A Book Chapter Example

Your Name

Contents

Chapter 1

An Introduction to Nonparametric Methods: Schistosomiasis

Using statistics is no substitute for thinking about the problem -Douglas Montgomery[1]

Randomization tests, permutation tests, and bootstrap methods are quickly gaining in popularity as methods for conduct statistical inference. Why? These nonparametric methods require fewer assumptions and provide results that are often more accurate than those from traditional techniques using well-known distributions (such as the normal, t, or F distribution). These methods are based on computer simulations instead of distributional assumptions and thus are particularly useful when the sample data are skewed or if the sample size is small. In addition, nonparametric methods can be extended to other parameters of interest, such as the median or standard deviation, while the well known parametric methods described in introductory statistics courses are often restricted to just inference for the population mean.

We begin this chapter by comparing two treatments for a potentially deadly disease called Schistosomiasis (shis-tuh-soh-mahy-uh-sis). We illustrate the basic concepts behind nonparametric methods by using randomization tests to:

- Provide an intuitive description of statistical inference.
- Conduct a randomization test by hand
- Use software to conduct a randomization test
- Compare one-sided and two-sided hypothesis tests

[1] Douglas Montgomery, Design and Analysis of Experiments, Fifth edition, Wiley, 2003, page 21.

- Making connections between randomization tests and conventional terminology

After working through the schistosomiasis investigation, you will have the opportunity to analyze several other data sets using randomization tests, permutation tests, bootstrap methods, and rank-based nonparametric tests.

1.1 Investigation: Can a New Drug Reduce the Spread of Schistosomiasis?

Schistosomiasis is a disease occurring in humans caused by parasitic flatworms called schistosomes (skis'-tuhsohms). Schistosomiasis affects about 200 million people worldwide and is a serious problem in sub-Saharan Africa, South America, China, and Southeast Asia. The disease can cause death, but more commonly results in chronic and debilitating symptoms, arising primarily from the body's immune reaction to parasite eggs lodged in the liver, spleen, and intestines.

Currently there is one drug, praziquantel (prā'zĭ-kwän'tĕl'), in common use for treatment of schistosomiasis; it is cheap and effective. However many organizations are worried about relying on a single drug to treat a serious disease which affects so many people worldwide. Drug resistance may have prompted a 1990s outbreak in Senegal, where cure rates were low. In 2007, several researchers published work involving a promising drug called K11777 that, in theory, might also treat schistosomiasis.

In this chapter, we will analyze data from this study where the researchers wanted to find out whether K11777 helps to stop schistosome worms from growing. In one phase of the study, 10 female laboratory mice and 10 male laboratory mice were deliberately infected with the schistosome parasite. Seven days after being infected with schistosomiasis, each mouse was given injections every day for 28 days. Within each sex, 5 mice were randomly assigned to a treatment of K11777 whereas the other 5 mice formed a control group injected with an equal volume of plain water. At day 49, the researchers euthanized the mice and measured both the number of eggs and the numbers of worms in the mice livers. Both numbers were expected to be lower if the drug was effective.

Table 1.1 gives the worm count for each mouse. An individual value plot of the data is shown in Figure 1.1. Notice that the treatment group has fewer worms than the control group for both females and males.

Table 1.1: Worm count data for the schistosomiasis study. Treatment is a regimen of K11777 injections from day 7 to day 35. Control is the same regimen, but with a water solution only.

Female		Male	
Treatment	Control	Treatment	Control
1	16	3.0	31
2	10	5.0	26
2	10	9.0	28
10	7	10.0	13
7	17	6.0	47
Mean 4.4	**12**	**6.6**	**29**

NOTE There is a difference between individual value plots and dotplots. In dotplots (such as Figures 1.3 and 1.4 shown later in this chapter), each observation is represented by a dot along a number line (x-axis). When values are close or the same, the dots are stacked. Dotplots can be used in place of histograms when the sample size is small. Individual value plots, as shown in Figure 1.1, are used to simultaneously display each observation for multiple groups. They can be used instead of boxplots to identify outliers and distribution shape, especially when there are relatively few observations.

Activity: *Describing the Data*

1. Use Figure 1.1 to visually compare the number of worms for the treatment and control groups for both the male and the female mice. Does each of the four groups appear to have a similar center and a similar spread? Are there any outliers (extreme observations that don't seem to fit with the rest of the data)?
2. Calculate appropriate summary statistics (e.g., the median, mean, standard deviation, and range) for each of the four groups. For the female mice, calculate the difference between the treatment and control group means. Do the same for the male mice.

The descriptive analysis in Questions 1 and 2 points to a positive treatment effect: K11777 appears to have

reduced the number of parasitic worms in this sample. But descriptive analysis is usually only the first step in ascertaining whether an effect is real; we often conduct a significance test or create a confidence interval to determine if chance alone could explain the effect.

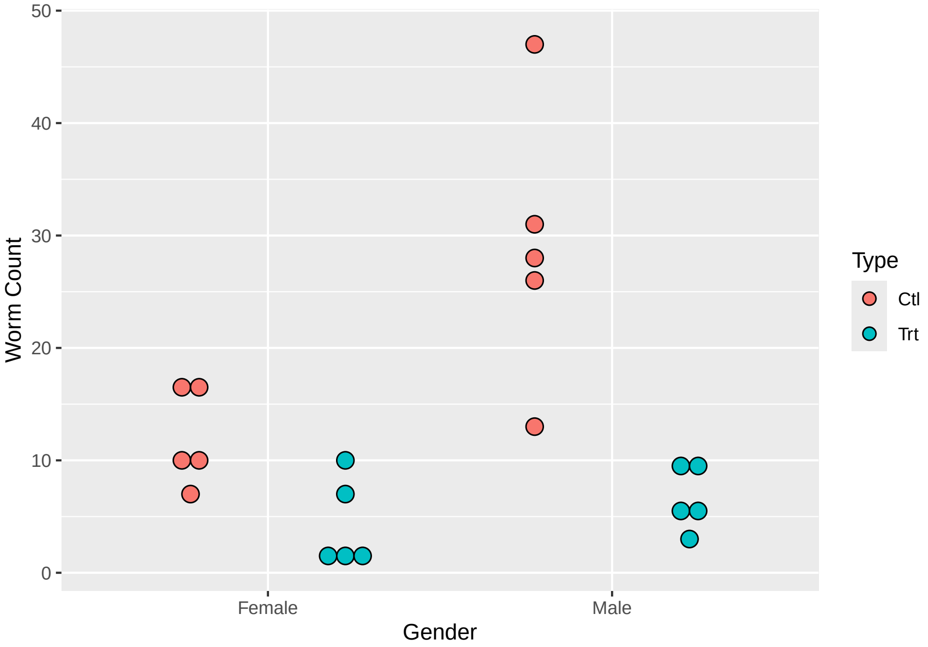

Figure 1.1: Individual value plot of the worm count data

Most introductory statistics courses focus on hypothesis tests that involve using a normal, t-, chi-square or F-distribution to calculate the p-value. These tests are often based on the central limit theorem. In the

schistosomiasis study, there are only five observations in each group. This is a much smaller sample size than is recommended for the central limit theorem, especially given that Figure 1.1 indicates that the data may not be normally distributed. Since we cannot be confident that the sample averages are normally distributed, we will use a distribution-free test, also called a nonparametric test. Such tests do not require the distribution of our sample statistic to have any specific form and are often useful in studies with very small sample sizes.

MATHEMATICAL NOTE: For any population with mean m and finite standard deviation s, the central limit theorem states that the sample mean x from an independent and identically distributed sample tends to follow the normal distribution if the sample size is large enough. The mean of x is the same as the population mean, m, while the standard deviation of x is s/1n, where n is the sample size.

We will use a form of nonparametric statistical inference known as a randomization hypothesis test to analyze the data from the schistosomiasis study. **Ran-**

domization hypothesis tests are significance tests that simulate the random allocation of units to treatments many times in order to determine the likelihood of observing an outcome at least as extreme as the one found in the actual study.

Key Concept: Parametric tests (such as z-tests, t-tests, or F-tests) assume that data come from a population that follows a probability distribution or use the central limit theorem to make inferences about a population. **Nonparametric tests** (such as randomization tests) do not require assumptions about the distribution of the population or large sample sizes in order to make inferences about a population.

We will introduce the basic concepts of randomization tests in a setting where units (mice in this example) are randomly allocated to a treatment or control group. Using a significance test, we will decide if an observed treatment effect (the observed difference between the mean responses in the treatment and control) is "real" or if "random chance alone" could plausibly explain the observed effect. The null hypothesis states that "random chance alone" is the reason for the observed effect. In this initial discussion, the alternative hypothesis will be onesided because we want to show that the true treatment mean (μtreatment) is less than the true control mean (μcontrol). Later, we will expand the discussion to consider modifications needed to deal with two-sided alternatives.

1.2 Statistical Inference Through a Randomization Test

Whether they take the form of significance tests or confidence intervals, inferential procedures rest on the fundamental question for inference: "What would happen if we did this many times?" Let's unpack this question in the context of the female mice in the schistosomiasis study. We observed a difference in means of 7.6 = 12.00 - 4.40 worms between control and treatment groups. While we expect that this large difference reflects the effectiveness of the drug, it is possible that chance alone could explain this difference. This "chance alone" position is usually called the null hypothesis and includes the following assumptions:

- The number of parasitic worms found in the liver naturally varies from mouse to mouse.
- Whether or not the drug is effective, there clearly is variability in the responses of mice to the infestation of schistosomes.

- Each group exhibits this variability, and even if the drug is not effective, some mice do better than others.
- The only explanation for the observed difference of 7.6 worms in the means is that the random allocation randomly placed mice with larger numbers of worms in the control group and mice with smaller numbers of worms in the treatment group.

In this study, the null hypothesis is that the treatment has no effect on the average worm count, and it

is denoted as | H_0: μcontrol = μtreatment Another way to write this null hypothesis is H_0: the treatment has no effect on average worm count

The research hypothesis (the treatment causes a reduction in the average worm count) is called the alternative hypothesis and is denoted H_a (or H_1). For example, H_a: mcontrol 7 mtreatment Another way to write this alternative hypothesis is Ha: the treatment reduces the average worm count Alternative hypotheses can be "one-sided, greater than" (as in this investigation), "one-sided, less-than" (the treatment causes an increase in worm count), or "two-sided" (the treatment mean is different, in one direction or the other, from the control mean). We chose to test a one-sided hypothesis because there is a clear research interest in one direction. In other words, we will take action (start using the drug) only if we can show that K11777 reduces the worm count.

Key Concept: The fundamental question for inference: Every statistical inference procedure (parametric or nonparametric) is based on the question "How does what we observed in our data compare to what would happen if the null hypothesis were actually true and we repeated the process many times?" For a randomization test comparing responses for two groups, this question becomes "How does the observed difference between groups compare to what would happen if the treatments actually had no effect on the individual responses and we repeated the random allocation of individuals to groups many times?"

Activity: *Conducting a Randomization Test by Hand*

3. To get a feel for the concept of a p-value, write each of the female worm counts on an index card. Shuffle the 10 index cards, and then draw five cards at random (without replacement). Call these five cards the treatment group and the five remaining cards the control group. Under the null hypothesis (i.e. the treatment has no effect on worm counts), this allocation mimics precisely what actually happened in our experiment, since the only cause of group differences is the random allocation. | Calculate the mean of the five cards representing the treatment group and the mean of the five cards representing the control group. Then find the difference between the control and treatment group means that you obtained in your allocation. To be consistent, take the control group mean minus the treatment group mean. Your work should look similar to the following simulation:

[[[Fig_CT]]]

4. If you were to do another random allocation, would you get the same difference in means? Explain.

5. Now, perform nine more random allocations, each time computing and writing down the difference in mean worm count between the control group and the treatment group. Make a dotplot of the 10 differences. What proportion of these differences are 7.6 or larger?

6. If you performed the simulation many times, would you expect a large percentage of the simulations to result in a mean difference greater than 7.6? Explain.

The reasoning in the previous activity leads us to the randomization test and an interpretation of the

fundamental question for inference. The fundamental question for this context is as follows: "If the null hypothesis were actually true and we randomly allocated our 10 mice to treatment and control groups many times, what proportion of the time would the observed difference in means be as big as or bigger than 7.6?" This long-run proportion is a probability that statisticians call the **p-value** of the randomization test. The p-values for most randomization tests are found through simulations. Despite the fact that simulations do not give exact p-values, they are usually preferred over the tedious and time-consuming process of listing all possible outcomes. Researchers usually pick a round number such as 10,000 repetitions of the simulation and approximate the p-value accordingly. Since this p-value is an approximation, it is often referred to as the **empirical p-value**.

Key Concept: Assuming that nothing except the random allocation process is creating group differences, the p-value of a randomization test is the probability of obtaining a group difference as large as or larger than the group difference actually observed in the experiment.

Key Concept:

The calculation of an empirical p-value requires these steps:

- Repeat the random allocation process a number of times (N times).
- Record, each time, whether or not the group difference exceeds or is the same as the one observed in the actual experiment (let X be the number of times the group difference exceeds or is the same as the one observed).
- Compute X/N to get the p-value, the proportion of times the difference exceeds or is the same as the observed difference.

NOTE: Many researchers include the observed value as one of the possible outcomes. In this case, N = 9999 iterations are typically used and the p-value is calculated as (X + 1)/(9999 + 1). The results are very similar whether X/10,000 or (X + 1)/(9999 + 1) is used. Including the observed value as one of the possible allocations is a more conservative approach and protects against getting a p-value of 0. Our observation from the actual experiment provides evidence that the true p-value is greater than zero.

1.3 Performing a Randomization Test Using a Computer Simulation

While physical simulations (such as the index cards activity) help us understand the process of computing an empirical p-value, using computer software is a much more efficient way of producing an empirical p-value based on a large number of iterations. If you are simulating 10 random allocations, it is just as easy to use index cards as a computer. However, the advantage of a computer simulation is that 10,000 random allocations can be conducted in almost the same amount of time it takes to simulate 10 allocations. In the following steps, you will develop a program to calculate an empirical p-value.

Activity: *Using Computer Simulations to Conduct a Hypothesis Test*

7. Use the technology instructions provided on the CD to insert the schistosomiasis data into a statistical software package and randomly allocate each of the 10 female worm counts to either the treatment or the control group.

8. Take the control group average minus the K11777 treatment group average.

9. Use the instructions to write a program, function, or macro to repeat the process 10,000 times. Count the number of simulations where the difference between the group averages (control minus K11777) is greater than or equal to 7.6, divide that count by 10,000, and report the resulting empirical p-value.

10. Create a histogram of the 10,000 simulated differences between group means and comment on the shape of the histogram. This histogram, created from simulations of a randomization test, is called an empirical randomization distribution. This distribution describes the frequency of each observed difference (between the control and treatment means) when the null hypothesis is true.

11. Based on your results in Questions 9 and 10 and assuming the null hypothesis is true, about how frequently do you think you would obtain a mean difference as large as or larger than 7.6 by random allocation alone?

12. Does your answer to Question 11 lead you to believe the "chance alone" position (i.e., the null hypothesis that the mean worm count is the same for both the treatment and the control), or does it lead you to believe that K11777 has a positive inhibitory effect on the schistosome worm in female mice? Explain.

Figure 1.2 shows a histogram resulting from the previous activity. A computer simulation of Question 9

resulted in a p-value of 281/10,000 = 0.0281. This result shows that random allocation alone would produce a mean group difference as large as or larger than 7.6 only about 3% of the time, suggesting that something other than chance is needed to explain the difference in group means. Since the only other distinction between the groups is the presence or absence of treatment, we can conclude that the treatment causes a reduction in worm counts.

We conducted four more simulations, each with 10,000 iterations, which resulted in p-values of 0.0272,

0.0282, 0.0268, and 0.0285. When the number of iterations is large, the empirical randomization distribution (such as the histogram created in Question 10) provides a precise estimate of the likelihood of all possible values of the difference between the control and treatment means. Thus, when the number of iterations is large, well-designed simulation studies result in empirical p-values that are fairly accurate. The larger the number of iterations (i.e., randomizations) within a simulation study, the more precise the p-value is.

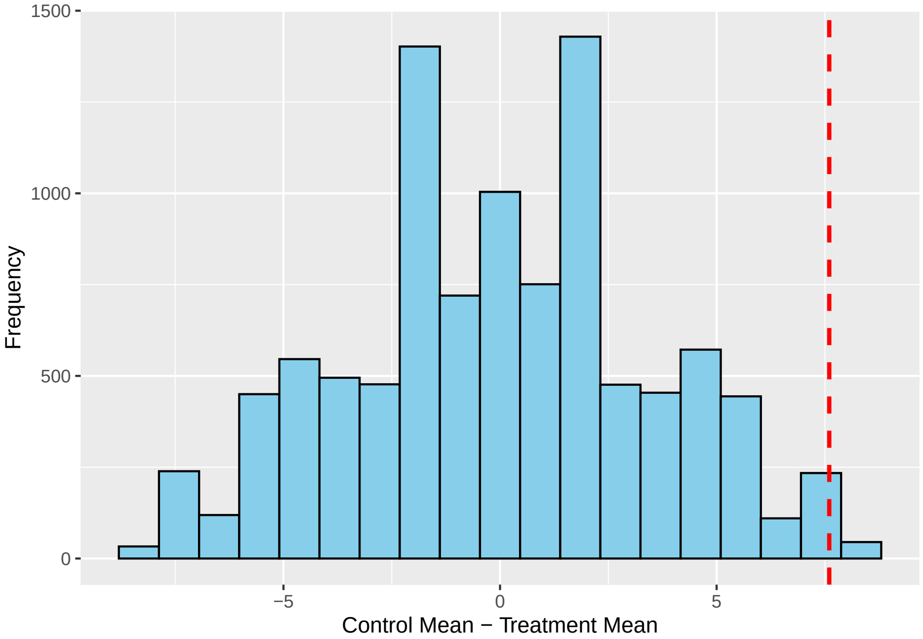

Figure 1.2: Histogram showing the results of a schistosomiasis simulation study. In this simulation, 281 out of 10,000 resulted in a difference greater than or equal to 7.6.

Because the sample sizes in the schistosomiasis study are small, it is possible to apply mathematical

methods to obtain an **exact p-value** for this randomization test. An exact p-value can be calculated by writing down the set of all possibilities (assuming each possible outcome is equally likely under the null hypothesis) and then calculating the proportion of the set for which the difference is at least as large as the observed difference. In the schistosomiasis study, this requires listing every possible combination in which five of the 10 female mice can be allocated to the treatment (and the other five assigned to the control). There are 252 possible combinations. For each of these combinations, the difference between the treatment and control means is then calculated. The exact p-value is the proportion of times in which the difference in the means is at least as large as the observed difference of 7.6 worms. Of these 252 combinations, six have a mean difference of 7.6 and one has a mean difference greater than 7.6 (namely 8.8). Since all 252 of these random allocations are equally likely, the exact p-value in this example is $7/252 = 0.0278$. However, most real studies are too large to list all possible samples. Randomization tests are almost always adequate, providing approximate p-values that are close enough to the true p-value.

CAUTION: Conducting a two-sample t-test on the female mice provides a p-value of 0.011. This p-value of 0.011 is accurate only if the observed test statistic (i.e., the difference between means) follows appropriate assumptions about the distribution. Figure 1.2 demonstrates that the distributional assumptions are violated. While the randomization test provides an approximate p-value "close to 0.0278," it provides a much better estimate of the exact p-value than does the two-sample t-test. Note that each of the five simulations listed gave a p-value closer to the exact p-value than the one given by the two-sample t-test. *Be careful not to trust a p-value provided by statistical software unless you are certain the appropriate assumptions are met.*

Key Concept: The larger the number of randomizations within a simulation study, the more precise the p-value is. When sample sizes are small or sample data clearly are not normal, a p-value derived from a randomization test with 10,000 randomizations is typically more accurate than a p-value calculated from a parametric test (such as the t -test).

Sometimes we have some threshold p-value at or below which we will reject the null hypothesis and

conclude in favor of the alternative. This threshold value is called a significance level and is usually denoted by the Greek letter alpha (α). Common values are $\alpha = 0.05$ and $\alpha = 0.01$, but the value will depend heavily on context and on the researcher's assessment of the acceptable risk of stating an incorrect conclusion. When the study's p-value is less than or equal to this significance level, we state that the results are statistically significant at level A. If you see the phrase "statistically significant" without a specification of α the writer is most likely assuming $\alpha = 0.05$, for reasons of history and convention alone. However, it is best to show the p-value instead of simply stating a result is significant at a particular α-level.

1.4 Two-Sided Tests

The direction of the alternative hypothesis is derived from the research hypothesis. In this K11777 study, we enter the study expecting a reduction in worm counts and hoping the data will bear out this expectation. It is our expectation, hope, or interest that drives the alternative hypothesis and the randomization calculation. Occasionally, we enter a study without a firm direction in mind for the alternative, in which case we use a two-sided alternative. Furthermore, even if we hope that the new treatment will be better than the old treatment or better than a control, we might be wrong—it may be that the new treatment is actually worse than the old treatment or even harmful (worse than the control). Some statisticians argue that a conservative objective approach is to always consider the two-sided alternative. For a **two-sided test**, the p-value must take into account extreme values of the test statistic in either direction (no matter which direction we actually observe in our sample data)

Key Concept: The direction of the alternative hypothesis does not depend on the sample data, but instead is determined by the research hypothesis before the data are collected.

We will now make our definition of the p-value more general to allow for a wider variety of significance

testing situations. The **p-value** is the probability of observing a group difference as extreme as or more extreme than the group difference actually observed in the sample data, assuming that there is nothing creating group differences except the random allocation process.

Activity: *A Two-Sided Hypothesis Test*

13. Run the simulation study again to find the empirical p-value for a two-sided hypothesis test to determine if there is a difference between the treatment and control group means for female mice.
14. Is the number of simulations resulting in a difference greater than or equal to 7.6 identical to the number of simulations resulting in a difference less than or equal to -7.6? Explain why these two values are likely to be close but not identical.
15. Explain why you expect the p-value for the two-sided alternative to be about double that for the onesided alternative. Hint: You may want to

look at Figure 1.2

16. Using the two-sided alternative hypothesis, the two-sample t-test provides a p-value of 0.022.[2] This p-value would provide strong evidence for rejecting the assumption that there is no difference between the treatment and the control (null hypothesis). However, this p-value should not be used to draw conclusions about this study. Explain why.

For the above study, a simulation involving 100,000 iterations provided an empirical p-value of 0.0554.

Again, because this particular data set is small, all 252 possible random allocations can be listed to find that the exact two-sided p-value is 14/252 = 0.0556.

1.5 What Can We Conclude from the Schistosomiasis Study?

The key question in this study is whether K11777 will reduce the spread of a common and potentially deadly disease. The result that you calculated from the one-sided randomization hypothesis test should have been close to the exact p-value of 0.0278. This small p-value allows you to reject the null hypothesis and conclude that the worm counts are lower in the female treatment group than in the female control group. In every study, it is important to consider how random allocation and random sampling impact the conclusions.

Random allocation: The schistosomiasis study was an **experiment** because the units (female mice)

were randomly allocated to treatment or control groups. To the best of our knowledge this experiment controlled for any outside influences and allows us to state that there is a cause and effect relationship between the treatment and response. Therefore, we can conclude that K11777 did cause a reduction in the average number of schistosome parasites in these female mice.

Random sampling: Mice for this type of study are typically ordered from a facility that breeds and raises lab

mice. It is possible that the mice in this study were biologically related or were exposed to something that caused their response to be different from that of other mice. Similarly, there are risks in simply assuming that male mice have the same response as females, so the end-of-chapter exercises provide an opportunity to conduct a separate test on the male mice. Since our sample of 10 female mice was not selected at random from the population of all mice, we should question whether the results from this study hold for all mice.

More importantly, the results have not shown that this new drug will have the same impact on humans as it does on mice. In addition, even though we found

[2] When we do not assume equal variances Minitab uses 7 degrees of freedom providing a p-value of 0.022 while R uses 7.929 degrees of freedom resulting in a p-value of 0.0194.

that K11777 does cause a reduction in worm counts, we did not specifically show that it will reduce the spread of the disease. Is the disease less deadly if only two worms are in the body instead of 10? Statistical consultants aren't typically expected to know the answers to these theoretical, biological, or medical types of questions, but they should ask questions to ensure that the study conclusions match the hypothesis that was tested. In most cases, drug tests require multiple levels of studies to ensure that the drug is safe and to show that the results are consistent across the entire population of interest. While this study is very promising, much more work is needed before we can conclude that K11777 can reduce the spread of schistosomiasis in humans.

A Closer Look: Nonparametric Methods

1.6 Permutation Tests versus Randomization Tests

The random allocation of experimental units (e.g., mice) to groups provides the basis for statistical inference in a randomized comparative experiment. In the schistosomiasis K11777 treatment study, we used a significance test to ascertain whether cause and effect was at work. In the context of the random allocation study design, we called our significance test a randomization test. | In **observational studies**, subjects are not randomly allocated to groups. In this context, we apply the same inferential procedures as in the previous experiment, but we commonly call the significance test a **permutation test** rather than a randomization test.[3] More importantly, in observational studies, the results of the test cannot typically be used to claim cause and effect; a researcher should exhibit more caution in the interpretation of results.

NOTE: The permutation test does not require that the data (or the sampling distribution) follow a normal distribution. However, the null hypothesis in a permutation test assumes that samples are taken from two populations that are similar. So, for example, if the two population variances are very different, the p-value of a permutation test may not be reliable. However, the two-sample t-test (taught in most introductory courses) allows us to assume unequal variances.

Key Concept: Whereas in experiments units are randomly allocated to treatment groups, observational studies do not impose a treatment on a unit. Because the ran-

[3] This text defines a randomization test as a permutation test that is based on random allocation. Some statisticians do not distinguish between permutation tests and randomization tests. They call simulation studies permutation tests, whether they are based on observational studies or experiments.

dom allocation process protects against potential biases caused by extraneous variables, experiments are often used to show causation.

Age Discrimination Study

Westvaco is a company that produces paper products. In 1991, Robert Martin was working in the engineering department of the company's envelope division when he was laid off in Round 2 of several rounds of layoffs by the company.3 He sued the company, claiming to be the victim of age discrimination. The ages of the 10 workers involved in Round 2 were: 25, 33, 35, 38, 48, 55, 55, 55, 56, and 64. The ages of the three people laid off were 55, 55, and 64.

Figure 1.3 shows a comparative dotplot for age by layoff category. This dotplot gives the impression that

Robert Martin may have a case: It appears as if older workers were more likely to be laid off. But we know enough about variability to be cautious.

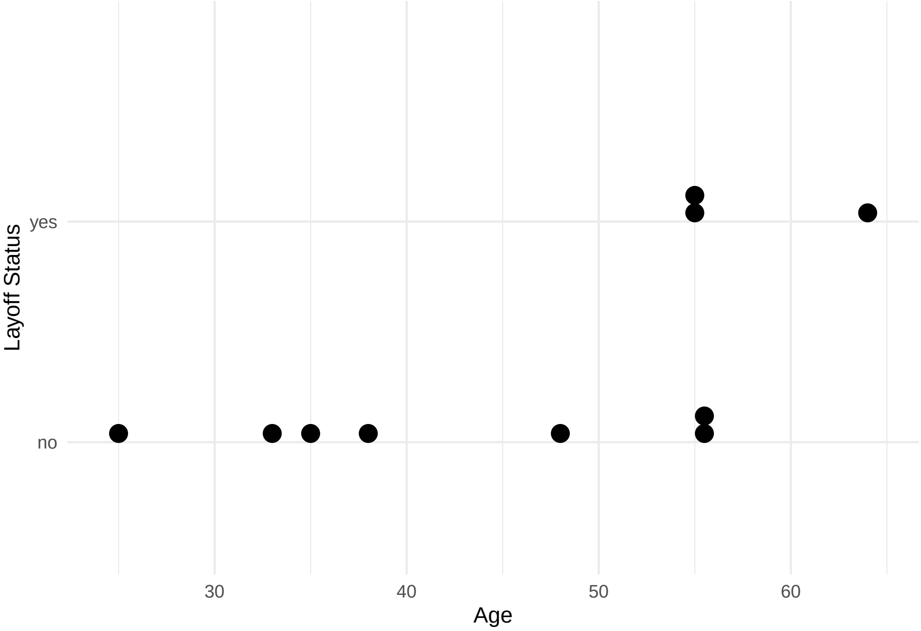

Figure 1.3: Dotplot of age in years of worker versus layoff (whether he or she was laid off)

Extended Activity: *Is There Evidence of Age Discrimination?*

Data set: Age 17. Conduct a permutation test to determine whether the observed difference between means is likely to occur just by chance. Use Age as the response variable and Layoff as the explanatory variable. Here we are interested in only a one-sided hypothesis test to determine if the mean age of people who were laid off is higher than the mean age of people who were not laid off.

18. Modify the program/macro you created in Question 17 to conduct a one-sided hypothesis test to determine if the median age of people who were laid off is higher than the median age of people who were not laid off. Report the p-value and compare your results to those in Question 17.

Since there was no random allocation (i.e., people were not randomly assigned to a layoff group),

statistical significance does not give us the right to assert that greater age is *causing* a difference in being laid off. The null hypothesis in this context becomes "The observed difference could be explained as if by random allocation alone." That is, we proceed as any practicing social scientist must when working with observational data. We "imagine" an experiment in which workers are randomly allocated to a layoff group and then determine if the observed average difference between the ages of laid-off workers and those not laid off is significantly larger than would be expected to occur by chance in a randomized comparative experiment. | While age could be the cause for the difference—hence proving an allegation of age discrimination— there are many other possibilities (i.e., extraneous variables), such as the educational levels of the workers, their competence to do the job, and ratings on past performance evaluations. Rejecting the "as if by random allocation" hypothesis in the nonrandomized context can be a useful step toward establishing causality; however, it cannot establish causality unless the extraneous variables have been properly accounted for. | In the actual court case, data from all three rounds of layoffs were statistically analyzed. The analysis showed some evidence that older people were more likely to be laid off; however, Robert Martin ended up settling out of court.

1.7 Permutation and Randomization Tests for Matched Pairs Designs

The ideas developed in this chapter can be extended to other study designs, such as a basic two-variable design called a matched pairs design. In a matched pairs design, each experimental unit provides both measurements in a study with two treatments (one of which could be a control). Conversely, in the completely randomized situation of the schistosomiasis K11777 treatment study, half the units were assigned to control and half to treatment; no mouse received both treatments.

Music and Relaxation

Grinnell College students Anne Tillema and Anna Tekippe conducted an experiment to study the effect of music on a person's level of relaxation. They hypothesized that fast songs would increase pulse rate more than slow songs. The file called Music contains the data from their experiment. They decided to use a person's pulse rate as an operational definition of the person's level of relaxation and to compare pulse rates for two selections of music: a fast song and a slow song. For the fast song they chose "Beyond" by Nine Inch Nails, and for the slow song they chose Rachmaninoff's "Vocalise." They recruited 28 student subjects for the experiment.

Anne and Anna came up with the following experimental design. Their fundamental question

involved two treatments: (1) listening to the fast song and (2) listening to the slow song. They could have randomly allocated 14 subjects to hear the fast song and 14 subjects to hear the slow song, but their more efficient approach was to have each subject provide both measurements. That is, each subject listened to both songs, giving rise to two data values for each subject, called a matched pairs. Randomization came into play when it was decided by a coin flip whether each subject would listen first to the fast song or the slow song.

NOTE: There are several uses of randomness mentioned in this chapter. The emphasis of this chapter is on the use of **randomization tests** for statistical inference. Most introductory statistics courses discuss random **sampling** from a population, which allows the results of a specific study to be generalized to a larger population. In experiments, units are **randomly allocated to groups** which allows researchers to make statements about causation. In this example, Anne and Anna **randomize the order** to prescribe two conditions on a single subject.

