Intermediate Statistical Investigations
Instructor's Guide

Overview

This text is intended as a second course for students who have completed an algebra-based introductory course. The course focuses on topics related to regression and analysis of variance, but as a unified theme (statistical models and explaining variation), with a minimum of mathematical detail. We make heavy use of free on-line applets that allow students to visualize and explore key ideas. Instructors can pair the text with standard statistical software (e.g., R, JMP, Minitab), but be cautious as some material may appear in a non-traditional order (e.g., you may want to wait until the middle of the course to introduce these tools which often display output we don't discuss in the first few chapters). For example, we discuss the standard error of the residuals early on, but do not show traditional regression tables until Ch. 5. We intend for the Preliminaries to be a quick exposure to multivariable thinking. Many of the ideas and contexts will be returned to later in the course. Chapter 1 begins formalizing the theme of explaining variation (e.g., including a Sources of Variation Diagram), though a statistical model and comparing models. The "separate means" model extends to more than two groups, and then to more than two variables. Ch. 2 begins with randomized block designs and the idea of "adjusting" for other variables in the model, ending with observational studies. Ch. 3 begins with factorial designs, introducing interactions, also ending with observational studies. Ch. 4 Introduces quantitative explanatory variables and multi-category categorical variables, and Ch. 5 expands to multiple quantitative variables. Ch. 6 gives a brief overview of logistic regression, starting with chi-square tests. Ch. 7 then focuses on issues of messy data and variable selection. In the following instructor's guide, we outline some strategies we have used in teaching with these materials. We hope to convey a tremendous amount of flexibility in presentation and amount of time spent on different sections (e.g., guided class activity vs. student exploration vs. out-of-class reading). The time estimates below may be on the conservative side and you may want to require more out of class work as well. On the textbook webpages, you can access example syllabi, data files and applets, errata in the first edition, and RMarkdown files that can be used with many of the explorations.

Preliminaries

Chapter Overview

The purpose of the Preliminaries chapter is to review some basic concepts from the first statistics course and to introduce students to *multivariable thinking*, and the general idea of explaining variation. Review concepts include observational units, variables, association and confounding, as well as basic one and two-variable graphs (histograms, mosaic plots, box plots) and summary statistics. In the introduction to multivariable thinking we ask students to consider how other variables may impact the relationship between the variables of interest, and how explaining variation in a response will reduce prediction error. Students are asked to brainstorm different sources of variation in a response and to consider how confounding variables will impact variable relationships in observational studies.

Example P.A: Graduate School Admissions at Berkeley

Overview

Example P.A considers an observational study on graduate school admission (binary response variable) at Berkeley. Explanatory variables are sex (binary) and program applied to (binary). This example is used to illustrate how the addition of a second explanatory variable (program) changes the original relationship and conclusions about the relationship between admission and sex. This study is a classic example of Simpson's Paradox.

Approximate class time

- One to two 50-75 minute class periods:
    - One 50-75 minute, in-class period for a guided discussion of Example P.A. (Example P.A could also be used as a pre-class reading.)
    - One 50-75 minute, in-class period for Exploration P.A. (Exploration P.A could also be assigned as homework.)

Implementation Tips

After introducing the study, give students an opportunity to identify the observational units, response and explanatory variables, and variable types. Provide a bar chart of the response variable (admission: yes/no) and discuss how this plot shows variability in the categorical response variable. The bar chart of the actual study data can be contrasted with hypothetical bar charts that show no variability in the response (e.g., 100% admitted or 100% not admitted), or the most variability in the response (e.g., 50% admitted and 50% not admitted). Ask students to brainstorm some possible reasons for this variability in admission. If no one suggests sex, add it to the list.

Next, show a mosaic plot of admission vs. sex. Ask students whether there appears to be an association between admission and sex. If they are having trouble answering this question, show examples of mosaic plots where there is non association (e.g., 50% admitted and 50% not admitted for both females and males; 80% admitted and 20% not admitted for both males and females). Also show an example or two where the association is stronger, including an example of a perfect association.

The mosaic plots provide an opportunity to review conditional proportions, so in addition to the plot, provide students with the contingency table and have them compute the proportions that are shown in the mosaic plot.

Next, show a mosaic plot of admission vs. program (binary) so that students can see that program is also associated with admission. Then provide the mosaic plot showing all three variables. Ask students to describe the association between admission in sex within each program, and how they differ from the overall association. This is an example of Simpson's Paradox.

The mosaic plot is helpful in explaining why Simpson's Paradox happens, because it shows the relative sample sizes within each program (higher proportion applied to program A) and the relative sample sizes of males and females within each program (much larger proportion of program A applicants are male, whereas applicants to program F are about 50/50 male/female). It is the imbalance of the male/female sample sizes within each program which causes Simpson's Paradox. Show students that when the admission rates of males for program A and program F get averaged together, the overall admission rate for males is closer to the program A admission rate (which is easier to get into) due to the larger sample size in that group. A similar thing happens for females, except the overall female admission rate is closer to their admission rate in program F (which is harder to get into). This creates the situation where overall females have a lower acceptance rate. (The idea of a weighted average will also recur later in the course.)

A key take-away from this example which will be very useful in later chapters is the idea of a confounding variable. Stress for students that there are many sources of variation in the response variable. (These sources can be placed into a Sources of Variation diagram.) When one (or more) of these sources is also associated with the explanatory variable of interest (sex, in this study) this extraneous source of variability becomes confounded with the explanatory variable. Thus, to be a confounding variable, a variable must be associated with both the response and explanatory variables. Give students an opportunity to explain how Simpson's paradox arises in their own words. For example, one reason for the higher overall acceptance rate for males is that males tend to apply to program A, and program A has a higher acceptance rate, while females tend to apply to program F and program F has a lower acceptance rate. Finally, because of the confounding relationship between sex and program, comparing the acceptance rates of males and females within program is one simple descriptive analysis method for taking the confounding variable into account.

Finally, this example introduces the Six Steps of a Statistical Investigation. Each example and most explorations will use these steps. While Step 4 is an important step, explain to students that statistical significance is not of interest in the Preliminaries (or Sections 1.1 and 1.2).

Technology

- No specific technology is used in this section, however, statistical software could be used to have students generate graphs and summary statistics.

Example P.B: Predicting Birth Weights

Overview

Example P.B considers another observational study, but this time with a quantitative response variable (birthweight) and both a quantitative (mother's BMI) and a categorical (whether or not the baby was carried to full term) explanatory variable. Students are introduced to the idea of a statistical model (e.g., single mean, separate means, regression) and residuals as a measure of left-over variation.

Approximate class time

- One to two 50-75 minute class periods if everything is done in class.
    - One 50-75 minute class period for a guided discussion of Example P.B. Example P.B can be assigned as a pre-class reading.
    - One 50-75 minute class period for Exploration P.B. This exploration could also be assigned as homework.

Implementation Tips

Begin Example P.B with a histogram and summary statistics of the response variable. Ask them what would be the best prediction of birthweight based on these results. Place the mean into a statistical model, then use the standard deviation of birthweights as a measure of the “typical” prediction error. Next, have brainstorm potential sources of variation in the response variable and fill in a Sources of Variation diagram. Make sure the *length of term* is a variable in the list if students don’t come up with it on their own.

Show side-by-side boxplots or histograms of birthweights for the two categories of term length and ask whether length of term is associated with birthweight. Then, ask students whether our statistical model would be improved (i.e., could we make better predictions) using information about length of term. Show students how residuals can be computed, and then how those residuals are used (generally) to compute the standard error of residuals, the amount of left-over variation for this model. The SE of residuals for the separate means model is smaller than the overall standard deviation, which indicates the separate means model is useful in explaining variation in birthweight (i.e., birthweights are closer to their group mean, than the overall mean).

Finally, show students, graphically, what it looks like when we add another variable to the model (mother’s BMI). Using a scatterplot, show how birthweight is related (weakly) to mother’s BMI, then show students how we can take both length of term and mother’s BMI into account simultaneously by color-coding and fitting separate regression lines for the two categories of term length. The SE of residuals for this model is less than the separate means model, illustrating how adding an additional variable can further increase the amount of explained variation.

Technology

- Exploration P.B uses the Multiple Variables applet.

Chapter 1: Sources of Variation

Chapter Overview

In Chapter 1 we focus on studies that have a quantitative response variable and a single categorical explanatory variable. In Sections 1.1, 1.2, and 1.3 the explanatory variable is binary. In Sections 1.4 and 1.5, the explanatory variable has more than two levels. In Section 1.6 we discuss statistical power. Continuing with the idea of explaining variation we begin to transform Sources of Variation diagrams into statistical models. Here we also introduce sums of squares as a way to capture the amount of total, explained, and unexplained variation in the response variable laying the foundation for the ANOVA *F*-test in later sections of Chapter 1 (focusing on those distinctions more than the calculation details). The chapter begins with a discussion of designed experiments, reviewing the importance and impacts of random assignment, as well as other important considerations in study design such as inclusion criteria and blinding, and concludes with an example of an observational study for contrast. Simulation-based inference is used as a conceptual framework for understanding statistical inference, followed by the theory-based pooled two-sample *t*-test and ANOVA *F*-test. Because this chapter is both setting the stage for the over-arching themes of the course (e.g., explained variation) but also, to some extent, reviewing content from previous courses, you may find that your students will understand this content at varying speeds based on the type of prior statistics course and the time since your students took the course.

Section 1.1: Sources of Variation in an Experiment

Overview

The goal of this section is to review some concepts students have seen in their previous statistics courses, as well as to continue the theme of identifying and thinking about sources of variation in the context of a real-world experiment on the effects of scent on consumer behavior. A large portion of this section is spent on Step 2 of the six-step investigative process, designing a study and collecting data. Here we review important study design principles such as inclusion criteria, blinding, and random assignment, as well as introduce terminology specific to designed experiments (e.g., experimental units and treatments). In Step 3 of the six-step investigative process, exploring data, we continue the theme of explaining variation by presenting graphs and summary statistics for both the overall consumer ratings and the ratings for each scent group, introducing both the single mean and separate means statistical models. Finally, we connect the unexplained sources of variability (as captured by the standard error of the residuals in the separate means model), back to the purpose of random assignment.

Approximate class time

- One to two 50-75 minute class periods:
    - One 50-75 minute, in-class period for a guided discussion of Example 1.1: Scent and Consumer Behavior. (Example 1.1 could also be used as a pre-class reading.)
    - One 50-75 minute, in-class period for Exploration 1.1: Memory Study, or assign as homework after collecting the data in class. We do see advantages to having students participate together in the data collection process.

Implementation tips

The six-step process provides a roadmap not only for the design and analysis of a statistical study, but also for the discussion of the study in the classroom. Getting students accustomed to this framework early, in the context of a simple statistical study, will make the discussion of more complicated studies in later chapters much easier. The first two steps provide opportunities for students to practice identifying the response and explanatory variables and whether the study is observational or an experiment, and to facilitate discussion of inclusion criteria, blinding, study protocol, as well as experimental design language of experimental units and treatments.

To facilitate a discussion of the sources of variation, first present a graph of the overall consumer ratings with summary statistics. Here, you can review shape, center, spread and outliers in a quantitative distribution. When discussing the center, ask students what they would predict the consumer rating would be for a new student and introduce the single mean model. As part of the model we need an estimate of precision, so turn the discussion to variation - both how it’s measured and perhaps more importantly, why the ratings vary. Have students brainstorm sources or reasons for this variability in consumer ratings, reminding them that whether or not the students were exposed to the pleasing scent is what the researchers hypothesize will cause variation in the response and is the purpose for running the study.

Next, present the graphs of the consumer ratings for each of the scent groups. Give students the opportunity to discuss whether the variable scent exposure explains some of the variability in the consumer ratings. This may be a difficult idea for students to grasp, so it helps to show an example of graphs in which the means of the two groups are the same and/or the within group variation is not smaller than the overall variation. It can also be helpful to ask students to update their prediction of consumer rating taking into account the scent exposure. Here, we want them to see that because the means of the two scent groups are different, we can make a more informed (better) prediction by using the mean of the scent group. As a follow-up discussion, you could present an example in which the scent group means are even further apart and ask whether the scent groups explain more or less variability than in the original data. At this point, introduce the separate means model, reminding students that this model too, needs a measure of precision or uncertainty.

After understanding that the explanatory variable explains some of the variation in the response, it's important students recognize that this variable does not explain all of the variation in the response. Ask the students whether scent exposure explain all of the variation in the consumer ratings. Most students will respond no, but may not clearly understand why. Ask them what would have to be true of the two graphs of scent exposure if scent exposure did explain all of the variation in the consumer ratings. Finally, provide the standard error of residuals, a sort of "average" of the two group standard deviations, as the measure of uncertainty of the separate means model.

To summarize, give students the opportunity to fill in the Sources of Variation diagram from the brainstormed list of sources or reasons for variability in the consumer ratings. Be sure to include at least one inclusion criteria and one potential source of variability that was held constant by design. In the unexplained sources column, add an "unknown" source. Then summarize how the Sources of Variation diagram relates to the separate means model.

Students may be in a hurry to get to statistical significance, but that's not the point of this section. The goal here is to focus on summary statistics, statistical models, and thinking about variation in the response. We'll get to statistical significance in Section 1.3.

Technology

- Exploration 1.1 requires the Multiple Variables applet.

