that there is not a significant difference among the diet means. Thus the use of litter as blocking in the RCB analysis in part c was effective in enabling us to see differences among treatments. This conclusion is consistent with the significant litter effect in the RCB analysis.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Diet	3	0.01160000	0.00386667	2.07	0.1576
Error	12	0.02240000	0.00186667		
Corrected Total	15	0.03400000			

Section 6.3. A Random Effects Model

In all of the ANOVA examples we have considered, the random error has been an essential part of the analysis. The variance of the random error is estimated by MSE, and this is used in the denominator of the F statistics used to determine if factors and/or interactions are significant. In some studies, however, randomness can occur in more ways than just through random error. These sources of randomness are called random effects.

When developing a plan for an experiment, researchers often pre-determine the levels of the factors that are going to be used in the experiment. When this occurs, we say that the levels of the factor are “fixed” and their effects are called fixed effects. In some cases, the levels of a factor are selected at random from some larger population, and it is the larger population that the researcher is interested in making inferences about. These effects are called random effects.

The simplest form of a random effects model has a single factor, and this factor is a random effect. Within each level of the random factor, measurements are taken independently on multiple experimental units. There are two questions of interest:

1. Is the variance of the random factor equal to 0 (or is it greater than 0)?
2. Assuming the variance of the random factor is not 0, what is the variance and how does it compare to the variance of the random error?

To explain the concepts of a random effect model, we will use the following example.

6.3.1. Example: Corn chip data

A manufacturer of corn chips receives raw material (corn) in batches. The batches are subdivided and processed to produce bags of corn chips. In terms of consistency of the product, the manufacturer would like to know which contributes more to the variability of final product: the batches or the bags. How would the manufacturer use this information? If the batches are the primary contributor to variability, then they should look for a more reliable supplier of the raw material. On the other hand, if the bags are the primary contributor to variability, then they should look at their own production facility to identify reasons for the variability.

Suppose that the manufacturer randomly selects four batches of raw material and within each batch randomly selects three bags upon which to measure the quality of the product. Each bag is an experimental unit, as shown in Figure 6.9. There are a total of 12 bags, so there will be 12 observations in the data set.

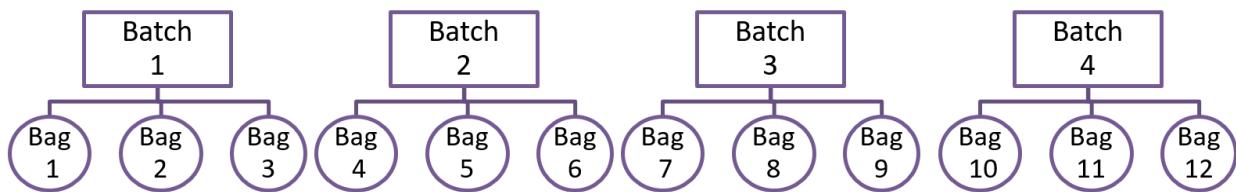


Figure 6.9. Bags and batches of corn chips

The researcher would like to know if there is a significant batch-to-batch variability and if so, how does it compare the variance of the bags within a batch. If the four batches were the only ones the researcher cared about, then this would be a typical one-way ANOVA setup. We would have the effects of the batches as fixed effects and the random variability that occurs in making measurements on the bags (the experimental units) would be the random error. There are two major reasons why this is not a typical ANOVA setup:

- The batches were selected randomly.
- The researcher is interested in the population of ALL batches of raw material, not just the four batches that were selected for this study.

These two conditions force the factor Batch to be a random effect, and the model that incorporates Batch is a random effects model.

When we have an equal number of experimental units for each of the levels of the random factor, the computations for a random effects model are the same as those for a fixed effects model (i.e., as in a one-way ANOVA). In addition, the sources of variability and degrees of freedom are the same. The test statistic for testing whether or not the variance of the random factor is 0 is the same as the test for equality of means when the factor is fixed. This is because, if the means are all the same then the variability between the means is 0. If the means are not all the same, then the variability between the means must be greater than 0.