Specifically, as determined by coin flips, half the subjects experienced the following procedure:

[one minute of rest; measure pulse (prepulse)] > [listen to fast song for 2 minutes; measure pulse for second minute (fast song pulse)] > [rest for one minute] > [listen to slow song for 2 minutes; measure pulse for second minute (slow song pulse)].

The other half experienced the procedure the same way except that they heard the slow song first and

the fast song second. | Each subject gives us two measurements of interest for analysis: (1) fast song pulse minus prepulse and (2) slow song pulse minus prepulse. In the data file, these two measurements are called `Fastdiff` and `Slowdiff`, respectively.

Figure 1.4 shows a dotplot of the 28 `Fastdiff`-minus-`Slowdiff` values. Notice that positive numbers

predominate and the mean difference is 1.857 beats per minute, both suggesting that the fast song does indeed heighten response (pulse rate) more than the slow song. We need to confirm this suspicion with a randomization test.

To perform a randomization test, we mimic the randomization procedure of the study design. Here,

the randomization determined the order in which the subject heard the songs, so randomization is applied to the two measurements of interest for each subject. To compute a p-value, we determine how frequently we would obtain an observed difference as large as or larger than 1.857.

```
#> Bin width defaults to 1/30 of the range of the data. Pick
#> better value with `binwidth`.
```

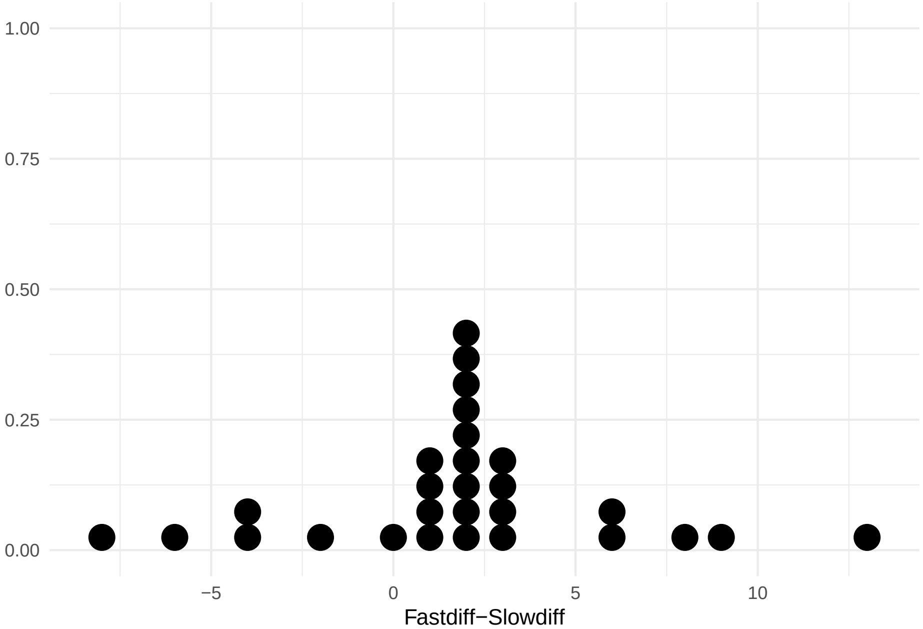

Figure 1.4: Dotplot of the difference in pulse rates for each of the 28 subjects.

Extended Activity: *Testing the Effect of Music on Relaxation*

Data set: `Music`

19. Before they looked at the data, Anne and Anna decided to use a one-sided

test to see whether fast music increased pulse rate more than slow music. Why is it important to determine the direction of the test before looking at the data?

20. Create a simulation to test the Music data. Use the technology instructions provided to randomly multiply a 1 or a -1 by each observed difference. This randomly assigns an order ('Fastdiff - Slowdiff' or 'Slowdiff - Fastdiff'). Then, for each iteration, calculate the mean difference. The p-value is the proportion of times your simulation found a mean difference greater than or equal to 1.857.

    (a) Create a histogram of the mean differences. Mark the area on the histogram that represents your p-value.

    (b) Use the p-value to state your conclusions in the context of the problem. Address random allocation and random sampling (or lack of either) when stating your conclusions.

CAUTION: The type of randomization in Question 20 does not account for extraneous variables such as a great love for Nine Inch Nails on the part of some students or complete boredom with this band on the part of others (i.e., "musical taste" is a possible confounder that randomizing the order of listening cannot randomize away). There will always be a caveat in this type of study, since we are rather crudely letting one Nine Inch Nails song "represent" fast songs.

1.8 The Bootstrap Distribution

Bootstrapping is another simulation technique that is commonly used to develop confidence intervals and hypothesis tests. Bootstrap techniques are useful because they generalize to situations where traditional methods based on the normal distribution cannot be applied. For example, they can be used to create confidence intervals and hypothesis tests for any parameter of interest, such as a median, ratio, or standard deviation. Bootstrap methods differ from previously discussed techniques in that they sample **with replacement** (randomly draw an observation from the original sample and put the observation back before drawing the next observation). | Permutation tests, randomization tests, and bootstrapping are often called **resampling techniques** because, instead of collecting many different samples from a population, we take repeated samples (called resamples) from just one random sample.

Extended Activity: *Creating a Sampling Distribution and a Bootstrap Distribution*

Data set: `ChiSq`

21. The file ChiSq contains data from a highly skewed population (with mean 0.9744 and standard deviation 1.3153). a. Take 1000 simple random samples of size 40 and calculate each mean (x). Plot the histogram of the 1000 sample means. The distribution of sample means is called the sampling distribution. b. What does the central limit theorem tell us about the shape, center, and spread of the sampling distribution in this example? c. Calculate the mean and standard deviation of the sampling distribution in Part A. Does the sampling distribution match what you would expect from the central limit theorem? Explain.

22. Take one simple random sample of size 40 from the ChiSq data. a. Take 1000 resamples (1000 samples of 40 observations with replacement from the one simple random sample). b. Calculate the mean of each resample (x*) and plot the histogram of the 1000 resample means. This distribution of resample means is called the bootstrap distribution. c. Compare the shape, center, and spread of the simulated histograms from Part B and Question 21 Part A. Are they similar?

23. Instead of using the sample mean, create a sampling distribution and bootstrap distribution of the standard deviation of the ChiSq data using a sample size of 40. Compare the shape, center, and spread of the simulated histograms and compare the mean and standard deviation of the distributions.

Key Concept: The bootstrap method takes one simple random sample of size n from a population. Then many resamples (with replacement) are taken from the original simple random sample. Each resample is the same size as the original random sample. The statistic of interest is calculated from each resample and used to create a bootstrap distribution.

In many real-world situations, the process used in Question 21 is not practical because collecting more

than one simple random sample is too expensive or time consuming. While the approach in Question 22 is computer intensive, it is simple and convenient since it uses only one simple random sample. The key idea behind bootstrap methods is the assumption that the original sample represents the population, so resamples from the one simple random sample can be used to represent samples from the population, as is done in Question 22. Thus, the bootstrap distribution provides an approximation of the sampling distribution.

Most traditional methods of statistical inference involve collecting one sample and calculating the sample

mean. Then, based on the central limit theorem, assumptions are made about the shape and spread of the sampling distribution. In Question 22 we used one sample to calculate the sample mean and then used the bootstrap distribution to estimate the shape and spread of the sampling distribution.

The central limit theorem tells us about the shape and spread of the sample mean. A key advantage of

the bootstrap distribution is that it works for any parameter of interest. Thus, the bootstrap distribution can be used to estimate the shape and spread for any sampling distribution of interest.

CAUTION: When sample sizes are small, one simple random sample may not represent the population very well. However, with larger sample sizes, the bootstrap distribution does represent the sampling distribution.

Figure 1.5 shows the sampling distribution and the bootstrap distribution when a sample size of 10 is used to estimate the mean of the `ChiSq` data. Notice that the spreads for both histograms are

roughly equivalent. The central limit theorem tells us that the standard deviation of the sampling distribution (the distribution of $\bar{x}$) should be $\sigma/\sqrt{n} = 1.3153/\sqrt{10} = 0.4159$. The standard deviation of the bootstrap distribution is 0.4541, which is a reasonable estimate of the standard deviation of the sampling distribution. In addition, both graphs have similar, right-skewed shapes. The strength of the bootstrap method is that it provides accurate estimates of the shape and spread of the sampling distribution. In general, histograms from the bootstrap distribution will have a similar shape and spread as histograms from the sampling distribution.

```
#> [1] "1.465279886"
```

[[[Fig1,5]]]

The bootstrap method does not improve our estimate of the population mean. The mean of the sampling distribution in Question 21 will typically be very close to the population mean. But the mean of the bootstrap distribution in Question 22 typically will not be as accurate, because it is based on only one simple random sample. Ideally, we would like to know how close the statistic from our original sample is to the population parameter. A statistic is biased if it is not centered at the value of the population parameter. We can use the bootstrap distribution to estimate the bias of a statistic. The difference between the original sample mean and the bootstrap mean is called the **bootstrap estimate of bias**.

Key Concept: The estimate of the mean (or any parameter of interest) provided by the bootstrap distribution is not any better than the estimate provided by the observed statistic from the original simple random sample. However, the shape and spread of the bootstrap distribution will be similar to the shape and spread of the sampling distribution. The bootstrap technique can be used to estimate sampling distribution shapes and standard deviations that cannot be calculated theoretically.

1.9 Using Bootstrap Methods to Create Confidence Intervals

A **confidence interval** gives a range of plausible values for some parameter. This is a range of values surrounding an observed estimate of the parameter—an estimate based on the data. To this range of values we attach a level of confidence that the true parameter lies in the range. An alpha-level, α, is often used to specify the level of confidence. For example, when $\alpha = 0.05$, we have a 100(1 - α), = 95% confidence level. Thus, a 100(1 - α), confidence interval gives an estimate of where we think the parameter is and how precisely we have it pinned down.

1.10 *Bootstrap t Confidence Intervals{-}

If the bootstrap distribution appears to be approximately normal, it is typically safe to assume that a t-distribution can be used to calculate a 100(1 - α), confidence interval for μ, often called a bootstrap t confidence interval:

$$\bar{x} \pm t^* \left(S^*\right) \tag{1.1}$$

where S^* is the standard deviation of the bootstrap distribution and t^* is the critical value of the t-distribution with n - 1 degrees of freedom.

The one simple random sample of size n = 10 used to create the bootstrap distribution in Figure 1.5b has a mean of $\bar{x} = 1.238$ and a standard deviation of s = 1.490. The bootstrap distribution in Figure 1.5b has a mean of $\bar{x}^* = 1.249$ and a standard deviation of $S^* = 0.4541$. Notice that Formula (1.1) uses the mean from the original sample but uses the bootstrap distribution to estimate the spread. If we *incorrectly assume* that the sampling distribution in Figure 1.5 is normal, a 95% bootstrap t confidence interval for μ is given by

$$\bar{x} \pm t^* \left(S^*\right) = 1.238 \pm 2.262(0.4541) \tag{1.1}$$

where $t^* = 2.262$ is the critical value corresponding to the 97.5th percentile of the t-distribution with n - 1 = 9 degrees of freedom. Thus, the 95% confidence interval for μ is (0.211, 2.265).

MATHEMATICAL NOTE: The bootstrap t confidence interval is similar to the traditional one-sample t confidence interval. The key difference is that the bootstrap distribution estimates the standard error of the statistic with S* instead of s/$\sqrt{n}$. When the data are not skewed and have no clear outliers, parametric tests are very effective with relatively small sample sizes (10–30 observations may be enough to use the t-distribution). The following formula uses the t-distribution to calculate a 100(1 - α), confidence interval for the mean of a normal population:

$$\bar{x} + t^* \left(\frac{s}{\sqrt{n}}\right) \tag{1.2}$$

where s/$\sqrt{n}$ is the standard error of x and t* is the critical value of the t-distribution with n - 1 degrees of freedom. Using the original sample of size 10 with mean 1.238 and standard deviation 1.490, we find that a 95sample means only when the sampling distribution is approximately normal. If the data are skewed, even sample sizes greater than 30 may not be large enough to make the sampling distribution appear normal.

With skewed data or small sample sizes (if the original data are not normally distributed), parametric methods (which are based on the central limit theorem) are not appropriate. In Figure 1.5 we see that the sampling distribution is skewed to the right. *Thus, with a sample size of 10, neither the traditional onesample t confidence interval nor the bootstrap t confidence interval is reliable in this example.* However, with a sample size of 40, the histograms in Questions 21 and 22 should tend to look somewhat normally distributed.

Bootstrap Percentile Confidence Intervals

Bootstrap percentile confidence intervals are found by calculating the appropriate percentiles of the bootstrap distribution. To find a 100(1 - α) confidence interval, take the $\alpha/2 * 100$ percentile of each tail of the bootstrap distribution. For example, to find a 95% confidence interval for μ, sort all the observations from the bootstrap distribution and find the values that represents the 2.5th and 97.5th percentiles of the bootstrap distribution. The 2.5th percentile of the bootstrap distribution in Figure 1.5b is 0.546, and the 97.5th percentile is 2.282. Thus, a 95% confidence interval for μ is (0.546, 2.282). Notice that the percentile confidence interval is not centered at the sample mean. Since the bootstrap distribution is right skewed, the right side of the confidence interval (2.282 - 1.238 = 1.044) is wider than the left side of the confidence interval (1.238 - 0.546 = 0.692). This lack of symmetry can influence the accuracy of the confidence interval.

Key Concept: A bootstrap percentile confidence interval contains the middle 100(1 - a)If the bootstrap distribution is symmetric and is centered on the observed statistic (i.e., not biased), percentile confidence intervals work well.

When to Use Bootstrap Confidence Intervals

Bootstrap methods are extremely useful when we cannot use theory, such as the central limit theorem, to approximate the sampling distribution. Thus, bootstrap methods can be used to create confidence intervals for essentially any parameter of interest, while the central limit theorem is limited to only a few parameters (such as the population mean).[4] However, bootstrap methods are not always reliable.

Small sample sizes still produce problems for bootstrap methods. When the sample size is small, (1) the sample statistic may not accurately estimate the population parameter, (2) the distribution of sample means is less likely to be symmetric, and (3) the shape and spread of the bootstrap distribution may not accurately represent those of the true sampling distribution.

In addition, bootstrap methods do not work equally well for all parameters. For example, the end-ofchapter

exercises show that bootstrapping often provides unreliable bootstrap distributions for median values because the median of a resample is likely to have only a few possible values. Thus, confidence intervals for medians should be used only with large ($n \geq 100$) sample sizes.

[4] Theoretical methods allow distributional tests for more than just the population mean. However, for purposes of this text it is sufficient to understand that distributional methods tend to be more complicated and are limited to testing only a few parameters that could be of interest.

It is not easy to determine whether bootstrap methods provide appropriate confidence intervals. The bootstrap t and bootstrap percentile confidence intervals are often compared to each other. While the percentile confidence interval tends to be more accurate, neither of the two should be used if the intervals are not relatively

close. If the bootstrap distribution is skewed or biased, other methods should be used to find confidence intervals. More advanced bootstrap methods (such as BCa and tilting confidence intervals) are available that are generally accurate when bias or skewness exists in the bootstrap distribution.[5]

Extended Activity:***Estimating Salaries of Medical Faculty***

Data set: `MedSalaries`. The file `MedSalaries` is a random sample of n = 100 salaries of medical doctors who were teaching at United States universities in 2009.

24. Create a bootstrap distribution of the mean by taking 1000 resamples (with replacement). Create a bootstrap t confidence interval and a bootstrap percentile distribution to estimate the mean salaries.
25. Create a bootstrap distribution of the standard deviation by taking 1000 resamples (with replacement). Create a bootstrap t confidence interval and a bootstrap percentile distribution to estimate the population standard deviation.
26. Use Formula (1.2) to create a 95
27. Explain why Formula (1.2) cannot be used to create a 95

1.11 Relationship Between the Randomization Test and the Two-Sample t-Test

R.A. Fisher, perhaps the preeminent statistician of the 20th century, introduced the randomization test in the context of a two-group randomly allocated experiment in his famous 1935 book, *Design of Experiments*.[6] At that time he acknowledged that the randomization test was not practical because of the computational intensity of the calculation. Clearly, 1935 predates modern computing. Indeed, Efron and Tibshirani describe the permutation test as "a computer-intensive statistical technique that predates computers."[7] Fisher went on to assert that the classical two-sample t-test (for independent samples) approximates the randomization test very well. Ernst cites references to several approximations to the randomization tests using classical and computationally tractable methods that have been published over time.[8]

If you have seen two-sample tests previously, it is likely to have been in the context of what Ernst calls the population model, which he distinguishes from the randomization model. In a **population model**, units are selected at random from one or more populations. Most observational studies are population models. One simple case of a population model involves comparing two separate population means. In this case, we can take two independent simple random samples and use the classic two-sample t-test to make the comparison.

In a ***randomization model**, a fixed number of experimental units are randomly allocated to treatments. Most experiments are randomization models. In randomization models such as the schistosomiasis example,

the two samples are formed from a collection of available experimental units that are randomly divided into two groups. Since there are a fixed number of units, the groups are not completely independent. For example, if one of the 10 male mice had a natural resistance to schistosomiasis and was randomly placed in the treatment group, we would expect the control group to have a slightly higher worm count. Since the two groups are not completely independent, the assumptions of the classic two-sample t-test are violated. Even if the sample sizes in the schistosomiasis study were much larger, the randomization test would be a more appropriate test than the two-sample t-test. However, empirical evidence has shown that the two-sample t-test is a very good approximation to the randomization test when sample sizes are large enough. We are fortunate that, in this age of modern computing, we no longer have to routinely compromise by using the t-test to approximate the randomization test.

Key Concept: Historically, the two-sample t -test was used to approximate the p-value in randomization models because randomization tests were too difficult to compute. However, now that computers can easily simulate random assignment to groups, randomization tests should be used to calculate p-values for randomization models, especially if sample sizes are fairly small.

1.12 Wilcoxon Rank Sum Tests for Two Independent Samples

The **Wilcoxon rank sum test**, also called the two-sample **Mann-Whitney test**, makes inferences about the difference between two populations based on data from two independent random samples. This test ranks observations from two samples by arranging them in order from smallest to largest.

Focusing on ranks instead of the actual observed values allows us to remove

Table 1.2: Randomly selected pitchers and first baseman from 2005 National League baseball teams.

Team	Position	Name	Salary(\$)
Milwaukee Brewers	Pitcher	Obermueller, Wes	342000
Houston Astros	Pitcher	Backe, Brandon	350000
Atlanta Braves	Pitcher	Sosa, Jorge	650000
Atlanta Braves	Pitcher	Thomson, John	4250000
Cincinnati Reds	First Baseman	Casey, Sean	7800000
Arizona Diamondbacks	First Baseman	Green, Shawn	7833333
San Diego Padres	First Baseman	Nevin, Phil	9625000
New York Mets	Pitcher	Glavine, Tom	10765608
Colorado Rockies	First Baseman	Helton, Todd	12600000
Philadelphia Phillies	First Baseman	Thome, Jim	13166667

Table 1.3: Ranking the 10 randomly selected 2005 National League baseball players.

Position	Pr	Pr	Pr	Pr	FB	FB	FB	Pr	FB	FB
Salary	342	350	650	4250	7800	7833	9625	10766	12600	13167
Rank	1	2	3	4	5	6	7	8	9	10

assumptions about the normal distribution. Rank-based tests have been used for many years. However, rank-based methods (discussed in this section and the next section) are much less accurate than methods based on simulations. In general, randomization tests, permutation tests, or bootstrap methods should be used whenever possible.

The following example examines whether pitchers and first basemen who play for National League baseball teams have the same salary distribution. The null and alternative hypotheses are written as,

H_0: the distribution of the salaries is the same for pitchers and first basemen
H_a: the distribution of the salaries is different for pitchers and first basemen

Table 1.2 shows the salaries of five pitchers and five first basemen who were randomly selected from all National League baseball players. Table 1.3 ranks each of the players based on 2005 salaries.

Note that if two players had exactly the same salary, standard practice would be to average the ranks of the tied values.

For the Wilcoxon rank sum test, we define the following terms:

- n_1 is the sample size for the first group (5 for the pitcher group in this example)

- n_2 is the sample size for the second group (5 for the first baseman group in this example)
- N= n_1 + n_2
- W, the Wilcoxon rank sum statistic, is the sum of the ranks in the first group (1 + 2 + 3 + 4 + 8 = 18)

If the two groups are from the same continuous distribution, then W has a mean,

$$\mu_W = \frac{n_1(N+1)}{2} = \frac{5(11)}{2} = 27.5 \tag{1.3}$$

and standard deviation[9]

$$\sigma_W = \sqrt{\frac{n_1 n_2 (N+1)}{12}} = \sqrt{\frac{(5)(5)(11)}{12}} = 4.787 \tag{1.4}$$

If W is far from μ_W, then the Wilcoxon rank sum test rejects the hypothesis that the two populations have identical distributions—that is, rejects H_0 (no difference in distribution of salaries) in favor of H_a (salary distributions are different based on position). The p-value is the probability of observing a sample statistic, W, at least as extreme as the one in our sample. Since 18 is less than the hypothesized mean, 27.5, the p-value for the two-sided test in this example is found by calculating 2 * P(W $\leq$ 18).

MATHEMATICAL NOTE: Computer software such as R, S-plus, or SAS tends to use the exact distribution of W, though Minitab uses a normal approximation for this test. If the data contain ties, the exact distribution for the Wilcoxon rank sum statistic changes and the standard deviation of W should be adjusted. Statistical software will typically detect the ties and use the normal distribution (using the adjusted standard deviation) instead of an exact distribution.[10]

Extended Activity: *Wilcoxon Rank Sum Tests*

Data set: `NLBB Salaries`

25. Using a software package, conduct the Wilcoxon rank sum test to determine if the distribution of salaries is different for pitchers than for first basemen.

26. Find 2 X P(W $\leq$ 18) assuming W N(27.5, 4.787). How does your answer compare to that fromQuestion 25?

Table 1.4: Randomly selected catchers from 2005 National League baseball teams.

Team	Position	Name	Salary(\$)
Pittsburgh Pirates	Catcher	Ross, David	338500
Los Angeles Dodgers	Catcher	Phillips, Jason	339000
Atlanta Braves	Catcher	Perez, Eddie	625000
Washington Nationals	Catcher	Bennett, Gary	750000
Pittsburgh Pirates	Catcher	Santiago, Benito	2150000

27. Use a two-sided two-sample t-test (assume unequal variances) to analyze the data. Are your conclusions the same as in Question 25? Create an individual value plot of the data. Are any distributional assumptions violated? Which test is more appropriate to use for this data set?

At first it may seem somewhat surprising that first basemen tend to make more than pitchers. However, in 2005 there were 19 first basemen and 215 pitchers in the National League. Many pitchers did not play much and got paid a low salary, whereas all 19 first basemen were considered quite valuable to their teams.

1.13 Kruskal-Wallis Test for Two or More Independent Samples

The ***Kruskal-Wallis*** **test** is another popular nonparametric test that is often used to compare two or more independent samples. Like ANOVA, a more common parametric test that will be discussed in later chapters, the Kruskal-Wallis test requires independent random samples from each population. When the data clearly deviate from the normal distribution, the Kruskal-Wallis test will be more likely than a one-way ANOVA to identify true differences in the population. The null and alternative hypotheses for the Kruskal-Wallis test are:

H_0: the distribution of the response variable is the same for all groups H_a: some responses are systematically higher in some groups than in others

The Kruskal-Wallis test is also based on ranks. The ranks are summed for each group, and when these group sums are far apart, we have evidence that the groups are different. While the calculations for the Kruskal-Wallis test statistic are provided here, we suggest using statistical software to conduct this significance test. Continuing the baseball salaries example, Table 1.4 displays salaries of five randomly selected catchers from 2005 National League baseball teams.

For the Kruskal-Wallis test, we define the following terms:

- n_1 is the sample size for the first group (5 for the pitcher group)
- n_2 is the sample size for the second group (5 for the first baseman group)
- n_3 is the sample size for the third group (5 for the catcher group)
- N = n_1 + n_2 + n_3
- R_i is the sum of the ranks for the ith group (R_1 = 35, R_2 = 62, and R_3 = 23)

The Kruskal-Wallis test statistic is calculated as,

$$H = \frac{12}{N(N+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i} - 3(N+1) = \frac{12}{(15)(16)}(\frac{35^2}{5} + \frac{62^2}{5} + \frac{23^2}{5}) - 3(16) = 7.98 \tag{1.5}$$

The exact distribution of H under the null hypothesis depends on each ni, so it is complex and time consuming to calculate. Even most statistical software packages use the chi-square approximation with I - 1 degrees of freedom to obtain p-values (where I is the number of groups).

NOTE: When the chi-square approximation is used, each group should have at least five observations.

Extended Activity: *Kruskal-Wallis Test*

Data set: `NLBB Salaries`

28. Using a software package, run the Kruskal-Wallis test (use all three groups with samples of size 5 per group) to determine if the distribution of salaries differs by position. Create an individual value plot of the data. Do the data look normally distributed in each group?

Mathematical Note: If the spread of each group appears to increase as the center (mean or median) increases, transforming the data—such as by taking the log of each response variable—will make the data appear much more normally distributed. Then parametric techniques can often be used on the transformed data. In the baseball salary example, the data are highly right skewed in at least two groups. While a log transformation on salaries is helpful, there is still not enough evidence that the transformed salaries are normally distributed. Thus, nonparametric methods are likely the most appropriate approach to testing whether there is a difference in the distribution of salaries based on position.

Nonparametric tests based on rank are usually less powerful (less likely to reject the null hypothesis) than the corresponding parametric tests. Thus, you are less likely to identify differences between groups when they really exist. If you are reasonably certain that the assumptions for the parametric procedure are satisfied, a parametric procedure should be used instead of a rank-based nonparametric procedure. Many introductory texts suggest that, in order to conduct a parametric test, you should have a sample size of 15 in each group and no skewed data or outliers.

1.14 Multiple Comparisons

In introductory texts, statistical inference is often described in terms of drawing one random sample, performing one significance test, and then stating appropriate conclusions—analysis done, case closed. However, there are many situations where inference is not that simple. Performing multiple statistical tests on the same data set can create several problems.

Using a significance level of $\alpha = 0.05$ (i.e., rejecting H_0 in favor of the alternative when the p-value is less than or equal to 0.05) helps to ensure that we won't make a wrong decision. In other words, one time out of 20 we expect to incorrectly reject the null hypothesis. But what if we want to do 20 or more tests on the same data set? Does this mean that we're sure to be wrong at least once? And if so, how can we tell which findings are incorrect? The following activities explore how researchers can protect themselves from drawing conclusions from statistical findings that could be the result of random chance.

Extended Activity:*Comparing Car Prices*

29. Open the 'Car1' data set and conduct three two-sided hypothesis tests to determine if there is a difference in price. Compare the means: Pontiac versus Buick (test 1), Cadillac versus Pontiac (test 2), and Cadillac versus Buick (test 3). Provide the p-value for each of these three tests. Which tests have a p-value less than 0.05?

30. Assuming the null hypotheses are true, each of the three tests in Question 29 has a 5% chance of inappropriately rejecting the null hypothesis. However, the probability that at least one of the three tests will inappropriately reject the null hypothesis is 14.26%. Assuming that the null hypothesis is true and that each test is independent, complete the following steps to convince yourself that this probability is correct.

    (a) Each test will either reject (R) or fail to reject (F). List all eight possible outcomes in the table below.

Case	Test_1	Test_2	Test_3	Probability
1	F	F	F	
2	F	F	R	
3	F	R	F	
4				
5				
6				
7				
8				

(b) The probability that each test rejects is $P(R) = 0.05$, and the probability that each test fails to reject is $P(F) = 0.95$. For example, the probability that all three tests fail to reject is $0.95^3 = 0.8574$. The probability that the first two fail to reject and the third does reject is $0.95 \times 0.95 \times 0.05 = 0.0451$. Complete the table. Verify the probabilities sum to 1. The probability that at least one test rejects is $1 - 0.8574 = 0.1426$.

31. Repeat Question 30 using ($\alpha = 0.10$). What is the probability that at least one of the three tests will inappropriately reject the null hypothesis?

32. To compare all four car makes, six hypothesis tests will be needed. List all six null hypotheses. Assuming independence and that the null hypothesis is true, what is the probability that at least one of the six tests will inappropriately reject the null hypothesis at $\alpha = 0.05$?

Extended Activity: *The Least-Significant Differences Method and the Bonferroni Method*

Data set: `Car1`

When the significance level is controlled for each individual test, as was done in Question 29, the process is often called the **least-significant differences method (LSD)**. Notice that using a = 0.05 for all tests has some undesirable properties, especially when a large number of tests being conducted. If 100 independent tests were conducted to compare multiple groups (and there really were no differences), the probability of incorrectly rejecting at least one test would be 1 - 0.95^{100} = 0.994. Thus, using α= 0.05 as a critical value for 100 comparisons will almost always lead us to incorrectly conclude that some results are significantly different.

One technique that is commonly used to address the problem with multiple comparisons is called the ***Bonferroni method**. This technique protects against the probability of false rejection by using a cutoff value of α/K, where K is the number of comparisons. In Question 29, there are three comparisons (i.e.,three hypothesis tests). Thus, a cutoff value of 0.05/3 = 0.01667 should be used. In other words, when there are three comparisons as in Question 29, the Bonferroni method rejects the null hypothesis when the p-value is less than or equal to 0.01667. Using the least-significant differences method ($\alpha = 0.05$), as was done in Question 29, we would conclude that the prices of Buicks and Chevrolets are significantly different, but using the Bonferroni method we would fail to reject in all three tests.

33. Repeat Question 30 using the Bonferroni cutoff value of $0.05/3 = 0.016667$ instead of $\alpha = 0.05$. Find the probability that at least one of the tests rejects.
34. Using all four groups of cars and $\alpha = 0.05$ (cutoff of 0.05/6), do any of the six tests reject the null hypothesis with the Bonferroni method?
35. If there were seven groups, 21 hypothesis tests would be needed to compare all possible pairs. Using $\alpha = 0.05$ and the Bonferroni's method (reject H_0 if the p-value is less than $0.05/21 = 0.00238$) what is the probability that at least one of the tests would reject?

MATHEMATICAL NOTE: Other terms that are commonly discussed with multiple comparisons are **familywise type I error** and ***comparisonwise type I error**. Bonferroni's method is an example of a technique that maintains the familywise type I error. With the familywise type I error 0.05, assuming that there really is no difference between any of the K pairs, there is only a 5differences method is used to maintain a comparisonwise type I error rate: Assuming that a particular null hypothesis test is true, there is a 5

Choosing a Critical Value

The α-level represents the probability of a **type I error**. A type I error can be considered a false alarm: Our hypothesis test has led us to conclude that we have found a significant difference when one does not exist. However, it is important to recognize that it is also possible to make a **type II error**, which means our hypothesis test failed to detect a significant difference when one exists. In essence, a type II error can be thought of as an alarm that failed to go off.

Notice that if the Bonferroni method is used with all six tests, the critical value for each individual test is $0.05/6 = 0.00833$. Thus, this method often fails to detect real differences between groups, leaving us open to a high rate of type II error while protecting us against type I errors.

Neither the least-significant differences nor the Bonferroni method is ideal. Caution should be used with both techniques, and neither technique should be used with numerous comparisons. The key is to recognize the benefits and limitations of each technique and to properly interpret what the results of each technique tell us. Some researchers suggest limiting the number of tests, using both techniques, and letting the reader decide

which conclusions to draw. Both techniques are commonly used when there are fewer than 10 comparisons. However, a researcher should always decide which comparisons to test before looking at the data.

Chapter Summary

This chapter described the basic concepts behind randomization tests, permutation tests, bootstrap methods, and rank-based nonparametric tests. **Parametric tests** (such as z-tests, t-tests or F-tests) assume that data follow a known a probability distribution or use the central limit theorem to make inferences about a population. *__Nonparametric tests__ do not require assumptions about the distribution of the population or the central limit theorem in order to make inferences about a population.

