## Contents

**Instructor’s Guide**  

- Unit A ................................................................. 4  
- Unit B ................................................................. 15  
- Unit C ................................................................. 29  
- Individual Instructor Experiences ................................. 46
**Instructor’s Guide**

**Introduction**

*All models are wrong, but some are useful.* —George Box

If all models are wrong then why did we write a book about modeling? Our emphasis is on “useful.”

We think that STAT1—our label for the usual, one-semester, introductory course in statistics—is a great course that touches on many parts of statistics. We know that the statistical inference part of STAT1 is challenging for many students and a challenge for many of us to teach, so we understand the penchant among instructors to show general ideas (sampling distributions, p-values, confidence intervals, etc.) and methods (t-tests, chi-square tests, etc.) without adding the framework of statistical modeling. Nonetheless, we see modeling as being at the heart of data description and statistical inference. Our goal for this book is to inculcate in students a model(ing) attitude so that it becomes second nature to them to build, assess, and use a wide variety of powerful models as they conduct statistical work.

Professional statisticians are constantly building, critiquing, and using models. Sometimes the use of a model is clear, as in the case of fitting a regression line. At other times a model is implied but may not be openly stated, as when a two-sample t-test indicates that one population mean is greater than another. When we ask “How much greater?” “How (un)certain are the results of our predictions?”, or “What other variables are important?”, we find that explicit modeling is an invaluable tool.

We want to open up the student to the world of statistics by having them see that world through the lens of modeling. In this book we show students how to build models and use them, how to compare models, and how to assess whether a model can be trusted. The main units of the book give the student exposure to three commonly used types of models—regression, analysis of variance, and logistic regression—but our ultimate goal is not limited to these classes of models. By studying modeling techniques, by practicing the use of models, by comparing models, and by becoming comfortable with statistical computing we hope that students will feel empowered to build models of many kinds.

In each chapter we provide guidance on general approaches to the analysis of differing kinds of data. We present the paradigm Choose, Fit, Assess, Use as a way to give structure to the analysis of data, and we point to helpful diagnostic methods that steer us away from foolishness, but we don’t want to be overly prescriptive. Any interesting dataset can be modeled in more than one way. The goal is not to find the model but to find a useful model, to know what it tells us and what it cannot tell us, to use the model to better understand the world and to make predictions, but never to be satisfied that...
we have found eternal truth and can summarize in one small space all that there is to know about a scientific problem. Two statisticians shown the same set of data will usually construct models that agree in their main features but not in every detail. That’s OK; indeed, that is good: When the two statisticians compare their models they both come away knowing more than they learned on their own.

This is not to say that all models are equally good; George Box would agree. Some models are hopelessly flawed, whereas others can be believed only up to a point. Some models are more valuable than others and some are very powerful and illuminating, but no model is perfect. Thus, as you set out to teach a course on statistical modeling, be open to a range of possibilities. Be willing to explore and to accept more than one “right answer” to a modeling question. Let the textbook provide guidance and advice, but don’t be constrained by the vision of the authors. (We would say the same thing to you no matter what statistical modeling book you were using, but we know that this advice applies to our book in particular.) We think that this book provides a good introduction to statistical modeling but (like any model) the book is useful, not perfect. Our goal is to help your students develop good skills, intuition, and habits, not to give them prescriptive advice that will allow them to automate their work. We lean heavily on computing in this book, which could not have been written if there didn’t exist powerful and flexible statistics software, but we have no fear of being replaced by computers because statistical modeling is a blend of science and art. Good statistical work involves creativity, flexibility, interpretation of graphs, and communication of results, so we have no fear that computers will replace us.

We expect that most instructors have experience working with and teaching regression, so for Unit A we provide rather brief, to-the-point advice in this instructor’s guide. Many have less experience with analysis of variance than with regression, so for Unit B we offer a bit more overall guidance. We expect that instructors have limited experience teaching logistic regression and will benefit from our more expansive comments on that topic in Unit C.

As for teaching the units, we expect everyone to start with Chapter 0, but after that there are many paths through the book. We present regression (Unit A) first, and there are chapters later in the book that assume coverage of Chapters 1 and 2, so we expect instructors to start with Chapters 0, 1, and 2. Many will continue with the rest of Unit A and then move to Unit B, so that they present the two units that handle a quantitative response together, saving logistic regression, the most challenging unit, for the end of the course.

However appealing alphabetical order might be, it is not necessary to adhere to the ABC arrangement of main topics. One alternative that some of us have used with success is to move from Unit A, regression with a quantitative response, to Unit C, regression with a binary response. Students with interests in mathematics, economics, or political science might enjoy this sequencing, but it can be used with any group of students.
A third option that might appeal to students in psychology or related fields is to cover Unit B after Chapter 2, returning to Chapter 3 (multiple regression) after a thorough exploration of ANOVA. Analysis of variance can be taught to students without any background in regression, and we have written Unit B in a way that allows for this option, although we think that the use of statistical models in the first part of Unit A (Chapters 1 and 2) is good preparation for Unit B.

One other option would be to cover Chapter 0 followed by the Chapters 1, 2, 5, and 9, studying regression, ANOVA, and logistic regression when there is a single predictor variable before dealing with multiple predictors.

Chapter 0—What Is a Statistical Model?

The label of “Chapter 0” might make you think that this is just a review of prerequisite material, but that is not the case. There are two purposes for this chapter, the first being to get everyone “on the same page” regarding terminology, notation, and the like. However, the second function is more important: We use Chapter 0 to introduce the use of statistical models. We expect that many students will have seen a simple regression model in STAT1 but few, if any, of them will have thought about the t-test as an application of modeling. Thus, we use the t-test setting, which should be familiar, to introduce fundamental ideas.

We find it helpful to start the course with a simple example of data collected from the students in the class, but the weight loss data of Example 0.3 could be used for class discussion. The important thing is to get the students talking about parameters vs. statistics, experimental or observational units, explanatory vs. response variables, and other terminology in Section 0.1. A written survey of the students to learn about their backgrounds is helpful, but discussion in class can be even more illuminating.

Later in the course students will be seeing material that is entirely new to them, but Section 0.2 provides a chance to solidify understanding of things they should know from STAT1 under the guise of being shown something new—the modeling approach that pervades the book. Here we introduce the four-step process of Choose (a form of the model), Fit (the model to the data), Assess (how well the model describes the data), and Use (the model to answer questions and make predictions). There are other ways to think about organizing one’s work, and we don’t want students to worry too much about whether they are in the Assess step or are in the Use step, for example, but we do want to provide some structure to the student’s work; we hope that the four-step process achieves that. In Example 0.6 we begin with a graph, which should always be the starting point of an analysis, and work our way through the four steps.
Instructor’s Guide - 4

Unit A

Unit A presents regression (i.e., ordinary, least-squares linear regression). We start with regression for many reasons. First, many students will have been introduced to the idea of regression in STAT1, at least as a descriptive tool. In contrast, most will not have seen analysis of variance (Unit B) and even fewer, if any, will have seen logistic regression (Unit C).

Just as important is the fact that regression is the one topic from STAT1 in which students may have seen the use of models, including an explicit error term. This means that students will be more likely to embrace the general modeling approach to data analysis that we are advocating throughout this text.

Finally, we note that teaching both logistic regression and ANOVA after linear regression has several benefits. In particular, logistic regression builds naturally on linear regression, with plentiful parallels in structure and process so that a good foundation in least squares regression prepares the student for studying logistic regression. We rely quite heavily on this parallel in our presentation of logistic regression in Unit C. And students generally find the modeling approach to data analysis more natural in the regression setting than in the analysis of variance setting, so waiting to introduce ANOVA until after they are comfortable with the idea of modeling is helpful. ANOVA can also be tied to regression modeling via the use of indicator/dummy variables, so study of regression before ANOVA can have some advantages as well.

Like all units in this book, we start with the simple case before proceeding to the multivariate case. This is followed by a chapter of interesting topics that you may wish to cover, but are not required for any other material in the course. Specifically, we break up the simple case into two chapters. In Chapter 1 we introduce the idea of the simple linear regression model along with the conditions required for the model. Chapter 2 continues the discussion of the simple linear model and addresses inference. Multiple regression is taken up in Chapter 3. The “additional topics” in Chapter 4 are optional and can be covered in any order. Indeed, topics from Chapter 4 can be sprinkled into Chapters 1–3, rather than being put off until the end of the unit.

Chapter 1—Simple Linear Regression

We call this chapter Simple Linear Regression, but that doesn’t mean that students will find it to be simple. We hope that many of the ideas here will be at least somewhat familiar to students from STAT1 but we do not assume such familiarity. Moreover, some students will have seen different notation or terminology, so there is value in working through everything carefully even for those for whom it is largely review. In fact, it is likely that this is the one area of the course where you may have students with varying degrees of background knowledge. We have found that it is important to
present the material in such a way as to level the playing field of all students—
introducing the ideas carefully to those who haven’t seen regression before, while
keeping those student who have seen it engaged. It will be easy for that latter group of
students to decide that they already know the material and therefore not pay attention,
but, in fact, there are subtleties that are new and important.

Section 1.1—The Simple Linear Regression Model. The primary example in this chapter
relates the price of a used Porsche to the number of miles on the car. In the Choose step
of modeling we make a scatterplot and see what we expect to see: an inverse
relationship with a clear linear trend. We define terms such as “residual” and “sum of
squared residuals” before proceeding to the Fit step, where we rely on technology to
find the equation of the least squares line.

Issues to keep in mind:
- If you did not cover Chapter 0, this is the student’s first introduction to the four-
  step process. In this section we cover the first two steps: Choose and Fit. It
  works best if you outline the whole process for the student, in general terms, at
  the beginning of this section and then focus on the first two steps as they are
discussed.
- It is helpful to have students practice interpreting the sample slope and intercept
  as well as calculating a residual or two for selected cars.

Exercises that can be assigned at this point: 1 – 3, 6, 8, 9, 32

Section 1.2—Conditions for a Simple Linear Model. In this section we discuss the
conditions associated with the linear model and present the regression standard error
computation and interpretation.

Issues to keep in mind:
- Maybe people use the word “assumptions” where we use the word “conditions.”
  Sometimes a condition must be assumed, for example, that the data were
  collected at random, but other conditions can and should be checked and cannot
  be simply asserted (assumed) as one would in a mathematics course.
- Students often have difficulty with the interpretation of the regression standard
  error in context. We encourage you to have them practice this with the Porsche
data set and exercises.

Exercises that can be assigned at this point: 4, 5, 24, 25

Section 1.3—Assessing Conditions. In this section we present examples of residual
plots (residuals vs. fitted values) to check for curvature and other patterns in the data
and for non-constant variance. We then show examples of normal quantile plots and
normal probability plots. We present the Assess step of the Porsche analysis here and,
having convinced ourselves that the model is reasonable, we show the Use step by
predicting the price of a used Porsche. Students should be able to make similar predictions using the fitted model.

Issues to keep in mind:

- The distinction between normal quantile plots and normal probability plots is not an important topic, other than the fact that different software products produce one or the other of these two plots. What is important is how to interpret them: a linear pattern supports the normality condition.
- Because this is the first time that students may have seen the last two steps of the four-step process (Assess and Use), it is helpful to highlight these two stages for them.
- By the end of this section, students should be able to follow the four-step procedure for simple linear regression when the relationship is, indeed, linear. They have seen all four steps, and this is a good time to take them through a new data scenario from choosing the model through to predicting a value of the response from the fitted model.

Exercises that can be assigned at this point: 7, 10, 18 – 23, 26

**Section 1.4—Transformations.** We discuss transformations, using a couple of new datasets. We feature square root and natural log transformations, using “log” to mean natural log. Other transformations are possible; there is not one “correct” transformation to use in a given situation. One of the many ways in which two competent statisticians might arrive at different analyses of the same data set is that they may use different transformations, or one may transform a variable while the other does not. Often a well-chosen transformation will enlighten an analysis, but this is not to say that other analyses are wrong.

Students will often struggle with these ideas—wanting there to be one right or, at least best, answer.

Exercises that can be assigned at this point: 11 – 15, 30, 31

**Section 1.5—Outliers & Influential Points.** We discuss outliers and influential points in this section. We also define standardized residuals and studentized residuals (also known as deleted-t residuals). We give as a guideline that standardized or studentized residuals might be compared to ±2.

Issues to keep in mind:

- A word of caution is in order: Many students become enamored with the idea of identifying a point as an outlier and then deleting it, as if it were never collected in the first place. We encourage students to check for outliers and influential points so that they become aware of how these points affect the fitted model. A good rule for data analysis is this: If a point looks suspicious, then investigate the
point (e.g., there may have been a data entry error) and fit the model with and without the suspicious point. If the qualitative answer to the important analysis question changes depending on whether or not the point is included, then state this in the report. In any event, do not delete a point without documenting and explaining why you are deleting the point.

- Note that terminology can differ among books and software packages so one needs to be careful when talking about “standardized residuals.”
- Some instructors may want to jump to Section 4.3 at this point for a more in-depth discussion of unusual points in regression.

Exercises that can be assigned at this point: 16, 17, 27 – 29

Comments on exercises

The Conceptual exercises can be completed without the use of a computer, but we expect the student to use computer software when doing most of the exercises (in this chapter and in the entire book). Many exercises appear in pairs, with a short answer given in the back of the book to the odd-numbered exercise. Exercise 1.11 is a bit surprising: taking logs straightens the scatterplot, showing that in log scale there is a strong linear effect (as scientific theory would predict), but the residual plot shows a remaining pattern that is hard to see in the scatterplot.

Chapter 2—Inference for Simple Linear Regression

Section 2.1—Inference for Regression Slope. We discuss inference for a regression slope with the discussion predicated on the error term having a normal distribution. We return to the Porsche dataset to present a t-test and a t-based confidence interval.

Issues to keep in mind:

- Students may need to be reminded (or taught) how to find the appropriate t* for the confidence interval for the slope.
- It can be helpful to conduct a simulation along the following lines: (1) Fix a set of x values, say x = 1, 2,..., 10; (2) chose a population intercept and slope, say 2 and 7, so that the true model is \( y = 2 + 7x \); (3) generate a set of random errors from a normal distribution with mean zero and standard deviation 4, say; (4) let \( y_i = 2 + 7x_i + \text{error}_i \); (5) feed this set of “data” into a regression program and get beta1-hat; repeat steps (3) – (5) many times (say, 1000 times) and collect the sample slopes (the beta1-hats); make a histogram of the results, which should be bell-shaped and centered at the true slope (which is 7 in this case). Here the students can see that the normality of the error term induces normality in the sampling distribution of beta1-hat, which justifies the t-test and confidence interval.
- Sections 4.5 and 4.6 provide tools for inference in the absence of normality and could be covered at this time.
Exercises that can be assigned at this point: 3, 11, 13, 14, 16

**Section 2.2—Partitioning Variability.** In this section we present analysis of variance for regression, partitioning variability into the model part and the error part. The resulting \( F \)-ratio is the square of the corresponding \( t \)-test ratio.

Issue to keep in mind:
- Partitioning variability is often a difficult topic for students to grasp, as they may not have focused on variability in their first course. Using a dotplot with just the response variable to show overall variability, a scatterplot of the response versus the explanatory variable to show the linear pattern, and a dotplot of the residuals to show the variability leftover may help students visualize what is going on.

Exercises that can be assigned at this point: 2, 4

**Section 2.3—Regression and Correlation.** In this section we remind students about the correlation coefficient, define the coefficient of variation, and relate the two statistics. We end the section by discussing inference for the correlation and the relationship between this inference method, the \( F \)-test for the model, and the \( t \)-test for the slope.

Issues to keep in mind:
- Many students have trouble interpreting \( r^2 \) as the percentage of variability in \( Y \) that is explained by the model (i.e., by \( X \)), so some practice is helpful. Misstatements such as “Mileage explains 79.5% of the price of a Porsche” or “Roughly 80% of the price of a car is due to its mileage” are common.
- We hope that students will benefit from studying how \( r^2 \) can be found as a ratio of part of the variability in \( Y \) to the total variability in \( Y \). The connection between \( r \) and \( \beta_1 \)-hat is also shown here, which gives us three ways to test for a linear relationship.

Exercises that can be assigned at this point: 1, 7 – 9, 12, 17, 19, 23, 25, 26, 28 – 35, 37 – 39.

**Section 2.4—Intervals for Predictions.** Many students will have seen confidence intervals for a slope in STAT1. Fewer will be familiar with prediction intervals, which are presented in this section, despite the fact that arguably prediction intervals, in general, are more useful and important than confidence intervals. Although software will (again) do the work of number crunching, we want students to understand how and why confidence intervals and prediction intervals differ.
Issues to keep in mind:

- A student should be able to explain why the confidence and prediction bands in Figure 2.2 have the shapes that they do.
- Some students will have difficulty distinguishing between the two types of intervals, so having them practice interpreting the intervals several times may help.