	Batch 1			Batch 2			Batch 3			Batch 4		
Bag	1	2	3	4	5	6	7	8	9	10	11	12
Quality	90	92	98	84	88	91	89	92	93	75	80	70

Table 6.15. Sample data for corn chip example

Suppose that the measured quality of each bag is as shown in Table 6.15. If we consider the batches to be fixed effects, we can run a typical one-way ANOVA on these data and generate the following ANOVA table.

Source	DF	Sum of Squares	Mean Square	F value	Pr>F
Model	3	609.67	203.22	13.78	0.0016
Error	8	118.00	14.75		
Corrected Total	11	727.67			

Table 6.16. Fixed effects ANOVA table for corn chip data

This analysis shows that the means are not all the same ($p = 0.0016$). In other words, the variance of the population of batches is greater than 0. While this information is important, we would like to have an estimate of the variance of the random factor Batch. To get this, we need to introduce the idea of expected mean squares.

The expected mean square of an MS in the ANOVA table is average value we would get if we could repeat the experiment many times under the same condition. To define these, we need some notation.

Let

MSA = mean square of the random factor A

MSE = mean square for error

σ_A^2 = true variance of the random factor

σ_E^2 = true variance of the error (this is what we have been calling σ^2)

r = number of observations per level of the random factor

With these definitions, the expected mean square for error is σ_E^2 and the expected mean square for the random factor A is Expected MSA = $\sigma_E^2 + r \cdot \sigma_A^2$. Substituting, we have

$$\text{Expected MSA} = (\text{Expected MSE}) + r \cdot \sigma_A^2 ,$$

so that

$$\sigma_A^2 = \frac{\text{Expected MSA} - \text{Expected MSE}}{r}$$

To get an estimate for σ_A^2 , we use the estimated mean squares from the ANOVA table. In the corn chip data, we have $r = 3$, so the estimated variance of the random factor is

$$\hat{\sigma}_A^2 = \frac{203.22 - 14.75}{3} = 62.82 .$$

This is our estimate of the batch-to-batch variability. We compare this to the MSE of 14.75, which is the estimated bag-to-bag variability. Since 62.82 is larger than 14.75, most of the variability in the corn chips is due to the variability among the batches, not the bags. To reduce the variability in the final product, the manufacturer should concentrate on the supplier of the raw material (i.e., the batches).

Note: The true variance of the random factor can never be negative. (No variance can ever be negative.) However, it is possible for the estimated value to be negative. If this occurs, we should adjust the estimate so that it is a reasonable number. There are several options, but the simplest is to change the estimate to 0.

6.3.2. SAS code for random effects

PROC GLM does not accommodate random effects. Instead, we must use PROC MIXED. The syntax for PROC MIXED is very similar to PROC GLM, with these minor modifications:

- The MODEL statement contains only the fixed factors.
- A RANDOM statement identifies the random factors.

PROC GLM does not have a RANDOM statement because it does not allow random factors.

The SAS code for analyzing the corn chip data is as follows.

```
PROC MIXED DATA=cornchips;
  CLASS Batch;
  MODEL Quality = ;
  RANDOM Batch;
  RUN;
```

Note that the MODEL statement does not contain any predictor variables because this example does not have any fixed factors. The MODEL statement is still required because it identifies the response variable.

When there are no fixed effects in the MODEL statement, the MIXED output will produce estimates of the variance components, as shown in Table 6.17. These are the same estimates (within roundoff error) that we calculated by hand using the PROC GLM output.

Covariance Parameter Estimates	
Cov Parm	Estimate
Batch	62.8241
Residual	14.7500

Table 6.17. Estimates of variance components from PROC MIXED

When there is more than one factor, it is possible to have combinations of fixed and random effects. This is called a mixed model, and PROC MIXED can be used to analyze this as well as many other types of experimental designs. Mixed models are discussed in the next section.

6.3.3. Example questions

Question 1

Four elementary school classes have 10 students each. The school administrator obtained the standardized test scores for the four classes and performed an analysis of variance as shown below. Assuming that class effects and student effects are random, which contributes more to the variability of test scores: the classes or the students?