The *__null hypothesis__, denoted H_0, states that in a study nothing is creating group differences except the random allocation process. The research hypothesis is called the **alternative hypothesis** and is denoted H_a (or H_1). The p-value is the likelihood of observing a statistic at least as extreme as the one observed

from the sample data when the null hypothesis is true. A threshold value, called a **significance level**, is denoted by the Greek letter alpha (α). When a study's p-value is less than or equal to this significance level, we state that the results are **statistically significant at level α**. Exact p-values are often difficult to calculate, but *__empirical p-values__ can often be simulated through a randomization or permutation test. The empirical p-value will become more precise as the number of randomizations within a simulation study increases.

The steps in a **randomization test** are as follows:

- An experiment is conducted in which units are assigned to a treatment and an observed sample statistic is calculated (such as the difference between group means).
- Software is used to simulate the random allocation process a number of times (N iterations).
- For each iteration, the statistic of interest (difference between group means) is recorded, with X being the number of times the statistic in the iteration exceeds or is the same as the observed statistic in the actual experiment.
- X/N is computed to find the p-value, the proportion of times the statistic exceeds or is the same as the observed difference.

A *__permutation test__ is a more general form of the randomization test. The steps in both tests are identical, except that permutation tests do not require random allocation. Randomization tests and permutation tests can provide very accurate results. These tests are preferred over parametric methods when the sample size is small or when there are outliers in a data set. Since real data sets tend not to come from exactly normal populations, it is important to recognize that even p-values from parametric tests are approximate (but typically accurate as long as the sample sizes are large enough, the data are not skewed, there are no outliers, and the data are reasonably normal). A graph such as a boxplot or individual value plot should always be created to determine if parametric methods are appropriate. Randomization tests are gaining popularity because they require fewer assumptions and are just as powerful as parametric tests.

Bootstrap methods take many (at least 1000) resamples with replacement of the original sample to create

a bootstrap distribution. If the bootstrap distribution is symmetric and unbiased, bootstrap t or bootstrap percentile confidence intervals can be used to approximate 100(1 - α)%, confidence intervals.

The steps in creating **bootstrap confidence intervals** are as follows:

- One sample of size n is taken from a population and the statistic of interest is calculated.
- Software is used to take resamples (with replacement) of size n from the original sample a number of times (N iterations). For each iteration, the statistic of interest is calculated from the resample.
- The **bootstrap distribution**, which is the distribution of all N resample statistics, is used to estimate the shape and spread of the sampling distribution.
- A **bootstrap t confidence interval** is found by calculating $\bar{x} \pm t^*(S^*)$ where S^* is the standard deviation of the bootstrap distribution and t^* is the critical value of the t(n - 1) distribution with 100(1 - α)%, of the area between - t* and t*.
- A 100(1 - α), bootstrap percentile confidence interval is found by taking the α / 2 * 100 percentile of each tail of the bootstrap distribution.

Bootstrap confidence intervals based on small samples can be unreliable. The bootstrap t or percentile confidence interval may be used if,

- the bootstrap distribution does not appear to be biased,
- the bootstrap distribution appears to be normal, and
- the bootstrap t and percentile confidence intervals are similar.

Simulation studies can easily be extended to testing other terms, such as the median or variance, whereas most parametric tests described in introductory statistics classes (such as the z-test and t-test) are restricted to testing for the mean. Simulation studies are an extremely useful tool that can fairly easily be used to calculate accurate p-values for research hypotheses when other tests are not appropriate.

Before computationally intensive techniques were easily available, rank-based nonparametric tests, such

as the **Wilcoxon rank sum** test and the **Kruskal-Wallis test**, were commonly used. These tests do not require assumptions about distributions, but they tend to be less informative because ranks are used instead of the actual data. Both the Mann-Whitney test and the Kruskal-Wallis test assume that sample data are from independent random samples whose distributions have the same shape and scale. Each sample in the Kruskal-Wallis test should consist of at least five measurements. Rank-based nonparametric tests tend to be less powerful (less likely to identify differences between groups) than parametric tests (when assumptions do hold) and resampling methods. When the sample sizes are small and there are reasons to doubt the normality assumption, rank-based nonparametric tests are recommended over parametric tests. Randomization tests and permutation tests are typically preferred over parametric and rank-based tests. Their p-values are often more reliable, and they are more flexible in the choice of parameter tested.

One final note of caution: Even though it is possible to analyze the same data with a variety of parametric

and nonparametric techniques, statisticians should never search around for a technique that provides the results they are looking for. Conducting multiple tests on the same data and choosing the test that provides the smallest p-value will cause the results to be unreliable. If possible, determine the type of analysis that will be conducted before the data are collected.

Exercises

E1. Is it important in the schistosomiasis study for all 20 mice to come from the same population of mice? Why or why not?

E2. Assume the researchers in this study haphazardly pulled the female mice from a cage and assigned the first five to the treatment and the last five to the control. Would you trust the results of the study as much as if five mice were randomly assigned to each group?

E3. A recent study in the northwest United States found that children who watched more television were more likely to be obese than children who watched less television. Can causation be inferred from this study?

E4. What is the difference between a random sample and a randomized experiment?

E5. Explain the difference between a population model and a randomization model.

E6. Explain how the independence assumption of the two-sample t-test is violated in a randomization model.

E7. If the sample size is large, will the histogram of the sample data have a shape similar to that of the normal distribution? Explain.

E8. If the sample size is large, will the sample mean be normally distributed? Explain.

E9. Why should boxplots or other graphical techniques be used to visualize data before a parametric test is conducted?

E10. Suppose that in our study of schistosomiasis in female mice the p-value was 0.85. Would you be able to conclude that there was no difference between the treatment and control means?

E11. **Using Other Test Statistics**
Data set: 'Mice'. One major advantage of randomization/permutation tests over classical methods is that they easily allow the use of test statistics other than the mean.

1. Modify the program/macro you created in Question 9 to measure a difference in group medians instead of a difference in means for the female mice. Report the p-value and compare your results to those for Question 9.

2. You might also wonder if there is a difference in the variability in the groups. Modify the macro you created in Question 9 to test whether the variances of the female groups are equal. Report the p-value and state your conclusions

E12. **Testing Male Mice**
Data set: 'Mice'.

1. Using the data for the male mice, run a simulation to decide whether K11777 inhibits schistosome viability (i.e., reduces worm count) in male mice. Describe the results, including a histogram of the simulation results, the p-value, and a summary statement indicating your conclusion about the research question of schistosome viability.

2. Modify the program/macro you created in Part A to measure a difference in group medians instead of a difference in means for the male mice. Report the p-value and compare your results to those for Part A.

3. You might also wonder if there is a difference in the variability in the groups. Modify the macro you created in Part A to test if the variances of each male group are equal. Report the p-value and state your conclusions.

E13. **Bird Nest Study**

Data set: ‘Birdnest‘. This data set was collected in the spring of 1999 for a class project by Amy Moore, a Grinnell College student. Each record in the data set represents data for a species of North American passerine bird. Passerines are “perching birds” and include many families of familiar small birds (e.g., sparrows and warblers) as well as some larger species like crows and ravens, but do not include hawks, owls, water fowl, wading birds, and woodpeckers. Moore took all North American passerines for which complete evolutionary data were available, which comprised 99 of the 470 species of passerines in North America (part of her study used this evolutionary information). One hypothesis of interest was about the relationship of body size to type of nest. Body size was measured as average length of the species, nest type was categorized as either closed or open. Although nests come in a variety of types (see the ‘Nesttype‘ variable), in this data set “closed” refers to nests with only a small opening to the outside, such as the tree-cavity nest of many nuthatches or the pendant-style nest of an oriole. “Open” nests include the cup-shaped nest of the American robin.

1. Moore suspected that closed nests tend to be built by larger birds, but here we will treat the alternative as two-sided, since her suspicion was based on scanty evidence. Use comparative dotplots or boxplots and summary statistics to describe the relationship between average body length and nest type (the ‘Closed‘ variable). (Note: ‘Closed‘ = 1 for closed nests; ‘Closed‘ = 0 for open nests.) Does it appear that Moore’s initial suspicion is borne out by the data? of the simulation results, the p-value, and a summary statement indicating your conclusion about the research question of schistosome viability.

2. Run a permutation test using a two-sided alternative to determine if type of nest varies by body length and interpret your results. Be sure to state your conclusions in the context of the problem and address how random allocation and random sampling (or lack of either) impact your conclusions.

E14. **Twins Brain Study**

Data set: ‘Twins‘. In a 1990 study by Suddath et al., reported in Ramsey and Schafer,[12] researchers used magnetic resonance imaging to measure the volume of various regions of the brain for a sample of 15 monozygotic twins, where one twin was affected with schizophrenia and the other was unaffected. The twins were from North America and comprised eight male pairs, and seven female pairs ranging in age from 25 to 44 at the time of the study. The sizes in volume (cm^3) of the hippocampus are in the file called ‘Twins‘.

1. Should the data be analyzed as match pairs or be treated as if there were two independent samples?

2. Use appropriate graphics and summary statistics to describe the difference in brain volume for affected and unaffected twins.

3. Use the appropriate permutation test to ascertain if the difference in brain volume described in Part B is the result of schizophrenia or if it could be explained as a chance difference. Report your p-value and summarize your conclusion.

E15. **Comparing Parametric and Nonparametric Tests**
Data set: ‘Birdnest‘and ‘Music‘.

1. Using a t-test, compute the two-sided p-value for the bird nest study in Exercise E.13. and compare the results to what you found with the randomization test.

2. Using a t-test, compute the one-sided p-value for the music study in Question 20 and compare the results to what you found with the randomization test.

E16. **Means versus Medians in Rank-Based Tests**
Data set: ‘SameMean‘. Rank-based nonparametric tests do not answer the same question as the corresponding parametric procedure. Many people assume that these nonparametric tests are testing for group medians. This is not always true. Rank-based tests can be interpreted as testing for the median only if the shapes and scales of the populations are the same. The following exercise illustrates this point by providing an example where the medians and the means are identical but nonparametric tests will reject the null hypothesis. | Use the ‘SameMean‘ data to conduct the Kruskal-Wallis test. Calculate the mean and median for each group. What conclusions can you draw from the data?

E17. **Rank Based Bird Nest Tests**
Data set: ‘Birdnest‘.

1. Use the Wilcoxon rank sum test to conduct a significance test for the bird nest study discussed in Exercise E.13.

2. Use the Kruskal-Wallis test to conduct a significance test for the bird nest study. Determine whether the distribution of bird size (response is Length) is the same for each nest type. Note that when the chi-square approximation is used, each group should have at least five observations. You may need to create an “other” group to combine all nest types with sample sizes less than five.

E18. **Bootstrap Confidence Intervals**
Data set: ‘ChiSq‘. Take a simple random sample of size 40 from the ‘ChiSq‘ data file.

1. Create a bootstrap distribution of the mean (or use the distribution you created in Question 22). Calculate a 95
2. Create a bootstrap distribution of the mean (or use the distribution you created in Question 22). Calculate a 95percentile confidence intervals for the mean reliable?
3. Create a bootstrap distribution of the standard deviation (or use the distribution you created in Question 23). Calculate a 95
4. Create a bootstrap distribution of the standard deviation (or use the distribution you created in Question 23). Calculate a 95deviation. Are the bootstrap t and percentile confidence intervals for the standard deviation reliable?

E19. **Medians and Trimmed Means in Bootstrap Confidence Intervals** Data set: ‘ChiSq‘.

1. Take a simple random sample of size n = 40 from the ChiSq data. Create a bootstrap distribution of the median by taking 1000 resamples (with replacement). Describe the shape of the bootstrap distribution and explain why bootstrap confidence intervals are unlikely to be reliable.
2. Take a second simple random sample of size n = 40 from the ChiSq data. Create a second bootstrap distribution of the median by taking 1000 resamples (with replacement). Describe the shape of the second bootstrap distribution. With a sample size of 40, why are bootstrap distributions of medians unlikely to be normal?
3. Bootstrap distributions for medians are unlikely to be normally distributed, and means tend to be influenced by outliers. The trimmed mean is a common measure of center that tends to better represent the average value with bootstrap methods. Trimmed means are calculated by first trimming the upper and lower values of the sample. For example, the 25of the middle 50

   | Take a simple random sample of size n = 40 from the ChiSq data. Create a bootstrap distribution of the 25for each resample calculate the mean of the middle 20 observations (remove the smallest 10 and largest 10 values in each resample). Create a histogram of the 1000 trimmed means and describe the shape of this bootstrap distribution. Create a bootstrap t confidence interval and a bootstrap percentile confidence interval to estimate the 25

E20. **Medians and Trimmed Means in Bootstrap Confidence Intervals** Data set: ‘MedSalaries‘.

1. The file ‘MedSalaries‘ is a random sample of salaries of medical doctors who were teaching at United States universities in 2009. Create a

bootstrap distribution of the median by taking 1000 resamples (with replacement). Describe the shape of the bootstrap distribution. Is it appropriate to create a bootstrap t confidence interval or a bootstrap percentile confidence interval for the median?

2. Create a bootstrap distribution of the 25(with replacement). In other words, calculate the mean of the middle 50 observations from each resample. Describe the shape of the bootstrap distribution. Is it appropriate to create a bootstrap t confidence interval or a bootstrap percentile confidence interval for the 25
3. Create a bootstrap distribution of the 5replacement). In other words, calculate the mean of the middle 90 observations from each resample. Describe the shape of the bootstrap distribution. Is it appropriate to create a bootstrap t confidence interval or a bootstrap percentile confidence interval for the 5
4. Calculate a bootstrap t confidence interval and bootstrap percentile confidence interval for each of the preceding parts of this exercise if the bootstrap distribution indicates that it is appropriate.

E21. **Multiple Comparisons**
Data set: ‘NLBB Salaries‘.

1. Conduct a permutation test to determine if there is a difference in mean salaries between pitchers and first basemen. Report the p-value and your conclusions based on an individual α-level of 0.05.
2. Conduct a permutation test to determine if there is a difference in mean salaries between pitchers and catchers. Report the p-value and your conclusions based on an individual α-level of 0.05.
3. Conduct a permutation test to determine if there is a difference in mean salaries between first basemen and catchers. Report the p-value and your conclusions based on an individual a-level of 0.05.
4. If each of the previous three tests uses an a-level of 0.05, what is the true probability that at least one of the tests will inappropriately reject the null hypothesis?
5. What is the individual critical value if you use the Bonferroni method with an overall (familywise) α-level of 0.05? Do any of your previous conclusions in the preceding parts of this exercise change if you test for an overall (familywise) comparison? Explain.

Chapter 2

Making Connections: The Two-Sample t-Test, Regression, and ANOVA

In theory, there's no difference between theory and practice. In practice, there is.
-Yogi Berra[1]

Statistics courses often teach the two-sample t-test, linear regression, and analysis of variance (ANOVA) as very distinct approaches to analyzing different types of data. However, this chapter makes connections among these three techniques by focusing on the statistical models. Statistical software has made it easy to calculate statistics and p-values. But without understanding the underlying model assumptions, it is easy to draw incorrect conclusions from the sample data. As studies become more complex, models become fundamental to drawing appropriate conclusions. In this chapter, a simple student experiment involving games and several additional studies are used to do the following:

- Compare the underlying statistical models for the two-sample t-test, linear regression, and ANOVA
- Discuss the model assumptions for each of these three tests
- Create and interpret normal probability plots
- Transform data in order to better fit the model assumptions
- Discuss the mathematical details of each hypothesis test and corresponding confidence interval

[1] Yogi Berra was an American League Baseball player and manager. This quote has also been attributed to computer scientist Jan L. A. van de Snepscheut.

2.1 Investigation: Do Distracting Colors Influence the Time to Complete a Game?

In 1935, John Stroop published a paper presenting his research on the reaction time of undergraduate students identifying ink colors.2 He found that students took a longer time identifying ink colors when the ink was used to spell a different color. For example, if the word "yellow" was printed in blue ink, students took longer to identify the blue ink because they automatically read the word "yellow." Even though students were told only to identify the ink color, the automatized behavior of reading interfered with the task and slowed their reaction time.[2] *Automatized behaviors* are behaviors that can be done automatically without carefully thinking through each step in the process. Stroop's work, demonstrating that automatized behaviors can act as a distracter for other desired behaviors, is so well known that the effect is often called the *Stroop effect*.

Several students in an introductory statistics class wanted to develop a final project that would test the impact of distracters. They decided to conduct a study to determine if students at their college would perform differently when a distracting color was incorporated into a computerized game. This game challenges people to place an assortment of shaped pegs into the appropriate spaces as quickly as possible. Before any data were collected, these students developed a clear set of procedures.

- 40 students would be randomly selected from the college.[3]
- 20 students would be assigned to the standard game and 20 would be assigned to a game with a color distracter. The student researchers would flip a coin to randomly assign subjects to a treatment. Once 20 subjects had been assigned to either group, the rest would automatically be assigned to play the other game.
- Subjects would see a picture of the game and have the rules clearly explained to them before they played the game. An example of both games is shown in Figure 2.1.
- Subjects would play the game in the same area with similar background noise to control for other possible distractions.
- The response variable would be the time in seconds from when the participant pressed the "start game" button to when he or she won the game.

> **NOTE** It is important to recognize that each subject in this study was assigned to exactly one treatment, either the standard game or the color distracter game. Some researchers may point out that a

[2] Note that many psychologists would call this procedural knowledge instead of automatized behavior. Both are processes that can be done without conscious thought, but automatized behaviors are processes that cannot be slowed down, do not decline with age, and show no gender differences.

[3] Since it was not possible to force college students to be involved in this study, these researchers randomly selected students from an online college directory until they had 40 students who were willing to play the game.

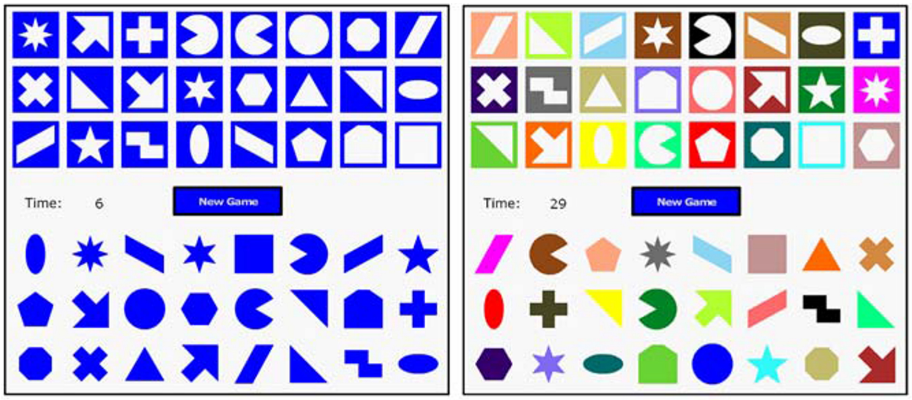

Figure 2.1: An image of the electronic Shapesplosion game with and without color distracters. The instructions for the game were to click and drag each peg to the space with the matching shape.
(#fig:fig2.1)

paired design (where each subject was assigned to both treatments) might have been more efficient. However, for the purposes of this chapter, this study will be treated as the students originally designed it: a study comparing two independent samples.

2.1.1 Understanding the Study Design

1. For this study, identify the units, the population for which conclusions can be drawn, the explanatory variable, and the response variable.

2. Is this study an experiment or an observational study? Explain.

3. The researchers hoped to determine if distracting colors influenced college students' response times when playing a computerized game. Write out in words and symbols appropriate null and alternative hypotheses. Let μ_1 represent the true mean response time of the color group and μ_2 the true mean response time of the standard group. Use a two-sided alternative hypothesis for this question.

4. Create an individual value plot or a boxplot of the Games1 data from this study. Describe the graph. For example, does it look as if the groups have equal means or equal standard deviations? Are there any unusual observations in the data set? Calculate the mean and standard deviation of the color distracter responses, $\bar{y}_1$ and s_1 , as well as the mean and standard deviation of the standard game responses, $\bar{y}_2$ and s_2.

observed		mean		error			
value		response		term			
(random)		(not random)		(random)			
↓		↓		↓			
$y_{1,j}$	=	μ_1	+	$\epsilon_{1,j}$			for $j = 1, 2, ... ,n_1$

2.2 The Two-Sample t-Test to Compare Population Means

2.2.1 The Statistical Model

Generally, **statistical models** have the following form:

observed value = mean response + random error

The statistical model describes each observed value in a data set as the sum of a mean response for some subgroup of interest (often called a group mean) and a random error term. The mean response is fixed for each group, while the random error term is used to model the uncertainty of each individual outcome. The random error term for each individual outcome cannot be predicted, but in the long run there is a regular pattern that can be modeled with a distribution (such as the normal distribution).

The key question in this study is whether or not the two types of games have different average completion times. The two-sample t-test starts with the assumption that the two group means are equal. This is often written as the null hypothesis $H_0 : \mu_1 - \mu_2 = 0$ or, equivalently, $H_0 : \mu_1 = \mu_2$.

The underlying model used in the two-sample t-test is designed to account for these two group means (μ_1 and μ_2) and random error. The statistical model for the first population, the color distracter group, is:

where j is used to represent each observation in the sample from the first population. For example, $y_{1,9}$ represents the 9th observation in the first group (the color distracter group). In this data set, there were 20 observations taken from the first population; thus, $n_1 = 20$.

This model states that the color distracter game is expected to be centered at the constant value μ_1 . In addition, each observation is expected to have some variability (random error) that is typically modeled by a normal distribution with a mean equal to zero and a fixed variance s2. Similarly, each observation from the second group, the standard game, can be modeled as the sum of μ_2 plus a random error term, $\epsilon_{2,j}$:

$$y_{2,j} = \mu_2 + \epsilon_{2,j} \quad \text{for} \quad j = 1, 2, ..., n_2 \tag{2.1}$$

where $n_2 = 20$, μ_2 is the mean of the standard group, and the $\epsilon_{2,j}$ are random variables (typically from a normal distribution) with a mean equal to zero and variance σ^2 . Often, this statistical model is more succinctly written as:

$$y_{i,j} = \mu_i + \epsilon_{i,j} \quad \text{for} \;\; j = 1, 2 \text{ and } j = 1, 2, ..., n_2 \quad \text{where} \;\; \epsilon_{i,j} \sim N(0, \sigma^2) \quad (2.1)$$

> **MATHMATICAL NOTE** You may recall from your introductory statistics course that adding a constant to each random variable in a population does not change the shape or spread of the population. Since each mean response (μ_i) is fixed (i.e., a constant value), Equation **??** can be used to show that $y_{i,j} \sim N(\mu_i, \sigma^2)$.

This model has one assumption that you may not have made when previously conducting a two-sample t-test. Equation **??** states that all $\epsilon_{i,j}$ come from a normally distributed population with a mean of zero and variance s2 . This is called the equal variance assumption. Some introductory statistics courses discuss only a two-sample t-test that does not require the equal variance assumption. The equal variance assumption is made here because it makes sense for this experiment, the data support it (s_1 is close to s_2), and it allows a direct comparison to ANOVA and regression models.

In Equation **??**, the mean response of the model is the population mean (μ_1 or μ_2). Just as a sample mean, $\bar{y}_i$, is used to estimate the population means, μ_i, residuals are used to estimate the random error terms. **Residuals** are the difference between the observed response and the estimated mean response. For example, the random error term $\epsilon_{1,12} = +\bar{y}_{1,12} - \mu_1$ is estimated by $\hat{\epsilon}_{1,12} = +\bar{y}_{1,12} - \bar{y}_1$.

> **NOTE** A **statistic** is any mathematical function of the sample data. **Parameters** are actual population values that cannot be known unless the entire population is sampled. The mean response is based on population parameters. If a sample data set is used, we do not know the population parameters. Sample statistics (such as the sample mean, $\bar{y}$, and the sample standard deviation, s) are used to estimate population parameters (μ and σ). Statisticians often use a hat on top of a parameter to represent an estimate of that parameter. For example, an estimate of the population standard deviation is written $s = \hat{\sigma}$, and an estimate for a mean is written $\bar{y}_1 = \hat{\mu}_1$ or $\bar{y}_2 = \hat{\mu}_2$.

2.2.2 Statistical Models for the Two-Sample t-Test

5. Assume that we have two very small populations that can be written as $y_{1,1} = 15$, $y_{1,2} = 17$, $y_{1,3} = 16$, $y_{2,1} = 11$, $y_{2,2} = 9$, $y_{2,3} = 10$. Find μ_1, μ_2, $\epsilon_{1,1}$, $\epsilon_{1,3}$, and $\epsilon_{2,1}$.

Notice the double subscripts on the observed responses: $y_{1,1}$ is read as "y one one." The first subscript tells us that the observation was from the first group, and the second subscript tells us the observation number. For example, $y_{1,j}$ is the jth observation from the first group.

6. Use the game study and the data in the file Games1 to identify n_1, n_2, $y_{1,12}$, $y_{2,12}$, $\epsilon_{1,12}$, and $\epsilon_{2,12}$, where $y_{1,12}$ represents the 12th observation from group 1 (the color distracter group). Note that since this is a sample, not a population, we do not know μ_1 or μ_2 , but we can estimate them with $\bar{y}_1 = \hat{\mu}_1$ and $\bar{y}_2 = \hat{\mu}_2$.

2.2.3 Model Assumptions for the Two-Sample t-Test

Several implicit assumptions are built into the model for the two-sample t-test shown in Equation **??**:

- Constant parameters: The population values in this model (μ_1 , μ_2 , and σ) do not change throughout the study.
- Additive terms: The model described in Equation **??** shows that the observed responses are the sum of our parameters and error terms. For example, we are not considering models such as $y_{i,j} = \mu_i * \epsilon_{i,j}$.
- $\epsilon_{i,j} \sim N(0, \sigma^2)$. This assumption has many key components:
- The error terms are independent and identically distributed (iid).
- The error terms follow a normal probability distribution.
- The error terms have a mean of zero. This implies that the average of several observed values will tend to be close to the true mean. In essence, there is no systematic bias in the error terms.
- The population variance σ^2 is the same for both groups (color distracter and standard games) being tested.

The first assumption tells us about the mean response. The parameter estimate ($\bar{y}_i$) would not be meaningful if the true parameter value (μ_i) were not constant throughout the study. The second assumption simply states the types of models we are building. In later chapters with more complex models, we will discuss how to use residual plots to determine if the model is appropriate. In this chapter, we will focus on the assumptions about the error terms

> **MATHMATICAL NOTE** In later chapters, we will show that a curved pattern in a residual versus fit plot suggests that an additive model may not be appropriate. In this example, there are only two fitted values (i.e., expected values), so we cannot see any curved patterns. When the additive assumption is violated, residual plots may also indicate different standard deviations, a nonnormal distribution, or lack of independence. Transforming the data to a new scale can often make the additivity assumption (and several of the other assumptions) more appropriate.

The statistical model described in Equation **??** assumes that $\epsilon_{i,j}$ are modeled as

independent and identically distributed (iid) random variables. The independent error term assumption states that there is no relationship between one observation and the next. For example, knowing that the 8th subject in a group played the game more quickly than average does not provide any information about whether the 7th or 9th person in the group will be above or below the average.

The identically distributed assumption states that each error is assumed to come from the same population distribution. Thus, each subject from a particular group is from the same population. If any error term based on a particular observation comes from a different population, the two-sample t-test will not be valid. For example, elementary school students may have different expected completion times for the Shapesplosion game than college students. It would be inappropriate to include younger students in a study where the population was assumed to be college students

Model assumptions for the residuals should always be checked with plots of the data. The extended activities will describe normality tests in more detail, but in most situations a simple graph of the residuals will suffice. The two sample t-test actually requires only that the sample means (each $\bar{y}_{i,j}$) be normally distributed. The central limit theorem allows us to assume this is true if group sample sizes are similar and large ($n_1 \geq 15$ and $n_2 \geq 15$) and there does not appear to be any extreme skewness or outliers in the residuals.

Since residuals are defined as the difference between each observed value and the corresponding group mean, they should always sum to zero. Thus, we cannot check residuals to determine whether each of the error terms is centered at zero. The assumption that the error terms are centered at zero is really stating that there are no other sources of variability that may be biasing our results. In essence, the only difference between the two population means is explained by the mean response.

To check the assumption that the two populations have the same variance, an informal test can be used. If the ratio of the sample standard deviations is less than 2, we can proceed with the analysis.[4]

Informal Test for Equal Variances

$$\text{if} \quad \frac{\max(s_1, s_2)}{\min(s_1, 2)} < 2 \quad \text{or, equivalently, if} \quad \frac{\max(s_1^2, s_2^2)}{\min(s_1^2, s_2^2)} < 4$$

then we do not have enough evidence to conclude that the population variances are different.

[4] Some texts suggest rejecting the equal variance assumption when the ratio is greater than 3 instead of 2. If the ratio is close to 2 (or 3), many statisticians would suggest conducting a more formal F-test for equal variances.

Several key observations should be made about the individual value plot shown in Figure 2.2:

- The mean completion time is higher for the color distracter group than for the standard group.
- Neither group appears to have clear outliers, skewness, or large gaps.
- The spread (variance) of the two groups appears to be similar.

> **Key Concept** Every statistical hypothesis test has basic underlying conditions that need to be checked before any valid conclusions can be drawn.

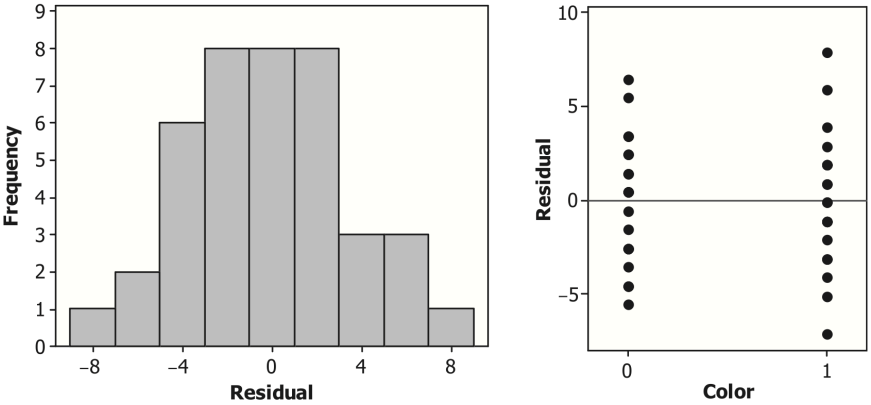

Figure 2.2: Individual value plot of the data from the color distracter and standard games.
(#fig:fig2.2)

Checking Assumptions for the t-Test

7. Calculate the residuals in the Games1 data. Plot a histogram of the residuals (or create a normal probability plot of the residuals). Do the residuals appear to be somewhat normally distributed?
8. Use the informal test to determine if the equal variance assumption is appropriate for this study.
9. The variable StudentID represents the order in which the games were played. Plot the residuals versus the order of the data to determine if any patterns exist that may indicate that the observations are not independent.
10. Use statistical software to conduct a two-sample t-test (assuming equal variances) and find the p-value corresponding to this statistic. In addition, use software to calculate a 95% confidence interval for the difference between the two means (μ_1 - μ_2). Equation **??** and the extended activities provide details on conducting these calculations by hand. If H0 : μ_1

$= \mu_2$ is true, the p**-value** states how likely it is that random chance alone would create a difference between two sample means ($\bar{y}_1 - \bar{y}_2$) at least as large as the one observed. Based on the p, what can you conclude about these two types of games?