Section 1.2: Quantifying Sources of Variation

Overview

The goal of this section is to quantify sources of variation. Continuing with the Scents and Consumer Behavior study in the example, we discuss how to compute the sum of squares total (same as sum of squares error for the single mean model), and the sum of squares model and sum of squares error for the separate means model. As part of the discussion of the sum of squares model, students are introduced to the idea of an effect of a treatment group. The formula for the standard error of residuals is also given. After defining the sums of squares, R^2 is introduced. The section ends with a discussion of practical significance. An Appendix after the Exploration provides some of the calculation details for the ANOVA table for those who are curious.

Approximate class time

- One to two 50-75 minute class periods:
    - One, 50-75 minute, in-class period for a guided discussion of Example 1.2: Scent and Consumer Behavior. (Example 1.2 could also be used as a pre-class reading.)
    - One 50-75 minute, in-class period for Exploration 1.2: Starry Nights, or assign as homework.

Implementation Tips

A sum of squares is not an intuitive idea for students and most students will not have seen it in an earlier course, but they have had quite a bit of exposure to standard deviations by this point in their study of statistics. Understanding a standard deviation as a typical deviation *from something* is helpful in understanding the sums of squares. Each of the sums of squares is based on a deviation of some quantity from the data from some other quantity from the data.

To introduce the sum of squares total, start with a dotplot or histogram of all 48 of the observed consumer ratings. Then remind students that the standard deviation of the distribution is the typical deviation of the consumer ratings from the overall mean. Providing the formula or asking the students to build the formula for the overall standard deviation based on the definition, point out the numerator as the SSTotal, and the denominator as the degrees of freedom total. The SSTotal is based on the *deviations of the data values from the overall mean*, or the residuals from the single mean model.

Then return to the Sources of Variation diagram from Example 1.1 and the separate means model. The Sources of Variation diagram gives the students a visual representation of the separate means model. Discuss how the SE of residuals captures the amount of unexplained variation, and review all of the sources that are contributing to the unexplained variation. Give students the opportunity to find several residuals from both the group exposed to the scent and the group not exposed to the scent, noting how the residuals are *deviations of the data values from the treatment means*. Summing the squares of the residual from the separate means model yields the SSError. The square root of the SSError divided by the degrees of freedom error gives the SE of residuals.

Finally, again using the Sources of Variation diagram from Example 1.1, the amount of explained variation is obtained by determining how far each of the treatment means is from the overall mean. These differences are the treatment effects, and the treatment effects measure *how far the treatment means deviate from the overall mean*. The SSModel for the separate means model is then found by weighting the effects squared by the sample sizes in the treatment group (when the sample sizes are the same).

Once the sums of squares are obtained, it can be shown that the SSModel and SSError partition the SSTotal using a pie chart. Then, the natural next step is to determine the fraction or proportion of the total variation that is explained by the explanatory variable in the separate means model, R^2. (Note: We try to always present R^2 as a decimal, and convert to a percentage for interpretation.)

This section closes with a brief discussion of practical significance in an attempt to give students a way to judge the size of R^2. In addition, the notion of effect size is introduced. This can be confusing to students because they have just been introduced to the treatment effects, and the effect size computation relies on the difference in the treatment means, rather than the effects. Giving students an opportunity to judge whether a difference is meaningful, or whether the fraction of variation explained is meaningful in the context of a study prior to assessing statistical significance is an important exercise.

Technology

- Exploration 1.2 uses the Multiple Variables applet.

Section 1.3: Is the Variation Explained Statistically Significant?

Overview

The goal of Example 1.3 is to re-introduce the idea of statistical significance. Having discussed sources of variation, quantified the explained vs. unexplained variation from the separate means model, and thought about how meaningful the differences in means is with respect to the unexplained variation (i.e., practical significance), we consider possible explanations for the study results. Could the results have arisen simply by random chance, or are the results indicative of a real difference in the population means? To answer this question we use simulation-based inference and then introduce the theory-based two-sample t-test.

Approximate class time

- One to three 50-75 minute class periods:
  - One to two 50-75 minute, in-class periods for a guided discussion of simulation-based inference and the pooled two-sample *t*-test and confidence interval in Example 1.3: Scent and Consumer Behavior. (Example 1.3 could also be used as a pre-class reading.)
- One 50-75 minute, in-class period for Exploration 1.3: Starry Nights, or assign as homework.

Implementation Tips

Example 1.1 was discussed using the framework of the six-step investigative process. Students were told that Step 4: Draw inferences beyond the data would be covered in Example 1.3.

The 3S Strategy is used to assess the statistical significance of the observed study results (e.g., difference in means, difference in medians, R^2 statistic, *t*-statistic, etc.). Using simulation and the Comparing Groups applet, students can see the re-randomization process of the observed consumer ratings to the two scent groups and the resulting null distribution of the statistic. The null distribution is then used to obtain an approximate p-value for the observed statistic. Because some students may not have been exposed to simulation-based inference in their first statistics course, the exploration in this section walks them through a tactile simulation using cards to reinforce the re-randomization process.

The *t*-statistic is introduced as a standardized statistic which takes into account sample size and the standard error of residuals. Thus, this statistic is more informative than the difference in means or R^2, because it is sensitive to the size of the sample and the amount of unexplained variation. In addition, because of its standardized nature, the *t*-statistic, by itself, can be used to judge statistical significance, as values greater than |2| tend to large enough to produce a p-value less than 0.05. The section closes with a discussion of the two-sample pooled *t*-test and *t*-interval. The *t*-distribution is shown to be a good approximation of the simulated null distribution when the validity conditions are met. (We start with the pooled *t*-test to parallel the ANOVA approach to come, but there is flexibility in how much you want to focus on this at this time. HW Exercise # asks them to compare the results.)

Technology

- Exploration 1.3 uses the Comparing Groups applet (similar capabilities as the Multiple Variables applet but assumes one explanatory variable and allows the simulation of the re-random assignment process)

Section 1.4: Comparing Several Groups

Overview

Section 1.4 extends the ideas of explaining variation and statistical models to a categorical explanatory variable with more than 2 levels, returning to an observational study on fish consumption and the level of omega-3 fatty acid in the blood. The example of this section follows the same progression as that of Example 1.3, beginning with simulation-based inference and the 3S Strategy, and proceeding to the theory-based *F*-test.

Approximate class time

- One to three 50-75 minute class periods:
    - One to two 50-75 minute, in-class periods for a guided discussion of simulation-based inference, the theory-based F-test and impact of sample size and MSError in Example 1.4: Fish Consumption and Omega-3. (Example 1.4 could also be used as a pre-class reading.)
    - One 50-75 minute, in-class period for Exploration 1.4: Golden Squirrels or assigned as homework.

Implementation Tips

Example 1.4 is about fish consumption and the level of omega-3 fatty acid in the blood. After the three previous examples which have focused on experiments, we return to an observational study as discussed in the Preliminaries. The five fish consumption classifications for the study participants were self-reported from a single survey question. Omega-3 was measured using a blood test. The example is presented again using the six-step process for statistical investigation, beginning with the research question (e.g., "Is there an association between fish consumption and omega-3?") and the study design, response, and explanatory variables. As part of Step 3: Exploring the data, students again see a histogram or dotplot of the overall data set, as well as a sources of variation diagram for the omega-3 fatty acid percentages. The statistical model, which now has five groups, the sums of squares, and R^2 are all natural extensions from the binary cases of the previous sections.

Students are pretty quick to figure out that the two-sample *t*-test can no longer be used, but we can again use R^2 to determine statistical significance using simulation-based inference. To make the transition into the *F*-statistic, we focus on the importance of having a standardized statistic, one that is sensitive to sample size. The *F*-statistic is derived from R^2 so that students may see how the sample size is incorporated into this standardized statistic. Exploration 1.4 gives students the opportunity to explore the impact of changing only the sample size on both R^2 and the *F*-statistic using the Comparing Groups applet. In this exercise students see that the value of R^2 does not change, but the shuffle-to-shuffle variation in the null distribution of R^2 decreases with increasing sample size. Whereas the value of the F-statistic gets much larger as sample size increases, as well as a decrease in the shuffle-to-shuffle variation of the null distribution. Like the *t*-statistic, in many cases the *F*-statistic can then be used on its own to judge strength of evidence (rough guideline: *F*-stat > 4) or in conjunction with the p-value. As with Example 1.3, the section closes with the theory-based *F*-test, including checking the validity conditions using the fish consumption group standard deviations and the histogram of the residuals.

In Step 5: Formulating conclusions and Step 6: Look back and ahead, students are reminded that this was an observational study so no causal conclusions may be made, as there could be confounding variables in the relationship we've observed between fish consumption and omega-3. In addition, because of the inclusion criteria and the fact that people were not randomly selected into the study, the conclusion about the association between fish consumption and omega-3 applies only to people similar to those who participated in the study. (You can decide how often you want students to use this "out" when defining a population to generalize to.)

Technology

- Exploration 1.4 uses the Comparing Groups applet.

Section 1.5: Confidence and Prediction Intervals

Overview

Section 1.5 continues with the fish consumption and omega-3 examples, following up on the significant *F*-test from Section 1.4 with the post-hoc analysis of the pairwise comparisons. The pairwise comparisons are done using confidence intervals for the difference in the means and are summarized with a Letters Plot. In addition, confidence intervals for a single mean are covered, as well as prediction intervals. This is a lot of material for one section, so you may want to adjust depending on how familiar your students are with the topics.

Approximate class time

- One to two 50-75 minute class periods:
  - One 50-75 minute, in-class period for a guided discussion of Example 1.5: Fish Consumption and Omega-3. (Example 1.5 could also be used as a pre-class reading.)
  - One, 50-75 minute, in-class period for Exploration 1.5 or assigned as homework. (Side-note: the post-hoc comparisons in Exploration 1.5 could be combined with Exploration 1.4 so that students see the follow-up pairwise comparisons immediately after finding the significant *F*-statistic in Exploration 1.4.)

Implementation Tips

With the statistically significant *F*-test in the fish consumption and omega-3 study, a natural next question to pose is which fish consumption groups have different mean omega-3? Confidence intervals for the difference in means are used to carry out the pairwise comparisons. The experiment-wise Type I error rate is *controlled* by only doing the follow-up comparisons in the case where the ANOVA *F*-test rejects. A Letters Plot is used to summarize the results of what can be potentially many pairwise comparisons when the number of treatment groups is large. After an introduction, students are reasonably good at interpreting a Letters Plot when they see one. However, going the other direction, creating a Letters Plot when provided with the confidence intervals can be difficult, particularly collapsing the letters, so as to use as few letters as possible. This task and learning how to succinctly summarize the pairwise comparisons requires plenty of practice.

The confidence intervals and prediction intervals of this section all use the standard error of the residuals as is common practice under the umbrella of an analysis of variance. Some time will need to be spent pointing out the differences/similarities between confidence intervals and prediction intervals.

Technology

- Exploration 1.5 uses the Comparing Groups applet.
- This might also be a good exploration to begin using statistical software such as R, JMP or Minitab.

Section 1.6: More Study Design Considerations

Overview

Section 1.6 turns to study design issues such as choosing sample size and how the sample sizes, as well as the within group variation and effect size affects the power of the study. This section presents formal definitions of Type II error and statistical power. Using the fish consumption and omega-3 study as a motivational example, we use simulation to explore how the distribution of the *F*-statistic changes as we change the sample sizes of the groups, or the within group variation, or the differences among the means. The example closes with a power analysis when designing a new study. This section is less critical for subsequent chapters in the book, but still provides some important information.

Approximate class time

- One to two 50-75 minute class periods:

- One, 50-75 minute, in-class period for a guided discussion of Example 1.6: Fish Consumption and Omega-3 (revisited). (Example 1.6 could also be assigned as a pre-class reading.)
- One, 50-75 minute, in-class period for Exploration 1.6 or assigned as homework

Implementation Tips and Tricks

It's helpful to begin the discussion by reminding students that the sample size for a study must be chosen and that the sample size is often limited by the budget of the study. Then, if we can only afford a certain number of individuals in the sample, is it "worth it" to carry out the study?

Students should be reasonably comfortable with the notion of the null distribution and what this distribution looks like for the *F*-statistic. They may not have seen the idea of a rejection region before, so it's worth the time to show how changing the significance level of the test, Type I error rate, affects the *F*-critical value. Under a particular set of conditions (given sample size per group, given group means, given standard error of residuals) the null distribution of the *F*-statistic can be simulated and the *F*-critical value determined for a particular significance level.

To determine the power of the test, a simulation can be carried out to determine the types of *F*-statistics we expect when the null hypothesis is false. Again for the particular set of conditions that were used to determine the *F*-critical, the alternative distribution of *F*-statistics may be simulated by taking random samples of the required size from a normal distribution with the given mean and standard error of residuals from each treatment group and determining the *F*-statistic. Students should see that this distribution of simulated *F*-statistics is shifted more to the right, and may not even *start* at 0. Using the *F*-critical value they identified from the null distribution, the power of the test may be determined from the simulated alternative distribution as the proportion of F-statistics that exceed the F-critical. In this way, we can investigate how the power and Type II error rate (proportion of F-statistics below the F-critical on the alternative distribution) change as we change the sample size or the differences between the means or the within group variation.