Exercises that can be assigned at this point: 5, 6, 10, 15, 18, 20 – 22, 24, 27, 36, 40, 43

Comments on exercises

Many exercises come in pairs so that the odd-numbered exercise can be assigned for practice and the even-numbered exercise collected and graded. Exercises 30 – 33 give the student practice with the idea that removing a point from a dataset can influence various summary statistics. Exercises 41 and 42 can be assigned at any point after finishing Section 2.1, though answers will be more complete if more material has been covered.

Chapter 3—Multiple Regression

In this chapter we extend the ideas from the previous two chapters to the setting in which more than one predictor is used. In a sense, this is the major move and purpose of the entire book—to be able to build and use more complicated models than those seen in the introductory course. A thorough understanding of multiple regression gives the student a great deal of modeling power, but developing that understanding can be a challenge as interrelationships among predictors can make model interpretation difficult.

Section 3.1—Multiple Linear Regression Model. We extend the model and notation of Chapters 1 and 2, letting \( k \) denote the number of predictors (other than the constant, \( \beta_0 \)). The first multiple regression model that we fit shows how winning percentage in professional football depends on PointsFor and PointsAgainst (offense and defense, if you will).

Issues to keep in mind:

- The number of predictors need not be the same as the number of variables, as one can create new predictors from existing variables (e.g., by creating interactions or by adding a quadratic term).
- Students will be tempted to look at the fitted model of \( \text{WinPct-hat} = 0.417 + 0.00177\text{PointsFor} - 0.00153\text{PointsAgainst} \) and say that WinPct increases by 0.00177 for every 1-unit increase in PointsFor \textit{holding PointsAgainst constant}. This is an admirable first attempt, but not a fully successful attempt at interpretation. Rather, we should say that WinPct increases (on average) by
0.00177 for every 1-unit increase in PointsFor allowing for simultaneous change in PointsAgainst. We focus on this in Section 3.5, but from the onset of the chapter we should bear in mind that predictors are generally related to one another so that if one of them changes the others tend to change as well.

Exercises that can be assigned at this point: 1 – 3

Section 3.2—Assessing a Multiple Regression Model. This section presents t-tests, confidence intervals, prediction intervals, ANOVA for multiple regression, and $r^2$, all of which are natural extensions of their Chapter 2 equivalents. One new topic is adjusted $r^2$, which is useful when comparing models of differing sizes.

Exercises that can be assigned at this point: 4, 11, 12 – 15, 22

Section 3.3—Comparing Two Regression Lines. Our first foray into the land of model comparisons occurs in this section when we ask whether a pair of parallel lines is needed or whether a single regression line gives a sufficient summary of a data set. We introduce the potent and flexible idea of indicator (dummy) variables in this section.

Issues to keep in mind:
- Indicator variables will show up many times in the remainder of the book, so it is worth taking time to assure that students understand how they work. For example, in Example 3.8 we create $I_{2000}$—an indicator variable for “Year = 2000”—but we could just as well have created $I_{1998}$ instead. However, we could not use both $I_{1998}$ and $I_{2000}$ in a multiple regression model unless we dropped the intercept term, $\beta_0$, from the model.
- Optional Topic 4.4: Coding Categorical Predictors could be covered here.

Exercises that can be assigned at this point: 20, 27

Section 3.4—New Predictors from Old. Here we tackle the concept of an interaction between predictors. In Example 3.10 we discuss the weights of fish, and one can think of the interaction of Length and Width as being equivalent to area. In other settings an interaction term generally will not have such a geometric anchor. We also present quadratic regression in this section and then extend to polynomial regression and the complete second order model.

Issues to keep in mind:
- Interactions are powerful but they can be difficult to interpret, so you can expect your students to struggle a bit here. Mathematically an interaction is the product of two variables, but thinking about how two variables interact in their effects on a response is a challenge.
- Students might ask why a quadratic regression model should be used rather than taking the square root of the response variable. Sometimes a square root
transformation does the job, but depending on the scale of the data it might work better to capture a quadratic trend seen in a scatterplot by fitting a quadratic model. But bear in mind that transforming the response variable means also changing the nature of the error term. If errors are additive and constant in the original scale, then we should fit a quadratic model in the original scale.

Exercises that can be assigned at this point: 6, 8, 9, 17 – 19, 23, 24, 26, 28 – 30; supplemental exercises 36 and 37

Section 3.5—Correlated Predictors. In this section we tackle the vexing (and ubiquitous) problem of working with predictors that are correlated with one another. The important point in this section is that predictors are often correlated, and this affects model building and interpretation. The variance inflation factor is introduced as a tool that can be used.

Issues to keep in mind:

- As we noted earlier, students will want to say things like “Beta2 is the effect that X2 has on Y, holding X1 constant.” But if X1 and X2 are correlated then it doesn’t make sense to say that X2 changes while X1 is held constant. Rather, we must talk about the effect on Y of a unit change in X2 allowing for simultaneous change in X1. This is a deep and difficult idea, so don’t be surprised if students struggle with it and slip into the “holding other variables constant” terminology. As you steer them back and correct their language you might want to quote the adage that “The temple of reason is entered through the courtyard of habit.”
- The variance inflation factor is a tool that can be used, but guard against thinking that automated tools will solve all problems— or any problems.

Exercises that can be assigned at this point: 5, 10, 16, 31

Section 3.6—Testing Subsets of Predictors. We extend the test of a single coefficient to testing several coefficients together using the nested models F-test. This technique builds on the partitioning of variability introduced in Section 2.2. We hope that it makes sense to the student that the F ratio compares sums of squares per degree of freedom: If the extra sum of squares picked up by the full model, per extra degree of freedom, is comparable to the sums of squares error per degree of freedom, then we have evidence that extending the reduced model to the full model is just an exercise in collecting random error—so the full model is unnecessarily complicated and the reduced model can be used instead.

Issue to keep in mind:

- It might help to show the students examples of pairs of models and ask in each case whether one of the models is nested inside of the other. Note: We don’t want to show this mathematically, but a nested models F-test can be used to
test a single coefficient and will give an F value equal to the square of the t-test ratio for that coefficient. The two tests are completely equivalent—although we would normally use the simpler t-test if we have only one coefficient to test.

Exercises that can be assigned at this point: 7, 21, 25, 32, 34

Section 3.7—Case Study: Predicting in Retail Clothing. The case study gives the student the chance to review ideas from the chapter within the Choose, Fit, Assess, Use framework. That scaffold tends to fade a bit into the background as we work through the book, which we think is OK. But at the end of a chapter packed with topics, it is helpful to use that familiar structure while applying many ideas.

Exercises that can be assigned at this point (general chapter exercises): 33 – 35

Chapter 4—Additional Topics in Regression

This chapter is a collection of optional topics from regression. Any of these topics could be taken up earlier, discussed after Chapters 1 – 3, or skipped altogether. We expect that most instructors will cover at most only a couple of these topics, which stand alone.

Topic 4.1—Added Variable Plots. The first topic explains added variable plots, which provide a graphical way to examine the effect that a predictor has on the response in the presence of other predictors. When using the Houses data from Example 4.1 in class we can start by fitting the multiple regression model and commenting on the two coefficients, 5.657 and 23.232. We then fit two simple regression models and save the residuals from each. The plot of these residuals against one another (Figure 4.3) rarely elicits any responses but there is an “aha” moment when a regression model is fit to those values and the slope turns out to be 23.232—the exact value of the Size coefficient from the multiple regression fit. The Houses example presented in the section can be used in class or replaced by any other example, although it is probably best to work with only a pair of predictors each of which has a moderate relationship with the response.

Topic 4.2—Techniques for Choosing Predictors. This section presents several “automatic” model selection techniques, ideas that can be controversial among statisticians. Among the methods that are available for model selection we present “best subsets,” “backward elimination,” and “stepwise regression” (which incorporates “forward selection”).

Issues to keep in mind:

- There are many statisticians who say that automated model selection procedures should never be used, out of fear that the procedures will be abused. Our response is that one might say this about all of statistics: In the hands of a
mindless automaton, statistical methods can lead to nonsensical results. The solution is not to abolish all of statistics but to use statistical ideas with care.

- In teaching the topic of model selection we encourage you to stress two points: (1) the methods in this section are secondary to thinking about the nature of the applied data analysis problem (i.e., your brain is more valuable than good computer software), and (2) generally there are many models that give quite useful summaries and predictions from a given data set.

- Students, particularly those of a mathematical bent, tend to favor a complicated model with a large $r^2$ over a more understandable model with a smaller $r^2$; we favor the opposite. Remind your students that the goal is not to find the model with the best $r^2$—if that were the goal then a computer could replace the human being doing the work—but to find a model that has good predictive qualities while being relatively simple, sensible, and understandable.

**Topic 4.3—Identifying Unusual Points in Regression.** Topic 4.3 puts mathematical detail behind the ideas first seen in Section 1.5. We find Cook’s distance to be a nice blend of “outlier-ness” and leverage, as it combines into a single measure two ways that a point can be unusual (roughly speaking: in the $y$ direction or in the $x$ direction). The material in this section can be handled with calculations or with graphing; we favor a mix of the two. The summary table at the end of the section might be presented even if the rest of the section is given only light treatment.

Issue to keep in mind:

- As we noted earlier, terminology can vary from book to book and amongst software packages, so students should be warned that this is an area in which there is not universal agreement on, for example, the definition of “standardized residual.”

**Topic 4.4—Coding Categorical Predictors.** This section extends the idea of an indicator variable (seen in Section 3.3) to the setting in which a predictor has more than two categories. This is only slightly more complicated than the use of a single dummy variable.

*The final two sections in Chapter 4 provide tools for inference when the usual conditions for the linear model are doubtful—or when one simply wants to take a different approach to inference.*

**Topic 4.5—Randomization Test for a Relationship.** The randomization test introduced in this section is in many ways a more natural inference procedure than the standard $t$-test. Some students may have seen randomization tests in STAT1, but for most this will be a new way of thinking about statistical significance. Instructors may opt to return to Chapter 0 and rework Example 0.6 using a randomization two-sample test (also known as a permutation test) before using randomization testing in the regression setting. We provide an R script for the Example 0.6 comparison immediately below; running this
script yields a p-value of about 0.0006, which is remarkably close to 0.00059, the t-test p-value (which was rounded off to 0.001 in Chapter 0).

R Script Begins

```r
Control <- c(12.5,12.0,1.0,-5.0,3.0,-5.0,7.5,-2.5,20.0,-1.0,2.0,4.5,
              -2.0,-17.0,19.0,-2.0,12.0,10.5,5.0)

Incentive <- c(25.5,24.0,8.0,15.5,21.0,4.5,30.0,7.5,10.0,18.0,5.0,
                -0.5,27.0,6.0,25.5,21.0,18.5)

mean(Control)  # gives [1] 3.921053
mean(Incentive) # gives [1] 15.67647

ObservedDiff <- mean(Control) - mean(Incentive)  # ObservedDiff
               # is -11.75542

fulldata <- append(Control,Incentive)
cards <- 1:36  # creates 36 cards, one "ID number" per observation
diffmeans <- NULL  # initializes a vector to hold the results

for(i in 1:10000){  # starts a loop that will be executed 10,000 times
  cards1 <- sample(cards,19,replace=FALSE)  # samples of 19 cards
  cards2 <- cards[-cards1]  # remaining cards in group 2
  group1 <- (fulldata[cards1])  # reads cards in group 1
  group2 <- (fulldata[cards2])  # reads cards in group 2
  diffmeans[i] <- mean(group1) - mean(group2)  # diff for this run
}

numgreater <- abs(diffmeans) >= abs(ObservedDiff)

mean(numgreater)  # proportion of randomizations with a "1"
```

R Script Ends

**Topic 4.6—Bootstrapping for Regression.** In this section one learns about bootstrapping.

**Issues to keep in mind:**

- As with randomization testing, the bootstrap will be a new idea for most students, although some may have seen it in STAT1 or perhaps elsewhere.
- The central idea of bootstrapping is “The bootstrap sample is to the sample as the sample is to the population.” Some students will find this to be a natural analogy, but others will question how it is possible to use the sample over and over again and learn anything about the population from this process. Remind them that we are sampling *with replacement* so a huge number of bootstrap samples are possible, of which we see only a small fraction.
- Students will probably find Method #2 for creating bootstrap confidence intervals to be most intuitive, but the other methods shown are often used in practice as well.
Comments on exercises

Exercises are divided by topic. Note that Exercise 4.8 introduces a new technique of cross validation. Before assigning this exercise it would be helpful to do a cross validation example in class.

Unit B

1. Which order for the units? The way we have written the book suggests that the three units be covered in ABC order, and that is how most of us have taught our Stat 2 courses. This order means that the two units with quantitative responses (A and B) come together. Unit C on logistic regression, the unit that many students find hardest, comes later in the semester.

It is possible instead to postpone B in order to go directly from A (ordinary regression) to C (logistic regression). The main difference between A and C is that for A the response is quantitative, for C the response is binary. One of us teaches Stat 2 in this order, an order that seems especially well-suited to students of economics and political science.

A third option is to get to Unit B as quickly as possible, immediately after Chapter 2.

This option might be attractive to students of psychology and experimental biology, for whom analysis of variance and design of experiments is important. (One of us has routinely taught a first statistics course that begins with ANOVA. In that course, students have not seen t-tests or regression at all.)

2. Analysis of variance (ANOVA) generalizes t-tests. ANOVA was created for data sets with a quantitative response and one or more categorical predictors. For students who have not seen ANOVA before, it can help to introduce the methods of ANOVA as generalizations of the two-sample and paired t-tests. Both ANOVA and t-tests have a quantitative response.

- One-way ANOVA (Chapter 5) generalizes the two-sample t-test (Chapter 0). The t-test has two groups and hence a binary predictor. One-way ANOVA has two or more groups and hence a categorical predictor with two or more values. In both cases the response variable is quantitative. For example, we might use a t-test to compare survival times for patients with stomach or bowel cancer. Example 5.8 uses one-way ANOVA to compare survival times for patients with cancers in five different kinds of organs.
• Two-way ANOVA (Chapter 6) generalizes the paired t-test. The paired t-test compares two groups using matched pairs of observational units. The two-way (additive) ANOVA model is used to compare two or more groups using matched sets of units. For example, you might use a paired t-test to compare the response of four human subjects to caffeine and a placebo, with each subject being measured twice, once with each drug. Example 6.1 uses a two-way ANOVA model to compare the response of four human subjects to caffeine, theobromine, and placebo.

3. ANOVA and design of experiments. On the surface, ANOVA may seem not to fit well with this book’s emphasis on modeling: For well-designed studies, the design dictates the model; by the time you have the data in hand, the model is already determined. Apparently, the Choose step is eliminated.

In fact, however, the Choose step is not only present, but is often far more pivotal for ANOVA than for regression. The reason: For regression, you typically choose a model that you hope will fit the data you already have. For ANOVA, you typically choose the model first, then produce data to fit the model you have chosen. The choice of model determines what data you will end up analyzing. The chart below oversimplifies, but illustrates the typical time reversal.

<table>
<thead>
<tr>
<th>Regression:</th>
<th>Research question</th>
<th>-&gt;</th>
<th>Data</th>
<th>-&gt;</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANOVA:</td>
<td>Research question</td>
<td>-&gt;</td>
<td>Model (design)</td>
<td>-&gt;</td>
<td>Data</td>
</tr>
</tbody>
</table>

In order to maintain the essential connection between ANOVA and experimental design, you may want to interweave sections from Chapter 8 on design with Chapters 5 and 6 on ANOVA, as shown in Table 1, which follows.

4. ANOVA and regression. In theory, ANOVA is just a special case of regression using only 0,1 indicator variables. (See Chapter 7, Section 7.5.) This view makes it possible to derive the theory for ANOVA tests and intervals just by applying specialized instances of the theory of regression. The resulting approach offers a satisfying theoretical unity and an economy of exposition as well.

In practice, however, ANOVA differs from the rest of regression in a number of important ways. ANOVA is closely tied to design of experiments in a way that regression in general is not. It is an oversimplification, but a useful guide nonetheless, to say “Regression for observational data, ANOVA for designed experiments.” If you want to serve students in psychology and experimental biology, it is important to stress links to the material on design in Chapter 8.

In addition, properly designed experiments of the sort considered in this unit have nice mathematical properties (balance, group symmetries, orthogonality) that set them apart from most regression data. These additional properties lead to simpler
ways of displaying and exploring ANOVA data—such as the two-way layouts and interaction plots of Chapter 6—and also simplify interpretation of results. For example, consider a plant nutrition experiment with three nutrients each at four concentrations. In all, there are 12 nutrient combinations. It is natural to show these 12 in a $3 \times 4$ table, with two categorical variables that correspond to rows and columns. The regression approach requires using an arbitrary set of eleven 0,1 variables.