Source	DF	Sum of Squares	Mean Square	F value	Pr > F
Class	3	410.83	136.94	5.28	0.0040
Error	36	933.11	25.92		
Total	39	1343.94			

Question 2

A metal alloy is produced in a high-temperature casting process. Each casting is broken down into smaller individual bars that are used in applications requiring small amounts of the alloy. The tensile strength of the alloy is critical to its intended future use. The casting process is designed to produce bars with an average tensile strength above minimum specifications. Some variation in tensile strength among the bars is acceptable when only a small proportion of bars do not meet specifications. However, excessive variation results in an unacceptable proportion of bars that do not meet specifications.

The data² are given in the file 'TensileStrengthData.txt' and are also shown below. Analyze these data in SAS, using both GLM and MIXED, and calculate the variance components. (Note: Some calculations will be done 'by hand' and some answers can be read from the SAS output.)

Casting											
1	Bar Strength	1	2	3	4	5	6	7	8	9	10
1	Bar Strength	88.0	88.0	94.8	90.0	93.0	89.0	86.0	92.9	89.0	93.0
2	Bar Strength	11	12	13	14	15	16	17	18	19	20
3	Bar Strength	85.9	88.6	90.0	87.1	85.6	86.0	91.0	89.6	93.0	87.5
1	Bar Strength	94.2	91.5	92.0	96.5	95.6	93.8	92.5	93.2	96.2	92.5

² "Design of Experiments: Statistical Principles of Research Design and Analysis", 2nd Edition, by Robert O. Kuehl.

6.3.4. Answers to example questions

Answer 1

We have $MS(class) = 136.94$, $MSE = 25.92$, and $r = 10$. Thus the estimate of the component of variance due to class is $(136.94 - 25.92)/10 = 11.10$. The component of variance due to error (i.e., to students) is 25.92. This tells us that there is more variability among students within a class than there is among the classes. Although these data are simulated, the results are typical of an educational situation. Students typically contribute more to variability in test scores than does the classroom, the variability of which is caused by such factors as teach, time of day in which a class is taught, and location of the classroom. (Note: This example is analogous to the corn chip example. The classes relate to the batches and the students relate to the bags.)

Answer 2

Here is the SAS code for GLM and MIXED.

```
DATA tensile;
  INPUT Casting Bar Strength @@;
  DATALINES;
  1 1 88.0   1 2 88.0   1 3 94.8   1 4 90.0   1 5 93.0
  1 6 89.0   1 7 86.0   1 8 92.9   1 9 89.0   1 10 93.0
  2 11 85.9   2 12 88.6   2 13 90.0   2 14 87.1   2 15 85.6
  2 16 86.0   2 17 91.0   2 18 89.6   2 19 93.0   2 20 87.5
  3 21 94.2   3 22 91.5   3 23 92.0   3 24 96.5   3 25 95.6
  3 26 93.8   3 27 92.5   3 28 93.2   3 29 96.2   3 30 92.5
  ;
PROC GLM DATA=tensile;
  CLASS Casting;
  MODEL Strength = Casting / SS3;
  RUN;
PROC MIXED DATA=tensile;
  CLASS Casting;
  MODEL Strength = ;
  RANDOM CASTING;
  RUN;
```

The ANOVA table from PROC GLM is shown here. Since there is only one factor in this model, the line for Model represents the values for Casting.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	147.8846667	73.9423333	12.71	0.0001
Error	27	157.1020000	5.8185926		
Corrected Total	29	304.9866667			

We see that the means for the factor “Casting” are not all the same because the p-value (0.0001) is less than 0.05. This tells us that the variance of the random effect “Casting” is not equal to 0.

The mean square for Casting is 73.942. The variability of the bars within each casting is what give the variance of error. Its mean square is 5.819. There are $r = 10$ metal bars for each casting. Thus the component of variance for casting is $(73.942 - 5.819)/10 = 6.81$. The contribution of the casting variance (6.81) is a bit larger than the contribution of the error variance (5.816).

The estimates of the variance components from PROC MIXED are shown below. “Residual” in PROC MIXED is the same as “Error” in PROC GLM. The estimates for the variance components are the same as what we calculated by hand from PROC GLM.