Students often misunderstand power as the probability that the F-statistic observed in the study is statistically significant, rather than the probability that the study will obtain an F-statistic that rejects H_o under a particular alternative hypothesis. To clarify, it may be helpful to interpret power as a long-run relative frequency. For example, if the power of a particular F-test is 0.80, we can interpret this as "over many, many studies where the sample size is n, and the means are x, y, z and the standard error of residuals is s, we can expect to collect data that will give us a large enough F-statistic to reject Ho in about 80% of the studies."

Technology

- Exploration 1.6 uses the Comparing Two Populations applet.
- It might also be helpful to show students a power analysis using a statistical software package such as R, JMP or Minitab.

Exercise Note

Some questions ask about "confidence interval for the effect." Depending on your software, you may wish to ask instead about a confidence interval for the group difference.

Chapter 2: Controlling Additional Sources of Variation

Chapter Overview

In Chapter 2 we dig into how to include a third variable, a second explanatory variable, into the design and analysis of an experiment, and into the analysis of an observational study. Section 2.1 covers matched pairs studies, adding another source of explained variation to the Sources of Variation diagram and introducing the idea of adjusting the observed response values for the effect of person. Section 2.2 extends the matched pairs study design to randomized complete block designs, using the ideas of adjusting the simulation to match the study design and adjusting the data for the block effects in the analysis. Section 2.3 continues with the idea of adjusting for a second explanatory variable in the context of observational studies.

Section 2.1: Paired Data

Overview

The goal of this section is to introduce students to *designing out* a potential source of variation in an experiment, and to illustrate how this can then reduce the unexplained variation present in the response. Although many students may have seen paired designs and the analysis using a one-sample *t*-test in their first statistics course, the analysis presented here extends the over-arching theme of explaining variation, leveraging the concepts of sums of squares and effects covered in Chapter 1.

Approximate class time

- One-to-two 50-75 minute class periods:
  - One 50-75 minute, in-class period for a guided discussion of paired designs. (Example 2.1 could also be used as a pre-class reading.)
  - One 50-75 minute, in-class period for Exploration 2.1: Chip Melting. This is another exploration in which students are given ownership of the data, by participating as the experimental units. At a minimum, allow 20-minutes in-class for data collection, then the remainder of the exploration could be assigned as homework.

Implementation Tips

In Example 2.1 we introduce the example study (Texts vs. Visual Distractions) without initially specifying that the design is paired. The purpose of this is to provide an opportunity to review the important concepts from Chapter 1 and to motivate the paired design through the lens of explaining variation. After giving the study background and without mentioning the design is paired, give students another opportunity to fill in a potential Sources of Variation diagram similar to those from Chapter 1, and to brainstorm sources of unexplained variation. You can also review the SSTotal, the SSModel from the one-variable analysis, and the SSError and point out again how these sums of squares map to the columns of the Sources of Variation diagram. In this example, the R^2 for the type of distraction is 0.077 so about 92.3% of the variation in braking reaction time is unexplained after accounting for the type of distraction. Using the Sources of Variation diagram, students can see that many of the unexplained sources of variability arise from a common source, the person. This can motivate the question of whether we can design a study which would account for much of this unexplained variation, which would enable us to move this person-to-person variation into the explained sources column. Walking through this process of moving the variation from the unexplained to the explained column can give students a nice visual reminder of how decreasing the number of sources of unexplained variation is likely to reduce the SSError. (You can also start previewing taking another "chunk" out of the SSE part of the pie chart.)

The Sources of Variation diagram also provides a nice path into the two-variable model, because we now have two explained sources of variability. To write out this model though, we need the person effects. Here is where we generalize the concept of an effect, showing how the effect of each person is computed in manner similar to computing the treatment effects. Once we have the person effects, the SSperson is computed, again in a manner similar to the SStreatment. At this point, return to the ANOVA table, to illustrate mathematically how the SSperson and degrees of freedom for person are removed from the SSError and degrees of freedom for error in the one-variable ANOVA. (You may want to emphasize for students that the way the study was designed because these calculations are completely "separable" and don't impact each other.) Finally, the standard error of residuals for the two-variable model can be computed and compared to the standard error of residuals from the one-variable model.

A couple of ideas that students may have trouble with initially include computing the degrees of freedom for person and the idea that the SSTotal stays constant. Using the pie chart when partitioning the variation can reinforce the idea of the fixed SSTotal. Also, it's important for students to understand that the treatment means, treatment effects, and SStreatment are not changed when we account for the person-to-person variation, only the SSError and SE of residuals change (and in fact, decrease, or at least can't increase).

Notes on Exploration 2.1

In Exploration 2.1: Melting Chips, students have another opportunity to participate in the data collection process and take ownership of the data. To carry out the data collection you'll need two types of baking chips (e.g., butterscotch and milk chocolate); playing cards, coins, or random number generator (computer, phone) for random assignment of order; and a timer. It's also helpful to set up a Google Form so that students can enter their melting times directly into a spreadsheet for you (you may need to encourage them to only enter numerical values). The current exploration requires both the stacked and unstacked versions of the data. For the stacked version, it's also helpful to have a column that 'identifies' the student. One way to do this and still maintain anonymity, is to have students enter their initials followed by any 4 digits in the Google Form.

Notes on Exercises 2.1.11 and 2.1.12

SSperson will need to be calculated manually.

Technology

- Example 2.1 has screen shots from the Matched Pairs applet and some of the homework exercises require this applet.
- Exploration 2.1 uses the Multiple Variables or Comparing Groups applet and a spreadsheet package like Excel.
- Statistical software could also be used.

Section 2.2: Randomized Complete Block Designs

Overview

Section 2.2 extends the matched pairs study to a randomized complete block design with three treatments. In Example 2.2, the design is repeated measures, blocking on person. In Exploration 2.2, the study blocks on a unit which is different from the experimental unit. To determine statistical significance of the factor we use two approaches: (1) simulation using within block re-randomization, and (2) adjusting the *F*-statistic by removing the variation due to blocks. Approach (1) reinforces the idea that the simulation should match the study design, and Approach (2) leads to the theory-based *F*-test.

Approximate class time

- Two to three 50-75 minute class periods if everything is done in class.
    - One-to-two 50-75 minute class periods: The first class period can be used for a guided discussion of the two approaches to analyzing a randomized complete block design. The second can be used to summarize and stress the important take-aways (see below). Example 2.2 can be assigned as a pre-class reading.
    - One 50-75 minute class period for Exploration 2.2. This exploration could also be assigned as homework.

Implementation Tips

Example 2.2 is a natural extension of the matched pairs study in Example 1.1 to three treatments. After giving a little background on the goals of the study, inclusion criteria and the treatments, have students fill in a blank Sources of Variation diagram focusing on the sources of unexplained variation. When reviewing the Sources of Variation diagram make sure you have at least 3-4 sources that come from the person, as well as measurement error and/or variation in the doses applied, order of application, etc. At this point, you could show the students the data, including the separate means model, R^2 for the stimulants, and the insignificant *F*-statistic for the stimulant. It's helpful also to walk them through the simulation focusing on the completely randomized nature of the re-randomization. The stimulants explain about 13% of the variation in finger tapping rates, but the *F*-statistic is small, 0.68, so a good question to ask is why do we not have a statistically significant *F*-statistic. Without too much prompting students will recognize that there is a lot of unexplained variation. (The pie chart of the sums of squares will help with this, a well.)

To address this unexplained variation, ask students how they would design the study in order to remove or account for as much unexplained variation as possible. Point out that many of the sources of unexplained variation come from the person and, following from the studies of Section 2.1, students usually arrive at the idea of blocking on a person. At this point, reveal that this is how the study was carried out, being sure to stress the importance of randomizing the order in which each participant undergoes each treatment. Finally, present an updated Sources of Variation diagram with the person-to-person variation removed from error and placed into the explained column.

The rest of this example then focuses on *how* to modify the analysis to reflect the change in the study design, i.e., to account for removing the person-to-person variation, using two approaches: (1) adjusting the simulation and (2) adjusting the *F*-statistic. The first approach is to adjust the simulation process. Using the *F*-statistic of 0.68, ask students how, in light of the actual study design, we should carry out a simulation in order to assess the statistical significance of 0.68. Pose this question using hypothetical index cards. Ask how many cards we would need, what should be written on the cards, and how we should shuffle and deal out the cards in a way that mimics the actual study design. The One Blocking Variable applet will demonstrate the within blocks re-randomization. Show the resulting null distribution and give students the opportunity to determine the approximate p-value for 0.68 from this distribution. This p-value can then be used to draw a conclusion about the stimulants, *after reflecting the restricted randomization used in the study design.*

An important take-away from the within blocks re-randomization is the reduction in the shuffle-to-shuffle variation in the null distribution of *F*-statistics. It can be helpful to show the two null distributions (completely randomized shuffling vs. within blocks or restricted re-randomization shuffling) side-by-side. Then discuss how the restricted/within blocks shuffling keeps the largest finger tapping rates, which come from Person D, from ending up in one treatment group and the smallest finger tapping rates, which tend to come from Person A, from ending up in another treatment group. Re-randomizing within a person/block, keeps the differences among the treatment means smaller, thereby reducing the shuffle-to-shuffle variation in the *F*-statistics. (This discussion requires that the students have a pretty good understanding of how the size of the *F*-statistic depends on how far apart the treatment means are. Weaker students may struggle with this.) An important idea to stress is that the "*F*>4" guideline is no longer applicable, because the null distribution has so much less variability than in completely randomized shuffling.

The second way we can remove the person-to-person variation from the error is to adjust the denominator of the *F*-statistic by taking the SSperson out of the SSError. This discussion is fairly straightforward as an extension of Example 2.1. Because the data set is so small, students can find the averages of each person, the person effects, and the SSperson without taking too much class time. Then walk through the removal of the SSperson from the SSError, and also the removal of the DF for person from the DF for error. Determine the adjusted *F*-statistic and then provide the theoretical *F*-distribution. The adjusted *F*-statistic, 7.88, is statistically significant, agreeing with the result of the modified simulation.

Another important take-away from the adjustment process is how the SSError can be obtained directly by removing the block effects from the observed finger tapping rates. The One Blocking Variable applet can be used to demonstrate how we subtract out the positive person effects and the negative person effects from each of the observed finger tapping rates, causing the values to shift towards their treatment means. The result is that the person-adjusted finger tapping rates have much less within treatment variation, but the treatment means have stayed the same. Stress the idea that the means of the treatments have stayed the same, because in Section 2.3 we return to observational studies where the means of the explanatory variable will change when we go through this adjustment process. A mosaic plot of the treatments vs. the study participants can be used to illustrate how these two variables are unrelated. A discussion about how the balanced sample sizes (each person undergoes each treatment only one time) and not the random assignment, leads to this lack of a relationship, is also important going into Section 2.3.

Notes on Exploration 2.2

The factor in Exploration 2.2 is storage method with three levels (air, modified air, and control). The experimental unit, the unit to which the treatment is applied, is a clamshell of strawberries. The goal of the study is to determine whether and how the storage method affects how fast the strawberries ripen. Students should recognize that the storage methods cannot all be applied to the same container of strawberries, so the experiment cannot block on the experimental unit, hence blocking on the extraneous source of variability, strawberry variety. The blocking unit and experimental unit are now different.

The One Blocking Variable applet can be used to carry out the pairwise comparisons of the means after adjusting for the blocks if you want to complete the analysis of the data after obtaining the statistically significant *F*-statistic.

In summarizing the important ideas from this exploration, point out that the researchers could have done this study on only one variety of strawberries. Using this more restricted inclusion criteria is likely to have had the same effect on the SSError as blocking on strawberry variety. The benefit of blocking on variety is the expanded scope of inference. Conclusions about the storage treatments apply to these five varieties, rather than just one variety, and in fact, the original study upon which this exploration is based, used many more than just 5 varieties. Also, show the mosaic plot of storage method vs. variety to help stress for students that the storage treatment means do not change when we adjust for variety because storage treatment and variety are not associated. This lack of association is guaranteed by the design in making the sample sizes balanced. This is not a result of random assignment (which guards against confounding with unmeasured variables).

Technology

- Example 2.2 and Exploration 2.2 both use the Comparing Groups and One Blocking Variable applets.
- Appendix: Calculation Details uses the Multiple Variables applet.
- If statistical software is used, we recommend using it in addition to the One Blocking Variable applet. The applet will allow students to see how the firmness values adjust when the variety effects are removed, i.e., shifting closer together around the treatment means, similar to a handicap in golf.

Section 2.3: Observational Studies

Overview

Section 2.3 covers two-variable models, adjusting for a confounding variable in observational studies. The big difference between this section and the previous two sections is that, when the explanatory variables are confounded, the adjustment process now has an effect on the mean response of the primary explanatory variable, rather than just a reduction in unexplained variability in the response. So, rather than designing out an extraneous source of variability, we adjust for a confounding variable, which leads to better, although not necessarily more significant, conclusions about the original explanatory variable. Please also note the Appendix at the end of the section.

Approximate class time

- One-to-two 50-75 minute class periods
    - One 50-75 minute class period for a guided discussion of adjusting for a confounding variable in an observational study. Example 2.3 could be assigned as a pre-class reading.
    - One 50-75 minute class period for Exploration 2.3. Exploration 2.3 could also be assigned as homework.

Implementation Tips

Example 2.3 returns to the salary discrimination study from the Preliminaries. In this example we use a subset of the data, n = 16,654 men who are in one of two possible education categories: below college and beyond college. The primary relationship of interest is still between wages and race (black, non-black).