**Parallels and Connections**

<table>
<thead>
<tr>
<th>Binary Predictor</th>
<th>Categorical predictor(s)</th>
<th>Experimental Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>two-sample t-test</td>
<td>One-way ANOVA (Ch. 5)</td>
<td>Completely randomized (8.1)</td>
</tr>
<tr>
<td>Paired t-test</td>
<td>Two-way additive</td>
<td>Randomized complete block (8.3)</td>
</tr>
</tbody>
</table>

**Table 1:** For those who want to teach design in parallel with the models and analyses, one possible order would be: 8.1, Chapter 5, 8.2 – 8.4, Chapter 6. Or another possible order would be 8.1, Chapter 5, 8.3, 6.1, 6.2, 8.4, 6.3, and 6.4.

**Chapter 5—One-way ANOVA**

This chapter introduces analysis of variance, leaning on data for 125 male fruit flies assigned to one of five groups in a study that sought to answer the question “Does reproductive behavior reduce longevity in fruit flies?” Section 5.1 presents the one-way model (the Choose and Fit steps), Section 5.2 addresses model conditions (the Assess step), Section 5.3 discusses the important topic of scope of inference (the Use step), in Section 5.4 we consider Fisher’s Least Significant Difference, finally Section 5.5 presents a summary. Depending on the backgrounds of the students, it might be helpful to use the two-sample t-test from Chapter 0 as a preview for this chapter, stressing that the familiar t-test can be considered a special case of ANOVA or, put the other way around, what the students know about comparing two groups will be extended in this chapter to comparing several groups.

**Section 5.1—The One-way Model: Comparing Groups.** This section begins with a set of dotplots, providing visual evidence in answer to the question of whether a simple model with a single mean is appropriate or whether a more complicated model with several means is needed. The section then moves ahead with a model, some discussion of the
conditions for that model (although formal checking of those conditions comes in the next section), and parameter estimates for the model.

Next we present decomposition of observations. Even students who have seen ANOVA in STAT1 will almost certainly not have seen decompositions, so expect to take a good deal of time on this concept, which ties directly to sums of squares and partitioning of variability. This “triple decomposition” is the base that supports all of ANOVA. We note here the difference between “stacked” and “unstacked” data. (These terms are the ones used by Minitab.) “Stacked” refers to the cases-by-variables format used for regression. “Unstacked” refers to using one row or column for the response values in each group. There is no example in the body of Chapter 5, so you may want to refer to Exercise 5.22, which lists data in the unstacked format, with one column for each group. The difference in format can be useful for pedagogy. The stacked format emphasizes connections to regression. The unstacked format emphasizes what it is that makes ANOVA different from regression. This difference will become even more important in Chapter 6 on two-way designs.

Issues to keep in mind:

- Students should get into the habit of always examining graphs early in any analysis; parallel dotplots or boxplots work well for ANOVA data.
- Formulas versus ideas. Books on ANOVA are heavy with subscripted formulas. These formulas are useful to some students, but not to others. It is quite possible to teach ANOVA without formulas, in the spirit of the introductory statistics book by Freedman et al.¹ For this approach, the unstacked display of data is all but essential. We will say more about this in the Guide for the next Chapter.

Exercises that can be assigned at this point: 1, 6, 7, 10, 15, 17, 31

Section 5.2—Assessing and Using the Model. There are three main ideas here: testing for a difference, checking conditions, and transforming to a new scale.

- Testing for a difference. Depending on what your students have seen, here are two possible approaches:
  1. Build from regression: The ANOVA F-test is a special case of the nested F-test, with full model allowing a different mean for each group, and null model requiring all means to be equal.
  2. Go for a visual or intuitive approach that emphasizes the rationale for ANOVA. The F-test compares two measures of variability—between groups and within groups. Dotplots allow you to eyeball the differences between group means and the variability within groups. The F-test does the same thing by arithmetic.

• Checking conditions. The main thing new here is that group standard deviations are required to be equal. (For the t-test, we use an approximate version that allows unequal SDs.) ANOVA is like the pooled t-test in its assuming a single SD to be estimated from all the residuals. Checking for constant variance is handled differently for ANOVA than for regression: we still talk about plotting residuals against fitted values but we also compute Smax/Smin and we mention Levene’s test in a footnote. Some instructors will want to present Levene’s test (Section 7.1) here.

We use normal probability plots to check the normality condition of the model, which will be familiar from Unit A.

• Transformations. When group SDs are unequal, it is often the case that the size of the SD is related to the size of the mean. For such data sets, there is often a change of scale that will make the SDs much more nearly equal.

Issues to keep in mind:
• One can think regression-like or ANOVA-like in presenting the F-test for ANOVA.
• Some instructors will want to present Levene’s test (Section 7.1) in checking the equal variance condition.
• Transformations can sometimes bring the data in line with the equal variance condition.

Exercises that can be assigned at this point: 4, 8, 9, 13, 16, 19 – 23, 25, 27 – 30, 32, 35

Section 5.3—Scope of Inference. Here we raise the question of scope of inference. We hope that students will have learned in STAT1 that experiments allow one to draw causal inferences about the treatments applied, in contrast to observational studies for which one can infer that populations differ but cannot make a clear statement about cause and effect. Moreover, even an inference about the populations being different requires that the observational units were selected at random. These ideas are reinforced in this section. We finish the section with a cautionary example about how complications in the way the sample is collected can limit our conclusions.

Issues to keep in mind:
• Cause-and-effect inferences are possible when units are randomly allocated to groups.
• Inferences from a sample to a population are possible when the sample units are randomly selected from the population.

Exercises that can be assigned at this point: 2, 3, 11, 12, 14
Section 5.4—Fisher’s Least Significant Difference (LSD). Here we dip one toe into the turbulent waters of multiple comparisons. The big idea is the difference between individual error rate and family-wise error rate. The topic has a high profile and a huge research literature, so we give it more detailed attention in Chapter 7, Section 2. You may want to look over this Guide for 7.2 at this point. Some of us see Sections 5.4 and 7.2 as alternatives: Do the short version now (5.4) and skip (7.2), or skip (5.4) and plan to cover (7.2) later on. Some statisticians—a minority—say you’ll do your students a favor if you skip both sections.

Issues to keep in mind:
- In one sense, the only thing here that is new to students is the use of mean squared error as the estimate of variance when computing the LSD to compare two means. Nonetheless, one cannot assume that this will be received as a simple variation on something already known.
- Topic 7.2 on multiple tests compares Fisher’s LSD, the Bonferroni method, and Tukey’s HSD, so some instructors will want to cover that section here.

Exercises that can be assigned at this point: 5, 18, 24, 26, 33, 34

Section 5.5—Chapter Summary. The summary of the chapter shows that there are quite a few definitions in this chapter. Fortunately, parallel structure across many components means that there are far fewer new concepts to master than there are new definitions and formulas.

Comments on exercises

Exercises 36 – 39 are more general and should probably be assigned after the whole chapter has been covered.

Chapter 6—Multifactor ANOVA

Two-way ANOVA is just one-way ANOVA with extra structure. That extra structure is typically pre-planned and built into the data collection process: It is in the planning stages that you choose your model. If you want to emphasize the links between choosing a model and designing an experiment, consider one of the orders of topics given in the caption to Table 1, orders that interlink Chapters 6 and 8.

Some of us choose to skip the entire chapter on design, which might be a good plan for a course that emphasizes observational data, regression, and applications to, for example, economics or political science. On the other hand, for students of biology and psychology, and for those headed to medical school, Chapter 8 is very important.
Section 6.1—The Two-way Additive Model (Main Effects Model). This section deals with the two-way additive model. Here “additive” means that the effects of the two factors are added together, with no interaction term. The model is most commonly used for data from complete block designs, the generalization of paired data. (See Section 8.3.) If students are used to the “cases-by-variables” or “stacked” format for data, it is important to make a transition to the two-way table for response values, with rows for one categorical variable and columns for the other. This layout makes the data decomposition easy: put the row averages at the right end of each row and the column averages at the bottom of each column. Subtract the grand average from the row and column averages to get the row and column effects. These are immediately meaningful in the context of the application. The null hypotheses are that the parameters that correspond to these effects are all zero.

Notation. With two-way data, the notation for ANOVA gets messy. We have reduced the usual three subscripts to two, but even so, many students will find the notation daunting and the formulas uninformative. We have included the standard formulas partly in deference to tradition, but also because some students do find them genuinely helpful. The subscripts correspond to the row-by-column format for showing the response values, with k for rows and j for columns. The dots (or periods) before or after subscripts indicate a row or column that has been “summed out” of the calculation. The mean for row k is \( \bar{y}_k \), a calculation that would require summing all values across the j subscript (that is, across columns). The mean for column j is \( \bar{y}_c \). The grand mean is \( \bar{y}_\cdot \) and the mean for the cell in row k, column j is \( \bar{y}_{kj} \). Row and column effects are \( \bar{y}_k - \bar{y}_\cdot \) and \( \bar{y}_j - \bar{y}_\cdot \) respectively.

For other students the notation just gets in the way of intuition. A rectangular format for the response values make it clear which averages we care about and how to compute them, without reliance on formulas. For two-way data this choice between the algebraic and the visual is a big one. If you opt for algebraic, you’ll want to focus on the boxed summaries and use the examples for illustration. If you opt for visual, you’ll want to ignore the boxed summaries, focus on the numerical examples, and encourage your students to talk about why the arithmetic makes sense in the context of the applications.

Issues to keep in mind:
- Get your students comfortable with the two-way table format for the data, even if you have heretofore emphasized the stacked view of data.
- The new ANOVA dot notation can be tricky for students so take time to explain it carefully.

Exercises that can be assigned at this point: 1 – 7, 11, 17, 23 – 27
Section 6.2—Interaction in the Two-way Model. Section 6.2 has four main ideas:

- What interaction is and why it matters
- The interaction graph
- Numerically, interaction is a difference of differences
- Visually, interaction is a difference of slopes

The key idea is that when two factors (two categorical variables) interact, their effect on the response is not what you would predict from the effects of the two variables taken individually. We recommend spending time on a discussion of interaction and its importance in science before moving on to the more technical aspects. We have included an extensive set of examples of interaction graphs, and we hope you and your students will find these useful for developing an intuitive sense of what these graphs mean.

Issue to keep in mind:

- For any two-way setting there are two options for the interaction plot: levels of Factor A can be put on the horizontal axis and levels of Factor B can be used for plotting symbols or the roles of A and B can be reversed; Figure 6.9 shows an example of this. It is a good idea to show students examples of this because the choice can make a difference in how one sees the story in the data.

Exercises that can be assigned at this point: 10, 13, 14, 19 – 21

Section 6.3—The Two-way Non-additive Model (Two-way ANOVA with Interaction).

Whereas Section 6.2 was packed with ideas and short on formulas, Section 6.3 is packed with formulas and short on ideas. For those who want to understand the cogs in the mechanism, the formulas are useful, although a deep understanding depends on linear algebra and is beyond the scope of this book. For everyone else, the section boils down to this: “The computer will give you a p-value that tells the strength of evidence against the null hypothesis of no interaction.”

Exercises that can be assigned at this point: 8, 9, 12, 15, 16, 18, 22, 28

Section 6.4—Case Study. The extended example in this section illustrates a number of data-analytic methods as part of an analysis of two-way data. Data transformation plays a big role. Not only does a new scale equalize standard deviations, it also largely eliminates interaction. (Interaction is scale-dependent.) Finally, replacing a categorical variable with linear and quadratic terms leads to a very simple model, with only two df for the model, as opposed to the original seven.

Comments on exercises

Exercise 29 is more general and should probably be assigned after the chapter is completed. Exercises 30 – 34 are supplementary and discuss how to figure out which
transformations may be best to deal with data that do not meet the necessary conditions for ANOVA.

Chapter 7—Additional Topics in Analysis of Variance

The five sections of Chapter 7 present optional topics. These can be covered in any order, or skipped altogether.

**Topic 7.1—Levene’s test.** The goal of Levene’s test is to tell whether there is strong evidence against the null hypothesis that group SDs for ANOVA are all equal. Levene’s test is a robust alternative to a more traditional F-test for variance equality, a test based on squaring that gives undue influence to outliers. Levene’s test relies on ANOVA applied to absolute differences from group medians: Compute a median for each group, subtract the median from the observed values in the group, and take absolute values. If group SDs are equal, we expect the mean absolute deviations to be equal. That’s what Levene’s ANOVA tests.

Issues to keep in mind:
- Warn students not to rely on Levene’s test as sole gatekeeper before doing an ordinary ANOVA. Remind them that if group means and group SDs are related by a simple pattern—for example, larger SDs go with larger means—a transformation may be called for. Students should also use parallel dotplots or boxplots as well as use the $SD_{max}/SD_{min} > 2$ check.
- Do not expect your students to find Levene’s test to be intuitive, but teach them that it does a good job helping check the equal variance condition for ANOVA.

**Topic 7.2—Multiple tests.** The key idea for this section is the difference between individual and family-wise error rates for tests and intervals. If you do a lot of tests or intervals each with a 5% error rate, the overall (family-wise) error rate will be much higher than 5%. If your goal is to keep the family-wise error rate at 5%, you need a smaller error rate for individual tests. This section shows ways to do this.

- You may find the following analogy useful to students: Imagine that each time the phone rings, there is a 5% chance the caller wants money from you. Suppose you decide to answer the next ten calls. What is the chance that at least one of the callers will want money?
- Each pairwise comparison in this section has the same form for its confidence interval: $(\text{mean}_1 - \text{mean}_2) \pm \text{margin of error}$. Each margin of error equals the product of a critical value and the estimated SD. Finally, the estimated SD equals $\sqrt{MSE \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$. Which critical value to use depends on the method of adjustment you choose to use.
• We present three choices for critical values: Fisher, Bonferroni, and Tukey. Fisher is simplest but more prone to false alarms—intervals that do not contain 0 even though the “true” difference is 0. Bonferroni is more prone to “misses”—intervals that contain 0 even when the true difference is not 0. Tukey is in between.

• You may want to point out to students that some statisticians see this particular enterprise as misguided. Here, briefly, are two main criticisms: (1) Should your conclusion about whether two treatments are different depend on whether there are other unrelated treatments in the study? If you think no, you should not adjust for multiple comparisons. (2) If you are doing a screening study to narrow down a large set of potentially useful interventions, a miss may be much worse than a false alarm. In such a situation it would be a mistake to use the methods of this section because they implement exactly the wrong priorities: They increase the miss rate in order to keep the family-wise false alarm rate low.

Topic 7.3—Comparisons and Contrasts. In the context of ANOVA, a comparison is a difference of two group means. A contrast generalizes the idea to compare weighted sums of group means. There are three main ideas in this section:

• What contrasts are and how to choose them from the applied context
• Estimating the standard error of a contrast
• Inference for contrasts: tests and intervals

A contrast is a difference of weighted sums of group means, for example:

\[
\frac{(\bar{y}_1 + \bar{y}_2)}{2} - \frac{(\bar{y}_3 + \bar{y}_4 + \bar{y}_5)}{3}
\]

The coefficients here, ½, ½, −1/3, −1/3, and −1/3, must add to zero.² Indeed, one possible point of confusion derives from the non-uniqueness of the choice of coefficients for a given contrast; so, in the above example the coefficients 3, 3, −2, −2, −2 would work equally well. The t-ratio for the two definitions would, of course, be the same because the factor of 6 in the contrast is matched by another factor of 6 in the denominator of the t, the SE part.

Note that a comparison between two means is a contrast with its only two non-zero coefficients equal to 1 and -1. So students will already have encountered contrasts in previous work, but one should not assume that this early experience will make the transition to the general case easy for all students.

² In a more mathematical context, a contrast is a coefficient vector orthogonal to the vector of all 1s. The SE of a contrast is proportional to its Euclidean length.
A sample contrast $\sum c_i \bar{y}_i$ estimates the corresponding parameter $\sum c_i \mu_i$. To test the null hypothesis that $\sum c_i \mu_i = 0$, we divide the contrast by its estimated standard error and compare to a $t$-distribution. For a confidence interval we use a margin of error equal to $t$ times the estimated standard error $SE = \sqrt{MSE} \sqrt{\sum c_i^2 / n_i}$.

**Topic 7.4—Nonparametric statistics.** Some instructors are fans of nonparametric methods because of the good operating properties that they have. Others will want to spend a good deal of time on the topic because these methods are widely used by researchers across many fields of application. Some will skip this topic because they covered Topic 4.5 (randomization tests) or 4.6 (the bootstrap) or both, whereas others will omit the topic simply because of time constraints. No matter what choice you make, we encourage you to at least make your students aware of the topic. In this section we present the Wilcoxon-Mann-Whitney test, which some may have seen in STAT1 as an alternative to the two-sample $t$-test, and then present the Kruskal-Wallis test as an alternative to ANOVA.

**Topic 7.5—ANOVA and Regression with Indicators.** This section shows how to treat ANOVA as a special case of regression for which every predictor is an indicator variable—a 0,1 variable that indicates group membership. The examples illustrate this fact first for one-way ANOVA and then for two-way.

Being able to apply regression results to ANOVA is important for theory. For mathematically inclined students the unity of the theory can be intellectually very satisfying as well. Finally, if you want a course that keeps its treatment of ANOVA to a minimum in order to spend more time on regression and logistic regression, presenting ANOVA as a special case of regression offers a way to save time.