2.3 The Regression Model to Compare Population Means

2.3.1 The Linear Regression Model

The simple linear regression model discussed in introductory statistics courses typically has the following form:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{for } i = 1, 2, ..., n \quad \text{where } \epsilon_i \sim N(0, \sigma^2) \tag{2.2}$$

A **simple linear regression** model is a straight-line regression model with a single explanatory variable and a single response variable. For this linear regression model, the mean response $(\beta_0 + \beta_1 x_i)$ is a function of two parameters, β_0 and β_1, and an explanatory variable, x. The random error terms, ϵ_i, are assumed to be independent and to follow a normal distribution with mean zero and variance σ^2.

In Equation **??**, we used double subscripts: $i = 1, 2$ was used to show that there were two distinct groups and $j = 1, 2, ..., n_i$ was used to identify each of the $n_1 = n_2 = 20$ items within the two groups. In the regression model, there is only one set of subscripts: $i = 1, 2, ..., n$, where $n = 40 = n_1 + n_2$. Instead of having two distinct means in the model (μ_1 and μ_2), as in the two-sample t-test, we have one regression model where the parameters, β_0 and β_1, are fixed. The categorical explanatory variable, x, indicates game type.

A procedure commonly used to incorporate categorical explanatory variables, such as the game type, into a regression model is to define **indicator variables**, also called **dummy variables**, that will take on the role of the x variable in the model. Creating dummy variables is a process of mapping the column of categorical data into 0 and 1 data. For example, the indicator variable will have the value 1 for every observation from the color distracter game and 0 for every observation from the standard game. Most statistical software packages have a command for automatically creating dummy variables.

> **NOTE** Typically an indicator variable is created for each category. Thus, there would be an indicator variable called Color equal to 1 for the color distracter game and 0 otherwise and another indicator variable called Standard equal to 1 for the standard game and 0 for all other categories. Notice that there is complete redundancy between the two indicator variables: Knowing the value of the Color variable automatically tells us the value of the Standard variable for each subject. Thus, only one of the indicator variables is needed in this model. Although this study has only two categories of games (color and standard), it is common for a categorical explanatory variable to have more than two categories. Chapter 3 provides the opportunity to use indicator variables when there are multiple categories.

Key Concept Indicator variables can be created to incorporate categorical explanatory variables into a regression model.

Calculating a Regression Model and Hypothesis Test for the Slope

11. Use the software instructions and the Games1 data to create indicator variables where $x = 1$ represents the color distracter game and $x = 0$ represents the standard game. Develop a regression model using Time as the response and the indicator variable as the explanatory variable.

12. Use statistical software to calculate the t-statistic and p-value for the hypothesis tests $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. In addition, construct a 95% confidence interval for β_1 . Based on these statistics, can you conclude that the coefficient, β_1, is significantly different from zero? Details for calculating these statistics by hand are provided in the extended activities.

13. Repeat the two previous questions, but use an indicator variable where $x = 1$ represents the standard game and $x = 0$ represents the color distracter game. Compare the regression line, hypothesis test, and p-value to those from the previous questions. When there are only two categories (color distracter and standard), does the choice of indicator variable impact your conclusions? Why or why not?

In the previous questions, we assigned x to be the dummy variable that indicates the type of game. Notice that the mean response is still a constant (nonrandom) value for each of the two game categories. In other words, when $x = 1$ the mean response is a fixed value, and when $x = 0$ the mean response is a fixed value. In addition, the "slope" coefficient (β_1) can be considered as an estimate of the average amount by which the response variable will change from the standard game ($x = 0$) to the color distracter game ($x = 1$).

Although the notation has changed, the regression model and the model used in the two-sample t-test are mathematically equivalent. When a subject is from the color distracter group, the mean response is μ_1 in the t-test and the mean response sets $x = 1$ in the regression model. Thus,

$$\mu_1 = \beta_0 + \beta_1(1) = \beta_0 + \beta_1 \tag{2.3}$$

When a subject is from the standard group, the mean response is μ_2 in the t-test and the mean response sets $x = 0$ in regression. Thus,

$$\mu_2 = \beta_0 + \beta_1(0) = \beta_0 \tag{2.4}$$

Equations **??** and **??** can be combined to show the relationship between the two-sample t-test and regression hypotheses.

$$\mu_1 - \mu_2 = (\beta_0 + \beta_1) - \beta_0 = \beta_1 \tag{2.5}$$

Thus, stating that $\mu_1 - \mu_2 = 0$ is equivalent to stating that $\beta_1 = 0$

> **Key Concept** In testing the difference in two population means, testing the null hypothesis $H_0 : \beta_1 = 0$ for a regression model is equivalent to testing the two-sample t-test hypothesis $H_0 : \mu_1 - \mu_2 = 0$ *when using the equal variance assumption.*

2.3.2 Model Assumptions for Regression

While no distributional assumptions are needed to create estimates of b0 and b1 , it is necessary to check the same model assumptions when conducting a hypothesis test for b1. Just as in the two-sample t-test, the model assumes that the parameters b0 , b1 , and σ^2 are constant. In addition, Equation **??** shows that our model consists of the mean response plus the error term. The regression model also assumes that $\epsilon_i \sim N(0, \sigma^2)$.

This expression represents the following four assumptions:

- The error terms are independent and identically distributed (iid).
- The error terms follow a normal probability distribution.
- The error terms have a mean of zero .
- The error terms in the regression model are assumed to come from a single population with variance σ^2 (i.e., the variance does not depend on x).

In regression, assumptions about the error terms are also checked by residual plots. Here, y_i represents each observed response and $\hat{y}_i = b_0 + b_1 x_i$ represents the estimated mean response. So the residuals are simply the observed value minus the estimated value: $\hat{\epsilon}_i = y_i - \hat{y}_i$

Figure 2.3 shows a histogram of the residuals and a plot of the residuals by type of game. The histogram shows that the residuals approximately follow the shape of a normal distribution. The residual versus game type graph shows that there are no obvious outliers and that the spread of both groups is roughly equivalent. Since residuals are just the mean response subtracted from the observed value, the center of the residual plots has shifted to zero. However, the spread of the residual versus game plot is identical to the spread of the individual value plot in Figure 2.2.

> **Key Concept** No assumptions are needed about the error terms to calculate estimates ($b_1 = \widehat{\beta}_1$ and $b_0 = \widehat{\beta}_0$) of the slope and intercept of the regression line. These estimates are simply well-known mathematical calculations. However, all the model assumptions should be satisfied in order to properly conduct a hypothesis test or create a confidence interval for β_1.

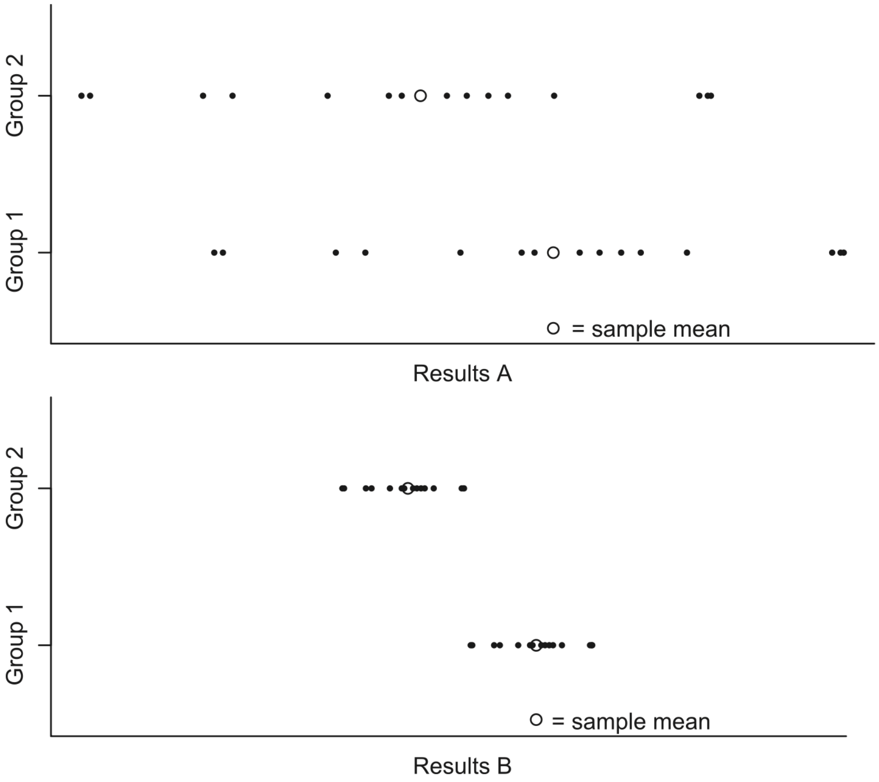

Figure 2.3: Histogram of residuals and plot of residuals versus color. (#fig:fig2.3)

Checking Model Assumptions

14. Calculate the residuals from the regression line in Question 11. Plot a histogram of the residuals (or create a normal probability plot of the residuals). In addition, create a residual versus order plot and use the informal test to determine if the equal variance assumption is appropriate for this study. Compare these plots to the residual plots created for the two-sample t-test. Why are these graphs so similar?

15. Create a scatterplot with the regression line in Question 11. Use the graph to give an interpretation of the slope and y-intercept, b1 and b0, in the context of the game study.

2.4 ANOVA to Compare Population Means

The term **ANOVA** is an acronym for **ANalysis Of VAriance**. ANOVA models often describe categorical explanatory variables in terms of factors and levels. The explanatory variable, also called a factor, in this study is the type of game; the two conditions, the two levels of the factor, are color distracter and standard.

2.4.1 The ANOVA Model

The ANOVA model for the game study can be written as

$$y_{i,j} = \mu + \alpha_i + \epsilon_{i,j} for i = 1, 2 and j = 1, 2, ..., n where \epsilon_{i,j} \sim N(0, \sigma^2) \qquad (2.6)$$

The mean response in the ANOVA model is $\mu + \alpha_1$ for the color distracter group and $\mu + \alpha_2$ for the standard group, where μ is the mean of all the completion times in the study. This overall mean is often called the grand mean or the benchmark value; α_1 is the **effect**, or **main effect**, of the color distracter group. **Effects** are a measure of differences between group means. The effect α_1 represents the change in the response from the grand mean to the color distracter group mean.[5]

To summarize, here is what the symbols in the model represent:

- $y_{i,j}$: observed completion time for subject j from group i
- μ: overall mean (the benchmark value)
- α_i: effect of group i (i = 1, 2)
- $\epsilon_{i,j}$: error for the jth subject ($j = 1, 2, ..., 20$) from the ith group ($i = 1, 2$)

Although the notation varies, the mean response for the ANOVA model is mathematically equivalent to the mean response in the t-test.

- $\mu_1 = \mu + \alpha_1$: population mean for the color distracter games
- $\mu_2 = \mu + \alpha_2$: population mean for the standard games

The ANOVA Model

16. Explain (or use equations to show) why the ANOVA hypothesis H0 : α_1 = α_2 is equivalent to the two- sample t-test hypothesis H0 : μ_1 = μ_2 . *In this text m is always considered the overall mean of the data. Also throughout this chapter, we are always assuming balanced data.

    Key Concept In the ANOVA model, there is the appearance that we are describing two means (μ_1 and μ_2) using three parameters (μ, α_1 , and α_2). Since it can be shown that $\alpha_2 = -\alpha_1$, there are actually just two parameters (μ and $alpha_1$) that are estimated.

[5] In this text m is always considered the overall mean of the data. Also throughout this chapter, we are always assuming balanced data.

Thus, the null hypothesis stating no effect size can also be written as $H_0 : \alpha_1 = \alpha_2 = 0$ or $H_0 : \mu_1 = \mu_2 = \mu$.

17. Write the proper ANOVA model [provide the appropriate ij subscripts as in Equation **??**] for the observation representing the 3rd subject from the color distracter group. Also give the notation for the observation representing the 20th subject from the standard group.

18. Why doesn't μ have any subscript in the ANOVA model?

After the data have been collected, the averages for all three meaningful groupings of the data can be calculated. The following mathematical notation is often used to represent the calculated sample averages:

- $\bar{y}_{..}$: **grand mean** (the overall average of the combined results)
- $\bar{y}_{1.}$: average for the color distracter game sample results
- $\bar{y}_{2.}$: average for the standard game sample results

  Note Throughout this chapter, $\bar{y}_{1.} = \bar{y}_1$ and $\bar{y}_{2.} = \bar{y}_2$. The dot notation is often used with more complex models to indicate that the average was taken over all values of that subscript. For example, $bary_{2.}$ averages over all $j = 1, 2, 3, ..., n_2$, observations from the standard game sample results.

The effect of the color distracter game, α_1 , can be estimated by $\hat{\alpha}_1 = \bar{y}_{1.} - \bar{y}_{..}$. Similarly, $\hat{\alpha}_2 = \bar{y}_{2.} - \bar{y}_{..}$ estimates the standard game effect, α_2. As in regression and the two-sample t-test, each residual $\hat{\epsilon}_{ij}$ is the difference between an observed value and the corresponding mean response.

$$\begin{aligned}\hat{\epsilon}_{ij} &= \text{observed} - (\text{grand mean} + \text{effect of group}_i) \\ &= y_{i,j} - [\bar{y}_{..} + \hat{\alpha}_i] \\ &= y_{i,j} - [\bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..})] \\ &= y_{i,j} - \bar{y}_{i.}\end{aligned}$$

Key Concept: Since the mean responses for the two-sample t-test, regression, and ANOVA are mathematically equivalent for this data set, the residual values are also identical for all three models.

Activity: Estimating the Model Values

19. Use the `Games1` data to calculate $\bar{y}_{..}$, $\bar{y}_{1.}$, and $\bar{y}_{2.}$.
20. Estimate the effect sizes for the color distracter game and the standard game.
21. The main effects are often visualized with a **main effects plot**. The main effects plot simply plots the average for each factor

level and, in this example, shows that the color distracter group exhibited a higher average completion time than the standard group. Main effect plots are not very informative with just one explanatory variable. However, in more complex data sets with several explanatory variables, main effect plots can be quite useful in comparing effect sizes across all the explanatory variables. Use statistical software to create a main effects plot.

22. Calculate the residual for the 20th observation from the standard group, $\hat{e}_{2,20}$.

Model Assumptions for ANOVA

The model assumptions for ANOVA are equivalent to those for the two previous tests. In fact, the assumptions discussed in this section are called the six *Fisher assumptions*, after Ronald Fisher, who developed the ANOVA and the corresponding F-test.

- The parameters (μ, each α_i, and σ^2) are constant throughout the study.
- Each term in the ANOVA model is added.
- The error terms are independent and identically distributed (iid).
- The error terms follow a normal probability distribution.
- The error terms have a mean of zero.
- Population variances within each factor level (each game type) are equal (i.e., the sample variances can be pooled).

The following questions provide an opportunity to use software to calculate an **F-statistic** (the test statistic for $H_0 : \alpha_1 = \alpha_2 = 0$ that is calculated using an ANOVA table) and corresponding p-value. In addition, you will use graphs to visualize the residuals to check the model assumptions. The extended activities will describe the ANOVA calculations in more detail.

Activity: Checking Assumptions

23. Use statistical software to calculate the F-statistic and find the p-value. Use the p-value to draw conclusions from this study.
24. How does the p-value in ANOVA compare to the p-value you found for the two-sample t-test and the regression model?
25. Take the square root of the F-statistic in the ANOVA table. Does this value look familiar? Explain.
26. Check the model assumptions by creating a histogram of the residuals, a plot of the residuals versus the type of game, and a plot of the residuals versus the order of the observations (the order in which data were collected). Are the residuals approximately normal? Are the residual variances similar for the two factor levels? Are there patterns in the residual plots that might

indicate that the residuals are not iid?

27. Compare the three statistical models. Describe the connections among the t-test, ANOVA, and regression. Why are the p-values the same for all three models?

2.5 Comparing Planned Variability to Random Variability

The statistical model (observed value = mean response + random error) assumes that there are only two types of variability that can occur in a study. The difference between subgroup means (i.e., the difference between mean response values) represents the **planned variability** in the study. For example, in the game study we plan to find that the mean for the color distracter group is different from the mean for the standard group. The random error term is used to model the uncertainty of each individual outcome, called the **random variability**.

All three test statistics described in this chapter are based on a ratio. Each hypothesis test is based on comparing the planned variability to the random variability. The numerator in every test represents differences between group means. The denominator is a measure based on the variability of the residuals. If the subgroup means are far apart compared to the random variability, the null hypothesis is rejected and we conclude that the two population means are different.

Figure 2.4 shows boxplots for two fictitious data sets, `Results A` and `Results B`. Notice that the differences between group means are identical. In other words, the numerator of the test statistic (difference between group means) is the same for both data sets.

Even though the difference between group means (planned variability as described by the mean response) is the same, the variability within each group (random variability represented by the error term) is different. The residual variation (the denominator) is much larger for `Results A` than for `Results B`. Thus, the `Results B` data set will correspond to a larger test statistic and a smaller p-value, and we are more likely to reject the null hypothesis. Thus, `Results B` provides much stronger evidence that the difference between group means is not due simply to chance, but due to real differences in the two population means.

Key Concept:
Random sampling and random allocation do not impact the type of statistical model or technique used, but they do impact the type of conclusions that can be drawn. When units are randomly sampled from a population, we can generalize the conclusions to that population. Well-designed experiments incorporate random allocation in a study and can be used to show causation.

Random sampling and random allocation can be used to convert unplanned systematic variability into random variability. For example, in the game study, the subjects' natural ability may bias the results if more talented subjects tend to play one game type over the other. However, if we randomly allocate subjects to a game type, we can expect each group to have an equivalent number of talented subjects. In addition, the variability in natural abilities now tends to look like the random variability that can be modeled with the error term.

In this chapter, we assume this was a well-designed study with no obvious biases. We focus on creating models and better understanding the random error term in order to determine if statistical techniques (two-sample t-test, regression, and ANOVA) are appropriate. Later chapters will discuss how to address extraneous variables and properly design studies.

NOTE:
Later chapters will explain how studies can be designed to control for the influence of extraneous variables that are suspected of potentially biasing the results. Extraneous variables can be controlled by *limiting the study to conditions that are as consistent as possible.* For example, the researchers could decide to have the subjects play all games with the same number of pegs and play all games in a quiet room at the same time of day. Extraneous variables can also be controlled by *incorporating a new variable into the mean response.* Instead of simply testing for the type of game (color or standard), the researchers could include a second explanatory variable in the study. For example, the researchers could test each student's ability before the study, group students into experienced and inexperienced groups, and then, within each experience group, randomly assign the type of game each student should play.

2.6 Random Sampling and Random Allocation

There is one more type of variability that is not included in the statistical model: **unplanned systematic variability**. This variability is caused by extraneous variables that can bias the results. **Extraneous variables** (such as time of day, prior computer game experience of the subject, number of pegs in the game, or amount of background noise) are not of interest in our study, but they may have an influence on the completion time of the game.

Essentially all studies have numerous extraneous variables that may be biasing the results. The problem is that we typically do not know all possible extraneous variables in a study or if they are biasing the results. **Random sampling** and **random allocation** are used to protect against the unwanted influence of extraneous variables:

- *Random sampling:* How was the sample collected? If the subjects in the sample were randomly selected from the population of interest, inferences can be drawn (generalized) to the entire population.

- *Random allocation:* How were units assigned to treatments? If the units were randomly allocated to treatment groups, a statistically significant result in a well-designed study shows that the treatment *causes* changes in the response variable.

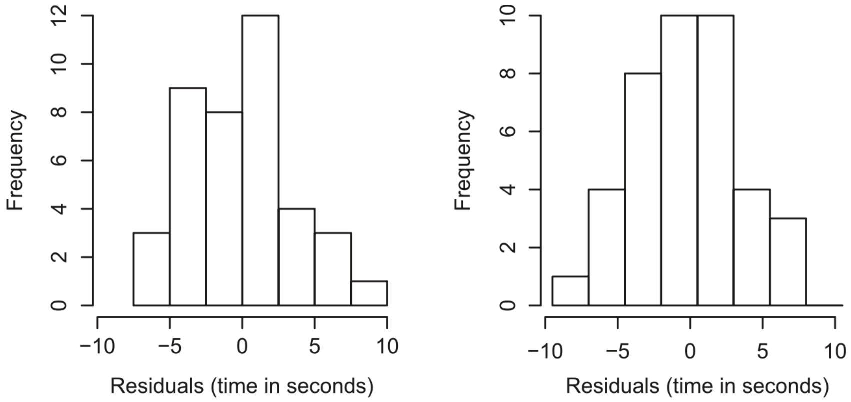

Figure 2.4: Dotplots representing data from two studies. The difference between the group means is the same in both data sets, but the random variation is not the same. The variability in the residuals is much larger for 'Results A' than for 'Results B'.
(#fig:fig2.4)

In the computer game study, students were "randomly" selected from the college. If the 40 students were truly a simple random sample of all students currently attending the college, the results of this study would hold for all students in the college. However, even if the researchers used a **sampling frame** (list of the population of all current students at their college) to randomly select 40 students, it would be unlikely that the first 40 subjects selected would agree to participate in the study. Thus, the population for the study would be all current college students who would agree to participate in the study. If the researchers' version of "random sample" meant a collection of friends who agreed to participate in their study, the conclusions would hold only for the 40 students who volunteered.

The key point is that it is often very difficult to collect a true simple random sample from the population. If the sample is not reflective of the entire population (an appropriate random sample is not collected), the result may contain biases which may invalidate the results.

Random allocation is much easier to do appropriately in this study. Simply flipping a fair coin is enough to randomly assign subjects to a particular type of game. Therefore, since the sample data led us to reject the null hypothesis, we can be quite certain that the type of game *caused* a difference in the average completion time.

2.7 What Can We Conclude from the Game Study?

Validation of model assumptions is essential before drawing conclusions from hypothesis tests. The residual plots created throughout this chapter appear to support the model assumptions. There are no clear trends or outliers in the residual plots. In general, the graphs do not give enough evidence to reject the assumption that the error terms are normally distributed with a mean of zero and a constant variance.

The p-value for all three hypothesis tests is 0.0279. When we assume the null hypothesis is true in an experiment, we are assuming that there is nothing creating group differences except the random allocation process. Under this assumption, a group difference at least as extreme as the one actually observed would occur only 2.79% of the time. This allows us to conclude that the type of game does cause a difference in completion times.

Under the conditions of this computer game study, we have shown that the statistical model and assumptions for the two-sample t-test (assuming equal variances), regression, and ANOVA models are mathematically equivalent. Thus, testing if there is a difference between means, if the regression slope is not zero, or if the factor effects are significant will lead to the same conclusion because they are exactly the same test.

The tests in this computer game study are identical because there were only two groups (two levels) of one explanatory variable and we assumed the variances of both groups were equivalent. Under these conditions, any of the three tests can be used to draw appropriate conclusions as long as the model assumptions are met. The extended activities and end-of-chapter exercises provide more details about the differences among the three tests.

A Closer Look: Statistical Models

2.8 Normal Probability Plots to Assess Normality

Figure 2.5 shows two histograms of the residuals calculated from the Games1 data. Both histograms use the same data and the same class widths (width of each bar); the only difference is that the bins start at different positions. Note that the histogram on the left looks somewhat skewed while the right graph is fairly symmetric.

These graphs are provided to illustrate that histograms are not always reliable for determining whether the residuals come from a normal distribution. Histograms are especially unreliable with small data sets, where the choice of class sizes can have a significant effect on the appearance of the graph.

An alternative to histograms is normal probability plots. A **normal probability plot** is a scatter-plot of observed data versus the corresponding percentiles of the normal distribution. If the scatterplot forms a straight line, the percentiles of observed data match the percentiles of a normal distribution and we make the assumption that the observed data could have come from a population with a normal distribution.

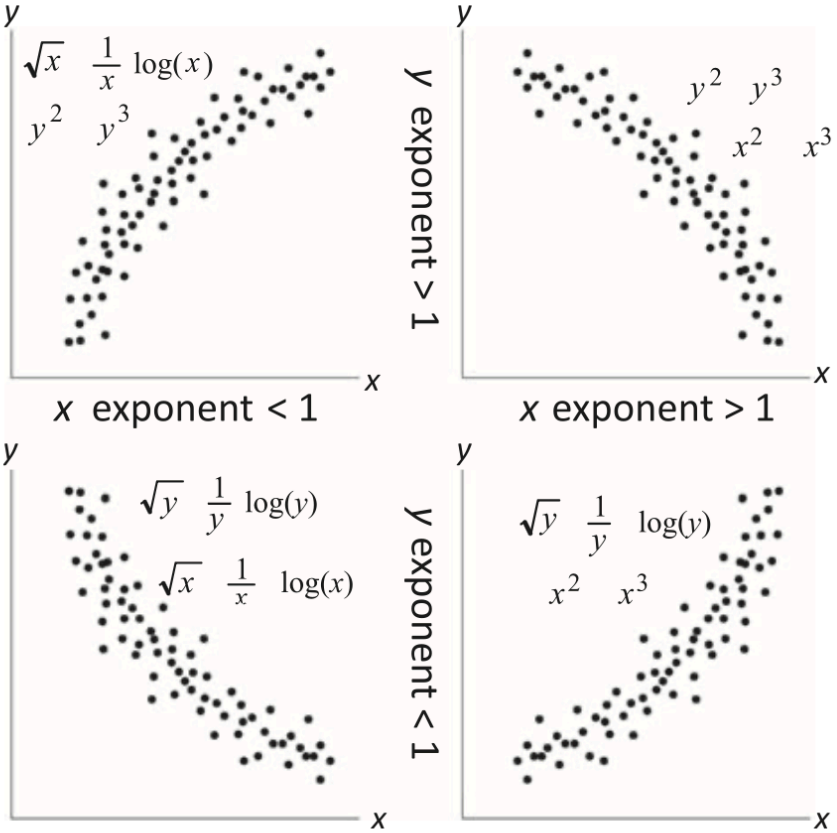

Figure 2.5: Two histograms of the computer game study residuals. (#fig:fig2.5)

Extended Activity: Creating Probability Plots

Data set: *Normal*

The following questions ask you to work through the process of calculating and interpreting probability plots. >28. **Calculating a Normal Probability Plot by Hand**
>Consider the following sample data set of $n = 5$ observations: 14, 11, 17, 15, 13. Complete the following steps to create a normal probability plot.
> a. Sort the data from smallest to largest. Use a subscript in parentheses, (i), to represent the ordered data. For example $y_{(1)} = 11$ is the smallest observation and $y_{(5)} = 17$ is the largest observed value.
> b. For each (i), calculate the $(i - 0.5)/n$ percentile of the standard normal distribution. For example, corresponding to $(i) = (1)$, the $(1-0.5)/5 = 10$th percentile of the standard normal distribution is -1.28, since $P(Z \leq -1.28) = 0.10$ when $Z \sim N(0, 1)$. For $(i) = (3)$, the $(3 - 0.5)/5 = 50$th percentile (i.e., the median) of the standard normal distribution is 0. Repeat this process for the other ordered values, $y_{(2)}, y_{(4)}$, and $y_{(5)}$.
> c. Make a normal probability plot by creating a scatterplot with the percentiles of the observed data along the x-axis and the percentiles of the standard normal distribution along the y-axis. If the data fall along a straight line, then the data are consistent with the hypothesis that they are a random sample from a population that is normally distributed.
> The data in this question are a little "heavier" toward the tails (the normal distribution has more observations in the center and fewer observations toward the tails than does this data set), so the probability plot has an S-shaped curve. With only five data points, the shape is not as clear as it would be for a data set with a larger sample size from a "heavy-tailed" population.
> d. If you standardized the data (subtracted the sample mean and then divided by the sample standard deviation), would you expect the shape of the normal probability plot to change?
> e. Does the shape of the normal probability plot change if you multiply each observation in the sample data set by 5?
> f. Does the shape of the normal probability plot change if you divide each observation in the sample data set by 3?
>29. **Plotting Normal Data** For this problem, use the Normal data set. The first column of data actually is a random sample from a normal distribution.
> a. Use software to create a histogram and normal probability plot of the first column of the Normal data set.
> b. Double the five largest observed values in the Normal data set. Create a histogram and normal probability plot of the "Largest 5 Doubled" data. Describe how the normal probability plot and the histogram change.
> c. Now, double the five smallest observed values in the original Normal data set. Create a histogram and normal probability plot of the "Smallest 5 Doubled" data. Describe how the normal probability plot and the histogram change.
> d. Draw (by hand) a picture of what a normal probability plot might look

like for a data set with *fewer* observations in both tails than you would expect from a normal distribution.
> e. Draw (by hand) a picture of what a normal probability plot might look like for a data set with *more* observations in both tails than you would expect from a normal distribution.

> **NOTE**
> As with any other hypothesis test, when we fail to reject H_0 we do not prove H_0 is true. Normal probability plots cannot be used to prove that the data came from a normal distribution (H_0), but they can be used to show that the data are consistent with data from a normal population.

Assessing whether or not data could have come from a normal population by examining a normal probability plot requires some experience and may seem subjective. After all, even when data do come from a normal population, sampling variability (random variability) will sometimes lead to a normal probability plot where the data do not lie along a straight line.

Extended Activity: Understanding Variability in Random Samples

Data set: *Games*1

30. If you have no experience with probability plots, it can be helpful to look at several sample data sets that actually do come from a normal distribution. Use software to draw several random samples of the same size from an actual normal population and create normal probability plots. These plots can be compared to the normal probability plot from the actual data. If the real data plot looks similar to the plots where you know that the population is normal, then your data are consistent with the null hypothesis (i.e., the data came from a normal population). If the real data plot is "extreme" (quite different from the plots coming from a normal population), then the differences are not likely due to chance and you can reject the hypothesis that the data came from a normal population.

   a. Create a normal probability plot of the residuals of the Games1 data from Question 26.

   b. Use software to draw a random sample of $n = 40$ from an actual normal probability distribution with mean 0 and standard deviation 1. Create a normal probability plot of the sample data.

c. Repeat the previous question eight more times, for a total of nine normal probability plots of "data" from an actual normal probability distribution. Does the plot in Part A resemble the nine plots with data sampled from a normal distribution? If you can't distinguish the Games1 residuals from the other plots, it would seem reasonable to assume that the Games1 residuals are normally distributed.

2.9 Transformations

Transformations for ANOVA

It is common for data to not follow a normal distribution or for subgroups to have dramatically different variances. For example, in biological studies it is common for subgroups with larger means to also have larger variances. Consider measuring the weights of various animal species. We expect the weights of mice to have less variability than the weights of elephants, as measurement instruments often get less precise (more variable) as the measurements get larger.

In these types of situations, the data can often be transformed to fit model assumptions. Transformations are monotonic mathematical operations that change the scale of the explanatory variable, the response variable, or both. When groups within a data set have unequal variances or when data are skewed to the right, a square-root or natural-logarithm transformation on the response variable can often change the data to a scale where the equal variance and normality assumptions are more closely satisfied. Then the transformed data can be analyzed using traditional techniques such as ANOVA or regression.

> **MATHEMATICAL NOTE**
> Monotonic functions preserve the order of the original data. A monotonic increasing function maintains the direction of the data: For any two data points, when $y_i > y_j$ then $f(y_i) > f(y_j)$. A monotonic decreasing function reverses the direction of the data: For any two data points, when $y_i > y_j$ then $f(y_i) < f(y_j)$. If the transformation is not monotonic over the range of sample data (i.e., if the data set contains zeros or negative numbers), simply add a constant to each number to make all numbers positive or nonzero before transforming the data.

Although an infinite number of transformations could be tried, it is best to focus on commonly used transformations such as the ones listed below:

The **square-root transformation** ($y^{1/2} = \sqrt{y}$) is commonly used when the response variable represents counts, such as a count of the number of observed species. Square-root transformations are also very useful when the variances are proportional to the means.

The **log transformation** is often used when the data represent size or weight measurements. In addition, it is useful when the standard deviations are proportional to the means. A common logarithm is based on the number 10 and written $\log_{10}(x)$. This log is defined as $\log_{10}(10^x) = x$. The natural logarithm, $\ln(x)$, is based on the number $e = 2.71828$, so $\ln(e^x) = x$. For statistical tests, it makes no difference whether you use log base 10 ($\log_{10}$) or natural logs (ln), because they differ only by a constant factor. The log base 10 of a number equals 2.303 times the natural log of the number. Log transformations are often preferred over other transformations because the results tend to be easier to interpret.