As in the previous two sections, begin by giving students the opportunity to fill in a Sources of Variation diagram for the study, brainstorming at least 5-6 unexplained sources of variation in wages (only Race should be in the explained sources column). Provide the one-variable analysis of Race, including boxplots, summary statistics, and the ANOVA, R^2, and SE of residuals. Ask students how this study differs from the previous studies of Chapter 2, and why confounding is now a concern. (Answer: This study does not use random assignment, and in fact, could not randomly assign employed men to the race groups!)

Because students often confuse extraneous variables and confounding variables, it's helpful to remind them that variables in the unexplained column are all extraneous sources. If random assignment to the explanatory variables or blocking? is used, it should keep these variables from becoming confounded with the explanatory variable. However, because random assignment is not used in this study, it's highly likely that one or many of these extraneous sources are confounded with the explanatory variable. Also, remind students that to be confounded with the explanatory variable, a variable like education must be both an extraneous source of variability in the response (i.e., it's associated with the response) and associated with the explanatory variable. So, as long as the study has data on that variable, we can check to see whether both of these relationships exist.

To determine whether education is a confounding variable, check for a relationship between education and wages, and between education and race. Show the one-variable analysis for wages vs. education, including the boxplots, summary statistics, ANOVA, and R^2. Point out that study participants with beyond a college education tend to make more money per week than those with below college education. Remind students that this means beyond a college education has a positive "effect" on weekly wages, whereas below college education has a negative effect on weekly wages. Then show a mosaic plot for race (x-axis) vs. education (y-axis). Have students summarize how the proportions of those with a beyond college education change with race, pointing out that, unlike in the strawberry or finger tapping studies, we cannot determine the impact on wages by changing only one variable at a time, as the distribution of education changes as we go from non-black to black. This establishes education as a confounding variable in the relationship between wages and race. It's helpful for students to practice explaining what that means in terms of the original association between wages and race. To help, ask students to provide an alternative explanation for the difference in mean wages between non-blacks and blacks. For example, another reason why the men in the non-black group make more per week, on average, than those in the black group is because the non-black group has a larger percentage of men with beyond a college education and those with a beyond a college education tend to make more money, on average.

Because education is a confounding variable, we would like to adjust or account for it in the analysis. If we adjust for education, and non-blacks still make significantly more per week, on average, than blacks, we have a stronger conclusion than if we ignore education. The adjustment process is going to be the same as that with strawberries. If the wage value is a beyond college education value, we'll subtract the positive education effect on wages, and if the wage value is a below college education value, we'll subtract the negative education effect on wages. In both the non-black and black groups, there is a higher proportion of below college values than beyond college values, so there are more wages values in both the non-black and black groups moving up than are moving down. However, because the black group has a higher proportion of its group in the below college education group, the net result is that the education-adjusted mean in the black group shifts up more than the education-adjusted mean in the non-black group. This brings the education-adjusted means for the two race groups closer together. The Multiple Variables applet can be used to show this adjustment process.

Errata

- Please check the Errata for an update to the end of chapter 2 exercise 2.CE.3.

Technology

- Exploration 2.3 uses the Multiple Variables applet.
- If statistical software is used, we recommend using it in addition to the Multiple Variables applet. The applet will allow students to see the adjustment of the responses, i.e., the shifting up and down of the acceleration times within the weight categories as the acceleration times are adjusted for the effects of low and high horsepower.

Appendix: Calculation Details

Overview

Often the appendices get overlooked. However, the demonstrations in this appendix are valuable for student understanding of the adjustment process in observational studies. The goal of this appendix is that students will be able to predict how a mean for a category of an explanatory variable will change based on how the explanatory variable is related to a confounder. A hypothetical data set based on the Salary Discrimination study is provided, **AdjustmentIntuition**. The data set has the wages of 100 individuals, 50 black and 50 non-black. A third variable is provided, sex (binary: male, female). There are four different columns in the data set labeled 'sex' with differing distributions of males and females within the black and non-black race categories. In all cases, males tend to make more (positive effect on wages) than females (negative effect on wages). The Multiple Variables applet is used to demonstrate how changing the way race and sex are confounded, changes the adjusted relationship between wages and race.

Approximate class time

- One 50-75 minute class period for guided discussion.

Implementation Tips

Begin by showing students the association between wages and race, noting the mean wages for blacks and non-blacks and the difference in these means. It's also helpful to provide the one-variable analysis, including SSRace, SSError, R^2 for race, and F-stat and p-value for testing the significance of race. Whether the test is significant is not important. We want to point out how the *F*-statistic, p-value, as well as the other summary values, will change as we adjust for sex.

Remind students that in order for sex to be a confounding variable in the relationship between wages and race, it has to be an extraneous source of variability in wages (i.e., associated with wages, or explains variation in wages) and it must be associated with race. So, show students first, using summary statistics and graphs that wages and SexAll are associated and stress that males tend to make more than females (i.e., males have a positive and females a negative effect on wages). Students may also notice that the mean wages for blacks and non-blacks are the same as the mean wages for females and males! Next, show students the relationship between race and the variable SexAll. This can be done in the applet using Race as the response and SexAll in the Subset By box, or using a separate mosaic plot. The important thing is for students to see the complete confounding between race and sex, that 100% of the non-black participants are male and 100% of the black participants are female. Finally, fit the two-variable model with SexAll above Race in the Explanatory box. The applet will not show any output for the statistical model or ANOVA because of the complete confounding. The variation in wages that is explained by race is the exact same variation explained by SexAll.

Next, use the variable Sex where most non-blacks are males and most blacks are females. Show students the relationship between Race and Sex, then fit the two-variable model with Sex above Race in the Explanatory box. Remind students that males tend to make more than females, then ask how the observed wage dots in the black group will shift (i.e., up or down) when we remove (i.e., subtract out) the positive effect on wages of being male and the negative effect on wages of being female. The blue dots should shift up and the red dots should shift down, but because there is a higher proportion of blue dots shifting up than red dots shifting down the net change in the mean wages of blacks after adjusting for sex is an increase. A similar discussion can be had for non-blacks, but the net change will be a decrease in the mean wages after adjusting for sex. In addition, the mean wages in the non-black group changes more, because that group has a greater imbalance between males and females.

The variable SexOther has a greater proportion of non-blacks who are female and a greater proportion of blacks who are male, and the variable SexEqual has a 50/50 distribution of males and females in both black and non-black groups (so Sex is not a confounding variable in the last case). Covering these two cases in this order, ask students to predict how the blue/red dots will shift and the net change in the mean wages for blacks and non-blacks after adjusting for sex. Then either demonstrate for the students, or have them use the applet themselves to see the adjustment process in action.

Besides being able to predict how the mean wages for blacks and non-blacks will adjust, students should be able to discuss whether (and possibly how) the SSrace, R^2 for race, and SSError will change in the two-variable model compared to the one-variable model containing only race when there is partial confounding or no confounding at all. Show students that adding sex to the model in all cases will reduce the SSError. Then go back and show how depending on the relationship between race and sex, adjusting for sex either brings the mean wages for blacks and non-blacks closer together (smaller SSRace, smaller R^2) or pushes them further apart (larger SSRace, larger R^2). These changes will then change the *F*-statistic and p-value for Race, but in the case of means getting closer together we may not be able to predict what will happen to the *F*-statistic, because the SSError is going down, as well.

Technology

- Multiple Variables applet

Chapter 3: Multi-factor Studies and Interactions

Chapter Overview

In Chapter 3 we consider the design and analysis of studies with more than one explanatory variable of interest, again covering both experiments and observational studies. Section 3.1 introduces factorial experiments with an example and exploration that do not have statistical interaction. In Section 3.2 we introduce statistical interaction in a completely randomized design and in Section 3.3 we discuss the importance of replication in considering interactions between a blocking variable and the factor(s) in block designs. The chapter closes with observational studies and interaction in Section 3.4.

Section 3.1: Multi-factor Experiments

Overview

The goals of this section are to design an experiment with more than one explanatory variable of interest and explore the benefits of a two-variable study where the levels of both variables are assigned by the researcher. The section introduces some design of experiments language including full factorial designs, balanced designs and main effects. Staying close to the ideas presented in the previous chapters, it continues to use the six-step process and focus on explaining variation, starting with the one-variable model using the treatments progressing to the two-variable model using both factors. Because there is not a statistical interaction between the factors, the variation explained by the one and two-variable models is almost the same.

Approximate class time

- One-to-two 50-75 minute class periods:
    - One 50-75 minute, in-class period for a guided discussion of factorial designs. (Example 3.1 could also be used as a pre-class reading.)
    - One 50-75 minute, in-class period for Exploration 3.1: Pig Growth. (Exploration 3.1 could also be assigned as homework.)

Implementation Tips

Example 3.1 is a two-factor study in a completely randomized design. The factors of interest are corporate credibility (negative vs. positive) and endorser of a women's running shoe (Florence Griffith Joyner vs. Rosanne Barr). The response variable is purchase intent which was measured using the sum of three 1-7 Likert scale questions, where higher sums mean higher likelihood of purchasing the product. A good way to motivate the design of factorial studies is to begin with some bad designs and have students identify the issue(s). For example, in the corporate credibility study we begin with a design that completely confounds corporate credibility with purchase intent, then give a design in which each participant is assigned to only one factor of interest. After discussing both of these bad designs, the appropriate full factorial design becomes reasonably intuitive.

As with other examples and explorations it's important to give students an opportunity to think about variation in the response, filling out the sources of variation diagram and brainstorming sources of unexplained variation, reminding them that the random assignment to treatments keeps the unexplained sources of variation from becoming confounded with the treatments/factors. Here too is an opportunity to motivate the balanced design, showing students that it is the equal sample sizes in the treatments that keeps the two factors from being confounded with each other. This can be done with either a 2x2 table of counts or a mosaic plot.

The analysis of the data can be introduced using a 2x2 table of the treatment means. From this table, have students determine the row means (e.g., corporate credibility means) and column means (e.g., endorser means) and the resulting main effects of corporate credibility and endorser. From here you could have students determine, by hand, the SStreatment, after finding the treatment effects, and the sums of squares for corporate credibility and endorser. Because there is not a statistical interaction between corporate credibility and endorser on the mean purchase intent, the SStreatment ≈ SScorpcred + SSendorser. Students should be able to work out the expected ANOVA tables for both the one-variable model using the treatments and the two-variable model using the two factors. The output from statistical software confirms their expectations.

Section 3.1 sets the stage for statistical interaction by first considering a factorial study that does not have a statistical interaction so that the sum of squares for the treatments is approximately equal to the sum of squares for the two-variable model using the factors. Point out to students that the decision of which model to use is based on the research question, not the adequacy of the model. Each model can be used to answer different research questions.

Technology

- Exploration 3.1 uses the Multiple Variables applet.
- Statistical software could also be used.

Section 3.2: Statistical Interactions

Overview

Section 3.2 introduces statistical interaction in a designed experiment. The 'difference in differences' statistic is used to both teach the concept of statistical interaction, but also to assess the strength of the interaction and the statistical significance of the interaction using simulation.

Approximate class time

- Two-to-three 50-75 minute class periods if everything is done in class.
    - One-to-two 50-75 minute class periods: The first class period can be used for a guided introduction to statistical interaction. The second can be used to assess the statistical

significance of the interaction and could also be used to investigate more thoroughly interaction plots. Example 3.2 can be assigned as a pre-class reading.

- One 50-75 minute class period for Exploration 3.2. This exploration could also be assigned as homework.

Implementation Tips

Example 3.2 is a designed experiment with two factors (temperature and air velocity) each at 2-levels in a completely randomized design. The experimental unit is a batch of pistachios and the response variable is the peroxide remaining after drying, measured as a percent. After introducing the example and working through the sources of variation diagram, one path through this example is to compare the predicted values from the two-variable model to the actual treatment means. Have students determine the predicted percentage peroxide remaining for all 4 treatment combinations, and identify the treatment that produces the most desirable outcome, the smallest mean percentage of peroxide remaining. Then compare this to what actually happened in the study. This comparison can be done using 2x2 tables of the predicted and actual treatment means, but should also be done using side-by-side interaction plots. The goal of this comparison is two-fold: (1) to show the additive nature of the two-variable model, and (2) to show that the two-variable model is not completely capturing the pattern that we see in the actual treatment means. Goal (2) motivates the need for a different model.

When thinking and talking about a statistical interaction, it's important for students to go beyond the idea that "the lines cross" or that "the lines are not parallel." Certainly, those ideas can be used to recognize, in an interaction plot, whether interaction exists. However, they do not convey what a statistical interaction is. To help students understand *why* the lines cross, frame the discussion in terms of the difference in the mean response between two-levels of one factor, changing depending on the level of the second factor. This provides an intuitive statistic for measuring the strength of the interaction, the "difference in differences" statistic. Give students some practice with this statistic by providing several hypothetical interaction plots, one that crosses and one that doesn't but still has interaction and one that shows parallel lines. Have the students discuss which exhibits the strongest interaction and why.

The difference in differences statistic can then be used in a simulation (Two-Variable ANOVA applet) to assess whether or not the statistical interaction is significant. The applet has the option to show both how the re-randomization re-assigns the observed responses to the treatments, and also how the interaction plot changes with each re-randomization.