In practice, however, using indicator variables to present ANOVA as regression has major disadvantages. Computer output tends to be harder to interpret. The cases-by-variables orientation of regression obscures or ignores the structures that make ANOVA different from regression. For example, two-way ANOVA has a natural two-way structure that the regression approach buries under multiple columns of 0s and 1s. Relying on the two-way layout with rows and columns for the two categorical explanatory variables can make two-way ANOVA quite intuitive, but this advantage is lost with the regression approach.

**Topic 7.6—Analysis of Covariance.** Analysis of covariance (ANCOVA) is a hybrid, a fancy version of ANOVA that allows you to compare groups after you first adjust the response for a quantitative predictor.

---

3 Instead of using indicators as predictors, there is an approach based on contrasts of the sort described in Section 3. This approach is useful for a number of areas of application, but it is beyond the scope of this book.
You may want to introduce the idea to your students along the following lines: “The basic idea of ANCOVA can be quite simple and intuitive, provided you keep your eye on the big picture and don’t let the details get in the way. Suppose you want to compare two drugs that claim to reduce blood pressure. It makes sense to use change in blood pressure, (Before-After), as your response. In effect, you adjust the After reading for Before. This is pretty much what ANCOVA does, except that ANCOVA uses regression to find the best-fitting adjustment. Keep in mind, as you work through the section, that ANCOVA is just a fancy version of a two-sample t-test on several groups of (Before-After) values.

Chapter 8—Overview of Experimental Design

As set out in this Guide’s introduction to Unit B, for ANOVA from a well-designed experiment the model choosing typically takes place before the data are collected. This is the chapter that deals with the choice of design that determines the model.

For the basics of design theory there are no formulas and no computations. Thinking is qualitative rather than mathematical, and mathematically inclined readers often find the material especially difficult if not downright frustrating. The challenge with design is not one of thinking abstractly, not one of understanding logical consequences deduced from abstract definitions; rather, the challenge is to recognize the abstractions “in the wild.”

Chapter 8 introduces three basic design principles: randomization (8.1), blocking (8.3), and factorial crossing (8.4). Corresponding to each principle there is a simplest possible design and model: one-way model (Chapter 5), two-way additive model (6.1), and two-way non-additive model (6.3). Along with the three principles and related designs, there is a small collection of key concepts: experimental units, treatments, blocks, factors, levels, and cells. A glossary at the end of the chapter gives definitions, but it is important to point out: For those coming to this subject for the first time, definitions are likely to prove annoyingly unhelpful at first.

The most efficient way to learn the concepts is through real examples. Read the description of an experiment: What are the experimental units? What is the treatment? How many levels? Are there blocks? If so, what are they? How are they created? How many factors? etc. If your students struggle or complain you might remind them that Plato considered abstract mathematics to be the easiest subject because it was the subject least obscured by the distracting details of context. In Plato’s sense, design is truly hard.
Section 2 on the randomization F-test is different in that it involves quantitative thinking. That section illustrates the logical link between design (random assignment) and inference (a probability model for computing p-values).

As you look ahead to the individual sections of Chapter 8, it may help to ask of each of Sections 1, 3, and 4: What is the design principle? What is the corresponding design? What are the key abstractions to learn to recognize? Are there any common pitfalls to avoid?

Section 8.1—Comparisons and Randomization. Here are the key ideas of this section.

**Design principle.** Random assignment: Assign treatments to units using a chance device. Randomizing (a) protects against bias, (b) permits conclusions about cause, and (c) provides a justification for using a probability model, as set out in Section 8.2.

**Design and model.** Completely randomized design and one-way ANOVA model (Chapter 5).

**Key abstractions.** Treatments and units. Informally, the treatment is what you do, the unit is what or whom you do it to. For the example of financial incentives for weight loss (Chapter 0) the treatment (“what you do”) is to give or not give a financial incentive. The unit (“whom you do it to”) is the experimental subject.

**Pitfall to avoid.** Don’t say “Treatment is what you assign.” Many books say this, but “assign” is ambiguous. You can assign treatments to people but you can also assign people to treatment groups.

Exercises that can now be assigned: 1 – 8, 25

Section 8.2—Randomization F-test. This section is different from the others in Chapter 8. You may prefer to defer or even skip this section in order to continue with the design Sections 3 and 4, which parallel and complement Section 1.

Section 2 makes good on an assertion from Section 1, namely, that random assignment provides a justification for a probability model that leads to p-values. If students are used to p-values by way of the normal distribution and its descendants, the logic of the randomization test will likely take some getting used to. The setting is one-way ANOVA, and the null hypothesis, as in Chapter 5, is that there are no treatment differences. The key idea: *If the null is true, and the treatments have no effect, then the response value for each individual would have been the same regardless of treatment.* (If I gave a reading of 10.5 under treatment A, I would have given a reading of 10.5 under any of the other treatments.) This means that the value of $F$ is completely determined by the random assignment. To see which values of $F$ are extreme, we repeat the random
assignment over and over, each time computing the value of $F$. The estimated p-value is the proportion of times we get an $F$ as big or bigger than the one from the actual data.

Exercises that can be assigned at this point: 16, 17, 30 – 32

**Section 8.3—Design Strategy: Blocking.** Here are the keys to this section.

*Design principle.* Create blocks of similar experimental units before you randomly assign treatments, one per unit within each block. (Abstract definition below.)

*Corresponding design and model.* Randomized complete block design and two-way additive (no-interaction) model (Section 6.1).

*Key abstract concept.* A block is a set of similar experimental units.

*Pitfall to avoid.* Don’t try to get to the abstract idea of a block too quickly. “Block” may well be the single hardest concept in basic design theory. Stay concrete, and work one at a time through a few examples of each of the three standard ways to create blocks: by matching similar units, by subdividing a larger entity to get units, and by repeated measures on the same individual. Eventually, but only after enough examples, students should be ready for a chart like this one:

<table>
<thead>
<tr>
<th>How to get blocks</th>
<th>What the block is</th>
<th>What the unit is</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matching units</td>
<td>Set of matched units</td>
<td>Individual</td>
</tr>
<tr>
<td>Subdividing</td>
<td>Individual or entity</td>
<td>A part of the entity</td>
</tr>
<tr>
<td>Reusing</td>
<td>Individual or entity</td>
<td>Time slot</td>
</tr>
</tbody>
</table>

Exercise that can be assigned at this point: 9 – 12, 15, 18 – 22, 24, 27 – 29

**Section 8.4—Design Strategy: Factorial Crossing.** We summarize the key ideas to factorial crossing.

*Design principle.* To study two factors in the same experiment, cross them: include at least one observed response value for each possible combination of levels of the two factors.

*Corresponding design and model.* Two-way factorial design and two-way ANOVA model, either additive (no interaction, Section 6.1) or non-additive (with interaction, 6.3 and 6.4).

*Key abstract concepts.* Factors, levels, crossing, cells, interaction. None of these are difficult in the way that treatments, units, and blocks can be, although the vocabulary itself may feel alien at first.
**Pitfalls to avoid.** In our experience there are no common pitfalls, but here are two things that are useful to emphasize.

Crossing can be presented visually and literally, by way of a two-way table. One factor corresponds to rows; each row is a level of the row factor. A second corresponds to columns; each column is a level of the column factor. Each row and column intersect (cross) in a cell. The two factors are completely crossed if there is an observed value for every cell. To measure interaction, you need more than one observation per cell.

Interaction can and should be understood both visually and numerically. Interaction is present if the slopes of the interaction plot are not the same. Numerically, interaction is a “difference of differences.” Start with a $2 \times 2$ example, such as Example 6.4 (pig feed) before going on to designs with more than two levels per factor.

Exercises that can be assigned at this point: 13, 14, 23, 26

---

**Unit C**

This unit is about data with a binary response and one or more predictors. These predictors can be either quantitative or categorical. A recurring example is the relationship between acceptance to medical school (Yes/No response) and predictors such as grade point average (quantitative) and sex (categorical).

Working with a binary response brings several new challenges for the student:

1. **Scatterplots are much less informative.** Every $y$-value is a 0 or a 1. For ordinary regression the fitted model is a line. Often that line is easy to approximate by eye. For logistic regression the fitted model is a logistic curve, and it is usually not possible to approximate the curve by eye. The bottom line is that the logistic model feels much more abstract, less “real.”

2. **The model is an equation for probabilities, not outcomes.** For ordinary regression, $Y = \beta_0 + \beta_1 X + \epsilon$. All of $\beta_0, \beta_1$, and the $\epsilon$s can be (and are) estimated. For logistic regression, what we observe for $Y$ is its value, 0 or 1, but what we model with an equation is not $Y$ itself but an unseen probability, $P(Y = 1)$. That probability cannot be observed, and there are no fitted $Y$s to compare with the observed $Y$s. Here again the bottom line is that the logistic model is remote from what you can see.

3. **The model equation relies on an unfamiliar transformation.** Ordinary regression uses a linear equation. Logistic regression fits a linear equation to the log odds $\log(\pi/(1 – \pi))$. Equivalently, logistic regression fits the probability $\pi$ using a logistic curve: $\pi = \exp(\beta_0 + \beta_1 X)/(1 + \exp(\beta_0 + \beta_1 X))$. Thus there are two forms of the logistic model, not just one as for ordinary regression.
A related complication is obtaining fitted values. For ordinary regression there is no need to “back-transform” because \( \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \). When fitting a logistic regression, you obtain estimates for the logodds, that is, \( \log(\text{odds}) = \hat{\beta}_0 + \hat{\beta}_1 X \). But often your interest centers on \( \pi = \Pr(Y = 1) \) so you’ll need to compute an estimate of fitted probabilities, use your coefficient estimates, \( \hat{\beta}_0, \hat{\beta}_1 \), to obtain and compute an estimate of \( \log(\pi/(1-\pi)) \), then with a little algebra you can estimate \( \pi = \Pr(Y = 1) \) for a given level of \( X \):

\[
\hat{\pi}(Y = 1) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 X)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 X)}
\]

4. *The method of fitting is harder.* For ordinary regression it is easy to visualize the fitted line and deviations from it. The idea that a good line will make the deviations small is reasonably intuitive, and so is the method of least squares. For logistic regression, the fitting is done by maximizing the likelihood. For many students working with unseeable probabilities is hard, all the more so because they are functions of unknown parameters.

For these and other reasons students are likely to find Unit C especially challenging. Unit C relies for its exposition on parallels between ordinary and logistic regression. As far as options for sequencing are concerned, Chapters 1 and 2 are all that are needed for Chapter 9. Chapters 1 – 3 are all that are needed for Chapter 10. Unit C does not use content from Unit B. Although it is possible to go directly from Unit A to Unit C, and some of us have done that, you may prefer to wait to start C until later in the term, after Unit B, to give students more experience before tackling what is likely to be the hardest unit.

Whenever you choose to cover C, here are three points to keep in mind:

1. Because of (1) – (4) above, it will likely be useful to rely heavily on applied context as a way to help students experience the models as having concrete meaning, rather than as abstract formulas.

2. There are two special cases that are easier:

   (a) Our beginning example (Losing Sleep) has several independent observations (y-values) for each x-value. For such data sets, it is possible to construct plots that make the two forms of the model easier to see. Plotting observed proportions \( \hat{p} = \#Yes/n \) versus \( X \) will give points that suggest the empirical logistic curve. Plotting empirical log odds \( \log(\frac{\hat{p}}{1-\hat{p}}) \) versus \( X \) will give points that suggest the linear fit.
(b) Section 2 of Chapter 9 deals with binary predictors. When both response and predictor are binary you can summarize the data with a familiar 2 x 2 table. The empirical logistic plot has only two points, which reduces fitting to drawing the line between them.

3. Students should practice back-transforming from fitted log odds to probabilities. Becoming at home with this equivalence is an essential part of making the models feel real. However, many times an analysis will not call for estimated probabilities explicitly. The coefficient associated with X in ordinary regression, the slope, can have meaning and be of interest in and of itself. Analogously, exponentiating the coefficient associated with X produce an odds ratio and it, too, can have meaning and be of interest aside from individual probabilities.

Chapter 9—Logistic Regression

Students are likely to find Chapter 9 to be the hardest of the three chapters in Unit C, and are likely to find Section 1, which introduces the logistic model, to be the hardest of the sections in Chapter 9. You may want to allow extra time for that first section. Once students have become familiar with the logistic model, the other three sections can be covered more quickly because they track what students will have seen already in connection with multiple regression in Unit A.

Section 9.1—Choosing a Logistic Regression Model. The logistic model is designed for data with a binary response \( Y = 0 \) or \( Y = 1 \). The model has both a chance part and a systematic part. It assumes (1) that for each value of \( X \) the values of \( Y \) are conditionally independent with the same probability \( p = P(Y = 1 | X) \), and (2) the log odds \( \log(p/(1 - p)) \) is a linear function of \( X \): \( \beta_0 + \beta_1 X \).

For all the reasons set out in the Unit C introduction, the logistic model is hard for students and takes time to learn. In this introductory section we have tried to present the model in small pieces while relying on applied examples.

- The model is designed for binary response data: We offer several examples. Students can suggest others based on these examples.
- If there are only two possible values for \( Y \), fitting a line to a scatterplot of \( Y \) versus \( X \) is almost never useful. Instead of fitting \( Y \) values we fit probabilities.
- Ordinary regression is not suitable for fitting probabilities because it can give values <0 or >1. Instead we work with probabilities transformed to log odds. Log odds can take on any real value.
- The model assumes that log odds, \( \log(p/(1 - p)) \), are a linear function of \( X \). (This is akin to the doctors example from Unit A, where the square root of the original response was a linear function of \( X \).)
- The model has two forms, linear for log odds, logistic for probabilities. (More below.) If we rewrite the model so that it is expressed in terms of \( \pi \) instead of \( \log(\pi/(1-\pi)) \), a graph of \( \pi \) as a function of \( X \) is not linear. (That’s why you see the logistic curve on these plots.)

**Odds and log odds.** Depending on the background of your students you may want to spend class time to drill first with odds and then with log odds. (“If \( p = 1/3 \), find the odds. If the odds = 1/3, what is the probability?”) Some students may find that visualizing a spinner helps: the probability is the ratio of the part to the whole; the odds is the ratio of the part to the other part. A sometimes-useful selling point for log odds is the symmetry: if \( p = 1/2 \), the log odds = 0, the natural midpoint of the real line. For any \( p \), the odds for \( p \) and for \( 1-p \) are reciprocals, so their log odds add to zero. Exercises 9.2 and 9.3 offer practice.

**Two forms of the model.** It can help to describe the models using both words and symbols:

- “Log odds are linear in \( X \)” \[ \log(\pi/(1-\pi)) = \beta_0 + \beta_1 X. \]

- “Probabilities follow a logistic curve” \[ \pi = \exp(\beta_0 + \beta_1 X)/(1 + \exp(\beta_0 + \beta_1 X)). \]

It can also help to use words to remind students how to back transform: To get from log odds to probabilities takes two steps. First exponentiate: “odds equals e-to-the-log-odds.” Then add 1 and form the ratio: probability = odds/(1 + odds).”

**Applied context.** Because the logistic model is so abstract, in the sense that what you fit is so remote from what you can see, it is important to spend time with concrete applied contexts. Ask students to “guesstimate” and then compute probabilities that correspond to various values of \( X \). The purpose is two-fold: first to make back-transforming feel routine and straightforward, and second to make the logistic model feel like a model that actually tells something useful about the data and context.

Exercises that can be assigned at this point: 1 – 3, 6, 8 – 11, 13, 14, 16

**Section 9.2—Logistic Regression and Odds Ratios.** This section is less demanding than 9.1. Covering the content will likely take less time. However, students may still be panting hard after Section 1, so a bit of a breather could be welcome.

There are two main ideas: what an odds ratio is and the fact that, for logistic models, the log of the odds ratio is constant and equals the slope. All this is introduced in the context of 2 \( \times \) 2 tables of counts. The response is binary and the predictor is binary. Such 2 \( \times \) 2 tables are a special case of logistic regression for which the analysis is simpler than for more complicated data sets. The simplicity makes it easier to pay more attention to the applied contexts, which in turn can help students who are still struggling with the abstraction of the logistic model.
Our lead example uses data on a treatment for headaches. The predictor $X$ is treatment (1 if treated, 0 if placebo). The response $Y$ is outcome (1 if the pain got better, 0 if not). For each value of $X$ we can compute the odds of improvement, and then form the odds ratio $= \frac{\text{odds for treated}}{\text{odds for placebo}}$. Note that for each $X$ the odds is already a ratio, which means that the odds ratio is a ratio of ratios. This can be a pitfall for the unwary.

If we plot the empirical log odds, $\log\left(\frac{\hat{\beta}}{1-\hat{\beta}}\right)$, versus $X$ for the two values of $X$, there are only two points, at $X = 0$ and at $X = 1$. The slope of the line through these two points is

$$slope = \frac{(\log \text{ odds at } X = 1) - (\log \text{ odds at } X = 0)}{1 - 0}$$

which is the log of the odds ratio. The same is true when the predictor is quantitative as long as the two $X$ values differ by 1. Students may find it useful to remember the phrase “e-to-the-slope equals the odds ratio when $X$ goes up by 1.”