The **reciprocal transformation** ($y^{-1} = 1/y$) is often useful when the data represent waiting times, such as time until death or time until a battery fails. If most responses are relatively close to zero but a few responses are larger, this transformation will reduce the effect of large response values.

The **arcsine transformation** ($\sin^{-1}(\sqrt{y})$) and **logit transformation** ($\log[y/(1-y)]$) are useful when measurements are proportions between 0 and 1. The arcsine transformation is often difficult to interpret and cannot be subjected to back transformation (described in the next section) to produce an informative interpretation. The logit function can be usefully interpreted and will be discussed in much more detail in Chapter 7.

Extended Activity: Transforming Emissions Data

Data set: Emission

31. The data set Emission provides hydrocarbon emission in parts per million (ppm) at idling speed for cars, based on the year each car was manufactured. These data were randomly sampled from a much larger study on pollution control in Albuquerque, New Mexico.

a. Create individual value plots or side-by-side boxplots of Emission versus Year. Compare the mean and standard deviation of each group. Do the data within each group look consistent with data from a normal population?

b. Transform the response by taking the log of Emission. Create individual value plots or side-by-side boxplots of log(Emission) versus Year. Compare the plot of the transformed data to the plot in Part A. Which plot shows data that better fit the model assumptions?

c. Calculate an ANOVA table, F-test, and p-value to determine if the average log(Emission) varies based on Year. Note that the end-of-chapter exercises and Section 2.9 show that ANOVA can compare more than two groups. In this question, $I = 5$ groups instead of 2 groups. However, the model and calculations are identical except that now $i = 1, 2, 3, 4, 5$ instead of $i = 1, 2$. The null hypothesis is $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ versus the alternative $H_a :$ at least one group mean is different from another.

d. Create residual plots to evaluate whether the model assumptions for the F-test are violated. Notice that although the log transformation was helpful, the data still have outliers. In addition, the equal variance and normality assumptions are still slightly violated. Some statisticians would consider the log-transformed data appropriate for the standard ANOVA. Others would try another transformation, such as taking the log of the transformed data again; this is called a log log transformation. Still others would suggest using a nonparametric test. (Nonparametric tests, such as the Kruskal-Wallis test, are described in Chapter 1.) Nonparametric tests do not require error terms to follow the normal distribution. While any of these analyses would be appropriate, it would not be appropriate to conduct several analyses on the same data and then report only the conclusions corresponding to the test that gave the smallest p-value. For example, if we tried three or four hypothesis tests each with an α-level $= 0.10$ and then simply picked the test with the smallest p-value, our chances of incorrectly rejecting the null hypothesis would actually be greater than 10%.

If there are very clear outliers, if data are skewed, or if the subgroup variances are clearly different, a transformation applied to the response variable may help the data fit the ANOVA model assumption. When the normality or equal variance assumption does not hold, the one-way ANOVA F-test still tends to be a fairly accurate method if there are equal sample sizes. The F-statistic is much less reliable when there are unbalanced sample sizes and one or more subgroups have a much larger variance than others.[3]

Back Transformations

Transformations are not simply a way of playing around with the data until you get the answer you want. It is important to recognize that there is no reason to believe that the original scale used for the measurements is better than other scales. For example, in testing for differences in lengths, should it matter if the original data were collected in meters or in feet? One scale is not better than the other; we transform data simply so that it is easier for our audience to interpret.

Some scales, such as pH levels,[6] are always presented using a logarithmic scale.

For testing for differences between groups, the p-values of transformed data are reliable as long as model assumptions are satisfied. However, other statistical results, such as confidence intervals or the slope coefficient, are typically best understood in the original units. Thus, it is often desirable to back transform the results. **Back transformations** do the opposite of the mathematical function used in the original data transformation. For example, if the natural log transformation was used, a back transformation is conducted by taking the exponent of the number. Unfortunately, it can be very difficult to interpret some statistical results in either the transformed or the back-transformed scale.

Consider conducting a t-test for the difference between the mean car emissions for the pre-63 and the 70–71 groups in Question 31. The standard deviation of the pre-63 group, 592, is more than double that of the 70–71 subgroup, 287.9. We will again assume equal variances in our t-test, but even if a different t-test were chosen, there would be clear evidence of nonnormality. Taking the natural log of Emission addresses the nonnormality problem and also makes the standard deviations very similar. The standard deviation is 0.57 for the transformed pre-63 group and is 0.678 for the transformed 70–71 group. The two-sided hypothesis test gives a p-value of 0.001, which provides strong evidence that there is a difference between group means. This test is valid since the model assumptions are met.

The 95% confidence interval for the transformed data is (−1.434, −0.411). However, this transformed confidence interval is not easy to interpret in terms of actual car emissions. The back-transformed confidence interval is

$$(e^{-1.434}, e^{-0.411}) = (0.238, 0.663)$$

Note that the confidence limits are no longer symmetrical. In addition, this confidence interval no longer is interpreted as the difference between two means, but now represents the confidence interval for the ratio between the two means. The end-of-chapter exercises provide additional examples of interpreting results on the transformed scale (and back-transformed scale).

> **CAUTION**
> The back-transformed data do not have the same meaning as the original raw data. For two log-transformed means, $\ln(\bar{y}_1) - \ln(\bar{y}_2) = \ln(\bar{y}_1/\bar{y}_2)$. Thus, back transforming the data ($e^{\ln(\bar{y}_1/\bar{y}_2)} = \bar{y}_1/\bar{y}_2$) results in the ratio of the two means. Results of back transformations based on the square-root, reciprocal, and arcsine transformations often have no practical interpretation.

> **Key Concept**
> It can be difficult to properly interpret slope coefficients or confi-

[6] pH is a measure of how acidic or how basic (alkaline) a solution is. It is measured as the negative logarithm (base 10) of the molar concentration of dissolved hydronium ions.

dence intervals using either transformed or back-transformed data. Hypothesis tests for differences between groups do not need to be back transformed.

Transformations for Regression

As with ANOVA, there are many situations in regression in which data are skewed, outliers exist, or the variability of the residuals tends to depend on the explanatory variable. Graphing the residuals is often the best way to identify the appropriate transformations. If the statistical model is correct, then no clear patterns (such as a strong curve or fan shape) should be seen in the plots. When there is no clear pattern in the residual plots, it is safe to assume that no statistical model based on the same explanatory variable will be a better fit for the data.

Extended Activity: Transforming Brain and Body Weight Data

Data set: $Weight$

32. The Weight data set contains the brain weights (y) and body weights (x) of 32 species.

a. Create a scatterplot of y versus x with a regression line ($\hat{y} = b_0 + b_1x$), a plot of residuals versus the explanatory variable, a plot of residuals versus predicted (or "fitted") values ($\hat{y}$), and either a normal probability plot or a histogram of the residuals.

b. Try various transformations of the explanatory and response variables to create a better linear regression model. Hint: Notice that since both the x and the y variable are right skewed and have outliers, both may need a transformation.

Choosing the Right Transformation

When a scatterplot of the data reveals a curved (nonlinear) shape, transformations are often used to straighten curved relationships so that a simple linear regression model will fit the data. In some cases, theoretical knowledge or previous studies can provide an indication of a suitable transformation. More formal methods, such as the Box-Cox method and the Box-Tidwell method,[4] can also be used to choose a transformation. However, the best indicators of an appropriate transformation are often found by viewing scatterplots and residual plots of the data.

Mosteller and Tukey introduced the ladder of powers and the bulging rule as a way to choose among the family of power transformations.[5] The following list of transformations is often referred to as the *ladder of powers* because power and logarithmic functions have a natural hierarchy:

$$\cdots,\ y^2,\ y^{-1},\ y^{-1/2},\ \log(y),\ y^{1/2},\ y,\ y^2,\ \cdots$$

Notice that $\log(y)$ replaces the transformation $y^0 = 1$, since setting everything to a constant value is not useful. Exponents greater than one will cause the function to increase at a faster rate. Exponents less than one (and the log) will cause the function to bend downward. The curves become progressively steeper (sharper) as the exponent moves away from one.

The bulging rule provides a visual method for determining appropriate transformations. Figure 2.6 shows four different curves (bulges) and indicates which powers of y and x would likely straighten the line. For example, the upper left quadrant of Figure 2.6 shows a curve that tends to become more linear if y is transformed to a power greater than one (such as y^2 or y^3) and x is transformed to a power less than one (such as $\sqrt{x}$, $\log(x)$, or x^{-1}).

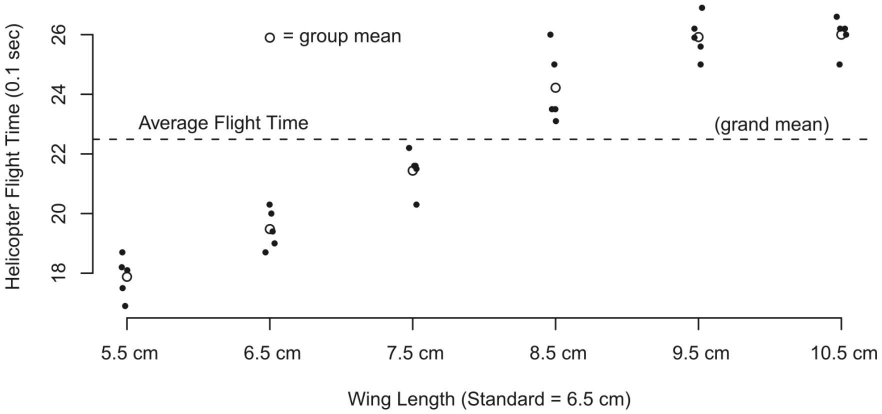

Figure 2.6: Bulge rule showing appropriate transformations to linearize curved data.
(#fig:fig2.6)

Performing a transformation to control problems with unequal variances can increase the nonlinearity between the explanatory and response variables. Transforming the response variable influences both the variation and the linearity, but transforming the explanatory variable influences only the linearity. Thus, it is best to transform the response variable first to deal with nonconstant variance and then consider additional transformations on the explanatory variable to make the model linear. The following steps are useful for choosing an appropriate transformation:

- Create a scatterplot of the original data.
- Use the ladder of powers or other methods to select a transformation for the explanatory variable, response variable, or both.
- Create a scatterplot of the transformed data. If the scatterplot is not linear, try a new transformation. If the scatterplot is linear, conduct the appropriate statistical analysis and create residual plots.
- If the residual plots are not random, try another transformation. If the residuals do appear random (the model assumptions about the error term are satisfied), then the statistical analysis is reliable.

Often there are no appropriate transformations that will satisfy all the model assumptions. Future chapters discuss more advanced techniques that can be used to allow for nonnormal residuals and for nonlinear relationships.

Extended Activity: Comparing Four (x, y) Data Sets

Data set: *RegrTrans*

33. Do the following for each of the four data sets:

    a. Create a scatterplot of y versus x with a regression line ($\hat{y} = b_0 + b_1x$), a plot of residuals versus the explanatory variable, a plot of residuals versus predicted (or "fitted") values ($\hat{y}$), and either a normal probability plot or a histogram of the residuals.

    b. By hand, sketch on the scatterplot a curve that would fit the data better than the regression line. Notice that the plot of residuals versus the explanatory variable emphasizes the patterns in the residuals much better than does the scatterplot of y versus x.

    c. Try various transformations of the explanatory and response variables to create a better linear regression model (as validated by graphical analysis of the residuals).

2.10 Calculating Test Statistics

This section is a rather terse description of the mathematical calculations behind the hypothesis tests and confidence intervals described in this chapter. Most introductory textbooks will dedicate an entire chapter to each of these techniques. The logic behind the calculations for regression and ANOVA will be described in more detail in later chapters of this text.

The Two-Sample t-Test with the Equal Variance Assumption

The two-sample t-test can be used to test whether two population means are equal. The null hypothesis about the population means ($H_0 : \mu_1 = \mu_2$) is rejected if the difference between the sample means, $\bar{y}_1$ and $\bar{y}_2$, is so large that it doesn't appear reasonable to assume that the groups have the same mean.

The test statistic for the two-sample t-test is

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \textit{where} \quad s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{2.7}$$

The above test statistic is a function of the following summary statistics from the sample data:

$$\bar{y}_1 = \bar{y}_{1.} = \frac{1}{n_1}\sum_{j=1}^{n_1} y_{1,j} \tag{2.2}$$

$$\bar{y}_2 = \bar{y}_{2.} = \frac{1}{n_2}\sum_{j=1}^{n_2} y_{2,j} \tag{2.3}$$

$$s_1 = \sqrt{\frac{1}{n_1 - 1}\sum_{j=1}^{n_1}(y_{1,j} - \bar{y}_{1.})^2} \tag{2.4}$$

$$s_2 = \sqrt{\frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(y_{2,j} - \bar{y}_{2.})^2} \tag{2.5}$$

The difference in population means ($\mu_1 - \mu_2$) is not known, but comes from the statement of the null hypothesis: $\mu_1 - \mu_2 = 0$ (or, equivalently, $\mu_1 = \mu_2$). Thus, the test statistic is simply a ratio of the distance between the two sample means to a measure of variation.

The **pooled standard deviation** (denoted s_p) uses a weighted average of the two sample variances in order to estimate the size of the variation in a typical random error term (i.e., σ, the common standard deviation for the two populations).

Probability theory can be used to prove that if the model assumptions are true, the t-statistic in Equation **??** follows a t-distribution with $(n_1 + n_2 - 2)$ degrees of freedom. If the t-statistic is large, the difference between the two means is large compared to the pooled standard deviation. We will reject the null hypothesis that the two means are equal (H_0: $\mu_1 = \mu_2$) in favor of H_a: $\mu_1 \neq \mu_2$ if the t-statistic is so large that it is unlikely to occur when $\mu_1 = \mu_2$. A large t-statistic corresponds to a small enough p-value, which is found with software or in a t-table.

Extended Activity: Calculating the Two-Sample t-Test

Data set: *Games*1

34. Use Equation **??** to calculate the test statistic (t) by hand (i.e., without statistical software) for the computer game study. Use software or a t-table with $(n_1 + n_2 - 2)$ degrees of freedom to find the p-value.

Regression

Introductory statistics textbooks describe how least squares techniques can be used to calculate the following statistics to estimate β_0 and β_1:

$$b_1 = \hat{\beta}_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}, \quad b_0 = \hat{\beta}_0 = \frac{\sum y_i - b_1 \sum x_i}{n} \tag{2.8}$$

where $n = n_1 + n_2$. In most introductory statistics texts, Equations **??** for the slope and intercept are simplified to

$$b_1 = r\frac{s_y}{s_x} \qquad \text{and} \qquad b_0 = \bar{y} - b_1 \bar{x}$$

where the sample correlation coefficient is

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{y_i - \bar{y}}{s_y} \right) \left(\frac{x_i - \bar{x}}{s_x} \right)$$

To test the null hypothesis $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$, it can be shown that the t-statistic for the slope coefficient is

$$t = \frac{b_1 - \beta_1}{\hat{\sigma}\sqrt{1/\sum_{i=1}^{n}(x_i - \bar{x})^2}} \quad \text{where} \quad \hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}[y_i - (b_0 + b_1 x_i)]^2}{n-2}} \tag{2.9}$$

Notice that $\hat{\sigma}$ is an estimate of the standard deviation of the random errors. If the sample statistic b_1 is far away from $\beta_1 = 0$ relative to the size of the estimated standard deviation, $\hat{\sigma}$, then the t-statistic will be large and the corresponding p-value will be small.

Probability theory can be used to prove that if the regression model assumptions are true, the t-statistic in Equation **??** follows a t-distribution with $n - 2 = n_1 + n_2 - 2$ degrees of freedom.

Extended Activity: Testing the Slope Coefficient

Data set: *Games*1

35. Without statistical software, use summary statistics and Equation **??** to calculate the test statistic under the null hypothesis that $\beta_1 = 0$. Use software or a t-table with $n-2 = n_1 + n_2 - 2$ degrees of freedom to find the p-value.
36. Compare the test statistic and p-values in Questions 34 and 35.

Analysis of Variance (ANOVA)

Several calculations will be made to test the hypothesis $H_0 : \mu_1 = \mu_2$ in ANOVA, but again the test statistic is a ratio of the spread between group sample means to the variability in the residuals.

If indeed $\mu_1 = \mu_2$ (i.e., $H_0 : \alpha_1 = \alpha_2 = 0$ is true), then we would expect the variation between the level means to be relatively small compared to the variability in the error terms. If the group means are relatively far apart, the F-statistic will be large and we will reject $H_0 : \mu_1 = \mu_2$ in favor of $H_a : \mu_1 \neq \mu_2$. While the logic is similar to that for the other tests described in this chapter, the test statistic for ANOVA, the F-statistic, requires many more calculations, as shown below.

Sums of squares (SS) are measures of spread, calculated in an ANOVA table like the one you saw in the software output for Questions 23 through 25. The three sums of squares calculated for the ANOVA table for the computer game study are described below.

Group sum of squares ($SS_{\textbf{Group}}$) measures the difference between group means (also called level means). Group sum of squares represents the variability we want in the model. For the computer game, SS_{Group} measures the spread between the two game type means, but ANOVA can be extended to more than just two groups.

Recall that the ith level mean is denoted as $\bar{y}_{i.}$, the grand mean is denoted as $\bar{y}_{..}$, and the corresponding level effect is $\hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..}$.

$$SS_{\text{Group}} = \sum(\text{each level effect})^2 = \sum_{i=1}^{I} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$$

(where \$I = \$ number of groups or levels)

$$= \sum_{i=1}^{2} 20 \times (\bar{y}_{i.} - \bar{y}_{..})^2 \qquad \text{for the computer game study}$$
$$= 20 \times (\bar{y}_{1.} - \bar{y}_{..})^2 + 20 \times (\bar{y}_{2.} - \bar{y}_{..})^2$$

Error sum of squares ($SS_{\textbf{Error}}$) measures the spread of the observed residuals. Recall that each residual is defined as an observed value minus the estimated mean response: $\hat{e}_{i,j} = y_{i,j} - \bar{y}_{i.}$. In any ANOVA model with one explanatory variable, the mean response is the level average.

$$SS_{\text{Error}} = \sum(\text{each residual effect})^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{i,j} - \bar{y}_{i.})^2$$
$$= \sum_{i=1}^{2}(n_i - 1) \times s_i^2 \qquad \text{since } s_i^2 = \frac{\sum_{j=1}^{n_i}(y_{i,j} - \bar{y}_{i.})^2}{n_i - 1}$$
$$= \sum_{i=1}^{2}(19) \times s_i^2 \qquad \text{for the computer game study}$$
$$= 19 \times s_1^2 + 19 \times s_2^2$$

Total sum of squares ($SS_{\textbf{Total}}$) measures the overall spread of the responses in the full data set.

$$SS_{\text{Total}} = \sum(\text{distance between each observation and the grand mean})^2 \quad (2.6)$$
$$= \sum_{i=1}^{I}\left[\sum_{j=1}^{n_i}(y_{i,j} - \bar{y}_{..})^2\right] \quad (2.7)$$
$$= (n-1) \times s^2 \qquad \text{since } s^2 = \frac{\sum_{i=1}^{I}\left[\sum_{j=1}^{n_i}(y_{i,j} - \bar{y}_{..})^2\right]}{n-1} \quad (2.8)$$
$$= (39) \times s^2 \qquad \text{for the computer game study} \quad (2.9)$$

Here, s^2 is the overall sample variance and $n = \sum_{i=1}^{I} n_i$ = total sample size.

It can be shown that $SS_{\text{Total}} = SS_{\text{Group}} + SS_{\text{Error}}$. While the specific formula for SS_{Error} is provided above, it is most easily calculated by subtracting SS_{Group} from SS_{Total}.

Degrees of freedom (df) for each sum of squares are calculated based on how many "free" pieces of information are summed. In this example, there are two levels of the game type factor. We can show that the weighted Type effects must sum to 0 ($n_1\hat{\alpha}_1 + n_2\hat{\alpha}_2 = 0$). This implies that knowing the color distracter game effect automatically forces a known effect for the standard game. Thus, SS_{Group} has only $I - 1 = 2 - 1 = 1$ df. SS_{Total}, like the usual one-sample variance, has $n - 1 = 40 - 1 = 39$ df. It can also be shown that $df_{\text{Total}} = df_{\text{Type}} + df_{\text{Error}}$: thus, $df_{\text{Error}} = df_{\text{Total}} - df_{\text{Type}} = (n-1) - (I-1) = n - I = n - 2 = 38$.

Mean squares (MS) are measures of "average" spread and are calculated by dividing a sum of squares (SS) by its associated degrees of freedom (df).

Group Mean squares ($MS_{\textbf{Group}}$) equals $SS_{\text{Group}}/df_{\text{Group}}$. MS_{Group} is a measure of variability between the levels of each factor and is often called **between-level variability**. It is actually just the variance of the level means.

Mean square error (MSE) equals $SS_{\text{Error}}/df_{\text{Error}}$. MSE is a pooled measure of the variability within each level, or the **within-level variability**. Remember that the variance is assumed to be the same for the responses within each level of the factor: $\sigma_1^2 = \sigma_2^2 = \sigma^2$. When $I = 2$, MSE is identical to the pooled variance (s_p^2) used in the two-sample t-statistic in Equation **??**.

If MS_{Group} is much larger than MSE, it is reasonable to conclude that there truly is a difference between level means and the difference we observed in our study was not simply due to chance variation (random error).

The F-statistic (MS_{Group}/MSE) is a ratio of the between-level variability to the within-level variability. If indeed $\mu_1 = \mu_2$ (i.e., $H_0 : \alpha_1 = \alpha_2 = 0$ is true), then we would expect the variation between the level means in our sample data to be about the same as the typical variation within levels and the F-statistic would be close to one. Larger values of the F-statistic would imply that the level means were farther apart than chance error alone could explain. These calculations are often summarized in an **ANOVA table**, as shown in Table 2.1.

[[[The table is not displaying. But it works when I knit on separate pdf]]]

Probability theory can be used to prove that if the model assumptions are true, the F-statistic follows an F-distribution with df_{Group} and df_{Error} degrees of freedom. The p-value gives the likelihood of observing an F-statistic at least this extreme (at least this large), assuming that the population means of the two game types are equal. Thus, when the p-value is small (e.g., less than 0.05 or 0.01), the effect size of the type of game is conventionally determined to be statistically significant.

Table 2.1: (#tab:tab2.1)Table 2.1 One-factor (”one-way”) ANOVA table (one factor with I levels).

Source	df	SS	MS	F-Statistic
Group	$I-1$	$\sum_{i=1}^{I} n_i(\overline{y}_i - \overline{y}_{..})^2$	$\frac{SS_{\text{Group}}}{df_{\text{Group}}}$	$\frac{MS_{\text{Group}}}{\text{MSE}}$
Error	$n-I$	$\sum_{i=1}^{I} (n_i - 1)s_i^2$	$\frac{SS_{\text{Error}}}{df_{\text{Error}}}$	
Total	$n-1$	$\sum_{i=1}^{I}\sum_{j=1}^{n_i} (y_{i,j} - \overline{y}_{..})^2$		

Extended Activity: Calculating an ANOVA Table

Data set: *Games*1

37. Use n_1, n_2, $\bar{y}_1$, $\bar{y}_2$, $\bar{y}_{..}$, s_1, and s_2 to calculate SS_{Type} (i.e., SS_{Group}), SS_{Error}, MS_{Type}, and MSE for the computer game study.
    Since the group for the computer game study is game type (the Type variable in the data set), we will use the more descriptive labels MS_{Type}, SS_{Type}, and df_{Type} instead of MS_{Group}, SS_{Group}, and df_{Group}. This is common practice and is similar to how statistical software reports results based on variable names.
38. Calculate the overall variance of the completion times in the entire data set and use it to find the total sum of squares (SS_{Total}) for the computer game study. Confirm that $SS_{\text{Total}} = SS_{\text{Type}} + SS_{\text{Error}}$ for the computer game study.
39. Calculate the F-statistic and use software or an F-distribution table with df_{Type} and df_{Error} degrees of freedom to find the p-value.

2.11 Confidence Intervals

In previous sections, hypothesis tests were used with sample data to assess the evidence for a claim about a population. They were used to determine if there was evidence to support the claim that the two population means were different. An alternative approach is to use confidence intervals to create an interval estimate of a population parameter (such as the difference between two population means) and provide a level of confidence in the interval estimate.

Each confidence interval discussed in this chapter has the following form:

estimate ± critical value × standard error of the estimate

The estimate is a sample statistic used to estimate the population parameter. In the computer game study, a confidence interval for the population mean μ_1 would have an estimate of $\bar{y}_1$. A confidence interval for $\mu_1 - \mu_2$ is estimated by $\bar{y}_1 - \bar{y}_2$.

A **confidence level** is the probability that the true parameter value will be captured by the confidence interval. In other words, a 95% confidence level ensures that the method used to calculate the confidence interval will successfully contain the true parameter value 95% of the time.

The **critical value** is a value from a distribution that is used to provide a confidence level for the interval. The critical values used for the two-sample t-test and regression are based on the t-distribution. The same model assumptions are used in both hypothesis tests and confidence intervals. Thus, the same distribution and degrees of freedom are used in the hypothesis test and confidence interval.

For the game study, a t-distribution with 38 degrees of freedom is used. The critical value t^*_{38} for a particular confidence level C is chosen so that C% of the area under the t-distribution is between $-t^*_{38}$ and t^*_{38}. For example, for a 95% confidence level ($C = 95$), $t^*_{38} = 2.02$, since 95% of the area under a t-distribution with 38 df is between -2.02 and 2.02.

If the confidence level were chosen to be 99% ($C = 99$), the confidence interval would be wider than before. A wider interval would have a higher probability of capturing the true mean. The critical value for a 99% confidence interval for the game study is $t^*_{38} = 2.71$.

The **standard error of the estimate** is a measure of the variability of the statistic. For example, a 95% confidence interval for μ_1 is

$$\bar{y}_1 \pm t^*_{19} \times \hat{\sigma}_{\bar{y}_1}$$

$$38.1 \pm 2.09 \times \frac{3.65}{\sqrt{20}}$$

$$(36.39, 39.81)$$

A 95% confidence interval for $\mu_1 - \mu_2$ is

$$\bar{y}_1 - \bar{y}_2 \pm t^*_{38} \times \hat{\sigma}_{\bar{y}_1 - \bar{y}_2}$$

$$38.1 - 35.55 \pm 2.02 \times \sqrt{\frac{193.65^2 + (19)3.39^2}{20 + 20 - 2}}$$

$$(0.29, 4.81)$$

Note that the "standard error of the estimate" in the confidence interval is identical to the denominator in the test statistic for the corresponding hypothesis test.

Extended Activity: Calculating a Confidence Interval for the Regression Coefficient

Data set: *Games1*

40. Note that the "standard error of the estimate" in the confidence interval is identical to the denominator in the test statistic for the corresponding hypothesis test. Use this information to *write out the formula* for a 95% confidence interval for β_1. Use the output from the corresponding t-test to find the estimate, the critical value, and the standard error of the regression coefficient. Use this information to calculate a 95% confidence interval for β_1.

MATHEMATICAL NOTE
The F-distribution used in ANOVA is not a symmetric distribution and all values are positive. Confidence intervals for the difference between two means are not calculated with an F-distribution. Note that the MSE in ANOVA $= s_p$ from the two-sample t-test and the square root of the critical value $\sqrt{F_{1,38}} = t^*_{38}$.

Key Concept
Statistics based on sample data are estimates of population parameters. A confidence interval allows us to calculate an interval that has probability C (often $C = 95\%$) of containing the true population parameter.

Chapter Summary

When there is only one explanatory variable (factor) with two levels, a t-test is typically used to analyze the data. While this chapter has shown that ANOVA or regression analysis techniques provide equivalent results in this setting, they are typically used with different study designs.

The tests used in ANOVA and regression are developed under the assumption that the variances of each group are equal, while a two-sample t-test could be used without this assumption. Two-sample t-tests are limited to comparing two groups, while ANOVA and regression techniques are often used to analyze data sets with multiple explanatory variables, each having many levels. All three techniques are used when the response variable is quantitative.

While the t-test for β_1 is appropriate, the scatterplot and regression line created in Question 15 show that a regression model does not accurately describe the data. For example, it would be meaningless to predict the expected completion time when $x = 0.5$. Simple linear regression models are typically used when the explanatory variables are quantitative.

Throughout this chapter, you were asked to evaluate **residual plots** to determine whether the model assumptions were met. When model assumptions are not met, the test statistics do not follow the corresponding t-distribution or F-distribution. Thus, the p-values may not be correct. No conclusions should be drawn from any statistical test without checking the appropriate assumptions.

The statistical model and corresponding hypothesis tests assume there are **no extraneous variables** that are biasing the study. Since it is typically impossible to identify all possible sources of bias during a study, **random sampling** and **random allocation** should be used.

In this chapter, we focused on testing the four assumptions about the **error terms** in each model. Table 2.2 shows the residual plots used in this chapter to check model assumptions.

[[[The table is not displaying. But it works when I knit on separate pdf]]]

Table 2.2: (#tab:tab2.2)Table 2.2 Plots that can be used to evaluate model assumptions about the residuals.

Assumption	Plot
Normality	Histogram or normal probability plot
Zero mean	No plot (errors will always sum to zero under these models)
Equal variances	Plot of original data or residual vs. fits
Independence	Residuals vs. order
Identically distributed	No plot (ensure each subject was sampled from the same population within each gr

If the model assumptions are violated, **transformations** can often be used to rescale the data to better fit some model assumptions. Several sophisticated mathematical tests are also available to test these model assumptions. While these tests are useful, plots are often just as effective and better assist in understanding the data. Plots should always be used to visualize the data before any conclusions are drawn. Later chapters will provide much more detail on checking assumptions for more complex models.

Exercises

E.1. Assume you are conducting a t-test to determine if there is a difference between two means. You have the following summary statistics: $\bar{x}_1 = 10$,

$\bar{x}_2 = 20$, and $s_1 = s_2 = 10$. Without completing the hypothesis tests, explain why $n_1 = n_2 = 100$ would result in a smaller p-value than $n_1 = n_2 = 16$.

E.2. If the test $H_0\colon \beta_1 = 0$ versus $H_a\colon \beta_1 \neq 0$ results in a small p-value, can we be confident that the regression model provides a good estimate of the response value for a given value of x_i? Provide an explanation for your answer.

E.3. What model assumptions (if any) need to be satisfied in order to calculate b_0 and b_1 in a simple linear regression model?

E.4. Explain why the model $y_i = \beta_0 + \beta_1 x_i$ is not appropriate, but $\hat{y}_i = b_0 + b_1 x_i$ is appropriate.

E.5. When there are only two levels (with equal sample sizes) being compared in an F-test, explain why $\alpha_1 = -\alpha_2$.

E.6. **Computer Games Again: Extending One-Way ANOVA to More Than Two Levels**

Data set: `Games2`

Assume that in the computer game study researchers were also interested in testing whether college students could play the game more quickly with their right or left hand. The data set `Games2` shows a column Type2 with four types of games, based on distracter and which hand was used.

1. Graph the data and compare the center and spread of the completion times for each of the four groups listed under Type2. Does any group have outliers or skewed data?
2. Conduct an ANOVA with one explanatory variable that has the four levels listed under Type2. Notice that this data set has $I = 4$ groups instead of 2 groups. However, the model and calculations are identical except that now $i = 1, 2, 3, 4$ instead of $i = 1, 2$ and now each group has a sample size of $n_i = 10$. The null hypothesis is $H_0\colon \mu_1 = \mu_2 = \mu_3 = \mu_4$ versus the alternative $H_a\colon$ at least one mean is different from another. Does the ANOVA show that a significant difference exists between group means?
3. Create residual plots. Are the model assumptions satisfied?
4. Are there any extraneous variables that might bias the results?
5. Assuming that the data for this game were collected in the same way as the `Game1` data, state your conclusions. Be sure to address random

sampling and random allocation.