Because the interaction is statistically significant, a two-variable model with interaction is more appropriate for these data. To estimate the interaction effects, return to the predicted values the students determined from the two-variable additive model. Find the differences between the observed treatment means and these predictions. These are the interaction effects, capturing how far off the additive model is from the interaction model. It's helpful to show an interaction plot with both the predictions from the additive model and the observed treatment means so that students can visually see the interaction effects. After obtaining the interaction effects, show students that the SSinteraction can be computed just like the sum of squares for any other model term. A nice discussion here, is to get students thinking about the size of the interaction effects when the interaction is stronger, or there is very little interaction between the two factors as this leads to the size of the SSinteraction and the subsequent theory-based F-test for the interaction.

A common student misconception is that it is the interaction that leads to the optimal treatment. In the pistachio study this is true, but it is not true in all studies. If the lines do not cross, but instead just converge, the additive model will still provide the optimal combination of the factors. The idea of adjusting the mean response at each level of a variable for a second variable and the interaction between two variables on the mean response can be confusing for students. As you progress into Section 3.4 (observational studies) you may want to spend some time on the differences in these two ideas. Another common trouble spot for students at this point in the curriculum is confusion between confounding, covariation and interaction. Because this is a balanced design, confounding and covariation are not present. However, in Section 3.4 when we come back to observational studies students will need to consider all three of these ideas, so spending some time in this section reminding students why there is not confounding between the two factors and hence no covariation will be helpful later.

Technology

- Example 3.2 and Exploration 3.2 both use the Two-variable ANOVA applet.
- If statistical software is used, we recommend using it in addition to the Two-variable ANOVA applet. The applet will allow students to see how the y-values are re-randomized to the treatments in a full-factorial design reinforcing the random assignment process used in the original study. It also allows students to work with the difference in differences statistic which reinforces conceptual understanding of statistical interaction.

Section 3.3: Replication

Overview

Section 3.3 covers generalized block and within-block factorial designs. Recall that in a randomized complete block design we cannot test for an interaction between the blocking variable and the factor due to a lack of replication. Because the generalized block has replicate experimental units assigned to each treatment within each block we are able to think about and test for interaction with the blocks. In a within-block factorial design, because there are more experimental units within each block to accommodate even one replicate of the full factorial, in addition to testing interactions among the factors, there are often enough degrees of freedom to test at least the two-way interactions between the factors and the blocking variable. In addition to defining and describing the advantages of these two types of block designs, an additional learning goal in Section 3.3 is to explain the benefits and challenges of replication.

Approximate class time

- One-to-two 50-75 minute class periods
  - One 50-75 minute class period for a guided discussion of the importance of replication, generalized block designs and within-block factorial designs. Example 3.3 can be assigned as a pre-class reading.
  - One 50-75 minute class period for Exploration 3.3. Exploration 3.3 could also be assigned as homework.

Implementation Tips

The generalized block design is motivated by the desire to test for interaction between the blocking variable and the factor(s). It's helpful to show students an example of a one-factor, randomized complete block design such as those they saw in Chapter 2. Provide the interaction plot of block x factor, and discuss whether it's possible to test for an interaction between the factor and the block. To help answer that question have students work out the degrees of freedom for the randomized complete block design if they wanted to test for the interaction between the blocking variable and the factor. Doing this results in 0 degrees of freedom for the error. If you are using statistical software, show students the analysis of both the two-variable model with and without interaction. Because there is no replication, any deviations from additivity in the interaction plot are assumed to be "random variation," and it's that variation that is used as the SSError. An important point to make here is that if the treatment means deviate substantially from additivity, and we believe there really is an interaction between the bock and factor, in the randomized complete block design we have no choice but to use the SSinteraction as the SSError. So a natural follow-up question is to then ask students how the design could be changed in a way that would allow for a measurement of the SSError that is independent of the SSinteraction, guiding students to the idea of replicating within the blocks.

The within-block factorial design is just an extension of the randomized complete block design to two (or more) factors run in a full-factorial. In this design, students should understand that replication is not necessary to test for the interaction(s) between/among the factors. In addition, because the sample size will naturally be larger to accommodate the full-factorial within each block, even with only one replicate per treatment within each block, there are often enough degrees of freedom to also test for at least the two-way interaction(s) between the blocking variable and the factor(s). To help students see this, have students work out the sources and degrees of freedom in the ANOVA table matching the different possible models which could be fit to the data. As an extension, show students an example where the generalized block design has two explanatory variables and now even the 3-way interaction between the blocking variable and both factors can be assessed.

In addition to the studies shown in the text, a nice motivating example for a generalized block design is a drug trial, blocking on sex (generally binary – male/female) or clinic (i.e., region of the country). It is common for studies done on newly developed medications to take place in various clinics across the country. Because of regional differences in health and diet, the design and analysis of the data blocks on clinic. This type of example can be used to show students the weaknesses in randomized complete block design, because in this type of study we could reasonably expect there to be an interaction between the clinic and the medication. Having only one individual on each medication within each clinic means we cannot test for the interaction between clinic and medication. Here is also where the small sample size argument really has merit since it's absurd to think any medication would be approved if it was only tested on one person within each clinic!

Technology

- Exploration 3.3 uses the Two-variable ANOVA applet and statistical software.

Section 3.4: Interactions in Observational Studies

Overview

Section 3.4 considers interactions with observational data. It continues with interpreting interaction plots and fitting two-variable models which include the interaction term, but now in the presence of covariation.

Approximate class time

- One-to-two 50-75 minute class periods.
    - One 50-75 minute class period for a guided discussion of interaction in observational studies. Example 3.4 can be assigned as a pre-class reading.
    - One 50-75 minute class period for Exploration 3.4. This exploration could also be assigned as homework.

Implementation Tips

This will be the second iteration through observational studies in which students have to deal with the impacts of confounding in the model. This comes as the last section of Chapter 3 in the hope that students have developed a clear understanding of interaction, that is separate from confounding/covariation, as both of these ideas will now be considered within the same model. For this reason, it is helpful to begin the discussions in this section with the additive model. Going from the one-variable model to the two-variable additive model, remind students how adding a second explanatory variable changes the relationship between the first explanatory variable and the response (e.g., adjusted means, adjusted SS). Contrast this with the designed experiments they have considered previously in this section. It is beneficial at this point to show the mosaic plot of the two explanatory variables in the observational study discussed in Section 3.4 to help explain why there is covariation in this model.

As in the previous sections, use the interaction plot as motivation for including the interaction term in the model. The residual plot for the two-variable model could also be used to motivate the inclusion of the interaction if the residuals show some curvature. This naturally lends itself to a discussion of model adequacy and which model is better (e.g., higher R^2, statistically significant interaction) and/or gives better fit to the data (residuals). Compare the interaction model output to that of the additive model also to discuss how including the interaction term has changed (again) the relationship between the original explanatory variable and the response. It's important here though to make sure students continue to be aware that they should be careful in interpreting main effects when interaction is present among the treatment means. It is also important to remind students that every time we add a variable to the model, more of the previously unexplained variation becomes explained. This can be reinforced by going back to the Sources of Variation diagram.

Technology

- Exploration 3.4 uses the Multiple Variables applet.
- Statistical software can also be used.

Chapter 4: Including a Quantitative Explanatory Variable

Chapter Overview

In Chapter 4 we consider one explanatory variable and two explanatory variable models which contain a quantitative explanatory variable. In Section 4.1 we explore linear relationships between two quantitative variables focusing on describing the linear association. Section 4.2 uses simulation-based inference, as well as the theory-based *t*-test to assess the strength of evidence of a linear association between two quantitative variables. Section 4.3 moves to a two-variable additive (no interaction) model using a quantitative and binary categorical predictor. Section 4.4 continues to use both a quantitative and binary categorical predictor, but now introduces the interaction. Finally, in Section 4.5, we consider two-variable models that include both a quantitative and categorical predictor where the categorical variable has more than two categories.

Section 4.1: Quantitative Explanatory Variables

Overview

Section 4.1 focuses on a one-variable model (one explanatory variable) where the explanatory or "predictor" variable is quantitative. The goals of this section are to describe the association between two quantitative variables numerically and graphically; interpret the least squares regression model between two quantitative variables; and compare and contrast the separate means model vs. the linear regression model. (Inference for regression comes in Section 4.2.) To transition students into the linear model leveraging the idea of explaining variation, the example considers a designed experiment which can be analyzed using either the separate means model or a linear regression model. The exploration considers an observational study in which either the separate means model or the linear regression model could be used.

Approximate class time

- One to two 50-75 minute class periods:
    - One 50-75 minute, in-class period for a guided discussion of describing linear associations. (Example 4.1 could also be used as a pre-class reading.)
    - One 50-75 minute, in-class period for Exploration 4.1: Fatty Acids and DNA. (Exploration 4.1 could also be assigned as homework.)

Implementation Tips

Example 4.1 is a one-factor study (one explanatory variable) in a completely randomized design. The factor of interest is percentage of ethanol used in a solvent in order to extract a particular polyphenol (proanthocyanidin, PC) from grape seed. As with the other studies considered so far, we recommend beginning with a Sources of Variation diagram. This helps the students focus on the variables of interest, whether this is an observational study or designed experiment, and unexplained sources of variation, as well as to begin to consider the limits on the scope of inference. To facilitate the transition to a quantitative variable, the example study was chosen so that it could initially be analyzed using the now familiar separate means model. Using dotplots or boxplots of the amount of PC recovered in each of the ethanol groups, ask students whether the ethanol % explains variation in the amount of PC recovered, and also whether it explains all of that variation, then show them the one-variable separate means model for a discussion of R^2 and statistical significance of the ethanol factor. Follow-up questions can focus on using the separate means model to make predictions. Next, again using a scatterplot (with the group means), ask students whether there are any interesting features in the relationship between the mean amount of PC extracted and the ethanol percentage. Without too much prodding students should see the increasing trend in the means. The linear regression model can then be introduced as the model we use to capture that trend, showing this model as an overlay on the scatterplot and providing the prediction equation. (You can also show the fairly "constant" increase in the means with each additional one-unit change in ethanol percentage.) Students should be able to express that like the separate means model, the linear regression model explains variation in the response variable. One way to get to this idea is to return to the separate means model and ask what sort of pattern we would expect in the means if the treatment variable did not explain variation in the response. Follow-up by asking what the linear trend would look like in this case. At this point, you could also show the students the ANOVA table and/or pie chart partitioning the variation explained by the regression model, paying particular attention to R^2 and its interpretation.

In the discussion of the linear regression model there are two key take-aways. First, students should be able to interpret both the *y*-intercept and the slope, in the context of the study. Emphasize that these interpretations are about the predicted (mean) response, not the individual response values. Second, the regression model is fit to the data using the method of least squares, i.e., by minimizing the SSError. Because students are already familiar with residuals, and the idea of wanting the SSError to be small, this is should be a reasonably quick discussion.

With two possible models we want to consider which model is better or which model should be used. In comparing the two models, remind students how the SSModel for the separate means model is computed, and then show them how there is an analogous computation for the SSModel for the regression model. An important point to make is that if the treatment means had a perfectly linear relationship, the SSModel values would be the same for both models. But, because the means do not fall on a perfect line, the separate means model will explain a greater proportion of variation in the response than the linear regression model. This may lead students to believe the separate means model is better. Three arguments against this are: (1) the linear regression model can be used to make predictions (interpolate) at values other than the three levels of ethanol % that were considered in the study, (but should not be used to extrapolate beyond these values of ethanol); (2) understanding and estimating the rate of change in the predicted response, i.e., the slope of the regression line, is often more informative than simply understanding that the treatment means are different; and (3) the linear regression model is a simpler model, costing only 1 degree of freedom to fit, which can be particularly helpful with small sample sizes. The first and second arguments may actually be goals of the study, and if so, the separate means model is definitely not the appropriate model. The third argument leads to a nice discussion of how the linear model can have a smaller standard error of residuals than the separate means model, when the R^2 is actually smaller for the linear model. Finally, as in any discussion of model comparison, a look at the residuals vs. predicted values plot (or explanatory variable) may help to shed some light on which model should be used. Do caution students that there is not always "one right model," but to focus on whether the model appears to fit the data/theory, and how to interpret it.

Technology

- Exploration 4.1 uses the Multiple Variables applet.
- Statistical software could also be used.

Section 4.2: Inference for Simple Linear Regression

Overview

Section 4.2 covers inference for simple linear regression using both simulation-based inference and theory-based tests. The validity conditions for the theory-based test are also covered.

Approximate class time

- One to two 50-75 minute class periods if everything is done in class.
    - One 50-75 minute class period for a guided discussion of simulation-based inference in a simple linear regression model, and the theory-based *t*-test (and *F*-test). Example 4.2 can be assigned as a pre-class reading.
    - One 50-75 minute class period for Exploration 4.2. This exploration could also be assigned as homework. Exploration 4.2 uses the same study as Exploration 4.1 and could easily be combined with Exploration 4.1 after covering both Sections 4.1 and 4.2.