Exercises that can be assigned at this point: 4, 5, 12, 15, 23

**Section 9.3—Assessing the Logistic Regression Model.** This section is about checking conditions for formal inference. There is more about other important elements of model assessment in the section “Assessing Logistic Regression Models” of Chapter 11. The focus here is on three conditions required for formal inference: linearity, randomness, and independence.

**Linearity.** The main check is the empirical logit plot, which groups data by $X$ values and plots group summaries to create a scatterplot that should look linear if the logistic model fits well. (In this context “linear” means “not curved”; moderate scatter around a line is allowed.) If your plot is curved, a transformation of $X$-values may give a linear plot.

**Randomness and independence** must be evaluated by thinking about how the data were produced. Except for randomized experiments and random samples, evaluation will call for judgment and opinions may differ. There are no commonly accepted guidelines, and so we have relied on a variety of examples to illustrate a range of possibilities. Keep in mind that a good statistical analysis need not, and often should not, result in a definitive yes-no conclusion.

Exercises that can be assigned at this point: 7, 17, 18
Section 9.4—Formal Inference: Tests and Intervals. The emphasis here is on (1) how to get p-values and intervals from computer printouts, and (2) how to interpret those p-values and intervals in applied contexts. It is important to keep in mind the checks of Section 9.3: whether and to what extent a data set was produced in a way that justifies using a probability model.

Getting p-values and intervals for logistic regression is a lot like what you do for ordinary regression, except that the particular statistics are different: z instead of t, chi-square instead of F. For a more detailed comparison, see the table at the end of the chapter summary. For both ordinary and logistic regression, p-values and intervals come from a standardized regression coefficient referred either to a t-distribution (ordinary regression) or the normal as a large-sample approximation (logistic regression). For ordinary regression, there is also a nested F-test based on the reduction in residual sum of squares. For logistic regression the residual deviance $-2\log L$ plays the role of residual SS, and the reduction in $-2\log L$ is referred to a chi-square distribution to get a p-value.

Exercises that can be assigned at this point: 19 – 22, 24 – 28

Chapter 9 Exercises

Exercises 29 – 33 are best assigned after the whole chapter has been covered.

Chapter 10—Multiple Logistic Regression

The role of Chapter 10 in Unit C is a lot like the role of Chapter 3 in Unit A. Just as Chapter 3 extends the regression model to multiple predictors, Chapter 10 does the same for the logistic regression model. Most topics in Chapter 10 will have appeared before, either in Chapter 9 for the logistic model or in Chapter 3 on multiple predictors. We have written Chapter 10 to rely on these connections.

The equation for the multiple logistic regression model has log odds on the left as in Chapter 9 and a linear combination of $X$-variables on the right as in Chapter 3. As in Chapter 9, there are two forms of the model; as in Chapter 9, finding fitted probabilities requires back-transforming; and as in Chapter 9, $z$ and $-2\log L$ are key summary statistics with normal and chi-square distributions. The right-hand side of the model equation is a copy of what we used in Chapter 3. As in Chapter 3 choosing a model involves deciding which predictors to include in the right-hand side and whether to transform any $X$-variables or to include interaction terms.

If students are comfortable with these topics from Chapters 3 and 9, Chapter 10 should go smoothly with no major new concepts. If students are still struggling with past material, Chapter 10 can serve as an opportunity for review and practice in a slightly new context.
Section 10.1—Overview. As set out above, we have tried to base our exposition on connections to topics from Chapters 3 and 9. The purpose of Section 1 is to call students’ attention to these many connections. We have deliberately gone into some detail in the hope that our preview will make it easier for students to recognize and rely on what they have already seen, and perhaps also, to deliver a “heads up” about topics that might be worth a review.

However (see the footnote in Section 10.1 of the book), we recognize that some readers may prefer to go directly to the examples of Section 10.2. Section 10.1 can be skipped, although we recommend returning to it later.

Exercises that can now be assigned: none

Section 10.2—Choosing, Fitting, and Interpreting Models. Section 10.2 introduces multiple logistic regression in the context of three examples based on two data sets. The first example is simple and straightforward. We analyze the relationship between a key U.S. Senate vote (Yes/No) and two explanatory variables: a Senator’s party affiliation (Democrat or Republican) and lifetime contributions from the oil industry. For this data set, the choice of model is clear-cut. We have only two predictors, empirical logit plots show linearity, and there is no evidence of interaction. We give the fitted model and use it to review back-transforming and the interpretation of the regression coefficients, along with a reminder, that as in Chapter 3, the interpretation carries an essential caveat about “allowing for simultaneous change in other variables.”

For many data sets, such as the data on admissions to medical school in Examples 10.2 and 10.3, choosing a model will not be straightforward. We have only 55 cases, half a dozen possible explanatory variables, some non-linearity, and evidence of some interactions. In such a situation, it is typically not possible to find a single model that is clearly and uniquely “best.” Examples 10.2 and 10.3 illustrate two approaches to model-searching: relying on pattern (10.2) and context (10.3). The first of the two examples uses a succession of empirical logit plots to explore relationships that might be useful in a model for the chance that a candidate will be accepted. The second uses context-based questions to choose pairs of models to be compared using a drop-in-deviance test. The question “If we know the grade-point averages of the applicants, do we improve the model by including the MCAT score?” leads to a pair of models that both include GPA, one with and one without the MCAT score.

The three examples of this section were juxtaposed in order to contrast them: A first, straightforward example where many if not most statisticians would end up with the same one model and a second data set for which no single model dominates all others; an empirical approach to model-searching based on plots; and a context-based approach using model pairs to address particular questions.
Exercises that can be assigned at this point: 1—6

**Section 10.3—Checking conditions.** The conditions you need to check, and the ways to check them, are essentially the same here as in the corresponding section of Chapter 9. You check linearity with empirical logit plots, and you check the probability part of the model (randomness and independence) by thinking carefully about how your data set was produced. Because none of this essential content is new, the section offers a chance for review and reinforcement.

At the same time, however, the section does push beyond the basics in two ways. First, we describe two families of data transformations that can be used to straighten curved relationships. Second, we illustrate an ad hoc analysis designed to check for independence in the data on the Senate vote. If time is short, either or both of these could be treated lightly or even skipped.

Exercises that can be assigned at this point: None

**Section 10.4—Formal Inference: Tests and Intervals.** Very little in this section is truly new, and so it offers an opportunity to review and reinforce the content that was introduced in Section 4 of Chapter 9. The two test statistics are the same here as in Chapter 9: (1) the standardized regression coefficient, which is roughly normal when samples are large enough, and (2) the drop in residual deviance, which is roughly chi-square when samples are large enough. These test statistics are computed, displayed in computer output, and used for inference in essentially the same way as in Chapter 9. Because the basics are the same as before, the section is devoted mainly to a set of five examples, 10.8 – 10.12, chosen either to review the basics, or to highlight issues that arise with multiple predictors.

- 10.8 illustrates a simple z-test and interval, with a focus on mechanics.
- 10.9 illustrates a possible misunderstanding of the z-test. When you have multiple predictors, the *p-value for a predictor does not tell you whether that predictor belongs in the model*. The p-value is conditional: it assumes that all the other predictors are in the model.
- 10.10 illustrates a simple use of the drop-in-deviance (nested likelihood ratio) test, with a focus on the mechanics.
- 10.11 shows that *sometimes the p-values from z and likelihood ratio tests may not agree*. Both p-values are based on approximations that work better when samples are larger. If samples are not large enough, the two approximations may give somewhat different answers.
- 10.12 shows how you can use a set of likelihood ratio tests to choose predictors to include in the model.

Exercises that can be assigned at this point: 7—18, 20 – 26; Note that 22 cannot be done in Minitab (the solution doesn’t converge). Also, 23(c) cannot be done in Minitab because Minitab does not compute Cook’s Distance for logistic regression.
Section 10.5—Case Study: Bird Nests. As with other case studies in the book, this one presents no new concepts or methods. Instead, it deploys the content of the chapter in service of an extended statistical analysis of a data set of scientific import.

Chapter 10 Exercises

Exercises 27 – 33 are best assigned after the entire chapter has been covered.

Chapter 11—Additional Topics in Logistic Regression

While you have been reassured repeatedly that much of what you will find in Unit C is really not all that new, in Chapter 11 you will find methods that appear to be new and foreign. But, in fact, the topics introduced in Chapter 11 have familiar objectives: fitting a regression model, assessing the conditions of the model, assessing its fit, and using the final fitted model. These were all done in earlier chapters and these will continue to be explored in this chapter in the context of logistic regression.

In addition, like Units A and B, a randomization test is presented for logistic regression. You will recognize the same fundamental ideas here and if you are feeling comfortable with the material from Chapters 9 and 10, understanding and operationalizing this test will be straightforward.

The section on analyzing $2 \times 2$ tables in Chapter 11 provides a bridge from more introductory level statistics to logistic regression. In many ways it illustrates a general theme of the STAT2 book: that many of the introductory methods encountered in beginning statistics courses as well as more sophisticated, advanced methods share a common modeling structure.

As with other additional topics sections in STAT2, each of these topics are stand-alone so that all, some, or none of these topics need to be included in a course.

Topic 11.1—Fitting the Logistic Regression Model. This section will allow you to get into the nitty gritty details of the new method introduced for fitting logistic regression models. This material will be challenging for most students, so if you plan to include maximum likelihood in all its glory, plenty of time will be needed. This section has two major parts which will be helpful in learning maximum likelihood. The first part provides a general introduction to maximum likelihood and the second part looks at maximum likelihood in the context of logistic regression. The notation and terms of the latter part can appear daunting and can obscure the ideas. Taking time to learn the fundamental ideas of maximum likelihood in the first part will ease the transition to understanding its application in logistic regression.
The notion of maximum likelihood is briefly mentioned in Chapter 9. The details of this approach to fitting models are presented in this section. With ordinary regression, least squares provides an intuitive approach for getting a best fitting line. At first glance, the method for fitting a logistic regression does not appear quite so intuitive. However, there is appealing intuition for the maximum likelihood principle, and a careful reading of this section will help you to gain an appreciation for this elegant approach to fitting a model.

**Model fitting in general.** Discussing the principle of maximum likelihood is a good time to reinforce the big picture of model fitting. When we model we observe data, and we use it to select values or estimates for unobservable parameters, here \( \beta_0 \) and \( \beta_1 \). There are all sorts of ways in which this can be done. Maximum likelihood is one such way. Least squares is another. In ordinary regression, we find estimates for our parameters, \( \beta_0 \) and \( \beta_1 \) (which specify a particular line), by minimizing the sum of the squared vertical distances between our data points and the fitted line. With logistic regression the parameters, \( \beta_0 \) and \( \beta_1 \), can be used to describe how likely the data are for different choices of \( \beta_0 \) and \( \beta_1 \). Our best estimates for the parameters, \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \), are those for which our data are most likely, that is, the maximum likelihood estimates. At this juncture it doesn’t hurt to once again remind students of the distinction between data and parameters, our reason for modeling, and the fact that different criteria can be used for finding parameter estimates. In fact, the principle of maximum likelihood is used in many more settings beyond logistic regression. And least squares estimates satisfy the principle of maximum likelihood! Finally, this business of model fitting is not restricted to the methods of least squares and maximum likelihood but extends to include many other approaches.

- **Likelihood:** Likelihood is tricky for students because when we begin our study of probability and statistics, we consider different data values and ask what is the chance of observing that data for a particular parameter such as \( \mu = 0 \). The data vary, the parameter, \( \mu = 0 \), remains the same. But with likelihood it is the parameters that vary and it is the data that remain the same. For example we may select a sample and summarize the data using \( \bar{X} \). With likelihood we take these data, \( \bar{X} \), and consider how likely it is for different possible parameter values. The data don’t change, the parameter values do.

- **The Tale of the Three Spinners:** This simple, concrete example is an effective way to introduce students to likelihood and the maximum likelihood principle. Have the students read this section carefully before class and then try out a similar set of questions together in class in pairs or small groups. There are sure to be many questions, but an in-class activity and discussion can be a good way to engage the students in learning this challenging material. *If students can use a set of real spinners and gather their own data at the start of class, this activity is likely to fix these maximum likelihood ideas.*

A wrap-up for this activity should highlight:
the distinction between the parameters and the data. Come up with an example or two that the spinner example simulates

- a description of the objective of maximum likelihood in the spinner example
- articulation of the meaning of a likelihood. What does the likelihood tell you?
- how parameter values change with likelihood, data do not
- that spins are independent so the factors in the likelihood can be multiplied
- the use of the calculated values for the likelihood tables to find maximum likelihood estimates
- graphing the likelihood
- using the likelihood graphs to find maximum likelihood estimates

- **MLE for logistic regression.** The three-spinner example provides a clean, relatively simple introduction to the principle of maximum likelihood: Choose those parameter estimates where your data are most likely. It is important to hang onto those ideas because the fundamentals remain the same when we use maximum likelihood in logistic regression, but it appears to be considerably more complicated. Returning to the three-spinner ideas will help.

Here are the steps in constructing the MLE for the logistic regression model:

- Each factor in a likelihood for logistic regression is a probability corresponding to each data point. Each observation in our dataset will contribute a factor to the likelihood much like each spin in the three spinner example contributed a factor to those likelihoods.

- Each data point \((X, Y = 0 \text{ or } 1)\) can be used to write a probability. You have seen the following form of the logistic regression model that writes \(\pi\) as a function of the parameters, \(\beta_0\) and \(\beta_1\), in numerous places in the text.

- If \(Y = 1\), the probability has the form:

  \[
  \pi = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}
  \]

- If \(Y = 0\), the probability has the form:

  \[
  1 - \pi = \frac{1}{1 + \exp(\beta_0 + \beta_1 X)}
  \]
When constructing the likelihood, we use one of these forms depending upon \( Y \) and we plug in the corresponding value \( X \) for each observation (\( X, Y = 0 \) or 1).

If you now look at your likelihood, there are no \( Y \)'s, there are no \( X \)'s. The only unknowns are the values for the parameters, \( \beta_0 \) and \( \beta_1 \). Just what we are looking for!

Conceptually we try different values for \( \beta_0 \) and \( \beta_1 \), and compare the likelihood for each combination. The combination of values for \( \beta_0 \) and \( \beta_1 \) where our data are most likely, where the likelihood is a maximum, would make good estimates of \( \beta_0 \) and \( \beta_1 \). These would be the maximum likelihood estimates.

**Computing an MLE:** We are grateful for the software programs that compute these estimates for us. Unlike ordinary regression, there is no set of formulas we can use to compute the estimates for ourselves.

There are a lot of moving parts when finding maximum likelihood estimates in logistic regression: complicated notation, multivariable calculus, numerical methods instead of closed form formulas. The two new critical concepts introduced here are the likelihood and maximum likelihood estimation. Take time to describe what a likelihood function does, how it takes possible values for our parameters and returns information on how likely our data are for those particular parameter values. Relate this to the three-spinner example where things are considerably simpler. It is not necessary to write out the likelihood although some students may be ready and capable of doing so. For maximum likelihood estimation, appeal to the students' intuition. What would be good choices for the parameter estimates? The concept of a likelihood function can direct students' attention to the intuitive choice of parameter estimates where our data are most likely.

**Topic 11.2—Assessing logistic regression models.** We begin with the sub-topic "assessing the utility and fit." Students should see the parallels here with tests used in ordinary least squares. Tests based on the difference in the likelihood are new in this unit, but they have intuitive appeal. Both types of assessments can be accomplished by comparing nested models.

**Utility**
- \( H_0 \): No predictors are needed
- \( H_1 \): Model with predictors
If you can't reject this null, the predictors have little utility.
Fit

- **H_0:** Model with your predictors
- **H_1:** Saturated Model (model with a predictor for every observation)

If you reject this null, it suggests that there are more predictors out there that may be helpful. Typically you are not interested in the saturated model because it is too finely defined. Furthermore, the saturated model cannot be constructed if there are no replicate observations at most levels of X. This approach to assessing the fit of a logistic regression works for the golf putting data where there are many observations for each distance, but it does not work for the Medical School data where there are few replicates per GPA.

We use the reduced-versus-full nested F-test to compare models, as follows:

**Likelihood Ratio Test**

\[ X^2 = -2\ln(L_{\text{Reduced}}) - (-2\ln(L_{\text{Full}})) \]

Note that the *deviance* for a model is

\[ -2\ln(L_{\text{Reduced}}) - (-2\ln(L_{\text{Saturated}})) \]

**Drop in Deviance Test**

\[ X^2 = -2\ln(L_{\text{Reduced}}) - (-2\ln(L_{\text{Saturated}})) - (-2\ln(L_{\text{Full}}) - (-2\ln(L_{\text{Saturated}}))) \]

\[ X^2 = \text{Deviance}_{\text{Reduced}} - \text{Deviance}_{\text{Full}} \]

The term \(2\ln(L_{\text{Saturated}})\) is the same for each term, so we can write the Likelihood Ratio Test as a Drop in Deviance.