Note that we could consider modeling the completion times as a function of two explanatory variables: the hand used (right or left) and the game type (standard or with color distracter). This would require a two-way ANOVA analysis (discussed in later chapters).

E.7. **Paper Towel Experiment: Comparing Three Factor Levels**

Data set: `PaperTowel`

As a final project in an introductory statistics class, several students decided to conduct a study to test the strength of paper towels. Television advertisements had claimed that a certain brand of paper towel was the strongest, and these students wanted to determine if there really was a difference. The students purchased rolls of towels at a local store and sampled 26 towels from 3 brands of paper towels: Bounty, Comfort, and Decorator.

Before any data were collected, these students determined that the following should be held as constant as possible throughout the study:

- 15 drops of water were applied to the center of each towel.
- Paper towels were selected that had the same size.
- The towels were held at all four corners by two people.
- Weights (10, 25, 50, or 100 grams) were slowly added to the center of each towel by a third person until it broke.
- The order in which the 26 paper towels were tested was randomized.

The file `PaperTowel` shows the breaking `Strength` of each towel. Breaking `Strength` is defined as the total weight that each towel successfully held. The next additional weight caused the towel to break.

1. For this paper towel study, identify the explanatory variable, the observational units (experimental units or subjects), and the response variable. Write out (in words and symbols) appropriate null and alternative hypotheses.
2. Graph the data and compare the center and spread of the breaking strength of each of the three brands. Does any group have outliers or skewed data?
3. Conduct an ANOVA. The null hypothesis is $H_0: \mu_1 = \mu_2 = \mu_3$ versus the alternative $H_a:$ at least one mean is different from another. Does the ANOVA show that a significant difference exists between brands?
4. Show that the equal variance assumption is violated in this study. Instead of using Strength as the response variable, use the natural

log of Strength to conduct an ANOVA. The null hypothesis is still stated as $H_0\colon \mu_1 = \mu_2 = \mu_3$ versus the alternative $H_a\colon$ at least one mean is different from another. Does the ANOVA show that a significant difference exists between brands?

5. Compare residual plots and the group standard deviations from the Strength and ln(Strength) F-tests. Which test should be used to state your conclusions? Explain.
6. Assume the students purchased one roll of Bounty paper towels and randomly selected 26 sheets from that one roll. The same holds true for other brands. What is the population for which the results of this study can hold?
7. Assume the students randomly purchased 26 rolls of Bounty paper towels from various stores and randomly selected one sheet from each roll. The same holds true for other brands. What is the population for which the results of this study can hold?

E.8. **Dr. Benjamin Spock**

Data set: `Jury`

Dr. Benjamin Spock was a well-known pediatrician who faced trial in 1968 for his activities as a Vietnam War protester. Specifically, he was charged with conspiring to violate the Selective Service Act by encouraging young men to resist the draft. As part of his defense, his counsel claimed that women were underrepresented on the jury. Women tend to be more sympathetic toward war protesters than men do. The defense counsel claimed that the judge had a history of choosing juries on which women were systematically underrepresented. At that time, jury members in Boston were chosen from a venire (a group of 30 to 200 individuals preselected from the population by the judge's clerk). By law, people were supposed to be selected for a venire at random. For Dr. Spock's trial, the judge's venire had 100 people and only 9 women, none of whom were selected to be on the actual jury.

Dr. Spock's defense counsel collected data on the percentages of women in venires from this judge's recent trials together with those of other judges in the Boston area.[6]

1. Graph the data and compare the center and spread of the percentages of women in the venires of each group (Judge). Does any group have outliers or skewed data?
2. Conduct an ANOVA with one explanatory variable and seven levels (judge). Notice that this data set now has groups with different sample sizes (i.e., $n_1 \neq n_2 \neq n_3$, etc.). The null hypothesis is $H_0\colon \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7$ versus the alternative $H_a\colon$

at least one mean is different from another. Does the ANOVA show that a significant difference exists between group means?

3. Dr. Spock's defense counsel collected the data. What questions would you ask the defense about the data collection process before you were convinced that the data could be used in court?

4. Assuming the data were properly collected, prepare a short statement advising the defense attorney whether the results of this study should be presented in court. If you suggest that the defense should use these results, provide an explanation of how these results are helpful. If you suggest that the defense should not use these results, provide an explanation of how these results are detrimental to the defense. It will be helpful to include a graph with a clear interpretation.

E.9. Tread Durability: Comparing Five Tire Brands

Data set: `Tires`

The data file `Tires` relates to five brands of tires chosen at random from local stores. Six tires of each brand were selected and placed in random order on a machine that tested tread durability in thousands of miles.

1. Graph the data and compare the center and spread of the durability measurements for each group (`Brand` of tire). Does any group have outliers or skewed data?

2. Conduct an ANOVA. Does the ANOVA show a significant difference exists between group means?

3. Create residual plots and compare the group standard deviations. Are the model assumptions satisfied?

4. Brand is used simply to identify different brands within the study. Explain why a simple linear regression model should not be used for the data, even though `Brand` could be treated as a quantitative variable.

E.10. Normal Probability Plots

Data set: `Games1`

1. Create a normal probability plot and histogram of the residuals from the `Games1` data (Question 26) and comment on the normality assumption for the random error term.

2. Create a normal probability plot or a histogram of the observed responses (completion times y_i, where $i = 1, 2, \ldots, n$) from the computer game study.

3. Explain why residuals should be used instead of the observed responses to test the normality assumption.

E.11. **Comparing Normal Probability Plots**

1. Use software to draw five random samples, each of size $n = 25$, from an actual normal probability distribution with mean 0 and standard deviation 1. Create a normal probability plot and histogram for each "sample."
2. Use software to draw five random samples, each of size $n = 50$, from an actual normal probability distribution with mean 0 and standard deviation 1. Create a normal probability plot and histogram for each "sample."
3. Describe how changing sample size impacts our ability to determine if a sample is truly from a normal distribution.
4. Use software to draw six random samples, each of size $n = 50$, from an actual normal probability distribution. Use means of -20 and 15 and standard deviations of 0.1, 3, and 300. Create a normal probability plot and histogram for each "sample." Explain why the mean and standard deviation of a population do not impact the normal probability plots.

E.12. **Transforming ANOVA Data**

Data set: `Hodgkins`

The data set `Hodgkins` contains plasma bradykininogen levels (in micrograms of bradykininogen per milliliter of plasma) in three types of subjects (normal, patients with active Hodgkin's disease, and patients with inactive Hodgkin's disease). The globulin bradykininogen is the precursor substance for bradykinin, which is thought to be a chemical mediator of inflammation.

1. Create individual value plots or side-by-side boxplots of the bradykininogen levels for each group. Compare the mean and standard deviation of each group. Do the data within each group look consistent with data from a normal population?
2. Calculate the ANOVA model estimates in order to create a normal probability plot of the residuals. Do the error terms look consistent with data from a normal population?
3. Transform the data by taking the log of the response. Create individual value plots or side-by-side boxplots of the transformed responses. Compare the mean and standard deviation of each group. Do the

transformed data within each group look consistent with data from a normal population?

4. Calculate the ANOVA model estimates using the transformed data in order to create a normal probability plot of the residuals. Do the error terms look consistent with data from a normal population? You may have to try out more than one transformation before you are satisfied with your answer to this part and Part C.

5. Using the transformed data, calculate an ANOVA table, F-test, and p-value to determine if the three patient types differed in their mean bradykininogen levels.

E.13. Transformations in Regression

Data set: `Weight`

In addition to brain weights and body weights, the `Weight` data set contains information on lifespan, gestation, and hours of sleep per day.

1. Create a scatterplot of lifespan versus body weight with a regression line ($\hat{y}_i = b_0 + b_1 x_i$), a plot of residuals versus the explanatory variable, a plot of residuals versus predicted (or "fitted") values ($\hat{y}_i$), and either a normal probability plot or a histogram of the residuals.

2. Try various transformations of the explanatory and response variables to create a better linear regression model.

3. Repeat Parts A and B using gestation as the response variable.

4. Repeat Parts A and B using sleep per day as the response variable.

E.14. Computer Games Again: Multiple Comparisons[7]

Data set: `Games2`

1. Conduct a t-test to determine if there is a difference in mean completion time between ColorLeft and StandardLeft. Report the p-value and your conclusions based on an individual α-level = 0.05.

2. Conduct a t-test to determine if there is a difference in mean completion time between ColorLeft and ColorRight. Report the p-value and your conclusions based on an individual α-level = 0.05.

3. Conduct a t-test to determine if there is a difference in mean completion time between ColorRight and StandardLeft. Report the p-value and your conclusions based on an individual α-level = 0.05.

[7] Multiple comparisons are discussed in the extended activities in Chapter 1.

4. List three other t-tests that could test for differences in mean completion time for different levels of Type2. In other words, what combinations of ColorRight, StandardRight, ColorLeft, and StandardLeft have not yet been tested?

5. Assume all six hypothesis tests were compared. If each of these tests are independent and each of the tests used an α-level = 0.05, what is the true probability that at least one of the tests would inappropriately reject the null hypothesis?

6. What is the individual critical value if you use the Bonferroni method with an overall (familywise) α-level = 0.10. Do any of your previous conclusions in Parts A through C change if you test for an overall (familywise) comparison? Explain.

E.15. **Interpreting Back Transformations**

Data set: `Skinfold`[7]

Celiac disease results in an inability to absorb carbohydrates and fats. Crohn's disease is another chronic intestinal disease in which the body's immune system attacks the intestines. Both Crohn's disease and celiac disease often result in malnutrition or impaired growth in children. A skinfold thickness measurement is a simple technique assessing body fat percentages by pinching the skin near the biceps and then using a calipers to measure the skin thickness.

1. Transform the original data into $\sqrt{\text{Thickness}}$, ln(Thickness), $\log_{10}$(Thickness), and the reciprocal (1/Thickness). For each transformation, conduct two-sample t-tests for differences between the mean of the Crohn's disease and celiac disease groups. Create a table with p-values and confidence intervals for the difference between the two means.

2. Back transform the confidence intervals in Part A. For example, square the upper and lower bounds of the confidence interval created from the $\sqrt{\text{Thickness}}$ data. Conduct the appropriate back transformation on the other three confidence intervals as well. Create a table of the four back-transformed confidence intervals. What do these back-transformed confidence intervals tell you?

Confidence limits for the difference between means often cannot be transformed back to the original scale. When reciprocal transformations have very small bounds, the back transformation provides unreasonably large bounds. For example, a skinfold transformation of 1/0.022 = 45.5 mm is not realistic. In addition, if the lower bound of a confidence interval is negative when the square-root transformation is used, back transforming the results (by squaring the bounds) will result in a confidence interval that does not contain zero. Thus, there are not reasonable practical interpretations of the back-transformed scales.

The log (this includes the natural log and $\log_{10}$) is often preferable over other transformations because the back transformation has a practical interpretation. The $\log_{10}$ back-transformed confidence interval (0.89, 2.03) provides results that can be interpreted, but not in the original units (millimeters). Notice that the confidence interval does not contain zero, but the results are not significant. Recall from your introductory statistics class that if a two-sided confidence interval contains zero, we fail to reject the null hypothesis for the corresponding two-sided hypothesis test. This is a 95% confidence interval for the ratio of the means. Thus, a value of one represents no difference between group means.

E.16. **Helicopters Again: Building Polynomial Regression Models**

Data set: `WingLength2`

If you completed the helicopter research project, this series of questions will help you further investigate the use of regression to determine the optimal wing length of paper helicopters to maximize flight time. It is likely that your data from the project did not appear to lie along a straight line, but rather had a curved pattern. In this exercise, we will use a larger data set collected on six different wing lengths.

Figure 2.7 shows an individual value plot of the data, with five observations for each of the six wing length groups. The curved pattern is quite pronounced, and this makes sense. At some point, the wings are so long that the helicopter does not spin stably or is simply too heavy and falls faster. It appears that there is some optimal wing length around 9 or 10 cm.

1. Using the `WingLength2` data set, try several transformations of either the response or the explanatory variable to see if you can alleviate the problem of nonlinearity.
2. It is likely that you cannot successfully find a transformation to solve the nonlinearity problem. A closer look at the residuals (Figure 2.8) helps to explain why. The residuals from the regression follow a somewhat sinusoidal pattern: down, then up, then down again. This is a pattern that is seen in a typical third-degree polynomial ($y = ax^3 + bx^2 + cx + d$). To create the appropriate regression model, we will introduce a new method called *polynomial regression* (instead of simple linear regression). In polynomial regression, we can fit the model

$$y_i = \beta_0+\beta_1 x_i+\beta_2 x_i^2+\beta_3 x_i^3+\varepsilon_i \qquad \text{for } i = 1, 2, \dots, n \text{ where } \varepsilon_i \sim N(0, \sigma^2) \tag{2.10}$$

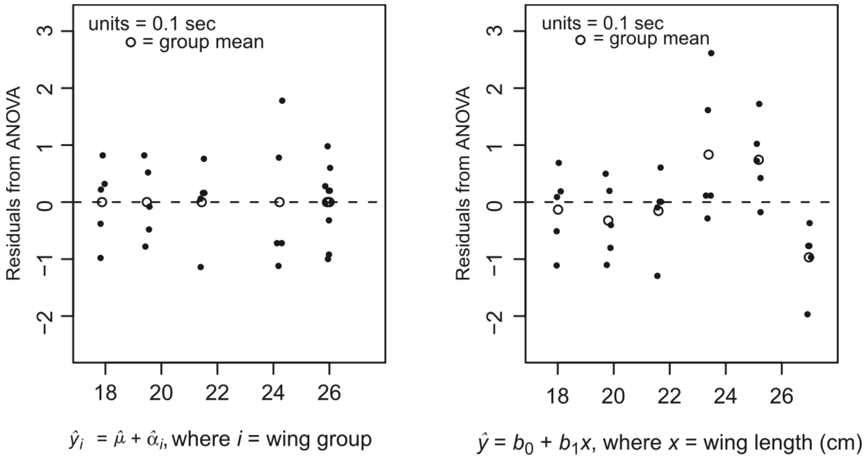

Figure 2.7: Flight times for paper helicopters when dropped from a height of 8 feet, each with a small paperclip attached at the bottom of the base of the helicopter.
(#fig:fig2.7)

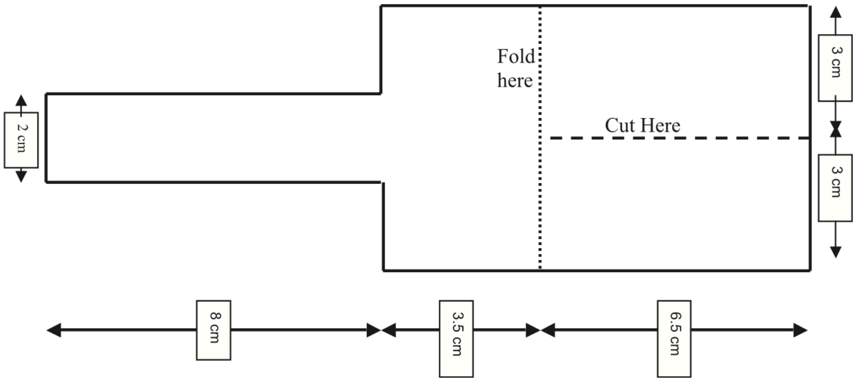

Figure 2.8: Residuals from an ANOVA and linear regression analysis of the flight times for paper helicopters from six groups: wing lengths 5.5 cm, 6.5 cm (standard), 7.5 cm, 8.5 cm, 19.5 cm, and 10.5 cm.
(#fig:fig2.8)

Using statistical computing software, fit the model in Equation **??** and report estimates for $\beta_0, \beta_1, \beta_2$, and β_3.

3. Look at a plot of the residuals from the polynomial regression versus the predicted values ($\hat{y}_i = b_0 + b_1 x_i + b_2 x_i^2 + b_3 x_i^3$) and a normal probability plot of the residuals and comment on the validity of the regression model assumptions for these data.

4. Given your answer to Part B and the fact that estimated flight times are considered to be a smooth (polynomial) function of wing length according to the relationship $\hat{y}_i = b_0 + b_1 x_i + b_2 x_i^2 + b_3 x_i^3$, estimate the optimal wing length that will lead to maximum flight time. *Note:* Use calculus or visual inspection to identify a maximum that occurs within a reasonable range of wing lengths.

Polynomial regression allows us to create a function relating the explanatory and response variables, and this can be useful for predicting responses for levels of the explanatory variable that were not actually measured as a part of the experiment. Of course, we should take care not to predict responses for levels of the explanatory variable that are quite far from those used in the experiment. The regression model may not extend past the domain we were able to analyze.

Endnotes

1. Yogi Berra was an American League Baseball player and manager. This quote has also been attributed to computer scientist Jan L. A. van de Snepscheut.

2. J. R. Stroop, "Studies of Interference in Serial Verbal Reactions," *Journal of Experimental Psychology*, 18 (1935): 643–662.

3. The following articles provide more details on transformations: N. R. Draper and W. G. Hunter, "Transformations: Some Examples Revisited," *Technometrics*, 11 (1969): 23–40; G. E. P. Box and D. R. Cox, "An Analysis of Transformations (with Discussion)," *Journal of the Royal Statistical Society B*, 26 (1964): 211–252; M. S. Bartlett, "The Use of Transformations," *Biometrics*, 3 (1947): 39–52; J. L. Dolby, "A Quick Method for Choosing a Transformation," *Technometrics*, 5 (1963): 317–326.

4. Details of these methods are described in D.C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, 3rd ed. (New York: Wiley, 2001).

5. F. Mosteller and J. W. Tukey, *Data Analysis and Regression* (Reading, MA: Addison-Wesley, 1977).

6. Data collected from H. Zeisel, "Dr. Spock and the Case of the Vanishing Women Jurors," *University of Chicago Law Review*, 37.1 (Autumn, 1969):

1–18.

7. Data from J. M. Bland and D. G. Altman, “Statistics Notes: The Use of Transformation When Comparing Two Means,” *BMJ*, 312 (1996): 1153.

8. G. E. P. Box, “Teaching Engineers Experimental Design with a Paper Helicopter,” *Quality Engineering*, 4 (1992): 453–459. Adapted with permission.

Chapter 3

Multiple Regression: How Much Is Your Car Worth?

Essentially, all models are wrong; some are useful.
-George E. P. Box[1]

Multiple regression is arguably the single most important method in all of statistics. Regression models are widely used in many disciplines. In addition, a good understanding of regression is all but essential for understanding many other, more sophisticated statistical methods.

This chapter consists of a set of activities that will enable you to build a multivariate regression model. The model will be used to describe the relationship between the retail price of 2005 used GM cars and various car characteristics, such as mileage, make, model, presence or absence of cruise control, and engine size. The set of activities in this chapter allows you work through the entire process of model building and assessment, including

- Applying variable selection techniques
- Using residual plots to check for violations of model assumptions, such as heteroskedasticity, outliers, autocorrelation, and nonnormality distributed errors
- Transforming data to better fit model assumptions
- Understanding the impact of correlated explanatory variables
- Incorporating categorical explanatory variables into a regression model
- Applying F-tests in multiple regression

[1] G.E.P. Box and N.R. Draper's Empirical Model-Building and Response Surfaces (New York: Wiley, 1987),p. 4 24.

3.1 Investigation: How Can We Build a Model to Estimate Used Car Prices?

Have you ever browsed through a car dealership and observed the sticker prices on the vehicles? If you have ever seriously considered purchasing a vehicle, you can probably relate to the difficulty of determining whether that vehicle is a good deal or not. Most dealerships are willing to negotiate on the sale price, so how can you know how much to negotiate? For novices (like this author), it is very helpful to refer to an outside pricing source, such as the Kelley Blue Book, before agreeing on a purchase price.

For over 80 years, Kelley Blue Book has been a resource for accurate vehicle pricing. The company's Website, http://www.kbb.com, provides a free online resource where anyone can input several car characteristics (such as age, mileage, make, model, and condition) and quickly receive a good estimate of the retail price.

In this chapter, you will use a relatively small subset of the Kelley Blue Book database to describe the association of several explanatory variables (car characteristics) with the retail value of a car. Before developing a complex multiple regression model with several variables, let's start with a quick review of the simple linear regression model by asking a question: Are cars with lower mileage worth more? It seems reasonable to expect to see a relationship between mileage (number of miles the car has been driven) and retail value. The data set Cars contains the make, model, equipment, mileage, and Kelley Blue Book suggested retail price of several used 2005 GM cars.

A Simple Linear Regression Model

Data set: *Cars* 1. Produce a scatterplot from the Cars data set to display the relationship between mileage (Mileage) and suggested retail price (Price). Does the scatterplot show a strong relationship between Mileage and Price?

2. Calculate the least squares regression line, $Price = b_0 + b_1(Mileage)$. Report the regression model, the R^2 value, the correlation coefficient, the t-statistics, and p-values for the estimated model coefficients (the intercept and slope). Based on these statistics, can you conclude that Mileage is a strong indicator of Price? Explain your reasoning in a few sentences.

3. The first car in this data set is a Buick Century with 8221 miles. Calculate the residual value for this car (the observed retail price minus the expected price calculated from the regression line).

   MATHMATICAL NOTE For any regression equation of the form $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ the hypothesis test for the slope of the regression equation (b1) is similar to other t-tests discussed in introductory textbooks. (Mathematical details for this hypothesis test are described in Chapter 2.) To test the null hypothesis that the slope coefficient is zero ($H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$), calculate the following test statistic:

where b1 is the estimated slope calculated from the data and $\hat{\sigma}_{b_1}$ is an estimate of the standard deviation of b1. Probability theory can be used to prove that if the regression model assumptions are true, the t-statistic in Equation **??** follows a t-distribution with n - 2 degrees of freedom. If the sample statistic, b_1, is far away from $b_1 = 0$ relative to the estimated standard deviation, the t-statistic will be large and the corresponding p-value will be small.

The t-statistic for the slope coefficient indicates that Mileage is an important variable. However, the R^2 value (the percentage of variation explained by the regression line) indicates that the regression line is not very useful in predicting retail price. (A review of the R^2 value is given in the extended activities.) As is always the case with statistics, we need to visualize the data rather than focus solely on a p-value. Figure 3.1 shows that the expected price decreases as mileage increases, but the observed points do not appear to be close to the regression line. Thus, it seems reasonable that including additional explanatory variables in the regression model might help to better explain the variation in retail price.

In this chapter, you will build a linear combination of explanatory variables that explains the response variable, retail price. As you work through the chapter, you will find that there is not one technique, or “recipe,” that will give the best model. In fact, you will come to see that there isn’t just one “best” model for these data.

Unlike in most mathematics classes, where every student is expected to submit the one right answer to an assignment, here it is expected that the final regression models submitted by various students will be at least slightly different. While a single “best” model may not exist for these data, there are certainly many bad models that should be avoided. This chapter focuses on understanding the process of developing a statistical model. It doesn’t matter if you are developing a regression model in economics, psychology, sociology, or engineering—there are common key questions and processes that should be evaluated before a final model is submitted.

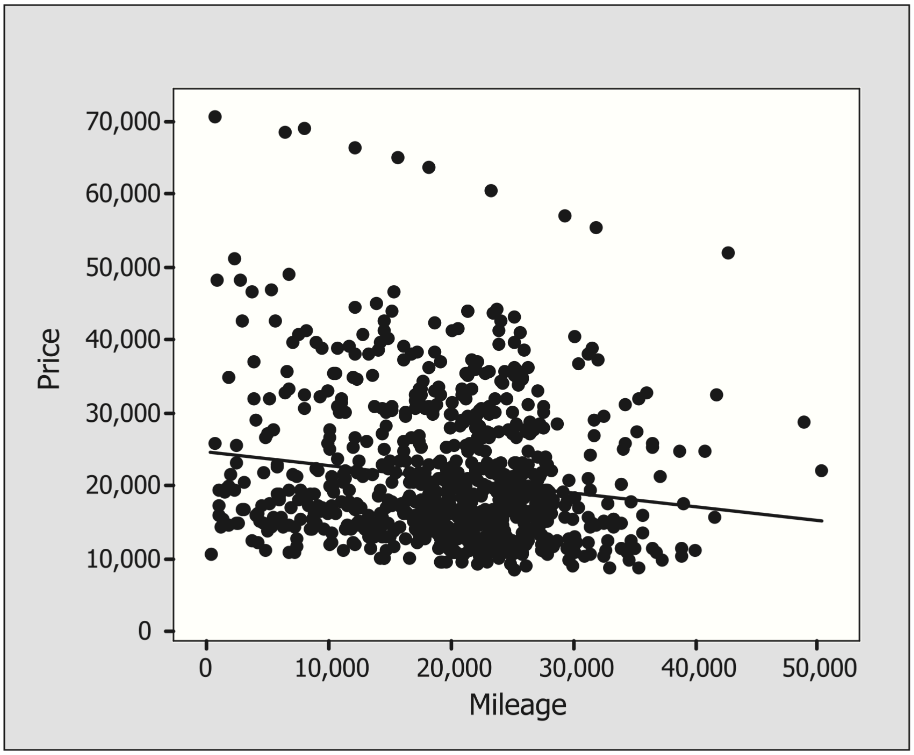

Figure 3.1: Scatterplot and least squares regression model: $Price = 24,765 - 0.1725(Mileage)$. The regression line shows that for each additional mile a car is driven, the expected price of the car decreases by about 17 cents. However, many points are not close to the regression line, indicating that the expected price is not an accurate estimate of the actual observed price.
(#fig:fig3.1)

3.2 Goals of Multiple Regression

It is important to note that multiple regression analysis can be used to serve different goals. The goals will influence the type of analysis that is conducted. The most common goals of multiple regression are to describe, predict, or confirm.

- **Describe**: A model may be developed to describe the relationship between multiple explanatory variables and the response variable.
- **Predict**: A regression model may be used to generalize to observations outside the sample. Just as in simple linear regression, explanatory variables should be within the range of the sample data to predict future responses.
- **Confirm**: Theories are often developed about which variables or combination of variables should be included in a model. For example, is mileage useful in predicting retail price? Inferential techniques can be used to test if the association between the explanatory variables and the response could just be due to chance. Theory may also predict the type of relationship that exists, such as "cars with lower mileage are worth more." More specific theories can also be tested, such as "retail price decreases linearly with mileage."

When the goal of developing a multiple regression model is description or prediction, the primary issue is often determining which variables to include in the model (and which to leave out). All potential explanatory variables can be included in a regression model, but that often results in a cumbersome model that is difficult to understand. On the other hand, a model that includes only one or two of the explanatory variables, such as the model in Figure 3.1, may be much less accurate than a more complex model. This tension between finding a simple model and finding the model that best explains the response is what makes it difficult to find a "best" model. The process of finding the most reasonable mix, which provides a relatively simple linear combination of explanatory variables, often resembles an exploratory artistic process much more than a formulaic recipe.

Including redundant or unnecessary variables not only creates an unwieldy model but also can lead to test statistics (and conclusions from corresponding hypothesis tests) that are less reliable. If explanatory variables are highly correlated, then their effects in the model will be estimated with more imprecision. This imprecision leads to larger standard errors and can lead to insignificant test results for individual variables that can be important in the model. Failing to include a relevant variable can result in biased estimates of the regression coefficients and invalid t-statistics, especially when the excluded variable is highly significant or when the excluded variable is correlated with other variables also in the model.[2]

[2] More details are provided in more advanced textbooks such as M. H. Kutner, J. Neter, C. J. Nachtsheim, and W. Li, Applied Linear Regression Models (New York: McGraw-Hill, 2004).

3.3 Variable Selection Techniques to Describe or Predict a Response

If your objective is to describe a relationship or predict new response variables, variable selection techniques are useful for determining which explanatory variables should be in the model. For this investigation, we will consider the response to be the suggested retail price from Kelley Blue Book (the Price variable in the data). We may initially believe the following are relevant potential explanatory variables:

- Make (Buick, Cadillac, Chevrolet, Pontiac, SAAB, Saturn)
- Model (specific car for each previously listed Make)
- Trim (specific type of Model)
- Type (Sedan, Coupe, Hatchback, Convertible, or Wagon)
- Cyl (number of cylinders: 4, 6, or 8)
- Liter (a measure of engine size)
- Doors (number of doors: 2, 4)
- Cruise (1 = cruise control, 0 = no cruise control)
- Sound (1 = upgraded speakers, 0 = standard speakers)
- Leather (1 = leather seats, 0 = not leather seats)
- Mileage (number of miles the car has been driven)

Stepwise Regression

When a large number of variables are available, **stepwise regression** is an iterative technique that has historically been used to identify key variables to include in a regression model. For example, forward stepwise regression begins by fitting several single-predictor regression models for the response; one regression model is developed for each individual explanatory variable. The single explanatory variable (call it X_1) that best explains the response (has the highest R^2 value) is selected to be in the model.[3]

In the next step, all possible regression models using X_1 and exactly one other explanatory variable are calculated. From among all these two-variable models, the regression model that best explains the response is used to identify X_2. After the first and second explanatory variables, X_1 and X_2, have been selected, the process is repeated to find X3. This continues until including additional variables in the model no longer greatly improves the model's ability to describe the response.[4]

[3] An F-test is conducted on each of the models. The size of the F-statistic (and corresponding p-value) is used to evaluate the fit of each model. When models with the same number of predictors are compared, the model with the largest F-statistic will also have the largest R^2 value.

[4] An α-level is often used to determine if any of the explanatory variables not currently in the model should be added to the model. If the p-value of all additional explanatory variables is greater than the α-level, no more variables will be entered into the model. Larger α-level (such as a = 0.2) will include more terms while smaller α-level (such as a = 0.05) will include

Backward stepwise regression is similar to forward stepwise regression except that it starts with all potential explanatory variables in the model. One by one, this technique removes variables that make the smallest contribution to the model fit until a "best" model is found.

While sequential techniques are easy to implement, there are usually better approaches to finding a regression model. Sequential techniques have a tendency to include too many variables and at the same timeso metimes eliminate important variables.3 With improvements in technology, most statisticians prefer to use more "global" techniques (such as best subset methods), which compare all possible subsets of the explanatory variables.

> **Note** Later sections will show that when explanatory variables are highly correlated, sequential procedures often leave out variables that explain (are highly correlated with) the response. In addition, sequential procedures involve numerous iterations and each iteration involves hypothesis tests about the significanceof coefficients. Remember that with any multiple comparison problem, an α-level of 0.05 means there is a 5% chance that each irrelevant variable will be found significant and may inappropriately be determined important for the model.

Selecting the "Best Subset" of Predictors

A researcher must balance increasing R^2 against keeping the model simple. When models with the same number of parameters are compared, the model with the highest R^2 value should typically be selected. A larger R^2 value indicates that more of the variation in the response variable is explained by the model. However, R^2 never decreases when another explanatory variable is added. Thus, other techniques are suggested for comparing models with different numbers of explanatory variables.

Statistics such as the adjusted R^2, Mallows' C_p, and Akaike's and Bayes' information criteria are used to determine a "best" model. Each of these statistics includes a penalty for including too many terms. In other words, when two models have equal R^2 values, each of these statistics will select the model with fewer terms. **Best subsets techniques** use several statistics to simultaneously compare several regression models with the same number of predictors.

The **coefficient of determination**, R^2 is the percentage of variation in the response variable that is explained by the regression line.

When the sum of the squared residuals $\sum_{i=1}^{n}(y_i - \hat{y})^2$ are small compared to the total spread of the responses $\sum_{i=1}^{n}(y_i - \bar{y})^2$, R^2 is close to one. $R^2 = 1$ indicates that the regression model perfectly fits the data.

fewer terms.