Implementation Tips

Example 4.2 uses the same study as Example 4.1, picking up with Step 4: Draw inferences beyond the data using the 3S Process. As in Chapter 2, one purpose of using simulation is to reinforce the randomness used in the study design. After providing the null and alternative hypotheses, or getting students to provide them, have students brainstorm possible statistics which could be used to measure the strength of the linear association in the observed data. To get students thinking about the simulation, have them discuss with each other, or write down individually, how to carry out one repetition of the simulation using index cards. Have them provide the following information: (1) number of cards required, (2) what is written on each card, (3) after shuffling, what is done with the cards, and (4) the statistic which should be determined from this one simulation and plotted to form the null distribution. The Two Quantitative Variables applet can be used to show students the simulation process using R^2, slope, and finally the *t*-statistic for the slope. As in earlier chapters, the conclusion will be the same regardless of the choice of statistic (although the p-value may vary slightly as these are not completely equivalent statistics).

One motivation for using the *t*-statistic, as was discussed in previous chapters, is the desire to have a standardized statistic. Showing students the formula for the *t*-statistic, will help them think about how the size of the *t*-statistic depends on both the sample size and the standard deviation of the explanatory variable. The Two Quantitative Variables applet (pasting in different datasets) can also be used to contrast the shuffle-to-shuffle variation when the sample size is small vs. large (all other things being the same), and/or when the standard deviation of the explanatory variable is small vs. large (all other things being the same). The idea of larger standard deviation in the explanatory variable reducing variation in the shuffled slopes will be initially counter-intuitive to students, but is a nice reminder to those designing their own studies.

Technology

- Example 4.2 and Exploration 4.2 both use the Two Quantitative Variables applet.

- If statistical software is used, we recommend using it in addition to the Two Quantitative Variables applet. The applet will allow students to see how the *y*-values are re-randomized to the existing explanatory variable values reinforcing the randomness used in the original experimental study. It also allows students to work with several different statistics, besides just the *t*-statistic, to assess the statistical significance of the linear association.

Section 4.3: Quantitative and Categorical Variables

Overview

Section 4.3 covers two-variable additive models with a quantitative and binary categorical variable in an observational study. Because the study is observational, one of the goals of the section is to illustrate again the way in which adding a variable into the model adjusts the relationship between the response and the original explanatory variable, which in this case is quantitative. Additionally, Section 4.3 covers the creation of indicator variables, using both effects and indicator coding, and uses residual plots to assess the validity of the fitted regression model.

Approximate class time

- One to two 50-75 minute class periods
    - One 50-75 minute class period for a guided discussion of adjusting the slope in a simple linear regression model for a binary categorical variable, discussion of the parallel lines model including effects and indicator coding, and assessing the fit of the model. Example 4.3 can be assigned as a pre-class reading.
    - One 50-75 minute class period for Exploration 4.3. Exploration 4.3 could also be assigned as homework.

Implementation Tips

To transition from Sections 4.1 and 4.2 which covered the simple linear regression model, begin the example study in Section 4.3 using a simple linear regression model, then add in a binary categorical variable. The Sources of Variation diagram is, again, a nice way to get students into the study, and to help them keep in mind the myriad of sources of variability that are unexplained. If using Example 4.3 the Michigan Housing Prices study, the primary relationship of interest is how the price of home relates to the size of the home. This linear relationship will be adjusted for the location of the home (near the lake or not), so make sure location is in the list of unexplained sources of variability.

In addition to the desire for good predictions and a small standard error of residuals, to motivate inclusion of location in the model, show students the scatterplot of price vs. size, and the residuals vs. predicted prices plot color coded by the location of the home. Students will recognize that the simple linear regression model underpredicts the prices of lake front homes and over predicts the prices of non-lake front homes, which is not a good thing for this model to do.

Before adding location to the model, consider plots and summary statistics of the relationships between price and location, and location and size, which will confirm that location and size are confounded. Remind students that because of this, they should expect the relationship (i.e., the slope) between price and size to change when we bring location into the model (i.e., adjust for location).

The Multiple Variables applet can be used to demonstrate how the prices shift (e.g., up or down) when adjusting for location and how the sizes shift (e.g., left or right) when adjusting for location. This discussion is very similar to that in the Appendix of Chapter 2. For example, because lakefront homes tend to be more expensive, when we adjust the prices for location (i.e., subtract out the effect of location) we expect the lakefront home prices to shift down and non-lakefront home prices to shift up. Similarly, we expect the sizes of lakefront homes to shift down and non-lakefront home sizes to shift up. (Because size is also quantitative, it now needs to be adjusted for location, just like price.) Adjusting both price and size for location will then change the slope of the regression line between price and size. The Multiple Variables applet will allow students to see the adjustment of the data values, and the resulting location-adjusted regression line compared to the original, unadjusted regression line. Encourage students to try to predict these changes before they check the box.

After adjusting the slope for location, show the regression lines (parallel) for each location. From here, students need to understand how statistical software will provide the equations for these two lines. Using effects coding is consistent with the effects version of the separate means model, and so is a nice transition from what students have already seen. Indicator coding allows for a *t*-test comparing the predicted price of a lakefront home to that of a non-lakefront home (about the intercept). In Example 4.3, we opt to show students both types of coding, but one or the other could be omitted. Whatever the decision, provide students with the regression table and give them practice writing out the full multiple regression model, and then simplifying that model to obtain a model for each location. It can also be helpful to show an annotated scatterplot with the distance between the two lines for the indicator coding model and another with the plus/minus effects from an overall line using the coefficients from the effects coding model. Also give students practice with interpreting the coefficients in the models, as well as drawing conclusions from the *t*-tests provided in the regression table. To keep the theme of explaining variation in the forefront, comparisons of the one and two-variable models can be done. Focus on how the SSsize changes when adjusting for location, as does the SSError.

A check of the residuals vs. predicted prices plot, color-coded by location, for the two-variable additive model will show curvature. This should be a sign to students that even though the residuals are generally smaller, the model fit is not adequate. This residual vs. predicted plot is the motivation for Section 4.4, including the interaction between location and size in the model.

Technology

- Example 4.3 and Exploration 4.3 use the Multiple Variable applet.
- If statistical software is used, we recommend using it in addition to the Multiple Variables applet. The applet will allow students to see how the data values in the scatterplot shift when they are adjusted for a categorical variable.

Section 4.4: Quantitative/Categorical Interactions

Overview

Section 4.4 considers interactions between quantitative and categorical predictors in an observational study. It continues with interpreting interaction plots and fitting two-variable models which include the interaction term, in the presence of covariation.

Approximate class time

- One to two 50-75 minute class periods.
  - One 50-75 minute class period for a guided discussion of interaction between a quantitative and categorical predictor. Example 4.4 can be assigned as a pre-class reading.
  - One 50-75 minute class period for Exploration 4.4. This exploration could also be assigned as homework.

Implementation Tips

The curvature evident in the residuals vs. predicted price plot from the two-variable additive model in Example 4.3 motivates the inclusion of the interaction term in the model. Because the two-variable additive model tends to underpredict the prices of large lakefront homes but overpredict the prices of large non-lakefront homes (visa versa for small homes), it suggests that the relationship between price and size depends on location. Show students a scatterplot of the home prices vs. sizes, color-coded by location and ask what it means for there to be an interaction between location and size on predicted price. Go back to the difference in differences idea. If we choose two different sizes, and if the difference in differences is not 0 (i.e., interaction is present) the regression lines should no longer be parallel. This amounts to fitting separate regression lines to the data for each location. Show students another plot, with the two separate regression lines fit to the data, along with the equations of those lines, as in Figure 4.4.3. From those two regression lines, students can determine the predicted price of a lakefront and non-lakefront home at two different sizes, confirming that the difference in differences is non-zero. The residual vs. predicted price plot for this model will look more random and might be a nice plot to include here to show students that fitting the regression lines with different slopes has dealt with the curvature seen in the previous residual plot.

As in Section 4.3, one strategy at this point is to show students the statistical output, including the regression table for the two-variable model with interaction. Have them write out the full multiple regression equation. (It may be easiest to choose either effects or indicator coding to use exclusively in Section 4.3 and Section 4.4.) From the full model, have students simplify to find the separate regression models for lakefront and non-lakefront homes, then practice interpreting the coefficients in the full regression model, as well as the *t*-tests for those coefficients. A helpful mantra is that indicator variables change intercepts and interaction terms change slopes.

The interpretation of the model coefficients can be difficult for students, particularly the idea that the adjusted slope for size is the change in the predicted price for a 1-square foot increase in size for the baseline or reference group only. To help students with interpreting the interaction coefficient, clearly show how the slope for the non-reference group has changed from the reference group slope by an amount equal to the interaction coefficient.

Because this is an observational study, it’s a good reminder to students to point out how the coefficients and sums of squares have changed in going from the two-variable additive model to the two-variable model with interaction, due to confounding that is now present not just between size and location, but between the interaction and size, and the interaction and location. One way to show students how/why the interaction is confounded with each predictor, is to show exactly how the interaction term is computed for the model, i.e., as the product of size and location.

Technology

- Exploration 4.4 uses the Multiple Variables applet.
- Statistical software can also be used.

Section 4.5: Multi-level Categorical Interactions

Overview

Section 4.5 extends Section 4.4, covering two-variable additive models and two-variable models with interaction when the categorical variable has more than two categories.

Approximate class time

- One to two 50-75 minute class periods.
  - One 50-75 minute class period for a guided discussion of interaction between a quantitative and categorical predictor. Example 4.4 can be assigned as a pre-class reading.
  - One 50-75 minute class period for Exploration 4.4. This exploration could also be assigned as homework.

Implementation Tips

The big idea for this section is the number of indicator terms necessary when fitting a two-variable model with a categorical variable with more than 2 levels. Because students should be reasonably comfortable fitting two-variable models, and the software (Multiple Variables applet or statistical software) does not care how many levels the categorical variable has, one option in teaching this section is to show students the regression table, using indicator coding in the two-variable additive model for example, then have them work out the simplified models for each level of the categorical variable. The idea is for students to discover that we only need *k*-1 indicator variables for a categorical variable with *k* levels. Then, as was done in Sections 4.3 and 4.4, give students the opportunity to practice interpreting the model coefficients, noting that all levels of the categorical variable have the same linear association between the response and the quantitative predictor.

In terms of statistical testing, this section introduces the partial *F*-test. Students will need to be reminded that taking out the categorical *variable* corresponds to removing all of the indicator terms at once. The need for the partial *F*-test test can be motivated by the desire to keep the Type I error rate under control. Instead of looking at the individual *t*-tests for the indicator variables, which either compare each clarity group to the overall relationship (effects coding) or compare each clarity group to the reference group (indicator coding), we want to decide first whether there are any differences among the clarity groups (as was done in Chapter 1 with the *F*-test). So, as much as possible, convey the idea that the partial *F*-test which compares the full model to the reduced model, is the same *F*-test students are used to seeing from earlier chapters.

The partial *F*-test is used again when fitting the two-variable model with interaction. Convey to students that in order to control the Type I error rate, we want to first consider whether any of the interaction terms is significant using the partial *F*-test. Then, if the answer is yes (i.e., we reject Ho), we then look at the *t*-tests for the individual interaction terms.

Some discussion of model building can be incorporated in this section. It is possible to collapse one or more of the clarity categories and/or one or more of the slopes. Discussions can be had about whether or not this would be of interest, as well as the general practice of keeping the main effects in the model when they are involved in a significant interaction.

Technology

- Exploration 4.5 uses the Multiple Variables applet.
- Statistical software can also be used.

Chapter 5: Multiple Quantitative Explanatory Variables

Chapter Overview

Chapter 5 covers multiple regression models when the predictors are quantitative. In Section 5.1 we consider two quantitative predictor variables in a designed experiment, both the two-variable additive and interaction models. In this section we also cover the process of standardizing the predictor variables to remove the linear association between the predictors and their interaction. In Section 5.2 we return to observational studies again looking at both the two-variable additive and interaction models, and standardizing when the predictors are linearly associated. Section 5.3 and 5.4 both cover nonlinear models; polynomial models in 5.3 and transformations in 5.4. Note that some of the material in Chapter 5 could be skipped, or touched on more briefly so that students may see what’s possible if they go on to take a class in regression.

Section 5.1: Experiments with Multiple Quantitative Explanatory Variables

Overview

Section 5.1 focuses on two-variable models, both additive and interaction, in a designed experiment. This section considers design issues with quantitative explanatory variables, as well as visualizations among three or more quantitative variables. After model fitting, students learn to interpret a “response surface,” and describe interactions between quantitative variables. Standardizing predictor variables is used to remove the linear association between the predictor variables and their interaction, as well as to make the *y*-intercept a more meaningful value in the context of the study.

Approximate class time

- Two to three 50-75 minute class periods:
  - One 50-75 minute, in-class period for a guided discussion of the experimental design when both predictors are quantitative and the two-variable additive model.
  - One 50-75 minute, in-class period for a guided discussion of the two-variable interaction model, including standardizing. (Example 5.1 could also be used as a pre-class reading.)

- One 50-75 minute, in-class period for Exploration 5.1. (Exploration 5.1 could also be assigned as homework.)

Implementation Tips

Example 5.1 returns to the Pistachio Bleaching study described in Example 3.2. The two factors of interest are drying temperature (60°, 75°, 90°F) and air velocity of the drying fan (1.5, 2 and 2.5 mph), now with all 3 levels for each. The experimental unit is a batch of pistachios and the response is the amount of peroxide remaining after drying measured as a percentage. The design of the study is completely randomized.

The research question for the study motivates the type of model(s) we will fit when analyzing the data. The goal of the study is to determine the combination of temperature and air velocity that will minimize the amount of peroxide remaining. That minimum value could occur at a combination of temperature and air velocity that was not actually run in the experiment. We will treat the explanatory variables as quantitative when fitting the statistical model so that we can interpolate, if necessary, to find the optimal combination.