These tests, particularly the Drop-in-Deviance Test, are used frequently in logistic regression modeling, so students should become familiar with them. It is a good idea to reinforce meaning for either kind of test. The likelihood provides a reading on how likely our data are under the particular model. We’d like to be able to settle on a model for which our data are most likely. Since we are dealing with \(-2*\log\) likelihood, we need to choose the model with the smallest \(-2*\log\) likelihood. Have students spend some time playing with the likelihood Excel sheet to develop a feel for this.

The deviance on the other hand is providing a reading on how different our observed values are from the predicted. The deviance for a model is made up of the sum of a deviance that is calculated for each observation. Given the choice, we’d like to pick the model with the smallest deviance.

We next introduce the subtopic of “assessing the conditions of a logistic regression model.” Here are some suggestions for that sub-topic.

The take home for **diagnostics** for logistic regression is that the methods for assessing models are significantly fewer and less powerful than the elaborate set of diagnostic tools that you learned about to assess ordinary least squares regression models. If the goal is inference, concerns investigated by finding out about data collection can be
much more critical than other kinds of diagnostics. As with many of the other models that you have encountered throughout the text, some reassurance that the observations are independent is desirable. If you have reason to believe the observations are not independent, there are other methods that are useful for analysis in this circumstance. More troublesome is data collection that has possibly introduced a bias of some kind, either by the sample frame employed or how the sampling was done or bias introduced by non-response.

**Example 11.5: Campaign contributions and votes** provides a useful example to discuss the assessing of conditions. We find that residual plots do serve a purpose. For example, a Pearson residuals plot with the this example is helpful. Unlike ordinary least squares regression, with logistic regression there are several different kinds of residuals that can be calculated.

1. **Deviance residuals**. Each observation has an individual deviance associated with it and when summed the individual deviances add up to the residual deviance. These can be plotted and examined for trends or extreme observations in order to assess the conditions for the model. Don’t expect these plots to adhere to the same expectations we have for residual plots for ordinary least squares regression. For example, issues related to equal variance do not apply here.

2. **Pearson residuals**. These residuals have the familiar form found in residuals in ordinary regression. The numerator is the difference in the observed count and the predicted count and the denominator scales the residuals using the model variance. While the form is like ordinary regression, don’t expect Pearson residuals to behave similarly in logistic regression. Nonetheless, these plots can be useful for spotting trends such as curvature or unusually large observations.

3. **Conditions related to the predictors**
   (a) **Leverages** refer only to the $X$ values and can be assessed in logistic regression just as they are assessed in ordinary regression.
   (b) **Cook’s Distance** measures the influence of a point when fitting a regression and depends upon a point’s leverage and standardized residual. Leverage doesn’t depend upon the observed values of $Y$; the standardized residual does and so this measure is not identical in meaning to Cook’s Distance in ordinary regression. It can still be helpful in identifying influential points.

The differences between these diagnostic measures in logistic regression and their counterparts in ordinary regression are subtle, and plots can be useful when looking for extreme observations or suspicious patterns. Students may need to be reminded, especially if they were well trained in residual analysis in ordinary regression, that these measures are not expected to be interpreted similarly for logistic regression.

A look at the R command for producing residuals reveals some of the options for creating residuals with logistic regression models:
residuals(object, type=c("ordinary", "score", "score.binary", "pearson", "deviance", "pseudo.dep", "partial", "dfbeta", "dfbetas", "dffit", "dffits", "hat", "gof", "lp1"), pl=FALSE, xlim, ylim, kint, label.curves=TRUE, which, ...)

“Object” is the name associated with the fit of the model. As far as type of residual goes, we have referred to “pearson” and “deviance” but there are a number of other options possible.

We next consider the sub-topic “assessing prediction.” Logistic regression can be used for describing relationships or for prediction. Given predictor values, a prediction based on a logistic regression model is a probability ($\hat{\pi}$), whereas the observed values for $Y$ are 0 or 1. There are a number of different ways in which to assess a logistic regression’s prediction. One that we introduce is the c-statistic. Consider all possible pairings of a success ($Y = 1$) and a failure ($Y = 0$) in a data set and determine the proportion of pairs where $\hat{\pi}$ is greater for the success in the pair than the prediction $\hat{\pi}$ is for the failure.

An even simpler way to assess prediction is to construct a 2 x 2 table of counts with actual observations (Failure or Success) cross-classified by the logistic regression prediction, where, say, a $\hat{\pi} > .5$ is a predicted success and a $\hat{\pi} < .5$ is a predicted failure. (One can use cutoff values different from .5.) A model that predicts well should produce a table with a high percentage of the counts in the (success, success) or (failure, failure) cells of the table.

By and large, Pseudo-$R^2$s are not to be encouraged.

Exercises: Note that Exercise 11.10 (b) and (d) can’t be done with R.

Section 11.3—Randomization Tests for Logistic Regression. If you have tried out the Randomization Tests for the other Units, you will find that Randomization Tests in logistic regression are quite similar. Students may wonder why we would need a Randomization Test if a test can be performed using logistic regression. Those logistic regression tests depend upon approximations and therefore require relatively large samples. In the absence of large samples, it is not a bad idea to check out Randomization Test results.

This section demonstrates how randomization tests can be used for discrete and continuous explanatory variables. It doesn’t matter if the randomly assigned explanatory variable is discrete or continuous, odd ratios can still be obtained using the logistic regression command and the observed odds ratio can be compared to those distributed randomly.

The archery data comes from an observational study, so no randomization has taken place. The randomization test can be more accurately described as a permutation test (where the values of the explanatory variables are permuted). Without true
randomization, the interpretation of the results cannot include a cause-and-effect statement.

Section 11.4—Analyzing Two-way Tables with Logistic Regression. Using logistic regression to analyze tables is yet another way in which we see that many methods introduced in first courses, here a chi-square test, fall under the general rubric of modeling. For tables where the response is binary—yes or no, success or failure, 0 or 1—and the explanatory variable is categorical with $K$ levels, we can present the data as a $2 \times K$ table and use logistic regression to investigate the nature of the relationship between the explanatory and response variables.

Structuring the data from two-way tables to use in a logistic regression can be a little tricky. Because the data are reported as counts there are two steps needed to perform the analysis.

1. **Enter the data as a table** into R or Minitab. See the appropriate software companion to see how to do this.

2. **Perform a binomial regression** using the number of responses that are successes and the total number of trials for each level of the explanatory variable.
   (a) Note that the number of coefficients estimated will be one less than the number of levels there is for the categorical explanatory variables. The missing level is the reference level. In R and Minitab, the reference level used is determined alphabetically, the level with a name closest to a. It is possible in R or Minitab to specify a different reference level.
   (b) It is important to remind students that the tests for these coefficients are making comparisons to the reference level. It may be that the individual $p$-values are not significant, but that does not imply that individual coefficients should be removed. It may even be that none of the individual $p$-values are significant, that still does not suggest that the explanatory variable is useless. To judge the utility of the explanatory variable, it is best to use a drop in deviance test. Think of all the individual levels as being “part of a package.” Resist the temptation to toss out individual levels.
   (c) The preceding assumes that you allow your software package to define the indicators for individual levels. Of course, you can create the indicators yourself and leave out the one you prefer to use for the reference level. This approach also has the advantage of letting you name each indicator, which can often be much more informative than the names R or Minitab will assign each level.
   (d) This flexibility in constructing binomial models for tables means that you will obtain different coefficients depending upon the reference level.

3. **Using the model** requires you to input 0s or 1s to obtain the **Odds Ratio** of interest.
4. **The drop in deviance test for the binomial regression**, which compares the model with no explanatory variable to the model that includes all of the levels of the categorical explanatory variable less one, is very close to the **chi-square test** taught early in a student’s statistical education.

5. **The advantage of the logistic regression** is that it is **easy to add continuous explanatory variables** to the model where that is not possible when analyzing tables with chi-square tests.
Individual Instructor Experiences Using STAT 2

We include here five essays that describe how five different instructors have used the textbook in an appropriate “Stat 2” course.

Essay #1: by Ann Cannon, Cornell College

When I taught the course Statistical Methods II from the Preliminary Edition of this book, I had a class of 13 students. All class meetings were in a computer lab with one computer per student. A third of the class was in their first two years of college, the other two-thirds were in their last two years of college. The majority of students were Psychology majors, but other majors accounted for included Biochemistry, Economics and Business, Physics, Mathematics, and Art History. The three first-year students had not yet declared a major. Several had taken the AP statistics course in high school, though most had taken the introductory statistics course at Cornell. All sections of the introductory course at Cornell College cover inference for one and two samples for means and proportions. Slightly fewer than half of all sections cover inference for the slope in a simple linear regression setting, another slightly fewer than half cover chi-square goodness of fit tests and chi-square tests for independence, and the remaining sections do not cover anything beyond two-sample tests. This means that some students in the STAT2 course had seen inference for the slope, some had seen chi-square, some had seen both (AP students), and some had seen neither.

The assessments for the course included 2 quizzes, 2 exams (a midterm and a final), 2 group projects, and one individual project. For the individual projects and one of the group projects, I provided data sets and asked the students to do an analysis and write a report. The first project required the use of multiple regression, the second used ANOVA. The second group project was a comprehensive data analysis project where students were required to come up with a question, collect data to answer that question, perform an appropriate analysis, write a report, and give an oral presentation to the class.

Because of the vast differences in backgrounds, and because the modeling idea is somewhat foreign to these students, I started with Chapter 0, spending about a week introducing the four-step plan and what modeling looks like for a two-sample means problem.

---

4 Cornell College actually uses the Block Plan where students take, and faculty teach, one course at a time (OCAAT) for 18 class days. I meet with students roughly 3.5 hours a day, 5 days a week (9 – 11 and 1:15 – 2:45). I have loosely translated the amount of time that I spend on topics from the OCAAT schedule to a semester schedule.
Next we spent about 4 weeks working through Chapters 1 and 2 covering Simple Linear Regression. I gave the first quiz after Section 2.1, covering Chapter 1 and Section 2.1 (this was about 2 weeks into Unit A). We finished the regression unit with about 3 weeks on the multivariate issues covered in Chapter 3. Because of timing issues I did not spend much time on Sections 3.5 and 3.6, though I later regretted not spending more time on 3.6 (nested F-tests), because it would have been helpful for the students to see this material before we got to logistic regression. I did not cover any topics from Chapter 4, and the mid-term exam covered all of Unit A.

We moved from Unit A straight into Unit B, covering Chapters 5 and 6 completely and Levene’s test (Topic 7.1), ANOVA and Regression with Indicators (Topic 7.5), and Analysis of Covariance (Topic 7.6) from Chapter 7. This took about 4 weeks. The second quiz covered Chapter 5.

We finished the course by spending about 3 weeks on Chapters 9 and 10. The final exam was comprehensive with a slightly heavier emphasis on logistic regression.

**Essay #2: by Jeff Witmer, Oberlin College**

I taught a course entitled Statistical Modeling in the spring semester of 2012 using a draft of the book; here is a summary of the semester. My students had mixed backgrounds: about one-third of them had AP Statistics as their preparation, others had taken a STAT1 course from my department, but some had other preparation (e.g., Research Methods from the Psychology Department). I began by collecting data from the students on the first day of classes, asking them (a) how many miles they were from home and (b) their sex. (I had the students use http://www.distance24.org/ as a tool for finding “miles from home.”) This gave me a dataset to use while teaching Chapter 0, where we reviewed the two-sample t-test, introduced the language of modeling, and introduced the software (I used R—which was new to most of the students). Some class meetings were held in a regular classroom but some were held in a computer lab.

After spending one week on Chapter 0 we spent six weeks on Unit A, which included covering all of the topics in Chapter 4. In fact, I sprinkled Sections 4.1 and 4.3 into earlier chapters in the unit, while saving some of Chapter 4 until after finishing Chapter 3.

In addition to weekly homework sets, I had the students complete two projects during Unit A, one on simple linear regression and one on multiple regression. The projects were built on the following idea: I had each student choose a city (a zip code actually) and find a sample of 25 or so homes recently sold in that location using the website zillow.com. For each home they recorded the sale price, the number of bedrooms, number of bathrooms, lot size, and square foot size of the home. For the projects in Unit A each student analyzed data from his or her chosen city. Near the end of Unit C I included a project in which the response variable was whether or not a home in the
student’s data set had at least two bathrooms. The project for Unit B combined home price data from all of the students with a pair of categorical predictor variables that I created: Region of the country (East, Midwest, South) and CitySize (Small, Medium, Large).

I gave an exam before the end of Unit A, with the students working in a computer lab to analyze data. A little over half of the exam required on-the-spot fitting of models, creation of graphs, etc., while the other half of the points could have been earned without using a computer.

I spent the last day before spring break having a data analysis contest. I divided the students into teams, gave them a set of data on birth weights of babies, and a question to answer: “Estimate the birth weight of a baby born to a mother with the following characteristics and characterize the uncertainty of your prediction.” I allowed them 45 minutes to come up with a model plus an answer to the question. The team with the best analysis won a few bonus points.

After Unit A I spent three weeks on Unit C, at the end of which I gave a second exam that included some Unit A material not covered on the first exam. I spent the final three weeks of the semester covering Unit B. The final exam emphasized Unit B, but included material from the entire semester.

A couple of times during the semester I spent the better part of a class presenting a case study. Near the end of Unit A I presented an analysis of stimulus package dollar allocation to congressional districts. The data set StimCapFinal.csv lists the amount of money from the 2009 economic stimulus package that went to each district together with a host of other variables including whether the member of the House of Representatives was a Republican and whether the vote in that district in the 2008 presidential election (“Margin”) was close. A cynical student might think that the Democratic majority in the House at the time would have directed extra money to districts held by Democrats or where the electorate was close to evenly split for Obama and McCain. One might think that poverty rate and unemployment rate should factor in, among other variables. The file Stimulus analysis.R has an R script that shows what I did in class. A first level of analysis shows that “Republican” has a strong negative coefficient, but “Margin” also has a clear negative effect. Analysis of residuals turns up an important fact: Districts with large positive residuals include state capitals. Upon reflection this is not surprising—and some students might speculate this in advance—because some stimulus money was allocated at the state level and was sent to state capitals. Adding “Capital” to the picture produces a model in which “Capital” has a huge effect, “Poverty” and a few other variables have large effects, whereas “Republican,” “Margin,” and “Tenure” (how long the member has been in Congress) have more modest effects.
Near the end of Unit C I presented a case study of the vote in the House of Representatives in March 2010 that created “Obamacare.” For this analysis I started with exploratory histograms and boxplots, looking at how the percentage of uninsured persons in a congressional district was related to whether a representative voted yes or no, for example. The file ObamaCare Vote.R has the R script that I used in class. A measure of political conservativeness, dwnom1, is included in the HealthCareVote.csv data file and has a strong relationship with votes cast on the bill. Plotting points in blue for Democrats and red for Republicans is helpful and shows that a couple of districts were unusual. I fit several logistic regression models. Along the way I noticed an example of the kind of thing that happens when predictors are correlated: If ObamaMargin and dwnom1 are in the model then neither Private ($p = 0.21$) nor Uninsured ($p = 0.82$) shows up as an important added predictor, but the pair of them is important ($p = 0.033$). The final model that I settled on shows effects for ObamaMargin, dwnom1, Private, and Uninsured plus a quadratic effect for ObamaMargin, meaning that the bigger the margin for Obama in 2008 in a district, the more likely the representative was to vote “yes.” But the rate of increase in the likelihood of voting yes drops off, which says that there is a limit to the effect that ObamaMargin had on the vote.

**Essay #3:** by Julie Legler, Saint Olaf College

I taught our Stat 272 “Statistical Modeling” course using an early version of the STAT2 text in the Spring of 2011. I was fortunate to have the R-based PowerPoint slides that Jeff Witmer and Robin Lock had used. They were extensive and fairly accurate by that time. Having the slides was particularly useful because I was using R and it was good to have commands available for me and the students (and instructor), and the R Companion was not yet completed.

In the past I had taught this course out of *The Statistical Sleuth* (Fred L. Ramsey and Daniel W. Schafer, Duxbury, second edition, 2002). It was readily apparent to me that the students found STAT2 much easier to understand and in a format that was much more familiar to them. I found the students comprehension with STAT2 to also exceed how students had done with the *Sleuth*.

I used handouts such as the one that I describe below to have students actively involved during the class period. I used some introductory R work to review concepts the students should be familiar with at the start of the class.

There were two exams each with an in-class and a take-home portion. Scores were higher than my previous experience with *Sleuth*. The syllabus also called for a project with a series of expectations laid out for the duration of the course. The projects were by and large well done, and my impression was that the text was a good reference for their work on the projects.
Description of Sample Handout #1: Chapter 0 Examples

1. Can we use the number of miles that a used car has been driven to predict the price that is being asked for the car? How much less can we expect to pay for each additional 1000 miles that the car has been driven? Would it be better to base our price predictions on the age of the car in years, rather than its mileage? Is it helpful to consider both age and mileage, or do we learn about as much about price by considering only one of these? Would the impact of mileage on the predicted price be different for a Honda as opposed to a Porsche?