Comparing Variable Selection Techniques

Dataset: Cars

4. Use the Cars data to conduct a stepwise regression analysis.

   a. Calculate seven regression models, each with one of the following explanatory variables: Cy1, Liter, Doors, Cruise, Sound, Leather, and Mileage. Identify the explanatory variable that corresponds to the model with the largest R^2 value. Call this variable X_1.

   b. Calculate six regression models. Each model should have two explanatory variables, X_1 and one of the other six explanatory variables. Find the two-variable model that has the highest R^2 value. How much did R^2 improve when this second variable was included?

   c. Instead of continuing this process to identify more variables, use the software instructions provided to conduct a stepwise regression analysis. List each of the explanatory variables in the model suggested by the stepwise regression procedure.

5. Use the software instructions provided to develop a model using best subsets techniques. Notice that stepwise regression simply states which model to use, while best subsets provides much more information and requires the user to choose how many variables to include in the model. In general, statisticians select models that have a relatively low Cp, a large R^2, and a relatively small number of explanatory variables. (It is rare for these statistics to all suggest the same model. Thus, the researcher much choose a model based on his or her goals. The extended activities provide additional details about each of these statistics.) Based on the output from best subsets, which explanatory variables should be included in a regression model?

6. Compare the regression models in Questions 4 and 5.

   a. Are different explanatory variables considered important?

   b. Did the stepwise regression in Question 4 provide any indication that Liter could be useful in predicting Price? Did the best subsets output in Question 5 provide any indication that Liter might be useful in predicting Price? Explain why best subsets techniques can be more informative than sequential techniques.

Neither sequential nor best subsets techniques guarantee a best model. Arbitrarily using slightly different criteria will produce different models. Best subset methods allow us to compare models with a specific number of predictors, but models with more predictors do not always include the same terms as smaller models. Thus, it is often difficult to interpret the importance of any coefficients in the model.

Variable selection techniques are useful in providing a high R^2 value while limiting the number of variables. When our goal is to develop a model to describe or predict a response, we are concerned not about the significance of each explanatory variable, but about how well the overall model fits.

If our goal involves confirming a theory, iterative techniques are not recommended. Confirming a theory is similar to hypothesis testing. Iterative variable selection techniques test each variable or combination of variables several times, and thus the p-values are not reliable. The stated significance level for a t-statistic is valid only if the data are used for a single test. If multiple tests are conducted to find the best equation, the actual significance level for each test for an individual component is invalid.

> **Key Concept** If variables are selected by iterative techniques, hypothesis tests should not be used to determine the significance of these same terms.

3.4 Checking Model Assumptions

The simple linear regression model discussed in introductory statistics courses typically has the following form:

For this linear regression model, the mean response$\beta_0 + \beta_1 x_i$ is a linear function of the explanatory variable, x. The multiple linear regression model has a very similar form. The key difference is that now more terms are included in the model.

In this chapter, p represents the number of parameters in the regression model $\beta_0 + \beta_1 x_{1,i} + ... + \beta_{p-1} x_{p-1,i}$ and n is the total number of observations in the data. In this chapter, we make the following assumptions about the regression model: - The model parameters *0 +* 1x*{1,i} + {…} +* {p-1}{x_{p-1,i} and σ^2 are constant. - Each term in the model is additive. - The error terms in the regression model are independent and have been sampled from a single population (identically distributed). This is often abbreviated as iid. - The error terms follow a normal probability distribution centered at zero with a fixed variance, σ^2. This assumption is denoted as $\epsilon_i \sim N(0, \sigma^2)$ for $ = 1, {…}, ~ n$.

Regression assumptions about the error terms are generally checked by looking at the residuals from the data: $(y_i - \hat{y}_i)$. Here, y_i are the observed responses and $\hat{y}_i$ are the estimated responses calculated by the regression model. Instead of formal hypothesis tests, plots will be used to visually assess whether the assumptions hold. The theory and methods are simplest when any scatterplot of residuals resembles a single, horizontal, oval balloon, but real data may not cooperate by conforming to the ideal pattern. An ornery plot may show a wedge, a curve, or multiple clusters. Figure 3.2 shows examples of each of these types of residual plots.

Suppose you held a cup full of coins and dropped them all at once. We hope to find residual plots that resemble the random pattern that would likely result from dropped coins (like the oval-shaped plot). The other three plots show patterns that would be very unlikely to occur by random chance. Any plot patterns that are not nice oval shapes suggest that the error terms are violating at least one model assumption, and thus it is likely that we have unreliable estimates of our model coefficients. The following section illustrates strategies for dealing with one of these unwanted shapes: a wedge-shaped pattern.

Note that in single-variable regression models, residual plots show the same information as the initial fitted line plot. However, the residual plots often emphasize violations of model assumptions better than the fitted line plot. In addition, multivariate regression lines are very difficult to visualize. Thus, residual plots are essential when multiple explanatory variables are used.

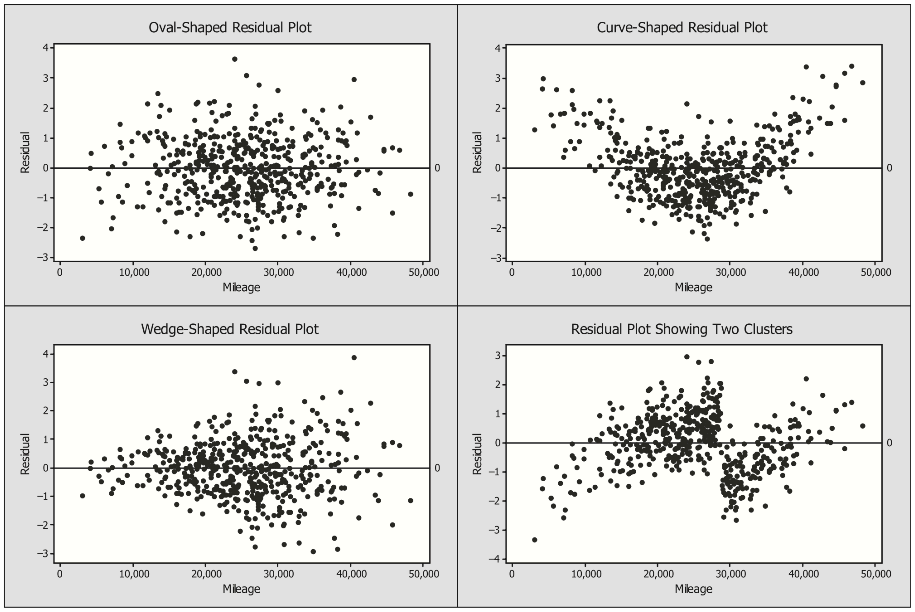

Figure 3.2: Common shapes of residual plots. Ideally, residual plots should look like a randomly scattered set of dropped coins, as seen in the oval-shaped plot. If a pattern exists, it is usually best to try other regression models. (#fig:fig3.2)

Heteroskedasticity

Heteroskedasticity is a term used to describe the situation where the variance of the error term is not constant for all levels of the explanatory variables. For example, in the regression equation $Price = 24,765 - 0.173(Mileage)$, the spread of the suggested retail price values around the regression line should be about the same whether mileage is 0 or mileage is 50,000. If heteroskedasticity exists in the model, the most common remedy is to transform either the explanatory variable, the response variable, or both in the hope that the transformed relationship will exhibit **homoskedasticity** (equal variances around the regression line) in the error terms.

7. Using the regression equation calculated in Question 5, create plots of the residuals versus each explanatory variable in the model. Also create a plot of the residuals versus the predicted retail price (often called a residual versus fit plot).

a. Does the size of the residuals tend to change as mileage changes

b. Does the size of the residuals tend to change as the predicted retail price changes? You should see patterns indicating heteroskedasticity (nonconstant variance).

c. Another pattern that may not be immediately obvious from these residual plots is the right skewness seen in the residual versus mileage plot. Often economic data, such as price, are right skewed. To see the pattern, look at just one vertical slice of this plot. With a pencil, draw a vertical line corresponding to mileage equal to 8000. Are the points in the residual plots balanced around the line $Y = 0$?

d. Describe any patterns seen in the other residual plots.

8. Transform the suggested retail price to log (Price) and 2Price. Transforming data using roots, logarithms, or reciprocals can often reduce heteroskedasticity and right skewness. (Transformations are discussed in Chapter 2. [[[add link]]]) Create regression models and residual plots for these transformed response variables using the explanatory variables selected in Question 5.

a. Which transformation did the best job of reducing the heteroskedasticity and skewness in the residual plots? Give the R^2 values of both new models.

b. Do the best residual plots correspond to the best R^2 values? Explain. While other transformations could be tried, throughout this investigation we will refer to the log-transformed response variable as TPrice.

Figure 3.3 shows residual plots that were created to answer Questions 7 and 8. Notice that when the response variable is Price, the residual versus fit plot has a clear wedge-shaped pattern. The residuals have much more spread when the fitted value is large (i.e., expected retail price is close to \$40,000) than when the fitted value is near \$10,000. Using TPrice as a response did improve the residual versus fit plot. Although there is still a faint wedge shape, the variability of the residuals is much more consistent as the fitted value changes. Figure 3.3 reveals another pattern in the residuals. The following section will address why points in both plots appear in clusters.

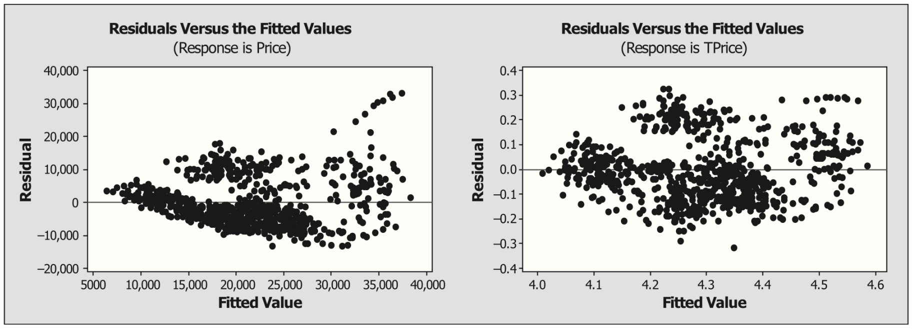

Figure 3.3: **Figure 3.3** Residual versus fit plots using Price and TPrice (the log10 transformation), as responses. The residual plot with Price as the response has a much stronger wedge-shaped pattern than the one with TPrice. (#fig:fig3.3)

Examining Residual Plots Across Time/Order

Autocorrelation exists when consecutive error terms are related. If autocorrelation exists, the assumption about the independence of the error terms is violated. To identify autocorrelation, we plot the residuals versus the order of the data entries. If the ordered plot shows a pattern, then we conclude that autocorrelation exists. When autocorrelation exists, the variable responsible for creating the pattern in the ordered residual plot should be included in the model.

9. Create a residual versus order plot from the TPrice versus Mileage regression line. Describe any pattern you see in the ordered residual plot. Apparently something in our data is affecting the residuals based on the order of the data. Clearly, time is not the influential factor in our data set (all of the data are from 2005). Can you suggest a variable in the Cars data set that may be causing this pattern?

10. Create a second residual versus order plot using TPrice as the response and using the explanatory variables selected in Question 5. Describe any patterns that you see in these plots.

    Note (order in which the data were collected) is perhaps the most common source of autocorrelation, but other forms, such as spatial autocorrelations, can also be present. If time is indeed a variable that should be included in the model, a specific type of regression model, called a time series model, should be used.

While ordered plots make sense in model checking only when there is a meaningful order to the data, the residual versus order plots could demonstrate the need to include additional explanatory variables in the regression model. Figure 3.4 shows the two residual plots created in Questions 9 and 10. Both plots show that the data points are clearly clustered by order. However, there is less clustering when the six explanatory variables (Mileage, Cyl, Doors, Cruise, Sound, and Leather) are in the model. Also notice that the residuals tend to be closer to zero in the second graph. Thus, the second graph (with six explanatory variables) tends to have estimates that are closer to the actual observed values.

We do not have a time variable in this data set, so reordering the data would not change the meaning of the data. Reordering the data could eliminate the pattern; however, the clear pattern seen in the residual versus order plots should not be ignored because it indicates that we could create a model with a much higher R^2 value if we could account for this pattern in our model. This type of autocorrelation is called taxonomic autocorrelation, meaning that the relationship seen in this residual plot is due to how the items in the data set are classified. Suggestions on how to address this issue are given in later sections.

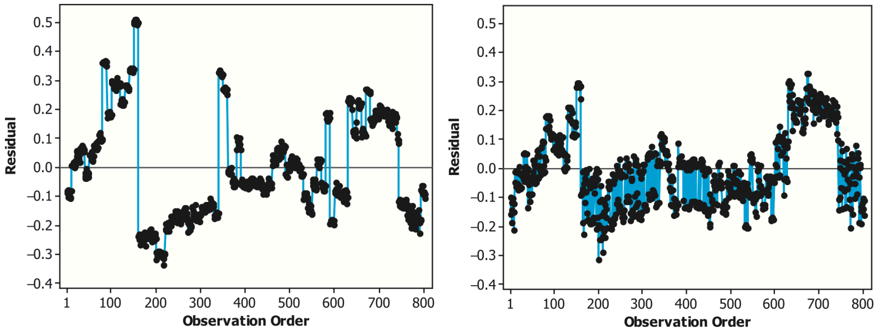

Figure 3.4: **Figure 3.4** Residual versus order plots using TPrice as the response. The first graph uses Mileage as the explanatory variable, and the second graph uses Mileage, Cyl, Doors, Cruise, Sound, and Leather as explanatory variables.
(#fig:fig3.4)

Outliers and Influential Observations

11. Calculate a regression equation using the explanatory variables suggested in Question 5 and Price as the response. Identify any residuals (or cluster of residuals) that don't seem to fit the overall pattern in the residual versus fit and residual versus mileage plots. Any data values that don't seem to fit the general pattern of the data set are called outliers.

a. Identify the specific rows of data that represent these points. Are there any consistencies that you can find?

b. Is this cluster of outliers helpful in identifying the patterns that were found in the ordered residual plots? Why or why not?

12. Run the analysis with and without the largest cluster of potential outliers (the cluster of outliers corresponds to the Cadillac convertibles). Use Price as the response. Does the cluster of outliers influence the coefficients in the regression line?

If the coefficients change dramatically between the regression models, these points are considered influential. If any observations are influential, great care should be taken to verify their accuracy. In addition to reducing heterskedasticity, transformations can often reduce the effect of outliers. Figure 3.5 shows the residual versus fit plots using Price and TPrice, respectively. The cluster of outliers corresponding to the Cadillac convertibles is much more visible in the plot with the untransformed (Price) response variable. Even though there is still clustering in the transformed data, the residuals corresponding to the Cadillac convertibles are no longer unusually large.

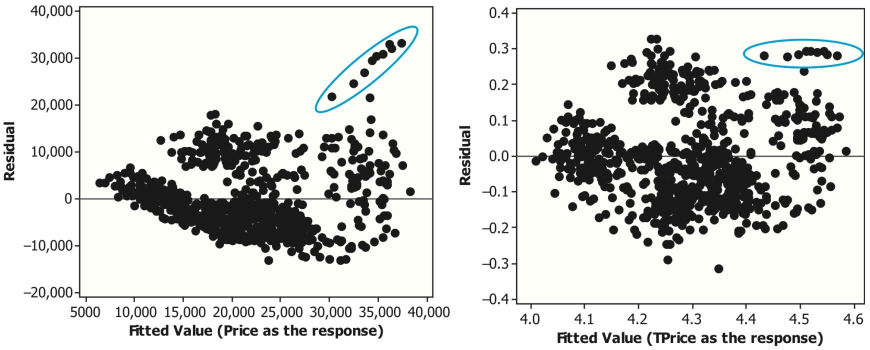

Figure 3.5: **Figure 3.5** Residual versus fit plots using Price and TPrice as the response. The circled observations in the plot using Price are no longer clear outliers in the plot using TPrice.
(#fig:fig3.5)

In some situations, clearly understanding outliers can be more time consuming (and possibly more interesting) than working with the rest of the data. It can be quite difficult to determine if an outlier was accurately recorded or whether the outliers should be included in the analysis.

If the outliers were accurately recorded and transformations are not useful in eliminating them, it can be difficult to know what to do with them. The simplest approach is to run the analysis twice: once with the outliers included and once without. If the results are similar, then it doesn't matter if the outliers are included or not. If the results do change, it is much more difficult to know what to do. Outliers should never automatically be removed because they don't fit the overall pattern of the data. Most statisticians tend to err on the side of keeping the outliers in the sample data set unless there is clear evidence that they were mistakenly recorded. Whatever final model is selected, it is important to clearly state if you are aware that your results are sensitive to outliers.

Normally Distributed Residuals

Even though the calculations of the regression model and R^2 do not depend on the normality assumption, identifying patterns in residual plots can often lead to another model that better explains the response variable.

To determine if the residuals are normally distributed, two graphs are often created: a histogram of the residuals and a normal probability plot. Normal probability plots are created by sorting the data (the residuals in this case) from smallest to largest. Then the sorted residuals are plotted against a theoretical normal distribution. If the plot forms a straight line, the actual data and the theoretical data have the same shape (i.e., the same distribution). (Normal probability plots are discussed in more detail in Chapter 2.)

13. Create a regression line to predict TPrice from Mileage. Create a histogram and a normal probability plot of the residuals.

a. Do the residuals appear to follow the normal distribution?

b. Are the ten outliers visible on the normal probability plot and the histogram?

Figure 3.6 shows the normal probability plot using six explanatory variables to estimate TPrice. While the outliers are not visible, both plots still show evidence of lack of normality. Simply plugging data into a software package and using an iterative variable selection technique will not reliably create a "best" model.

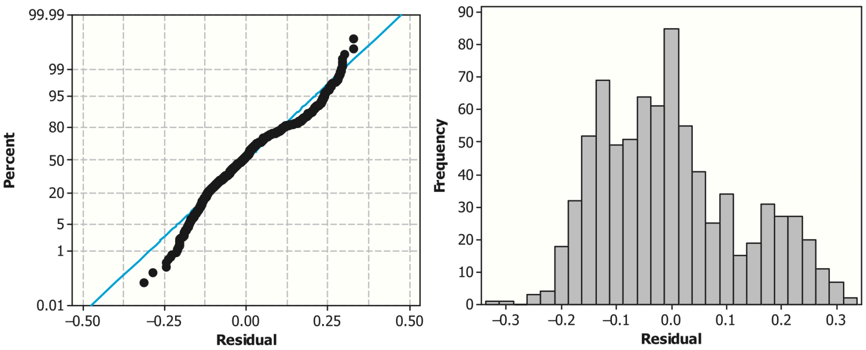

Figure 3.6: **Figure 3.6** Normal probability plot and histogram of residuals from the model using TPrice as the response and Mileage, Cyl, Doors, Cruise, Sound, and Leather as the explanatory variables.
(#fig:fig3.6)

> **Key Concept** Before a final model is selected, the residuals should be plotted against fitted (estimated) values, observation order, the theoretical normal distribution, and each explanatory variable in the model. Table 3.1 shows how each residual plot is used to check model assumptions. If a pattern exists in any of the residual plots, the R^2 value is likely to improve if different explanatory variables or transformations are included in the model.

[[[The table is not displaying. But it works when I knit on separate pdf]]]

Table 3.1: (#tab:tab3.1)Table 3.1 Plots that can be used to evaluate model assumptions about the residuals.

Assumption	Plot
Normality	Histogram or normal probability plot
Zero mean	No plot (errors will always sum to zero under these models)
Equal variances	Plot of original data or residual vs. fits
Independence	Residuals vs. order
Identically distributed	No plot (ensure each subject was sampled from the same population within each gr

Correlation Between Explanatory Variables

Multicollinearity exists when two or more explanatory variables in a multiple regression model are highly correlated with each other. If two explanatory variables X_1 and X_2 are highly correlated, it can be very difficult to identify whether X_1, X_2, or both variables are actually responsible for influencing the response variable, Y.

14. Create three regression models using Price as the response variable. In all three cases, provide the regression model, R^2 value, t-statistic for the slope coefficients, and corresponding p-values.

a. In the first model, use only Mileage and Liter as the explanatory variables. Is Liter an important explanatory variable in this model?
b. In the second model, use only Mileage and number of cylinders (Cyl) as the explanatory variables. Is Cyl an important explanatory variable in this model?
c. In the third model, use Mileage, Liter, and number of cylinders (Cyl) as the explanatory variables. How did the test statistics and p-values change when all three explanatory variables were included in the model?

15. Note that the R^2 values are essentially the same in all three models in Question 14. The coefficients for Mileage also stay roughly the same for all three models—the inclusion of Liter or Cyl in the model does not appear to influence the Mileage coefficient. Depending on which model is

used, we state that for each additional mile on the car, Price is reduced somewhere between \$0.152 to \$0.16. Describe how the coefficient for Liter depends on whether Cyl is in the model.

16. Plot Cyl versus Liter and calculate the correlation between these two variables. Is there a strong correlation between these two variables? Explain.

Recall that Question 4 suggested deleting Liter from the model. The goal in stepwise regression is to find the "best" model based on the R^2 value. If two explanatory variables both impact the response in the same way, stepwise regression will rather arbitrarily pick one variable and ignore the other.

> **Key Concept** Stepwise regression can often completely miss important explanatory variables when there is multicollinearity.

Most software packages can create a matrix plot of the correlations and corresponding scatterplots of all explanatory variables. This matrix plot is helpful in identifying any patterns of interdependence among the explanatory variables. An easy-to-apply guideline for determining if multicollinearity needs to be dealt with is to use the **variance inflation factor (VIF)**.

VIF conducts a regression of each explanatory variable (X_i) on the remaining explanatory variables, calculates the corresponding R^2 value (${R_i}^2$), and then calculates the following function for each variable Xi : $1/(1 - {R_i}^2)$. If the ${R_i}^2$ value is zero, VIF is one, and Xi is uncorrelated with all other explanatory variables. Montgomery, Peck, and Vining state, "Practical experience indicates that if any of the VIFs exceed 5 or 10, it is an indication that the associated regression coefficients are poorly estimated because of multicollinearity."[5]

Figure 3.7 shows that Liter and Cyl are highly correlated. Within the context of this problem, it doesn't make physical sense to consider holding one variable constant while the other variable increases. In general, it may not be possible to "fix" a multicollinearity problem. If the goal is simply to describe or predict retail prices, multicollinearity is not a critical issue. Redundant variables should be eliminated from the model, but highly correlated variables that both contribute to the model are acceptable if you are not interpreting the coefficients. However, if your goal is to confirm whether an explanatory variable is associated with a response (test a theory), then it is essential to identify the presence of multicollinearity and to recognize that the coefficients are unreliable when it exists.

> **Key Concept** If your goal is to create a model to describe or predict, then multicollinearity really is not a problem. Note that multicollinearity has very little impact on the R^2 value. However, if your goal is to understand how a specific explanatory variable influences the response, as is often done when confirming a theory, then multi-

[5] Details of these methods are described in D.C. Montgomery, E. A. Peck, and G. G. Vining, Introduction to Linear Regression Analysis, 3rd ed. (New York: Wiley, 2001).

collinearity can cause coefficients (and their corresponding p-values when testing their significance) to be unreliable.

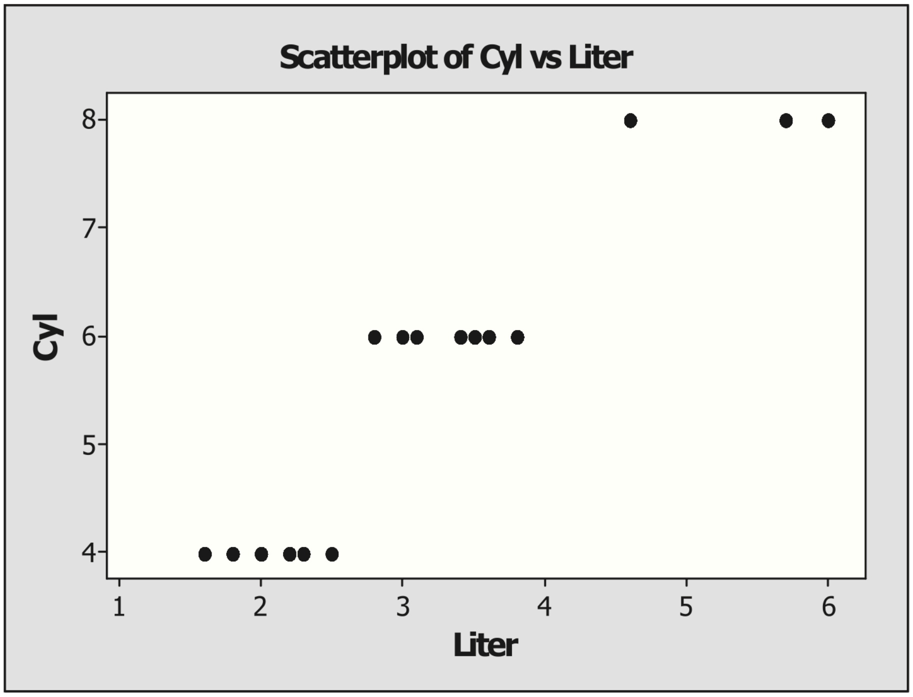

Figure 3.7: **Figure 3.7** Scatterplot showing a clear association between Liter and Cyl.
(#fig:fig3.7)

The following approaches are commonly used to address multicollinearity: - **Get more information**: If it is possible, expanding the data collection may lead to samples where the variables are not so correlated. Consider whether a greater range of data could be collected or whether data could be measured differently so that the variables are not correlated. For example, the data here are only for GM cars. Perhaps the relationship between engine size in liters and number of cylinders is not so strong for data from a wider variety of manufacturers. - **Re-evaluate the model**: When two explanatory variables are highly correlated, deleting one variable will not significantly impact the R^2 value. However, if there are theoretical reasons to include both variables in the model, keep both terms. In our example, Liter and number of cylinders (Cyl) are measuring essentially the same quantity. Liter represents the volume displaced during one complete engine cycle; number of cylinders (Cyl) also is a measure of the volume that can be displaced. - **Combine the variables**: Using other statistical techniques such as principal components, it is possible to combine the correlated variables "optimally" into a single variable that can be used in the model. There may be theoretical reasons to combine variables in a certain way. For example, the volume (size) and weight of a car are likely highly positively correlated. Perhaps a new variable defined as density = weight/volume could be used in a model predicting price rather than either of these individual variables.

In this investigation, we are simply attempting to develop a model that can be used to estimate price, so multicollinearity will not have much impact on our results. If we did re-evaluate the model in light of the fact that Liter and number of cylinders (Cyl) both measure displacement (engine size), we might note that Liter is a more specific variable, taking on several values, while Cyl has only three values in the data set. Thus, we might choose to keep Liter and remove Cyl in the model.

3.5 Interpreting Model Coefficients

While multiple regression is very useful in understanding the impacts of various explanatory variables on a response, there are important limitations. When predictors in the model are highly correlated, the size and meaning of the coefficients can be difficult to interpret. In Question 14, the following three models were developed:

The interpretation of model coefficients is more complex in multiple linear regression than in simple linear regression. It can be misleading to try to interpret a coefficient without considering other terms in the model. For example, when Mileage and Liter are the two predictors in a regression model, the Liter coefficient might seem to indicate that an increase of one in Liter will increase the expected price by \$4968. However, when Cyl is also included in the model, the Liter coefficient seems to indicate that an increase of one in Liter will increase the expected price by \$1545. The size of a regression coefficient and even the direction can change depending on which other terms are in the model.

In this investigation, we have shown that Liter and Cyl are highly correlated. Thus, it is unreasonable to believe that Liter would change by one unit but Cyl would stay constant. The multiple linear regression coefficients cannot be considered in isolation. Instead, the Liter coefficient shows how the expected price will change when Liter increases by one unit, after accounting for corresponding changes in all the other explanatory variables in the model.

> **Key Concept** In multiple linear regression, the coefficients tell us how much the expected response will change when the explanatory variable increases by one unit, after accounting for corresponding changes in all other terms in the model.

3.6 Categorical Explanatory Variables

As we saw in Question 9, there is a clear pattern in the residual versus order plot for the Kelley Blue Book car pricing data. It is likely that one of the categorical variables (Make, Model, Trim, or Type) could explain this pattern.

If any of these categorical variables are related to the response variable, then we want to add these variables to our regression model. A common procedure used to incorporate categorical explanatory variables into a regression model is to define indicator variables, also called dummy variables. Creating indicator variables is a process of mapping the one column (variable) of categorical data into several columns (indicator variables) of 0 and 1 data.

Let's take the variable Make as an example. The six possible values (Buick, Cadillac, Chevrolet, Pontiac, SAAB, Saturn) can be recoded using six indicator variables, one for each of the six makes of car. For example, the indicator variable for Buick will have the value 1 for every car that is a Buick and 0 for each car that is not a Buick. Most statistical software packages have a command for creating the indicator variables automatically.

Creating Indicator Variables

17. Create boxplots or individual value plots of the response variable TPrice versus the categorical variables Make, Model, Trim, and Type. Describe any patterns you see.
18. Create indicator variables for Make. Name the columns, in order, Buick, Cadillac, Chevrolet, Pontiac, SAAB, and Saturn. Look at the new data columns and describe how the indicator variables are defined. For example, list all possible outcomes for the Cadillac indicator variable and explain what each outcome represents.

Any of the indicator variables in Question 18 can be incorporated into a regression model. However, if you want to include Make in its entirety in the model, do not include all six indicator variables. Five will suffice because there is complete redundancy in the sixth indicator variable. If the first five indicator variables are all 0 for a particular car, we automatically know that this car belongs to the sixth category. Below, we will leave the Buick indicator variable out of our model. The coefficient for an indicator variable is an estimate of the average amount by which the response variable will change.

For example, the estimated coefficient for the Saturn variable is an estimate of the average difference in TPrice when the car is a Saturn rather than a Buick (after adjusting for corresponding changes in all other terms in the model).

> **Mathematical Note** For any categorical explanatory variable with g groups, only $g-1$ terms should be included in the regression model. Most software packages use matrix algebra to develop multiple re-

gression models. If all g terms are in the model, explanatory variables will be 100% correlated (you can exactly predict the value of one variable if you know the other variables) and the needed matrix inversion cannot be done. If the researcher chooses to leave all g terms in the model, most software packages will arbitrarily remove one term so that the needed matrix calculations can be completed.

It may be tempting to simplify the model by including only a few of the most significant indicator variables. For example, instead of including five indicator variables for Make, you might consider only using Cadillac and SAAB. Most statisticians would recommend against this. By limiting the model to only indicator variables that are significant in the sample data set, we can overfit the model.

Models are **overfit** when researchers overwork a data set in order to increase the R^2 value. For example, a researcher could spend a significant amount of time picking a few indicator variables from Make, Model, Trim, and Type in order to find the best R^2 value. While the model would likely estimate the mean response well, it unlikely to accurately predict new values of the response variable.

This is a fairly nuanced point. The purpose of variable selection techniques is to select the variables that best explain the response variable. However, overfitting may occur if we break up categorical variables into smaller units and then pick and choose among the best parts of those variables. (The Model Validation section in the extended activities discusses this topic in more detail. [[[add link]]])

Activity: Building Regression Models with Indicator Variables

19 Build a new regression model using TPrice as the response and Mileage, Liter, Saturn, Cadillac, Chevrolet, Pontiac, and SAAB as the explanatory variables. Explain why you expect the R^2 value to increase when you add terms for Make.

20. Create indicator variables for Type. Include the Make and Type indicator variables, plus the variables Liter, Doors, Cruise, Sound, Leather, and Mileage, in a model to predict TPrice. Remember to leave at least one category out of the model for the Make and Type indicator variables (e.g., leave Buick and Hatchback out of the model). Compare this regression model to the other models that you have fit to the data in this investigation. Does the normal probability plot suggest that the residuals could follow a normal distribution? Describe whether the residual versus fit, the residual versus order, and the residual versus each explanatory variable plots look more random than they did in earlier problems.