But, before getting into the analysis it's important to spend some time on the design of the study. The design is completely randomized, but some nice questions to ask students include questions about the treatments, what is meant by completely randomized, and whether the design should be balanced or not. The balance is important here because we do not want there to be an association (e.g., linear) between the two explanatory variables. Show students some example graphs of the design region (temperature vs. air velocity) of full-factorial designs when the sample sizes are not balanced to help students see why balance is helpful in keeping the two predictors from being linearly associated. And as always remind them to use random assignment to avoid confounding with unmeasured variables.

As in earlier sections, it's helpful to begin the analysis of the data with the one-variable models. Show students scatterplots of the amount of peroxide remaining vs. temperature and vs. air velocity, as well as the statistical models and ANOVA output. Important take-aways from these models are the R^2 values, SE of residuals, sums of squares, and the slope coefficients (and starting to think about how/whether these will change in the combined model).

Because the goal of the study is to find the optimal combination of temperature and air velocity, we need a model that includes both variables. A 3D scatterplot can be used to introduce the idea of fitting a plane (e.g., piece of paper) to the observed data, so as to minimize the sum of the squared residuals. Provide the regression table for the two-variable additive model. Have students simplify this model at the three different values of air velocity, for example, run in the study. This will help them see that regardless of the value of air velocity, the slope of the regression line between the amount of peroxide remaining and temperature stays the same (but at different intercepts). Point out that this is exactly the same as the models in Chapter 4, but now air velocity could take on any value between 1.5 and 2.5 mph, so we have many, many regression lines at these different values of air velocity all with the same slope coefficient on temperature. Stacking all of these regression lines forms a plane. Provide ample opportunity for students to practice interpreting the coefficients of the model.

An interesting question for students to consider is which variable is more important in predicting the amount of peroxide remaining. Quantities like R^2, the *t*-statistic and p-value can be used to answer this question, but the slopes cannot be directly compared because they apply to variables on very different scales. To compare the slopes, we can standardize the two predictors and used these standardized variables in our model. To help students with standardizing, use statistical software or the Multiple Variables applet to find the mean and standard deviation of temperature and air velocity. Then, have students find the *z*-score for several of the experimental values of temperature and air velocity. Either put these values into the original data table as new columns as students compute them, or if you are working in statistical software, use a formula to compute the standardized values of temperature and air velocity as new columns in the data table, but it's important for students to see this standardization process even though most packages will do it automatically.

Show students, using 3D scatterplots of temperature vs. air velocity and standardized temperature vs. standardized air velocity that the relationship between the predictors does not change as a result of standardizing, nor does the relationship between the response and the predictors. Then fit the two-variable additive model using the standardized variables. There are several important take-aways from this analysis. First, the sums of squares and R^2 values, and SE of residuals have not changed. Second, the interpretations of the model coefficients have changed. For the slope coefficients, a one "unit" change in standardized temperature or standardized air velocity is a one standard deviation change in temperature (or air velocity). The *y*-intercept is now the predicted amount of peroxide remaining at the mean temperature and mean air velocity. Third, assuming a one-standard deviation increase in the predictors can be considered of equal importance/feasibility, the slope coefficients of standardized temperature and standardized air velocity can be directly compared. Finally, as in the home prices study from Chapter 4, the residuals vs. predicted values plot can be used to motivate next fitting the two-variable model with interaction.

An additional benefit to standardizing the predictors is that it removes the linear association between each predictor and their interaction. Show students using pairwise scatterplots how temperature is linearly related to the temperature x air velocity interaction and similarly for air velocity, but how this is not the case for the standardized temperature and standardized air velocity. Remind students that this "trick" (standardizing to remove multicollinearity) only works with "product terms" like interactions (and later polynomial terms).

For the two-variable model with interaction, showing the 3D scatterplot with the response surface can help students see that the interaction adds curvature to the response surface. As was done previously with the additive model, select several values of standardized air velocity (e.g., -1, 0, 1) and have students simplify the model for standardized temperature to see the changing slope coefficient on temperature. Showing these three regression lines on a plot of predicted peroxide remaining vs. temperature can reinforce the interaction as these lines will not be parallel. Interpretations of the coefficients in the two-variable model with interaction can be onerous for students, particularly the interpretation of the coefficient on the interaction term. Give them plenty of opportunities to practice. For example, you can focus on comparing the slopes of x_1 at two values of x_2 and/or on the coefficient of x_1 as a function of x_2. For interpreting the initial slope of x_1 focus on setting (standardized) x_2 at its mean so the interaction term drops out.

Technology

- Exploration 5.1 uses statistical software.

Section 5.2: Observational Studies with Multiple Quantitative Variables

Overview

Section 5.2 covers two-variable models where both variables are quantitative in observational studies. The goals of the section include visualization of adjusted associations, creating and interpreting added variable plots, interpreting model coefficients, and adjusted sums of squares. In addition, the problems which can be encountered when using explanatory variables that are linearly related are explored.

Approximate class time

- One to two 50-75 minute class periods if everything is done in class.
    - One 50-75 minute class period for a guided discussion of multiple regression when the two predictors are linearly associated. Example 5.2 can be assigned as a pre-class reading.
    - One 50-75 minute class period for Exploration 5.2. This exploration could also be assigned as homework.

Implementation Tips

Example 5.2 looks at predicting IQ score (specifically, performance IQ, "PIQ") from brain size and height in data collected from 40 university student volunteers. Begin the analysis of the data by looking at the linear association between PIQ and brain size using a simple linear regression model and scatterplot. Note the R^2, sum of squares, SE of residuals, and the model equation, particularly the slope coefficient. Next, consider scatterplots and simple linear regression models of PIQ vs. height and height vs. brain size. PIQ and height have a slight negative, linear association, while height and brain size have a moderate, positive, linear association. These associations mean height and brain size are confounded in this study.

The Multiple Variables applet can be used to demonstrate for students the adjustment process of the observed data values for height. After asking students to predict what will happen to high PIQ values and low PIQ values when they get adjusted for height (and similarly for large and small brain sizes), use the applet to confirm for students the adjusted data values (some shifts are quite small), and then the height-adjusted slope between PIQ and brain size. In this example, height and brain size have a positive linear association, but PIQ and height have a negative linear association. Thus, when adjusting both PIQ and brain size for height, the relationship between PIQ and brain size becomes stronger. This can be an unintuitive result for students. We can think of the confounding variable in this case, as "masking" or "dampening" the original relationship between PIQ and brain size. If interested, an equation for the adjusted slope is provided. After discussing the process of adjusting for height, provide the two-variable additive model and discuss, giving students opportunities to interpret the model coefficients.

A natural extension to the two-variable additive model is the two-variable interaction model. Show students the scatterplots of each predictor vs. the interaction. In this model, not only are the predictors linearly related to each other, but they are each also related to the interaction. Compare the two-variable additive model side-by-side with the two-variable interaction model. Students should note that while the two-variable interaction model is significant, none of the individual predictors is statistically significant. In addition, the SSError went up (slightly) with the interaction model, and the standard error of the slope coefficient for height is now about 10 times larger, and that for brain size is 15 times larger! The key idea in these comparisons is that strange things are happening and these are a result of the strong linear associations between the predictors and the interaction term.

One solution to this is to re-fit the model using the standardized variables of height and brain size. After showing students this model output, it should be apparent that standardizing has fixed the strange results see in the unstandardized model. An important point to make here, is that while standardizing removes the linear association between height and the interaction of height and brain size, and between brain size and the interaction between height and brain size, it does not remove the linear association between brain size and height.

There are a couple of important reminders for students. First, a stronger association doesn't necessarily translate into a steeper slope. And second, whether/how two explanatory variables are associated with each other (i.e., confounded) is very different from whether/how two explanatory variables interact on the response.

Technology

- Example 5.2 and Exploration 5.2 both use the Multiple Variables applet.
- If statistical software is used, we recommend using it in addition to the Multiple Variables applet. The applet will allow students to see how the observed brain sizes and PIQ values are adjusted for height.

Section 5.3: Modeling Nonlinear Associations Part I – Polynomial models

Overview

Section 4.3 covers polynomial models for nonlinear associations. Both a quadratic and cubic model are fit to observational data. In addition to interpretations of model coefficients, the fit of the models is evaluated and compared.

Approximate class time

- One to two 50-75 minute class periods
  - One 50-75 minute class period for a guided discussion of fitting polynomial models. Example 5.3 can be assigned as a pre-class reading.
  - One 50-75 minute class period for Exploration 5.3. Exploration 5.3 could also be assigned as homework.

Implementation Tips

After introducing the example study, begin the discussion of the analysis by looking at a scatterplot of the sea ice extent vs. year. Ask students to comment on whether a linear regression model would be an appropriate model and why. Show the linear regression line fit to the data, then ask them to predict what the residual vs. predicted plot will look like. After confirming their prediction, ask what type of model would be more appropriate, one that would result in a more random looking residual vs.

predicted plot. (Students may say a model that includes interaction, based on what they have learned in earlier sections. Because we only have one predictor though, year, the interaction model doesn't apply.) You may also want to remind students that while there is some linear trend to the data, the residual plot reveals there is additional pattern above and beyond the linear trend that can be better explained by a more complicated model.

Show students the quadratic model, including the fitted regression model, regression table, ANOVA table. and residual plot. Comparing the residual plot of the quadratic model to the linear model, students should see that the quadratic model is a better fit to these data. When interpreting the model coefficients, remind students of what they learned in algebra about parabolas: when the coefficient on the quadratic term is negative, the parabola opens down, and thus, has a maximum value. This is an important idea contextually here, because a maximum means there were years when the sea ice extent was increasing and then has been decreasing since.

A consideration in this model, as with all other models, is whether the predictors are linearly associated. Show students the scatterplot of year vs. year^2 to confirm they most definitely are. Standardizing can again by used to remove this linear association with a product term, by shifting the *y*-intercept from year 0, to the mean year in the data of 1995.5. Showing students how this works using scatterplots can be very helpful (the association is still nonlinear, but the nonlinear part is now in our x-values of interest and having that curvature there is actually a good thing for improving our ability to predict the response). Standardizing makes the model a bit easier to interpret as well, since the *y*-intercept now provides the predicted sea ice extent in the mean year of 1995.5 rather than the non-sensical year 0.

For comparison, and to show students another polynomial model, fit the cubic model, using either unstandardized or standardized year. Compare the residual plot for this model to that of the quadratic model. There is not much difference in the residual plots and the cubic term is not statistically significant indicating the quadratic model is sufficient.

Technology

- Exploration 5.3 uses statistical software.

Section 5.4: Modeling Nonlinear Associations Part II - Transformations

Overview

Section 5.4 covers the use of transformations in order to meet the model conditions. Different transformations are used to correct for non-normality and/or unequal variance. In particular, transformations are considered as a way to capture nonlinear behavior.

Approximate class time

- One to two 50-75 minute class periods.
  - One 50-75 minute class period for a guided discussion of transformations. Example 5.4 can be assigned as a pre-class reading.
  - One 50-75 minute class period for Exploration 5.4. This exploration could also be assigned as homework.

Implementation Tips

Example 5.4 returns to the Salary Discrimination study from the Preliminaries and Chapter 2. This is an opportunity to fit a complicated multiple regression model which contains several predictor variables in addition to race (education, experience, region, type of metropolitan area). The residuals for this model, as well as the earlier model from Chapter 2, are extremely right skewed, violating the normality condition. In addition, there is some evidence of unequal variance in the residuals vs. predicted plot.

Show students that taking the log (or using a square root transformation) of the wages will correct for these validity condition violations. A key idea for students to remember is that we cannot compare the transformed and untransformed models using R^2 in any longer, because the SSTotal has not stayed constant.

Interpretations of model coefficients, predicted values and confidence intervals will all be on the transformed scale (e.g., log), which does not make sense contextually to most folks. Show students the back-transformation process to help them arrive at suitable interpretations. A key idea in the interpretation of the back-transformed endpoints on the confidence interval for the slope coefficient of race, for example, is the multiplicative effect. That is, the endpoints represent how many times smaller the median salary of blacks (with same education, experience, region/city) is compared to non-blacks.

Technology

- Exploration 5.4 uses statistical software.

Chapter 6: Categorical Response Variable

Chapter Overview

Chapter 6 focuses on research questions involving a binary response variable. The chapter begins in Section 6.1with a review of descriptive and inferential analyses of two-way tables, including relative risk and odds ratio. Section 6.2 introduces the logistic regression model with one explanatory variable, both quantitative and categorical. The chapter concludes by considering multiple logistic regression models in Section 6.3. This chapter serves as a broad introduction to logistic regression and does not go deeply into the topic. Instead, we aim to continue the emphasis on explained variation.

Section 6.1: Comparing Proportions

Overview

Section 6.1 is a review of descriptive and inferential techniques for comparing groups with a binary response variable that students most likely saw in their first statistics course. These techniques include a two-sample *z*-test and the chi-square test. Students are also introduced to relative risk and odds ratio. Simulation-based inference is used as well.