2. Do babies begin to walk at an earlier age if they engage in a regimen of special exercises? Or does any kind of exercise suffice? Or does exercise have no connection to when a baby begins to walk?

3. If we find a footprint and a handprint at the scene of a crime, are they helpful for predicting the height of the person who left them? How about for predicting whether the person is male or female?

4. Can we distinguish among different species of hawks based solely on the lengths of their tails?

5. Do students with a higher grade-point average really have a better chance of being accepted to medical school? How much better? How well can we predict whether or not an applicant is accepted based on his/her GPA? Is there a difference between male and female students’ chances for admission? If so, does one sex retain its advantage even after GPA is accounted for?

6. Can a handheld device that sends a magnetic pulse into the head reduce pain for migraine sufferers?

7. When people serve ice cream to themselves, do they take more if they are using a bigger bowl? What if they are using a bigger spoon?

8. Which is more strongly related to average score for professional golfers: driving distance, driving accuracy, putting performance, or iron play? Are all of these useful for predicting a golfer’s average score? Which are most useful? How much of the variability in golfers’ scores can be explained by knowing all of these other values?

Questions for each example:
1. What are the observational units?
2. What is the response variable, $Y$? Is it quantitative or categorical? Is it binary or not?
3. What is(are) the explanatory variable(s), $X$? Is it (are they) quantitative or categorical?
Is it (are they) binary or not?

4. CHOOSE: What kind of models might be useful for this example?
5. FIT: (We will later describe how to fit a model with R.)
6. ASSESS: (Later we will scrutinize the study design and use residual plots here.)
7. USE:
   (a) Is the goal of fitting a model in this example to make a prediction, to understand the underlying science of the situation, or to compare effects of various predictors on the response variable?
   (b) What kinds of inferences can be made in this example?
      - Where did the observational units come from? If randomly sampled, generalizations to a population should be possible.
      - How did the observational units get into groups being compared? If randomized, causal statements should be possible.

Description of Sample Handout #2:

A good first-day activity uses the web site of Shonda Kuiper at Grinnell, which engages students in electronic games. In this example students go to the web site www.cs.grinnell.edu/~kuipers/statsgames and played the game Memorathon. Students enter data about themselves and can select conditions of play, such as whether or not to have sound tones played as the various segments of the circle are lit up. The first class day we make two-sample comparisons like male – female or sound – no sound and use either a two-sample t-test or an R-based randomization test to do the inference, with me providing the R code for the randomization test.

Essay #4: by Brad Hartlaub, Kenyon College

I have taught Data Analysis (Math 206) out of previous drafts of this text twice, Spring Semester 2010 and Spring Semester 2011. Both times the course was taught in a computer-equipped classroom where students had access to personal computers and statistical software every class period. The course enrollments were 17 students in 2010 and 19 students in 2011. All of the students had completed our introductory statistics course (Math 106) or AP Statistics. The primary difference in the two courses was that SAS was used in 2010 and RStudio was used in 2011. Much to my surprise, the students embraced the change, and I did not hear any complaints about the start-up time needed to get familiar with R.

One of the pedagogical advantages of teaching in a computer-equipped classroom where Synchroneyes is available on the instructor’s machine is that students can easily share and discuss their models and analyses. The composition of the class was similar both years. The majors of the students in the class were Anthropology, Biology, Chemistry, Economics, Mathematics, Neuroscience, Political Science, Psychology, Physics, and Sociology. These varied student interests were extremely valuable and
provided for rich discussions when the students worked on group or individual projects, which were a major part of the class.

I based student course grades on the percentages in the table below.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Homework</td>
<td>20%</td>
</tr>
<tr>
<td>Small Group Project</td>
<td>15%</td>
</tr>
<tr>
<td>Exam 1</td>
<td>20%</td>
</tr>
<tr>
<td>Exam 2</td>
<td>20%</td>
</tr>
<tr>
<td>Final Project</td>
<td>25%</td>
</tr>
</tbody>
</table>

For the small group project, students were asked to solve a practical data analysis problem with at least one other member of the class. A written component (paper or poster) and/or an oral presentation to the class were required. For example, one semester many student groups analyzed data from a two-semester study at the University of Florida on the effect of an audience response system on student attitudes and achievement. These data were provided by Professor Megan Mocko. Other student groups developed models for the analysis of kinetic data from variants of a calcium-binding protein, HIV, Hardwood forest growth and mortality, and gender roles and relationships in first-year college students.

For the final project each student was asked to find a data set and apply an appropriate analysis. Ideally, the data set was collected by the student or obtained from a local resource. That is, I encouraged students to design and conduct their own experiments.

The coverage of the course was similar both semesters. We started with a review of statistical inference, expanding on some of the topics in Chapter 0. The students appreciated the review of simple linear regression in Chapter 1, especially since they were getting comfortable with new statistical software (SAS or RStudio). Chapters 2—4 in Unit A contained new material for almost all of the students in these classes. After completing all of the advanced topics in regression, we moved into ANOVA models (Unit B). One-way, two-way, and multifactor ANOVA models were chosen, fit, assessed, and used in numerous applied settings. Since many students were familiar with design of experiments from other courses or labs, we did not cover Chapter 8. Instead, we used the remaining time to focus on logistic regression models. We were able to spend more time with Chapter 10 (Multiple Logistic Regression) in Spring Semester 2011, and some of the students included the material from Chapter 11 (Additional Topics on Multiple Logistic Regression) in their final projects.
Essay #5: by Robin Lock, Saint Lawrence University

I taught our Math 213 course (Applied Regression Analysis—also known as Stat II) using the Preliminary edition of Stat2 in Fall 2011, after using earlier drafts of materials in previous semesters. The course met for three 60-minute sessions per week in a computer classroom, which capped the enrollment at 29 students. In the past I had used Minitab as the main software package, then a mixture of Minitab and R, and finally only R (through RStudio) in the latest iteration. Here is an outline of the timing and topics:

<table>
<thead>
<tr>
<th>Days</th>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1—2</td>
<td>Unit 0 Intro to Models and R/RStudio</td>
</tr>
<tr>
<td>3—5</td>
<td>Chapter 1 Simple Linear Model</td>
</tr>
<tr>
<td>6</td>
<td>Topic 4.3 Identifying Outliers/Influential Points</td>
</tr>
<tr>
<td>7—10</td>
<td>Chapter 2 Inference for Regression</td>
</tr>
<tr>
<td>11</td>
<td>Topic 4.5 Randomization Tests</td>
</tr>
<tr>
<td>12—19</td>
<td>Chapter 3 Multiple Regression (including Topics 4.1 &amp; 4.4)</td>
</tr>
<tr>
<td>20—21</td>
<td>Topic 4.2 Model Selection</td>
</tr>
<tr>
<td>22</td>
<td>Midterm on Unit A</td>
</tr>
<tr>
<td>23—24</td>
<td>Group project on regression/cross-validation</td>
</tr>
<tr>
<td>25</td>
<td>Topic 4.7 Bootstrap Intervals</td>
</tr>
<tr>
<td>26—29</td>
<td>Chapter 9 Logistic Regression (single predictor)</td>
</tr>
<tr>
<td>31—34</td>
<td>Chapter 10 Multiple Logistic Regression (including Topics 11.1 &amp; 11.2)</td>
</tr>
<tr>
<td>35</td>
<td>Topic 11.4 Logistic Regression and Chi-square Test for Tables</td>
</tr>
<tr>
<td>36</td>
<td>Exam on Unit C</td>
</tr>
<tr>
<td>37—39</td>
<td>Chapter 5 One-Way ANOVA (including Topics 7.2 &amp; 7.5)</td>
</tr>
<tr>
<td>40—42</td>
<td>Chapter 6 Multifactor ANOVA</td>
</tr>
</tbody>
</table>

In addition to the two exams listed above, a final exam, several quizzes, and a few sets of assigned problems, students worked on several data analysis projects identified below. Copies of the handouts follow this essay.

**Project #1—Simple linear regression.** Predicting used car prices based on age with individual, student-collected datasets.

**Project #2—Multiple regression.** Extend Project #1 to include mileage as a predictor.

**Project #3—Model building and cross-validation.** Using a dataset about red wines to predict quality. Groups worked on subsets of the data with an in-class competition to compare models for a holdout sample.

**Project #4—Binary logistic regression.** On student’s own choice of data (many chose sports themes to predict winners of competitions).

**Project #5—ANOVA for means.** Back to the car data from Projects #1 and #2 with each student getting data for another three students’ car models to give four groups to compare mean price.

A nice feature of these projects is that the results are unique for each student (or group), but still relatively consistent across the students to facilitate grading.
Math 213: Project #1    Due: Tue. 9/20/11
R. Lock    9/12/11    Data due: Fri. 9/16/11
Value = 25 points

Correlation & Simple Linear Regression
Applied to a Virtual Used Car Lot

SITUATION:
Suppose that you are interested in purchasing a used car. How much should you expect to pay? Obviously the price will depend on the type of car you get (the model) and how much it’s been used. For this project you will investigate how the price might depend on the age of the car (in years with 2011=1 year old). While collecting data, you’ll also want to find and record the mileage of the car. Code both the prices and mileage in thousands with one decimal place. Thus a 2003 Honda Civic with 75,670 miles that’s selling for $5,685 would be coded as age=9, miles=75.7, price=5.7. Please put the data in your dataset with the variable names age, miles, and price in that order for consistency.

DATA SOURCE:
To provide a convenient location to sample lots of cars, we'll do our “shopping” on the Internet. You may try one of the sites listed below or find one of your own. You should focus on a single car model (for example, a Dodge Caravan) when you search for price listings and try to choose a car model that’s been around for a while (so you get some variety in ages). Be careful if the results of your search are displayed in order. You should find prices for as random a sample as feasible with at least 25 cars. Be sure that you are getting prices for actual cars—not “blue book,” theoretical prices. Also, try to get a reasonable range of years.
Web sites to try: autobytel.com, autotrader.com, or, cars.com.

REPORT:
1. Start with an introduction that describes your data, source, and procedure for choosing the cars to include.
2. Use R to directly compute each of the following summary statistics. Include both the values and the R commands you use to find them.
   \[ \bar{x}, \ s, \ \bar{y}, \ s_y, \ SSX, \ SSXY, \ SSY \ (\text{same as SSTotal}), \ SSModel, \ SSE \]
3. Show how to calculate the least squares regression line that best fits your data using the values generated in (2)—you can confirm the result with R. Interpret (in context) what the slope estimate tells you about prices and ages of your used car model. Explain why the sign (positive/negative) make sense.
4. Again, using the values in (2) above, show how to estimate the standard deviation of the error term for your model.
5. Produce scatterplot of the relationship with the regression line drawn on it.
6. Produce residual plots and comment on how well your data appear to fit the conditions for a simple linear model. Don’t worry about doing transformations if there are problems with the conditions.
7. Find the car in your sample with the largest residual (in magnitude—positive or negative). For that car, use R to find its studentized residual, leverage, and Cook’s distance. Would any of these values be considered unusual? Specify the criteria you use for each measure.
8. Compute and interpret a 90% confidence interval for the slope of your model (show calculation).

9. Compute the value of \( r^2 \) two ways—as the square of the correlation and using the partitioned sums of squares from the ANOVA. Write a sentence that interprets the result as a percentage in context.

10. Test the strength of the linear relationship between your variables using each of the three methods discussed in class. Show hypotheses and the details for calculating the test statistic in each case. Indicate the reference distribution (t or F including degrees of freedom) and use technology to get any p-value(s). One conclusion should suffice for all the tests.
   \[ \Rightarrow \text{test for correlation} \]
   \[ \Rightarrow \text{test for slope} \]
   \[ \Rightarrow \text{ANOVA for regression} \]

11. Choose a particular value of age for which you are interested in predicting the price of a car. Show how to calculate each of the following quantities—by hand! You may confirm the values with R. Write sentences that carefully interpret each of the intervals (in terms of car prices) and show the distinction between them.
   \[ \Rightarrow \text{predicted value for } Y \]
   \[ \Rightarrow 90\% \text{ confidence interval for } \mu_Y. \]
   \[ \Rightarrow 90\% \text{ prediction interval for an individual } Y. \]

12. According to your model, is there an age at which the car should be free? If so, find this age and comment on what the “free car” phenomenon says about the appropriateness of your model.

13. Write a conclusion. Discuss your overall impressions of the linear model for describing your data. Point out any unusual data values, interesting phenomena, or obvious departures from regression assumptions.

Technology notes:

Note #1: Enter the data using a package (e.g., Excel, Fathom, or Minitab) that will let you save it as a comma separated (.csv) file. The first row should be the names of the three variables **age, miles, price**. You will need to upload that file to a folder in your RStudio workspace and then import the dataset.

Note #2: Your project report should be word-processed, with graphs, R commands, and output embedded along with interpretations. You do not need to typeset calculations—feel free to leave some space and write them in by hand.

Due dates:

**By Friday 9/16/11:** Please send me a copy of the .csv file with your data. In the text of the email message mention the model of car that you have chosen to use.

**By Tuesday 9/20/11:** Turn in a written copy of your report.
Multiple Regression Models for Car Prices

You will continue with the data you have collected for modeling the price of a certain model of used car (in $1,000’s), based on the age of the car (in years) and the number of miles driven (coded in 1,000’s). Please provide whatever computer output is needed as you answer each question. You do not need to replicate by hand any computations that the computer provides (unless specifically directed to do so).

Model #1: Use Age and Miles as predictors
(a) Run the model with two predictors (age and miles) for price as the response variable and provide the output (both the summary and the anova for the model).
(b) Find the residual for the first car in your sample. Show the actual computation for this part, based on your prediction equation and the data for that car.
(c) Assess the importance of each of the predictors in the model—be sure to indicate the specific value(s) from the output you are using to make the assessments. Include hypotheses and conclusions in context.
(d) Assess the overall effectiveness of this model (with a formal test). Again, be sure to include hypotheses and the specific value(s) you are using from the output to reach a conclusion.

Model #2: Polynomial models
One of the drawbacks of the linear model in the original project was the “free car” phenomenon that often occurred where the predicted price is eventually negative as the line decreases for older cars. Let’s see if adding one or more polynomial terms might help with this.
(a) Fit a quadratic model using age to predict price. Give the prediction equation and show a scatterplot of the data with the quadratic fit drawn on it.
(b) Does the quadratic model allow for some age where a car has a zero or negative predicted price? Justify your answer using a calculation or graph.
(c) What happens in the quadratic model for cars that are very old? Can you think of a plausible explanation for this, or is it a flaw of the quadratic model?
(d) Would the fit improve significantly if you also included a cubic term? Justify your answer.

Model #3: Complete second-order model
(a) Write down the complete second-order model for predicting a used car price based on age and miles. Note: This is before estimating the fit.
(b) Use R to estimate the coefficients in this model. Include the summary and anova output and write down the prediction equation.
(c) Show the details of a nested F-test for just the second-order terms (quadratic and interaction) that involve miles from this model. Include whatever output you need to get the information to complete the test. Be sure to show all the usual details (hypotheses, test statistic, p-value and an informative conclusion in context). In particular, show how the test statistic is computed by putting values into a formula.

Wrap-up: Based on the various models you have considered for predicting the price of a used car, which would you recommend using in practice? Give some justification for your answer.
**Math 213**: Project #3 (Group)  Due: Fri. 10/21/11  
R. Lock  10/17/11  Value: 20 points

**Red Wine Challenge**

*Can we predict the quality of red wine based on its chemical composition?*

That is your task for this assignment. Each team will get two data sets with information about 11 potential predictors and a quality rating (on a 0 – 10 scale) for a sample of red wines. You will use one of the datasets (n = 300) as training data to help build an initial model, reserving the second dataset (n = 200) as a holdout sample to test your model.

**Part 1. Build an initial model**

Use the data in `RedWineTrainX.csv` (where “X” corresponds to your team number) to select a set of predictors to include in your model. This training sample should have information for a sample of 300 red wines. Keep track of the process you use and decisions you make to arrive at an initial set of predictors. Your report should include a summary of this process. You don’t need to show all the output for every model you consider, but you should give a clear description of the path you took and the criteria you used to compare competing models. Also, do not rely on a single method to find a model (e.g., don’t just run backward elimination and take whatever model it gives you).

In addition to the commentary on model selection, please include the following information for your initial choice of a model:

- The `summary()` output for your model
- Comments on which (if any) of the predictors are not significant at a 5% level
- Comments on what the VIF values tell you about the individual predictors in your model.