Covariance Parameter Estimates	
Cov Parm	Estimate
Casting	6.8124
Residual	5.8186

Section 6.4. Mixed Effects Models

In a two-way ANOVA involving factors A and B, we have assumed that the levels of the factors are fixed by the researcher in advance of the study. We say that the factors are fixed effects. In this section, we consider cases in which the levels of one of the factors are fixed and the levels of the other factor are selected randomly from a population of such levels. This type of situation leads us to what we call mixed effects models.

For example, suppose a medical researcher wishes to compare the recovery times of two medical procedures and selects four hospitals at random from a population of hospitals as locations for doing the study. Each hospital does both procedures three times. It is possible that there is an interaction between the medical procedure and the hospital because not all the hospitals have the same patients or doctors. Thus the hospital effect is not additive as it would be if it were a block.

The hospitals are random effects because they have been selected at random from a larger population, but the medical procedures are fixed effects because they were predetermined for the study. (The procedures do not come from a random sample of all possible procedures.) Because both a fixed effect and a random effect are involved in the study, we will need to analyze this with a mixed effects model.

6.4.1. The mixed effects model

Suppose we have two factors A and B, where A is the fixed effects factor and B is the random effects factor. The mixed effects model looks similar to the two-way ANOVA model with interaction.

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \quad 6.1$$

where μ is the overall mean, α is the effect of factor A, β is the effect of factor B, γ is the interaction, and ε is the random error. In a fixed effects model, α, β , and γ are all fixed effects, and we generate estimates for each of their levels. When the factor B is random, the β 's and the γ 's are assumed to be selected randomly from normally distributed populations, that have mean 0 and variances σ_β^2 and σ_γ^2 , respectively. We still assume that the ε_{ijk} 's are normal, independent and identically distributed with mean 0 and constant variance, but we now denote the common variance as σ_E^2 .

When there are an equal number of observations for every A and B combination, the test statistics for testing the main effect of A, the main effect of B and the interaction are given by

- $F_A = \frac{MSA}{MS(A * B)}$
- $F_B = \frac{MSB}{MS(A * B)}$
- $F_{AB} = \frac{MS(A * B)}{MSE}$

Notice the change in the denominators. In the mixed effects analysis, the denominator of the F statistics for both the A and B main effects is the mean square for the interaction. In a fixed effects analysis, this denominator is the MSE.

For each of these test statistics, the numerator and denominator degrees of freedom for the F distribution are the degrees of freedom associated with the ratio used to calculate the test statistic. These are calculated the same way as if they were all fixed effects. If factor A has 'a' levels and factor B has 'b' levels, then $dfA = a - 1$, $dfB = b - 1$ and $dfAB = (a - 1)(b - 1)$. This results in the following degrees of freedom for each test statistic:

- For F_A : numerator degrees of freedom = dfA ; denominator degrees of freedom = $dfAB$
- For F_B : numerator degrees of freedom = dfB ; denominator degrees of freedom = $dfAB$
- For F_{AB} : numerator degrees of freedom = $dfAB$; denominator degrees of freedom = dfE

Whenever there are random effects in the data, PROC GLM does not calculate the test statistics correctly. This is because PROC GLM consider all effects to be fixed effects, and it uses the denominator MSE to calculate all the test statistics.

6.4.2. Example: Patient recovery times

Consider the patient recovery times shown in Table 6.18. Two procedures (a and b) were specifically chosen for the study, but the four hospitals were selected at random. Each procedure was performed three times at each hospital.

Patient Recovery Times		
Hospital	Procedure a	Procedure b
1	10, 12, 18	15, 20, 23
2	9, 11, 12	13, 15, 20
3	10, 12, 16	24, 26, 28
4	10, 11, 16	24, 28, 30

Table 6.18. Data for patient recovery times

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Hospital	3	163.125	54.375	5.30	0.0099
Procedure	1	590.042	590.042	57.57	<.0001
Hospital*Procedure	3	110.792	36.931	3.60	0.0368
Error	16	164.000	10.250		
Corrected Total	23	1027.958			

Table 6.19. ANOVA table for fixed effects analysis of the patient data

For these data, we fit a fixed effects model using PROC GLM and obtained the ANOVA table shown in Table 6.19. The test statistics and p-values for both Hospital and Procedure have been crossed out because they are incorrect. The test statistic for the interaction term is correct for both a mixed effect analysis and a fixed effects analysis, because they both use MSE in the denominator. Here are the correct test statistics for a mixed effects analysis:

- For Hospital: $F = 54.375 / 36.931 = 1.472$, df = (3, 3)
- For Procedure: $F = 590.042 / 36.931 = 15.977$, df = (1,3)

The correct analysis would have Hospital as a random effect and Procedure as a fixed effect. The interaction between these two factors will be a random effect because **any interaction that involves a random effect is always a random effect**. Here is the SAS code to generate the results we need.

```

PROC MIXED DATA = surgery;
CLASS Hospital Procedure;
MODEL Recovery = Procedure;
RANDOM Hospital Hospital*Procedure;
RUN;

```

Note that Procedure is the only factor in the MODEL statement because this is the only fixed effect.

Both Hospital and Hospital*Procedure are in the RANDOM statement because **any interaction that involves a random effect is always a random effect**.

The relevant parts of the SAS output are shown in Table 6.20. Note that both Hospital and the interaction are in the Covariance Parameter Estimates table because both of these are random effects. Both of these effects are assumed to follow a normal distribution with mean 0. The estimate variance for Hospital is 2.9074, and the estimated variance for the interaction is 8.8935. The estimated variance for the Residual is 10.25, which is the MSE from the fixed effects analysis.

The second table in the PROC MIXED output is for the fixed effects in the model. For the hospital data, the only fixed effect is Procedure. The information for fixed effects that is generated by PROC MIXED is interpreted exactly the way we would interpret the PROC GLM output. The test for Procedure is testing whether or not the two procedures have the same mean recovery times. Since there are only two procedures, the numerator degrees of freedom is $2 - 1 = 1$. The denominator degrees of freedom is the degrees of freedom for the Hospital*Procedure interaction, which is $(2 - 1)(4 - 1) = 3$. Note that the test statistic for Procedure is given as 15.98, as compared to 15.977 that we calculated by hand.

Covariance Parameter Estimates	
Cov Parm	Estimate
Hospital	2.9074
Hospital*Procedure	8.8935
Residual	10.2500

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Procedure	1	3	15.98	0.0281

Table 6.20. SAS output for mixed effects analysis of the patient data

Suppose the researcher asks: “Would these two procedures have different mean recovery times if they were performed across the population of hospitals?” In other words, the researcher is interested in the main effect of Procedure. Since the p-value for this test is 0.0281, we would conclude that the means are different.

The distinction between a mixed effect analysis and fixed effects analysis can sometimes be a bit obscure. In a mixed effects analysis, we are most often interested in the main effect of the fixed factor. We want to know how the fixed factor behaves when averages across the levels of the random factor. In a fixed effects analysis, we are interested in main effects only if there is not a significant interaction. Therefore, testing for interaction is often the most important thing in a fixed effects analysis.

In the hospital example, a regional administrator might like to make a recommendation to the population of hospitals. For this purpose, how well a procedure does “on average” across the population of hospitals would be important. This is what we would get out of a mixed effects analysis. In a fixed effects analysis, we would be interested only in the four hospitals in the study. Of particular importance would be whether or not the recovery time depends on the hospital in which the procedure is performed (i.e., is there an interaction?), and if so, which procedure to recommend for each hospital.

In studies involving blocks, may statisticians consider blocks as random effects. The random effects analysis assumes that the blocks are selected randomly from a hypothetical population of blocks. Think of it this way, the researcher can make inferences about how the treatments compare across a population of such blocks. Unfortunately, the population of blocks is often not well-defined so it is not always clear exactly what population of blocks the results might apply to.

If the design is a randomized complete block, with each treatment occurring exactly once in each block, then the fixed effects analysis and the mixed effects analysis give the same F statistics and p-values for comparing treatment means. This is because the MSE for the randomized complete block design is equal to the mean square for treatment*block. The fixed effect analysis uses MSE in the denominator, while the mixed effects uses MS for treatment*block. Since these are the same, the F statistics will be the same, and hence the p-values will be the same. If the blocks are not complete, or if there are multiple experimental units within each block, then the fixed effects and mixed effects analyses will not give the same test results for the treatment effects.