Approximate class time

- Two to three 50-75 minute class periods:

- One 50-75 minute, in-class period for a guided discussion of descriptive measures in two-way tables (e.g., difference in proportions, relative risk, odds ratio) and simulation-based inference.
- One 50-75 minute, in-class period for a guided discussion of the theory-based tests, two-sample *z*-test and the chi-square test. (Example 6.1 could also be used as a pre-class reading.)
- One 50-75 minute, in-class period for Exploration 6.1. (Exploration 6.1 could also be assigned as homework.)

Implementation Tips

Example 6.1 is a designed experiment studying whether the wording on a driver's license application has an impact on whether people decide to become organ donors. The response variable is binary: yes/no. The explanatory variable is the way the organ donation question is worded as either "opt-in" or "neutral." Study participants were randomly assigned to one of the two question wordings.

As in other studies throughout the book, begin with a graph and summary statistics for the response variable overall. The idea of variability in a binary response will be new to students, so show them several different hypothetical bar charts and ask them which one has more/less variability. Make sure to have an example with no variability (100% Yes, 0% No) and one with the most variability (50% Yes, 50% No). The observed data is somewhere in between. Once students grasp the idea of variability in a categorical response, have them brainstorm a Sources of Variation diagram for the study.

The next step in the descriptive analysis is to look at a segmented bar graph and/or mosaic plot showing the conditional distribution of the response within each question wording group. And, again, the question is whether the question wording explains variation in the response. As in the case of a quantitative response when the explanatory variable means differ, when the conditional distributions differ between the two levels of the explanatory variable, the explanatory variable explains some variation in the response. To help students with this idea, show examples of segmented bar charts and/or mosaic plots in which there is no association between the two variables. Be sure to show the case where the yes/no distribution is 50/50 for both opt-in and neutral, the case where the yes/no distribution is 70/30 for both opt-in and neutral, and then several cases where the explanatory variable will explain more or less variation (e.g., 45/55 opt-in and 60/40 neutral; 90/10 opt-in and 30/70 neutral).

There are several descriptive statistics that can be used to capture how different the groups are. These include the difference in conditional proportions, relative risk, and odds ratio. Because most students have not seen relative risk or odds ratio, spend time both on the computation of these statistics, as well as the interpretation.

The Analyzing Two-way Tables applet can be used to simulate the difference in conditional proportions, relative risk, odds ratio, as well as the chi-square statistic. This applet can then also be used to obtain the two-sample *z*-test p-value and the chi-square p-value.

Near the end of the Example 6.1 we offer some guidance on which statistics to use. This provides a nice opportunity to discuss with students the pros/cons of different statistics. There is flexibility in how much time you spend comparing the relative merits of these statistics. Some statisticians now argue that the simple difference in conditional proportions is not very meaningful. Others note that only odds ratios should be used in case control studies, but students will be least comfortable with odds ratios. Also, the p-value is the same no matter which statistic you use (based on one degree of freedom), the distinction only impacts interpretation and confidence intervals. The end of Example 6.1 also provides a segue into studies where the explanatory variable has more than two levels, and the chi-square test of homogeneity and chi-square test of independence as extensions to the material covered in the text.

Technology

- Example 6.1 and Exploration 6.1 both use the Analyzing Two-way Tables applet.
- If statistical software is used, we recommend using it in addition to the Analyzing Two-way Tables applet so that students can work through the simulation-based inference portion of the exploration.

Section 6.2: Introduction to Logistic Regression

Overview

Section 6.2 introduces logistic regression models using one explanatory variable (either quantitative or categorical).

Approximate class time

- One to two 50-75 minute class periods if everything is done in class.
  - One 50-75 minute class period for a guided discussion of a one-variable logistic regression model. Example 6.2 can be assigned as a pre-class reading.
  - One 50-75 minute class period for Exploration 6.2. This exploration could also be assigned as homework.

Implementation Tips

Example 6.2 considers an observational study on smoking and survival rates in the United Kingdom. The response variable is survival (yes/no) 20-years after the original survey was started. Smoking (yes/no) is the first explanatory variable of interest. A model using age group is also considered.

Like Example 6.2, begin the data exploration using a bar chart of the response variable only. Have students brainstorm a Sources of Variation diagram where smoking status is the explanatory variable of interest. Present the segmented bar chart and/or mosaic plot and have students discuss whether/how smoking status explains variation in 20-year survival. Either present or have students compute the odds ratio, and write out an interpretation of both the odds ratio and a confidence interval for the odds ratio.

There is a statistically significant association between smoking status and 20-year survival, however in the counter-intuitive direction and because this is an observational study, we should expect there to be confounding variables. Age group at baseline is potentially one such variable. Show students a two-way table of 20-year survival vs. age group (7 categories) and have them compute the conditional proportions within each age group who survived to 20-years. A chi-square test can be used to test for a significant association.

To establish age as a confounding variable in the relationship between survival and smoking status, there needs to be an association between age and smoking status. This is where Section 6.3 begins, however, you could show students this mosaic plot here, before getting into the logistic regression model.

The rest of Section 6.2 focuses on the logistic regression model by treating age group as a quantitative variable and modeling the trend. The motivation for the logistic regression model is that it allows us to incorporate more than one explanatory variable. Show students the scatterplot of the proportion alive at 20-years vs. age. Students should recognize that a simple linear regression model is not appropriate. Given what they learned in Chapter 5, some students may suggest a polynomial model, others may suggest a transformation. Using the logit transformation will linearize the S-shaped curvature in the scatterplot of the proportion alive vs. age. Have students determine the odds of survival for several of the age groups, and then the log-odds of survival (using natural log). Then plot these values against age.

To find the logistic regression equation, the estimated regression model for the linear relationship in the plot of log-odds vs. age, use statistical software. Once this model is obtained, spend time helping students with the interpretations of the *y*-intercept and slope of this model, including the back-transformation to a predicted probability. Emphasize that final interpretations should be in terms of probabilities or odds, not log-odds.

This section concludes with the logistic regression model using smoking status as the explanatory variable. Again, statistical software can be used to find the logistic regression model. Once the model is obtained, emphasize the interpretations of the model coefficients and statistical tests, drawing parallels to the work they did in earlier chapters.

Technology

- Exploration 6.2 uses statistical software.

Section 6.3: Multiple Logistic Regression Models

Overview

Section 6.3 covers logistic regression models using multiple categorical and/or quantitative explanatory variables and interactions. Some ideas of model fitting are also covered, including the correlation between observed and predicted probabilities and classification rate.

Approximate class time

- Two to three 50-75 minute class periods
    - One 50-75 minute class period for a guided discussion of fitting a multiple logistic regression model
    - One 50-75 minute class period for a guided discussion of model diagnostics. Example 6.3 can be assigned as a pre-class reading.
    - One 50-75 minute class period for Exploration 6.3. Exploration 6.3 could also be assigned as homework.

Implementation Tips

Example 6.3 is a continuation of Example 6.2. If not shown previously, or as a reminder, provide the mosaic plot of smoking status vs. age group to confirm that age group is a confounding variable in the relationship between 20-year survival and smoking status. Next, show the conditional odds ratios (students could compute one of these) between smoking status and survival within each age group. The idea here, is for students to see that the odds ratios change with age and that in most age groups the odds of survival are smaller for smokers than non-smokers, the opposite of what they saw if the age variable is ignored. This is a nice example of Simpson's Paradox. You can use the discussion as review of "adjusted vs. unadjusted" associations as in earlier chapters.

Present students first, with a scatterplot of the proportion alive at 20-years vs. age for smokers and non-smokers. Use statistical software to fit a two-variable logistic regression model where smoking status and age are the predictors. Then show students the predicted probability curves on the scatterplot, noting the lines are "parallel" as we would expect with the two-variable model (if shifted left/right the curves would overlay). Emphasize the interpretations of model coefficients.

Next, fit the two-variable model with the interaction term and again show students the scatterplot with the predicted probability curves. These curves cross, showing the presence of the interaction in the model. Again spend time on interpretation of the model coefficients, noting that the interaction term is not statistically significant.

Because the interaction is not statistically significant, we could drop it and go with the two-variable model. To check model performance we cannot use R^2 because the response is not quantitative. However, you can remind students that with quantitative variables R^2 = corr(y, $\hat{y}$) and we can look at how well the predicted probabilities agree with the observed proportion alive in each age group,. If there is strong agreement, the correlation between these values should be very close to 1. Alternatively, use the predicted probabilities to determine a predicted "yes, alive" or "no, not alive" and compare how well these predicted 1's and 0's match the actual 1's and 0's, as summarized by the correct classification rate. Typically, if the predicted probability is greater than 0.5, the observation is predicted as a "success." This is also an opportunity to show students a sensitivity analysis by changing the classification rule.

Technology

- Exploration 6.3 uses statistical software.

Chapter 7: Practical Issues

Chapter Overview

Chapter 7 addresses the practical issues that can often be encountered when collecting and analyzing data. Section 7.1 covers basic techniques for evaluating and handling missing data and outliers. Section 7.2 covers the basics of building robust multivariable statistical models.

Section 7.1: Dealing with the Messes Created by Messy Data

Overview

Section 7.1 illustrates how missing data and outliers can impact statistical analyses. It covers some basic techniques for exploring and handling missing data and outliers in statistical analyses.

Approximate class time

- One to two 50-75 minute class periods:
    - One 50-75 minute, in-class period for a guided discussion of outliers and missing data in statistical analyses. (Example 7.1 could also be used as a pre-class reading.)
    - One 50-75 minute, in-class period for Exploration 7.1. (Exploration 7.1 could also be assigned as homework.)

Implementation Tips

Example 7.1 covers the analysis for a research question comparing data from two observational studies, the Framingham Heart Study and a Public Health Screening Study. The goal is to compare levels of Omega-3 between the two studies, so the two sets of data will need to be merged together. The researchers are interested in adjusting for both age and sex (binary). Missing data and outliers are found in both studies.

Begin with a discussion of the impacts of missing data. It’s important to show students both how missing data can impact the effective sample size, and how it can affect the conclusions for a study. Both sets of data have missing data, so one strategy is to work through the investigation of missing data in the Framingham Study, then show students more quickly the results of this exploration within the Public Health Screenings study.

Show students the data table highlighting the rows with missing Omega-3 values along with counts and percentages of the total missing and non-missing. This is an opportunity to discuss with students that “missing” may not be only blank responses, but entries coded as “999”, or “NA”, or “-1”, etc. Because sex and age are covariates of interest in the study, show students how indicator variables can be used to explore differences in age (quantitative) and sex (binary) for respondents with missing and non-missing Omega-3 values. Key ideas here are for students to not only understand the patterns of missing data, but to make some attempt at understanding why the data are missing, and to understand that statistical software will drop the data for all respondents that do not have complete data for the variables in the analysis (e.g., omega-3, age, sex).

The Public Health Screenings study has two large outlying omega-3 values. The presence of these values provides an opportunity to discuss with students the importance of considering whether the outliers are clearly erroneous omega-3 values and should be dropped, or whether they are plausible values for this population and so should be kept in the analysis, but may impact subsequent comparisons of omega-3 between the two studies.

Once the issues of missing data and outliers have been addressed, the comparisons of omega-3 between the two studies adjust for both age and sex. This is a nice way to review previous chapters, giving students an opportunity to practice again with the ideas of confounding and adjusting for both a quantitative and categorical predictor.

Best Practices for dealing with missing data and outliers are provided at the end of Section 7.1.

Technology

- Exploration 7.1 uses a spreadsheet package (e.g., Excel), the Multiple Variables applet, and the One Proportion Inference applet.
- Statistical software may also be used.

Section 7.2: Multiple Regression with Many Explanatory Variables

Overview

Section 7.2 covers multiple regression models with many explanatory variables. Specifically, this section deals with understanding the impact of relationships among explanatory variables in multiple regression models. Best practices for model building are also presented.

Approximate class time

- One to two 50-75 minute class periods if everything is done in class.
    - One 50-75 minute class period for a guided discussion building multiple regression models. Example 7.2 can be assigned as a pre-class reading.
    - One 50-75 minute class period for Exploration 7.2. This exploration could also be assigned as homework.

Implementation Tips

Example 7.2 covers building predictive models in the context of a real estate study. The response variable is the asking price of a home. Sources of explained variation include number of bedrooms, bathrooms, garage stalls, the year the home was built, total rooms, square footage, lot size, age, and number of stories. The example begins with an exploratory data analysis and ends with model building.

Begin with an exploratory data analysis, including univariate graphs and descriptive statistics for each variable, as well as a missing data exploration. The investigation of the missing data is a good review of Section 7.1. Follow-up the univariate analysis by looking at all of the pairwise scatterplots and pairwise correlations. This will allow students to determine which variables appear to be the most strongly related to asking price, which variables are most strongly associated with each other, and also whether any transformations are necessary to linearize the relationship(s) with asking price.

To check for overfitting, we present the method of cross validation. This involves selecting a subset of the data to use for model building, while the remaining data is used for model validation. While the example data set is quite small, this method will give them insight into the types of methods used in much larger data sets of the variety seen in data science.

Using the discovery data, after describing briefly the different methods available for model building, choose one or two models to show as examples. Provide the computer output for each step in the model building process and have students determine what variable should be entered or removed at that step. Be sure to point out how the relationship with asking price of the variables already in the model changes when a new variable is entered (or removed). After arriving at the final model, apply the model to the validation data set. Emphasize that model selection is not an exact science and several models could be reasonable and ideally final selections will also involve subject matter knowledge.

Best Practices for model building are included at the end of Section 7.2

Technology

- Exploration 7.2 uses statistical software.