**Part 2. Residual analysis**

Do a residual analysis for the model you chose in Part 1. Include any plots relevant to checking model conditions—with interpretations. (Note: A residual vs. fits plot will probably show a series of sloping lines—this is due to the fact that the response variable, Quality, only takes on a few integer values—don’t worry about this as a “pattern” to try to avoid.) Also check whether any of the data cases are unusual with respect to studentized residuals, leverage, or influence (Cook’s D). Since there are a lot of data points don’t worry about the “mild” cases for studentized residuals or leverage, but please indicate what specific criteria you are using to identify “unusual” points.

Feel free to adjust your model (either the predictors included or data values that are used to fit it) on the basis of your residual analysis, but don’t worry too much about trying to get all conditions “perfect.” If you do refit something, be sure to include the new `summary()` output.

**Part 3. Cross-validation**

In some situations a model might fit the peculiarities of a specific sample of data well, but not reflect structure that is really present in the population. A good test for how your model might work on “real” red wines can be simulated by seeing how well your fitted model does at predicting wines that were NOT in your original sample. This is why we reserved an additional 200 cases as a holdout sample in `RedWinesTestX.csv`.

Import your test data and compute predicted *Quality* scores for each of the cases using your model after the initial fit and residual analysis in parts 1 and 2. There are two ways to do this in R:

1. If you have the model stored in a variable (e.g., `mymodel`), you can use it on all of the data in the holdout sample with `predict.lm`. 
Instructor’s Guide - 58

> predQ=predict.lm(mymodel, newdata=RedWineTestX)

OR

(2) You can calculate the values directly from the prediction equation (assuming you have detached your training sample and attached the holdout data so those values are used for the variables). For example,

> predQ=1.689-0.001629*AllSulfurO2+0.383924*Alcohol

Also, compute the residuals for the 200 holdout cases with a command like

> residQ=RedWineTestX$Quality - predQ

(a) Compute the mean and standard deviation of these residuals. Are they close to what you expect from the training model?

(b) Are there any holdout cases that are especially poorly predicted by the training model? If so, please identify by the row number(s). As a rough standardization you might divide the residuals by the standard deviation from (a)

(c) Compute the correlation between the predicted and actual values for the holdout sample. This is known as the cross-validation correlation. We don’t expect the training model to do better at predicting values different from those that were used to build it (as reflected in the original $R^2$), but an effective model shouldn’t do a lot worse at predicting the holdout values. Square the cross-validation correlation to get an $R^2$ value and subtract it from the original $R^2$ of the training sample. This is known as the shrinkage. We won’t have specific rules about how little the shrinkage should be, but give an opinion on whether the shrinkage looks OK to you or too large in your situation.

Part 4. Final model

You may make further adjustments to your model at this point, based on the cross-validation information or other information available from the holdout sample as well as your original training sample (but not using data from other teams). If you make any changes to the predictors included in the model or estimated coefficients, please summarize what you did and supply the computer output for your final model.

Also, for your final model please turn in:

(a) the final model summary sheet

(b) your prediction for the mean quality of all red wines with the characteristics shown below; include a 95% interval for this value

FixedAcid=9.5, VolatileAcid=0.50, Citric=0.60, Sugar=3.20, Chlorides=0.10, FreeSulfurO2=16, AllSulfurO2=40, Density=0.994, pH=3.70, Sulfates=0.64, Alcohol=11.8

(c) Send me (glock@stlawu.edu) an email message that contains your final prediction equation written in a form that can be copied and computed in R. For example, Team #11 might send

predict11=1.689-0.001629*AllSulfurO2+0.383924*Alcohol

Finally, on Friday we will have a competition between all of the teams’ models for predicting the quality with a different sample of 400 red wines. The “winner” will be the team with the smallest sum of squared errors (SSE) in their predictions for the contest sample.
Binary Logistic Regression

Your data for this project should consist of a binary response variable having just two categories and several (3 to 6) potential predictors (see some examples below). At least two of the predictors should be good quantitative variables—others can be 0/1 indicators or quantitative. Your goal will be to investigate logistic regression models to study how the chance of being in one of the two groups is related to the other variables. You’ll get more details about what’s required for the analysis later, but the project will be divided into two parts—one using simple logistic regression techniques to explore the relationship between your binary response and a single quantitative predictor; the other to consider models using multiple predictors.

Data ideas: You’ll probably want at least 50 data cases. For each case you’ll need a value for the binary response (coded as 0/1, yes/no or something more informative) and values for each of the predictors. R can work fine with values coded like “F” and “M” as well as 0/1.

Some examples:

1. Y = result of a hockey game (win/loss – ignore tie games)
   X = shots for, shots against, penalty minutes for, penalty minutes against, home? (1/0)

This would involve selecting a sample of box scores, randomly choosing either the winner or loser of each game and recording each of the data values. There are LOTS of possibilities along these lines.

2. Y = Red/Blue states in the 2008 Presidential election
   X = unemployment rate, % urban, per capita income (in $1,000’s), per capita federal aid

3. Y = Domestic/Foreign car model
   X = base price, city mpg, hwy mpg, weight, # cylinders, horsepower

Notes:

- Don’t use a relationship that is too obvious. If you know the number of runs scored by both teams in a baseball game you can “predict” the result with certainty!

- Avoid cases where you have a very lopsided proportion (very few “yes” or “no” responses)—like wins for this year’s men’s soccer team that is currently 16-0 (and only give a sample of n = 16).

- For some situations you can “manufacture” a binary response from a quantitative variable. For example, use a cutoff in per capita GDP to classify countries as “rich” or “poor” and then try to predict that status (but be careful not to use too obvious predictors).

- Stop in if you are having trouble finding a suitable data idea.

Data due date: A copy of an Excel (.csv file) with your data should be sent via e-mail attachment by Monday, November 7th. Include in the e-mail message a description of your data situation and variables. Be sure to document the source of your data (e.g., with a reference or a specific URL if from the web).

Final note: You may (and, in fact, are encouraged to) work on this project with a partner, in which case turn in a common dataset.
Math 213: Project #4  
Data due: Mon. 11/7/11
Write-up due: Mon. 11/14/11
Value: 20 points

**Binary Logistic Regression**

See the Project #4 preliminary handout for a description of the data requirements. You should have a binary response variable and several (3 to 6) potential predictors. Your goal will be to investigate logistic regression models for the binary response. After an appropriate introduction describing your data situation and variables, divide your main analysis into two parts:

**Single:** Choose a *single* quantitative predictor and run a logistic regression.

(a) Include the summary output using R to estimate the coefficients of this model.

(b) Interpret the slope coefficient (in terms of an odds ratio) and interpret the test for the slope. Be sure to do these in the context of your data situation.

(c) Compute a 95% confidence interval for your slope and use it to find a confidence interval for the odds ratio. Does your interval include the value 1? Why does that matter?

(d) Show (by hand) how to use the fitted model for predicting a couple of typical cases, being sure to explain what you are finding in terms of your data situation.

(e) Include a plot of the logistic fit (with commentary).

(f) Show how to compute the G-statistic and use it to test the effectiveness of your model. Be sure to indicate where the p-value for the test comes from.

**Multiple:** Choose your best model *using at least two predictors*. Try to balance getting a good fit with keeping the model simple (but use multiple predictors—even if one or all are not very effective).

(a) Explain briefly the process that led to your choice for a final model. Include summary output from R for your final model.

(b) As above, show (by hand) how to use the fitted model to predict one typical case.

(c) Comment on the effectiveness of each predictor in the model as well as the overall fit. Be sure to indicate what value(s) from the output lead to your conclusions.

**Prediction table:** Use either your single or multiple model for this part—whichever you think is better.

(a) Store the predicted probabilities for all of your data cases in a variable. Hint: Use `fitted()`.

(b) Classify each data point as being a predicted “success” (1) if the predicted \( \hat{\pi}_i \) is greater than 0.5, and a predicted “failure” (0) if \( \hat{\pi}_i \) is less than 0.5. Hint: You can do this fairly easily in R. Think about how we created the indicator variables.

(c) Look at the classifications for each of your data points and create a 2 x 2 table showing counts of how the data are classified (predicted success or predicted failure) versus their actual response values. Hint: The `table( , )` command in R should be useful here.

(d) Comment on the accuracy of the classifications for our data. Did you have many cases that were “misclassified” (i.e., predicted to be “success” when actually a “failure” or vice versa)?

If working with a partner, please turn in a common report.
Math 213: Project #5
R. Lock 11/28/11

Due: Tuesday 12/6/11  Value: 20 points

ANOVA in the Virtual Used Car Lot

This project will continue our earlier investigation of factors affecting used car prices. In Project #1, you each focused on a single car model and single predictor of price (the age of the car). In Project #2, you included a second predictor (miles) and worked with multiple regression. Now we will expand each of your datasets to include three more models of cars. To save you the trouble of collecting a lot more data, you can share each other’s data according to the chart on the back.

To compare car prices among your four car types:

1. Produce a set of side-by-side boxplots to compare the distributions of your four groups of cars. Comment on any obvious differences in the distributions. R hint: You can use `boxplot(Y~X)` much like you use `plot(Y~X)`.

2. Produce summary statistics (sample size, mean, and standard deviation) for each of the groups AND the entire sample of car prices. R hint: Use the `tapply()` function.

3. Based on just what you see in the comparative graph and summary statistics:
   (a) Comment on whether you think there are significant differences in the mean prices among your four car models.
   (b) Comment on any concerns you see about the conditions for the ANOVA for means model.

4. Run the ANOVA for difference in means using R. Include the computer output showing the ANOVA table; state hypotheses and provide a conclusion in the context of your data.

5. Show how to use the summary statistics in #2 to verify “by hand” (i.e., with a calculator) the various sums of squares terms (SSGroups, SSE, SSTotal) in the ANOVA model.

6. Produce a plot (or plots) and/or summary statistics to comment on the appropriateness of the following conditions for your data and model.
   (a) Normality of the residuals
   (b) Equality of the variances

7. Choose either (a) or (b) depending on the result of your ANOVA test.
   (a) If your ANOVA indicates there are significant differences among the car means: Discuss where the significant differences occur using both Fisher LSD and Tukey HSD methods. Interpreting appropriate computer output is fine, but please include the output.
   (b) If your ANOVA indicates there are not significant differences among the car means: Compute a 90% confidence interval for the difference in mean price between your car model and the next model in your group. Do this two ways: Using the Fisher LSD method and Tukey’s HSD adjustment (assuming all possible comparisons). Interpreting appropriate computer output is fine, but please include the output.

8. Recall that we can also handle a categorical predictor with multiple categories (like your four car models) using ordinary multiple regression if we create indicator variables for each category and include all but one of the indicators in the model.
(a) Run a multiple regression to predict Price using the Model variable. Since Model has text values, R will automatically create the required indicator variables and leave one out when fitting the model. Give the summary output and anova from this multiple regression model.

(b) Interpret each of the coefficients in the “dummy” regression in the context of prices (or mean prices) of your car models.

Optional challenge: One possible drawback of the analysis for this project is that different people might have chosen cars with quite different ages (or mileages) when collecting their samples. Thus an apparent “difference” between two car models might be due to one sample having a lot more older cars in it than another. Write down (and run) a regression model that allows you to quickly check for price differences between your original car model and the other three groups after accounting for variability that might be due to age or mileage of the cars. Explain how you use the output from the model to address this question.

The data are stored as .csv files in the rstudioshared>Math213ClassData>Project5 directory. Find your name under Group #1 and read across to find the other cars you should use.

<table>
<thead>
<tr>
<th>Group #1</th>
<th>Group #2</th>
<th>Group #3</th>
<th>Group #4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biles (CamryA)</td>
<td>Duggan (Corolla)</td>
<td>McKenna (CamryC)</td>
<td>Shorb (Accord)</td>
</tr>
<tr>
<td>Montgomery (BMW)</td>
<td>Rotich (Chevy)</td>
<td>Miller (Escalade)</td>
<td>Fingar (PorscheA)</td>
</tr>
<tr>
<td>VanHaaren (MonteCarlo)</td>
<td>Hart (Explorer)</td>
<td>Pidgeon (AudiTT)</td>
<td>Parham (PorscheB)</td>
</tr>
<tr>
<td>Slingluff (Jetta)</td>
<td>Richter (Civic)</td>
<td>Brechner (SantaFe)</td>
<td>DeBernardi (Acura)</td>
</tr>
<tr>
<td>DeBernardi (Acura)</td>
<td>Parham (PorscheB)</td>
<td>Sultzer (Volvo)</td>
<td>Roslevich (OutbackB)</td>
</tr>
<tr>
<td>Cook (Silverado)</td>
<td>McKenna (CamryC)</td>
<td>Biles (CamryA)</td>
<td>Duggan (Corolla)</td>
</tr>
<tr>
<td>Richter (Civic)</td>
<td>Slingluff (Jetta)</td>
<td>Horne (CamryB)</td>
<td>Pidgeon (AudiTT)</td>
</tr>
<tr>
<td>Parham (PorscheB)</td>
<td>Miller (Escalade)</td>
<td>Fingar (PorscheA)</td>
<td>Rotich (Chevy)</td>
</tr>
<tr>
<td>Miller (Escalade)</td>
<td>Parham (PorscheB)</td>
<td>Fingar (PorscheA)</td>
<td>Limoncelli (Mercedes)</td>
</tr>
<tr>
<td>Roslevich (OutbackB)</td>
<td>Ehrensbeck (OutbackA)</td>
<td>DeBernardi (Acura)</td>
<td>Richter (Civic)</td>
</tr>
<tr>
<td>Duggan (Corolla)</td>
<td>Biles (CamryA)</td>
<td>Cook (Silverado)</td>
<td>Buckingham (Durango)</td>
</tr>
<tr>
<td>Farrell (Camaro)</td>
<td>Buckingham (Durango)</td>
<td>Pidgeon (AudiTT)</td>
<td>Richter (Civic)</td>
</tr>
<tr>
<td>Quadri (Audi)</td>
<td>Pidgeon (AudiTT)</td>
<td>Richter (Civic)</td>
<td>Harty (Explorer)</td>
</tr>
<tr>
<td>Harty (Explorer)</td>
<td>VanHaaren (MonteCarlo)</td>
<td>Quadri (Audi)</td>
<td>Farrell (Camaro)</td>
</tr>
<tr>
<td>Limoncelli (Mercedes)</td>
<td>Miller (Escalade)</td>
<td>Cook (Silverado)</td>
<td>Montgomery (BMW)</td>
</tr>
<tr>
<td>Brechner (SantaFe)</td>
<td>Horne (CamryB)</td>
<td>Richter (Civic)</td>
<td>Shorb (Accord)</td>
</tr>
<tr>
<td>Geraghty (CherokeeC)</td>
<td>Schembri (F150)</td>
<td>McKenna (CamryC)</td>
<td>Duggan (Corolla)</td>
</tr>
<tr>
<td>Schembri (F150)</td>
<td>Geraghty (CherokeeC)</td>
<td>Cook (Silverado)</td>
<td>Biles (CamryA)</td>
</tr>
<tr>
<td>Rotich (Chevy)</td>
<td>Montgomery (BMW)</td>
<td>Parham (PorscheB)</td>
<td>Fingar (PorscheA)</td>
</tr>
<tr>
<td>Ehrensbeck (OutbackA)</td>
<td>Roslevich (OutbackB)</td>
<td>Sultzer (Volvo)</td>
<td>Richter (Civic)</td>
</tr>
<tr>
<td>Fingar (PorscheA)</td>
<td>Parham (PorscheB)</td>
<td>Miller (Escalade)</td>
<td>Montgomery (BMW)</td>
</tr>
<tr>
<td>Pidgeon (AudiTT)</td>
<td>Quadri (Audi)</td>
<td>Richter (Civic)</td>
<td>Slingluff (Jetta)</td>
</tr>
<tr>
<td>Sultzer (Volvo)</td>
<td>Shorb (Accord)</td>
<td>DeBernardi (Acura)</td>
<td>Harty (Explorer)</td>
</tr>
<tr>
<td>Buckingham (Durango)</td>
<td>Farrell (Camaro)</td>
<td>Pidgeon (AudiTT)</td>
<td>Richter (Civic)</td>
</tr>
<tr>
<td>McKenna (CamryC)</td>
<td>Cook (Silverado)</td>
<td>Geraghty (CherokeeC)</td>
<td>Viehe (CherokeeB)</td>
</tr>
<tr>
<td>Capinos (CherokeeA)</td>
<td>Viehe (CherokeeB)</td>
<td>Biles (CamryA)</td>
<td>McKenna (CamryC)</td>
</tr>
<tr>
<td>Viehe (CherokeeB)</td>
<td>Capinos (CherokeeA)</td>
<td>Duggan (Corolla)</td>
<td>Cook (Silverado)</td>
</tr>
<tr>
<td>Shorb (Accord)</td>
<td>Sultzer (Volvo)</td>
<td>VanHaaren (MonteCarlo)</td>
<td>Ehrensbeck (OutbackA)</td>
</tr>
</tbody>
</table>

To create a dataframe with all four datasets, import them each into an R session and use the rbind( ) command. If the dataframes are named A, B, C, and D (and have the same variable names) then

```r
> mydata=rbind(A,B,C,D)
```

will combine them into a single dataframe (called mydata).