The additional categorical variables are important in improving the regression model. Figure 3.8 shows that when Make and Type are not in the model, the residuals are clustered. When Make and Type are included in the model, the residuals appear to be more randomly distributed. By incorporating Make and Type, we have been able to explain some of variability that was causing the clustering. In addition, the sizes of the residuals are much smaller. Smaller residuals indicate a better fitting model and a higher R^2 value.

Even though Make and Type improved the residual plots, there is still clustering that might be improved by incorporating Model into the regression equation. However, if the goal is simply to develop a model that accurately estimates retail prices, the R^2 value in Question 20 is already fairly high. Are there a few more terms that can be added to the model that would dramatically improve the R^2 value?

To determine a final model, you should attempt to maximize the R^2 value while simultaneously keeping the model relatively simple. For your final model, you should comment on the residual versus fit, residual versus order, and any other residual plots that previously showed patterns. If any pattern exists in the residual plots, it may be worth attempting a new regression model that will account for these patterns. If the regression model can be modified to address the patterns in the residuals, the R^2 value will improve. However, note that the R^2 value is already fairly high. It may not be worth making the model more complex for only a slight increase in the R^2 value.

21. Create a regression model that is simple (i.e., does not have too many terms) and still accurately predicts retail price. Validate the model assumptions. Look at residual plots and check for heteroskedasticity, multicollinearity, autocorrelation, and outliers. Your final model should not have significant clusters, skewness, outliers, or heteroskedasticity appearing in the residual plots. Submit your suggested least squares regression formula along with a limited number of appropriate graphs that provide justification for your model. Describe why you believe this model is "best."

3.7 What Can We Conclude from the 2005 GM Car Study?

The data are from an observational study, not an experiment. Therefore, even though the R^2 value reveals a strong relationship between our explanatory variables and the response, a significant correlation (and thus a significant coefficient) does not imply a causal link between the explanatory variable and the response. There may be theoretical or practical reasons to believe that mileage (or any of the other explanatory variables) causes lower prices, but the final model can be used only to show that there is an association.

Best subsets and residual graphs were used to develop a model that is useful for describing or predicting the retail price based on a function of the explanatory

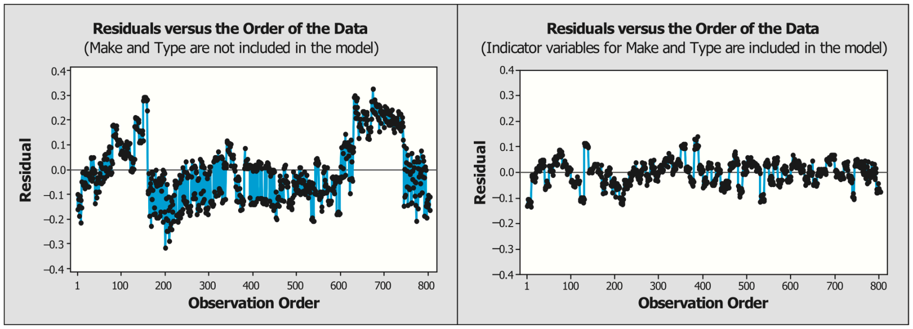

Figure 3.8: **Figure 3.8** Residual versus order plots show that incorporating the indicator variables into the regression model improves the random behavior and reduces the sizes of the residuals.
(#fig:fig3.8)

variables. However, since iterative techniques were used, the p-values corresponding to the significance of each individual coefficient are not reliable.

For this data set, cars were randomly selected within each make, model, and type of 2005 GM car produced, and then suggested retail prices were determined from Kelley Blue Book. While this is not a simple random sample of all 2005 GM cars actually on the road, there is still reason to believe that your final model will provide an accurate description or prediction of retail price for used 2005 GM cars. Of course, as time goes by, the value of these cars will be reduced and updated models will need to be developed.

3.8 F-Tests for Multiple Regression

Decomposition of Sum of Squares

Many of the calculations involved in multiple regression are very closely related to those for the simple linear regression model. Simple linear regression models have the advantage of being easily visualized with scatterplots. Thus, we start with a simple linear regression model to derive several key equations used in multiple regression.

Figure 1 shows a scatterplot and regression line for a subset of the used 2005 Kelly Blue Book Cars data. The data set is restricted to just Chevrolet Cavalier Sedans. In this scatterplot, one specific observation is highlighted: the point where $Mileage$ is 11,488 and the observed $Price$ is $y_i = 14{,}678.1$.

In Figure 1, we see that for any individual observation the total deviation $(y_i - \bar{y})$ is decomposed into two parts:

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \tag{3.7}$$

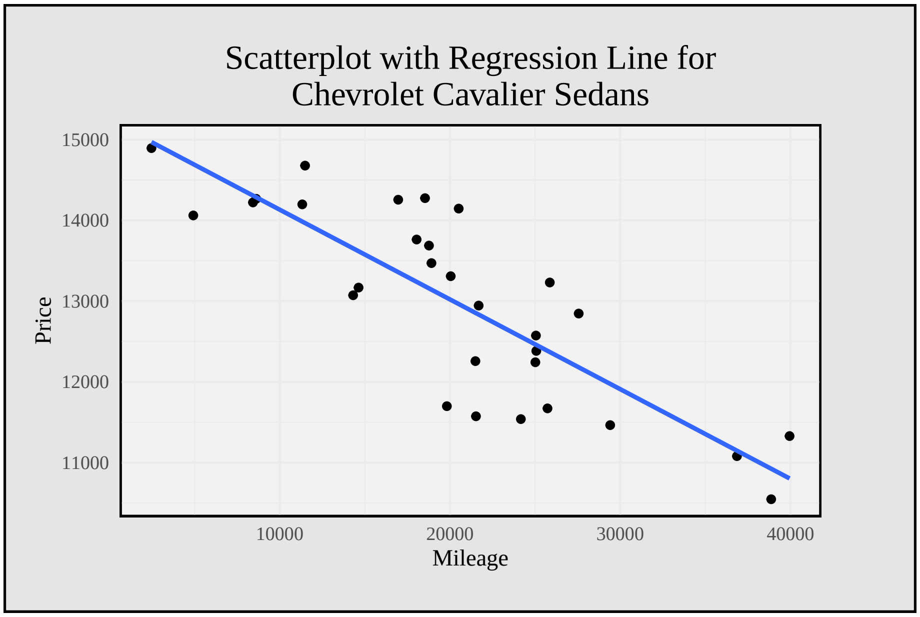

Figure 3.9: (#fig:df_1-plot)Scatterplot and regression line for Chevrolet Cavalier Sedans: Price = 15,244 - 0.111(Mileage).

Using our highlighted observation (11,488, 14,678.1), we see that

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \tag{3.7}$$

$$14{,}678.1 - 12{,}962 = (14{,}678.1 - 13{,}967.68) + (13{,}967.68 - 12{,}962)$$
$$1716.1 = 710.42 + 1005.68$$

Squaring both sides of Equation **??** and then summing over all observations results in

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) \\ &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \qquad (3.8)\end{aligned}$$

The key point of the previous calculations is to show that the total variability in the response, $\sum_{i=1}^{n}(y_i - \bar{y})^2$, can be decomposed into the following:

Total sum of squares (SST) = Residual sum of squares (SSE) + Regression sum of squares (SSR)

MATHEMATICAL NOTE
To show that Equation **??** is true, we can write

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

> **MATHEMATICAL NOTE**
> Recall that the sum of residuals, $\sum_{i=1}^{n}(y_i - \hat{y}_i)$, equals zero. In addition, it can be shown that the sum of the residuals, weighted by the corresponding predicted value, always sums to zero: $\sum_{i=1}^{n}\hat{y}_i(y_i - \hat{y}_i) = 0$. (See Questions 25 through 29.)

Extended Activity: A Closer Look at Least Squares Regression Equations

Data set: *Cavalier*
Note that calculus is required for Activity Questions 25 through 29.

22. Create a regression model to predict Price from Mileage for the Cavalier data. Calculate the total sum of squares (SST), residual sum of squares (SSE), and regression sum of squares (SSR). Verify that SST = SSE + SSR.

23. Show that $\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$ for the model given in the previous question.

24. Using your final model in Question 21, calculate the total sum of squares (SST), residual sum of squares (SSE), and regression sum of squares (SSR). Verify that SST = SSE + SSR.

25. Set the partial derivative of the residual sum of squares with respect to b_0 to zero, to show that $b_0 n + b_1\sum_{i=1}^{n}x_i = \sum_{i=1}^{n}y_i$.
26. Set the partial derivative of the residual sum of squares with respect to b_1 to zero, to show that $b_0\sum_{i=1}^{n}x_i + b_1\sum_{i=1}^{n}x_i^2 = \sum_{i=1}^{n}x_i y_i$.
27. The equations in Questions 25 and 26 are called the normal equations for simple linear regression. Use the normal equations to derive the least squares regression coefficients, b_0 and b_1.

Table 3.2: ANOVA table for a least squares regression model, where n is the number of observations and p is the number of terms in the model (including the constant term).

Source	df	SS	MS	F.Statistic
Regression	$p-1$	$SSR=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$	$MS_{Regr}=\frac{SSR}{df_{Regr}}$	$F=MS_{Regr}/MSE$
Error	$n-p$	$SSE=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	$MSE=\frac{SSE}{df_{Error}}=\hat{\sigma}^2$	
Total	$n-1$	$SST=\sum_{i=1}^{n}(y_i-\bar{y})^2$		

28. Use the fact that $\sum_{i=1}^{n}(y_i-\hat{y}_i)=0$ and $\hat{y}_i=b_0+b_1x_i$ to show that

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

29. Use Questions 26 and 28 to show that $\sum_{i=1}^{n}\hat{y}_i(y_i-\hat{y}_i)=0$.

The Analysis of Variance Table

The objective of regression is to create a model that best fits the observed points. Least squares regression models define a "best fit" as a model that minimizes the sum of squared residual values, $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$.

The coefficient of determination, R^2, is the percentage of variation in the response variable that is explained by the regression line:

> **KEY CONCEPT**
> The coefficient of determination, R^2, is a measure of the usefulness of the explanatory variables in the model. If the explanatory variables are useful in predicting the response, the residual sum of squares, $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$, is small compared to the total spread of the responses, $\sum_{i=1}^{n}(y_i-\bar{y})^2$. In other words, the amount of variability explained by the regression model, $\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$, is a large proportion of the total variability of the responses.

The sum of squares calculations are often summarized in an analysis of variance (ANOVA) table, as shown in Table 1.

[[[The table not displaying. But it works when I knit on separate pdf]]]

Testing the Significance of a Regression Model

Once a model has been developed, we are often interested in testing if there is a relationship between the response and the set of all explanatory terms in the model. To conduct an overall test of model adequacy, we test the following null and alternative hypotheses:

Notice that the β_0 term in our regression model is not included in the null or the alternative hypothesis. Table 1 provides the details for the calculation of the F-statistic:

$$F = \frac{MS_{Regr}}{MSE} = \frac{SSR/(p-1)}{SSE/(n-p)} \tag{3.11}$$

This statistic follows an $F_{p-1,\,n-p}$ distribution, where n is the number of observations and p is the number of terms in the model (including β_0). The same assumptions about the error terms, $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$, need to be checked before conducting the hypothesis test.

NOTE
There are no model assumptions needed about the error terms to calculate estimates of the coefficients. However, all the model assumptions should be checked before conducting a hypothesis test.

The Extra Sum of Squares F-Test

We are often interested in testing the contribution of a particular variable (or subset of variables) to the regression sum of squares. The **extra sum of squares F-test** can test the contribution of a specific set of variables by comparing the residuals of a full and a reduced model.

Suppose a model has been fit with k terms—we call this a **full model**. We may hypothesize that only $p < k$ terms really contribute to the regression model—we call this smaller model the **reduced model**. In this situation, we want to test whether

The previous ANOVA F-test can be modified to provide an F-test for this hypothesis. Notice that this hypothesis test makes no assumptions about the other terms, $\beta_0, \beta_1, \ldots, \beta_{p-1}$, in the model. In addition, *every term in the reduced model must also be in the full model.*

This statistic follows an F-distribution with $k-p$ and $n-k$ degrees of freedom. The extra sum of squares F-test determines whether the difference between the sum of squared residuals in the full and reduced models is so large that it is unlikely to occur by chance.

Extended Activity: Testing Multiple Coefficients

Data set: *Cavalier* Consider the Cavalier data set and the regression model $y = \beta_0 + \beta_1(Mileage) + \beta_2(Cruise) + \epsilon$
30. Submit the ANOVA table, F-statistic, and p-value to test the hypothesis $H_0 : \beta_1 = \beta_2 = 0$ versus H_a: at least one of the coefficients is not 0.
31. Conduct an extra sum of squares test to determine if $Trim$ is useful. More specifically, use the reduced model in the previous question and the full model

$$y = \beta_0 + \beta_1(\text{Mileage}) + \beta_2(\text{Cruise}) + \beta_3(\text{LS Sport Sedan 4D}) + \beta_4(\text{Sedan 4D})$$

to test the hypothesis $H_0 : \beta_3 = \beta_4 = 0$ versus H_a: at least one of the coefficients is not 0.

3.9 Developing a Model to Confirm a Theory

If the goal is to confirm a theoretical relationship, statisticians tend to go through the following steps to identify an appropriate theoretical model.

• Verify that the response variable provides the information needed to address the question of interest. What are the range and variability of responses you expect to observe? Is the response measurement precise enough to address the question of interest? • Investigate all explanatory variables that may be of importance or could potentially influence your results. Note that some terms in the model will be included even though the coefficients may not be significant. In most studies, there is often prior information or a theoretical explanation for the relationship between explanatory and response variables. Nonstatistical information is often essential in developing good statistical models.
• For each of the explanatory variables that you plan to include in the model, describe whether you would expect a positive or negative correlation between that variable and the response variable.
• Use any background information available to identify what other factors are assumed to be controlled within the model. Could measurements, materials, and the process of data collection create unwanted variability? Identify any explanatory variables that may influence the response; then determine if information on these variables can be collected and if the variables can be controlled throughout the study. For example, in the Kelley Blue Book data set, the condition of the car was assumed to be the same for all cars. The data were collected for GM cars with model year 2005. Since these cars were relatively new and the cars were considered to be in excellent condition, any model we create for these data would not be relevant for cars that had been in any type of accident.
• What conditions would be considered normal for this type of study? Are these conditions controllable? If a condition changed during the study, how might it impact the results?

After a theoretical model is developed, regression analysis is conducted one time to determine if the data support the theories.

KEY CONCEPT
The same data should not be looked at both to develop a model and to test it.

Extended Activity: Testing a Theory on Cars

Data set: *Cars*
Assume that you have been asked to determine if there is an association between each of the explanatory variables and the response in the *Cars* data set.

32. Use any background information you may have (not the *Cars* data set) to predict how each explana- tory variable (except $Model$ and $Trim$) will influence $TPrice$. For example, will *Liter* or $Mileage$ have a positive or negative association with $TPrice$? List each $Make$ and identify which will impact $TPrice$ most and in which direction.
33. Identify which factors are controlled in this data set. Can you suggest any factors outside the pro- vided data set that should have been included? If coefficients are found to be significant (have small p-values), will these relationships hold for all areas in the United States? Will the relationships hold for 2004 or 2001 cars?
34. Run a regression analysis to test your hypothesized model. Which variables are important in your model? Did you accurately estimate the direction of each relationship? Note that even if a variable is not significant, it is typically kept in the model if there is a theoretical justification for it.

3.10 Interaction and Terms for Curvature

In addition to using the variables provided in a data set, it is often beneficial to create new variables that are functions of the existing explanatory variables. These new explanatory variables are often quadratic (X^2), cubic (X^3), or a product of two explanatory variables ($X_1 * X_2$), called interaction terms.

An **interaction** is present if the effect of one variable, such as $Mileage$, depends on a second variable, such as Cyl. If an interaction exists, the influence of Cyl changes for different $Mileage$ values, and also the influence of $Mileage$ will depend on Cyl.

The data set $4 - 8Cyl$ includes several four- and eight-cylinder cars from the original Cars data. Figure 2 shows a scatterplot and regression line to predict *Price* using both $Mileage$ and Cyl. The regression model in Figure 2 has no interaction term. The parallel lines show that the estimated impact of changing cylinder size does not depend on mileage. Thus, for any given number of miles, when the number of cylinders changes from four to eight, we expect an increase in Price of $4 \times 3443 = 13{,}772$.

In the same way, the *Mileage* coefficient states that holding *Cyl* constant, we expect *Price* to decrease by 0.20 for each additional mile on the car.

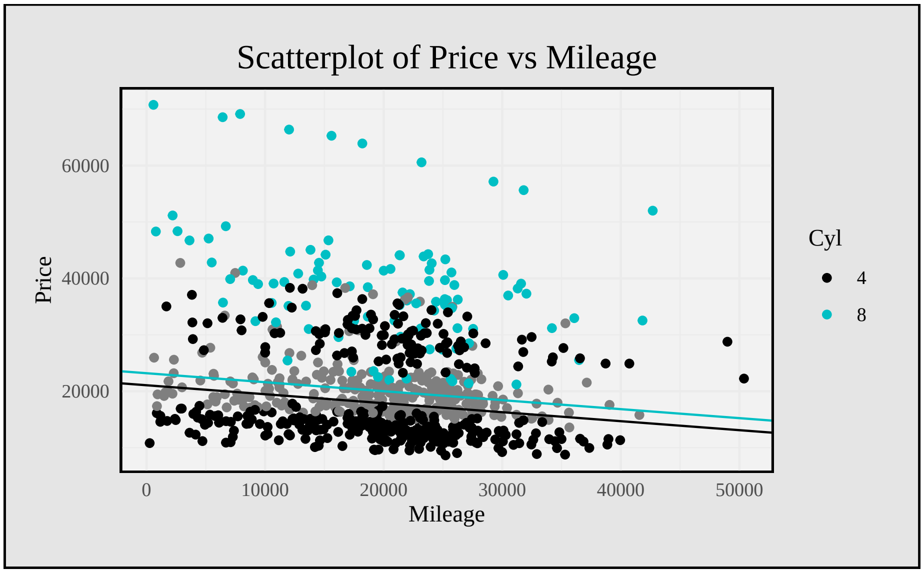

Figure 3.10: (#fig:df_2-plot)Scatterplot and least squares regression line: Price = 15,349 - 0.20(Mileage) + 3443(Cyl). For each cylinder size, an increase of one mile is expected to reduce price by $0.20.

Figure 3 shows a scatterplot and regression line to predict *Price* using *Mileage*, *Cyl*, and a $Mileage * Cyl$ interaction term (called *MileCyl*). The lack of parallel lines in the regression model $\text{Price} = 4533 + 0.340(\text{Mileage}) + 5431(\text{Cyl}) - 0.0995(\text{MileCyl})$ indicates an interaction effect.

Caution should be used in interpreting coefficients when interaction terms are present. The coefficient for *Mileage* can no longer be globally interpreted as reducing *Price* by 0.20 for each additional mile. Now, when there are four cylinders, *Price* is reduced by 0.058 $[0.340(1) - 0.0995(1 \times 4) = -0.058]$ with each additional mile. When there are eight cylinders, *Price* is reduced by 0.456 $[0.340(1) - 0.0995(1 \times 8) = -0.456]$ with each additional mile. Thus, an additional mile impacts *Price* differently depending on the second variable, *Cyl*.

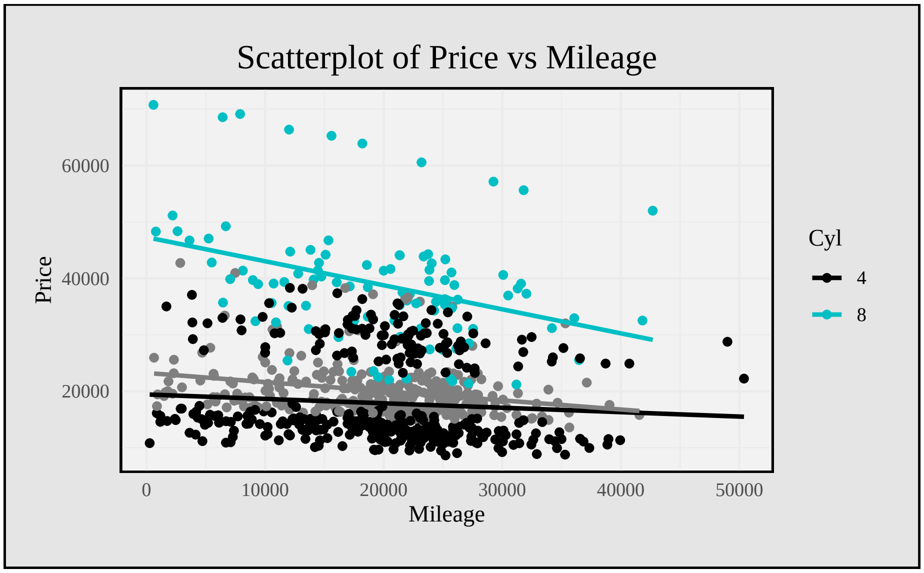

Figure 3.11: (#fig:df_3-plot)Scatterplot and and least squares regression line: Price = 4533 + 0.340(Mileage) + 5431(Cyl) + 0.0995(MileCyl). If the interaction term (MileCyl) is important, we expect to have regression lines that are not parallel.

Extended Activity: Understanding Interaction Terms

Data set: $4-8Cyl$

35. Use the $4-8Cyl$ data set to calculate the two regression equations shown in Figures 2 and 3.

a. Does the R^2_{adj} value increase when the interaction term is added? Based on the change in R^2_{adj}, should the interaction term be included in the model?
b. For both models, calculate the estimated price of a four-cylinder car when $Mileage$ = 10,000.
c. Assuming $Mileage$ = 10,000, for both models explain how increasing from four to eight cylinders will impact the estimated price.
d. Conduct an extra sum of squares test to determine if the $MileCyl$ interaction term is important to the model.

36. Use the $4-8Cyl$ data set to calculate the regression line $Price = \beta_0 + \beta_1(Mileage) + \beta_3(Cadillac) + \beta_4(SAAB)$. You will need to create indicator variables for $Make$ before calculating the regression line.

a. Create a scatterplot with $Mileage$ as the explanatory variable and $Price$

as the response. Overlay a second graph with $Mileage$ as the explanatory variable and $\hat{y}$ as the response. Notice that the predicted values (the $\hat{y}$ values) form two separate lines. Do the parallel lines (no interaction model) look appropriate?

b. Conduct one extra sum of squares test to determine if interaction terms ($MileCadillac$ and $MileSAAB$) are important to the model (i.e., test the hypothesis $H_0 : \beta_5 = \beta_6 = 0$ versus H_a: at least one of the coefficients is not 0, where β_5 and β_6 are the coefficients for the two interaction terms). Create a scatterplot with the full regression model to explain the results of the hypothesis test.

Quadratic and Cubic Terms

If a plot of residuals versus an explanatory variable shows curvature, the model may be improved by including a quadratic term. Is the relationship between mileage and retail price linear or quadratic for the Kelley Blue Book data? To test this, a quadratic term $Mileage * Mileage$ can be created and included in a regression model.

> **MATHEMATICAL NOTE**
> Even though models with quadratic (x^2) or cubic (x^3) terms are not linear functions of the original explanatory variables, the mean response is linear in the regression coefficients ($\beta_0, \beta_1, \beta_2, ...$). For example $y = \beta_0 + z_1\beta_1 + z_2\beta_2 + \epsilon$ would be considered a linear regression model when $z_1 = x$ and $z_2 = x^2$.

Extended Activity: Understanding Quadratic Terms

Data set: MPG The MPG data compare the miles per gallon of several cars against the speed the car was going as well as displacement. Displacement is a measure of the volume of all the cylinders within an engine. The larger the displacement, the more quickly fuel can move through an engine, giving the vehicle more power.

37. Use the MPG data to create a regression model to predict MPG from *Speed* and *Displacement*: MPG $= \beta_0 + \beta_1$(Speed) $+ \beta_2$(Displacement).

a. What are the regression equation and R^2 value?

b. Look at residual versus *Speed* and residual versus *Displacement* plots. Describe any patterns you see.

c. What does the residual normal probability plot show?

38. Create a regression model to predict MPG from Speed: MPG $= \beta_0 + \beta_1$(Speed).

a. What are the regression equation and R^2 value?
b. Look at residual versus *Speed* and residual versus *Displacement* plots. Describe any patterns in the residual plots.
c. Describe any patterns in the residual normal probability plot.
d. Is *Displacement* an important explanatory variable? Use the residual plots and R^2 to give an intuitive explanation.

39. Create a model using displacement to predict MPG: MPG = $\beta_0 + \beta_1$(Displacement).

a. What are the regression equation and R^2 value?
b. Look at residual versus *Speed* and residual versus *Displacement* plots. Describe any patterns in the residual plots.

40. Create a (Speed)^2 term (called $Speed_Sq$) and incorporate that term into your regression model to predict MPG: MPG = $\beta_0 + \beta_1$(Speed) + β_2(Displacement) + β_3(Speed_Sq).

a. What are the regression equation and R^2 value?
b. Look at residual versus *Speed* and residual versus *Displacement* plots. Describe any changes when (Speed)^2 is added to the model.
c. What does the residual normal probability plot show?

Extended Activity: Creating New Terms to Predict the Retail Price of Cars

Data set: *Cars* The potential outliers identified in Question 11 can provide an interesting demonstration of an interaction. Figure 3.12 shows that the slope to predict *Price* from *Mileage* for the ten Cadillac XLR-V8s is much steeper than the slope found when using the other cars. This shows that depreciation for these high-end cars is almost 50 cents a mile, as opposed to 15 cents a mile on average for all car types combined.

41. Create a quadratic mileage term. Create two models to predict $TPrice$, one with only $Mileage$ and another with both $Mileage$ and $(Mileage)^2$ (called $MileSq$).

a. How much does the R^2 value increase if a quadratic term is added to the model $TPrice = \beta_0 + \beta_1(Mileage)$?
b. Look at plots of residuals versus $Mileage$ in both models. Did the addition of the $MileSq$ term improve the residual plots?

42. Create an interaction term $Mileage*Cyl$ (called $MileCyl$). Use $Mileage$, Cyl, and $MileCyl$ to predict $Price$. Does this interaction term appear to improve the model? Use residual plots and R^2 to justify your answer.

    While there is no "best" model, many final models developed by students in Question 21 tend to include the terms *Cadillac*, *Convertible*, and *Liter*. Since each of these terms is related to the Cadillac XLR-V8,

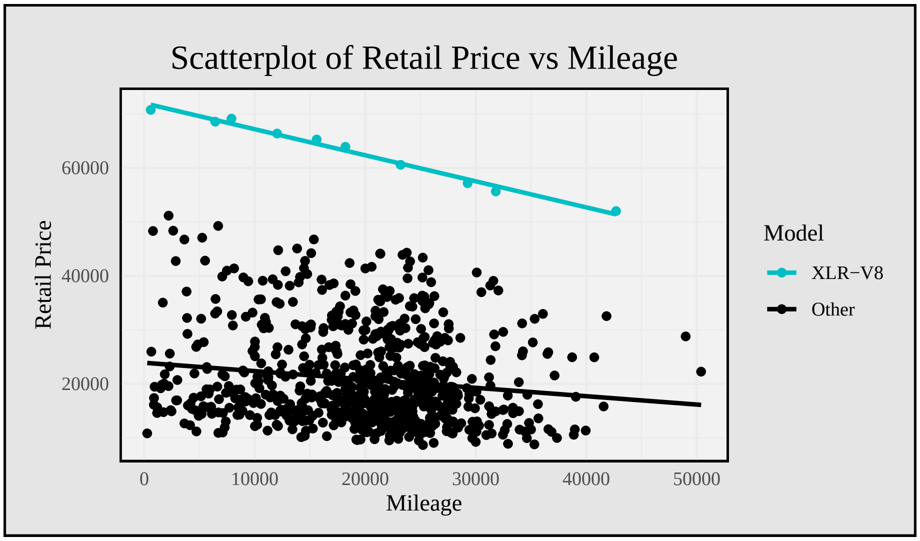

Figure 3.12: (#fig:df_4-plot)Scatterplot and regression lines: For the Cadillac XLR-V8, the regression line is Price = 71,997 - 0.4827(Mileage). This is a much steeper line than the regression line for all other cars: Price = 23,894 - 0.1549(Mileage).

it may be helpful to include an interaction term for $Mileage * Cadillac$, $Mileage * Convertible$, or $Mileage * Liter$. Other $Mileage$, $Make$, or $Type$ interactions may also be helpful additions to the model.

43. Develop additional quadratic and interaction terms. Determine if they improve the regression model in Question 42.

44. Submit a new regression model that best predicts $TPrice$. Does including quadratic or interaction terms improve your model from what was developed in Question 21?

Unless there is a clear reason to include them, researchers typically do not create interaction terms and test whether they should be included in a model. Most of the researcher's effort should be spent on determining whether the original explanatory variables provided in the data set are related to the response. If an interaction term $(X_i * X_j)$ is included in a final model, it is common practice to include each of the original terms (X_i and X_j) in the model as well (even if the coefficients of the original terms are close to zero).

3.11 A Closer Look at Variable Selection Criteria

The growing number of large data sets as well as increasing computer power has dramatically improved the ability of researchers to find a **parsimonious model** (a model that carefully selects a relatively small number of the most useful explanatory variables). However, even with intensive computing power, the process of finding a "best" model is often more of an art than a science.

As shown earlier, stepwise procedures that use prespecified conditions to automatically add or delete variables can have some limitations:

• When explanatory variables are correlated, stepwise procedures often fail to include variables that are useful in describing the results.
• Stepwise procedures tend to overfit the data (fit unhelpful variables) by searching for any terms that explain the variability in the sample results. This chance variability in the sample results may not be reflected in the entire population from which the sample was collected.
• The automated stepwise process provides a "best" model that can be easily misinterpreted, since it doesn't require a researcher to explore the data to get an intuitive feel for the data. For example, stepwise procedures don't encourage researchers to look at residual plots that may reveal interesting patterns within the data.

Adjusted R^2

While R^2 is useful in determining how well a particular model fits the data, it is not very useful in variable selection. Adding new explanatory variables to a regression model will never increase the residual sum of squares; thus, R^2 will increase (or stay the same) even when an explanatory variable does not contribute to the fit.

The **adjusted** R^2 (R^2_{adj}) increases only if the improvement in model fit outweighs the cost of adding an additional term in the model:

$$R^2_{adj} = 1 - \left(\frac{n-1}{n-p}\right)\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \left(\frac{n-1}{n-p}\right)(1 - R^2) \qquad (3.13)$$

where n is the sample size and p is the number of coefficients in the model (including the constant term).

> ##MATHEMATICAL NOTE##
> Intuitively, each additional term in a regression model means that one additional parameter value must be estimated. Each parameter estimate costs an additional degree of freedom. Thus, R^2_{adj} is an R^2

value that is adjusted for degrees of freedom and can be written as

$$R^2_{adj} = 1 - \frac{\text{MSE}}{\text{SST}/(n-1)}$$

R^2_{adj} measures the spread of the residuals using MSE, while R^2 measures the spread of the residuals using SSE.

Mallows' C_p

Another approach to variable selection is to use Mallows' C_p statistic:

$$C_p = (n-p)\left(\frac{\hat{\sigma}^2}{\hat{\sigma}^2_{Full}}\right) + (2p-n) = (n-p)\left(\frac{MSE}{MSE_{Full}}\right) + (2p-n) \quad (3.14)$$

where n is the sample size, p is the number of coefficients in the model (including the constant term), $\hat{\sigma}^2$ estimates the variance of the residuals in the model, and $\hat{\sigma}^2_{Full}$ estimates the variance of the residuals in the full model (i.e., the model with all potential explanatory variables in the data set).

If the current model lacks an important explanatory variable, $\hat{\sigma}^2$ is much larger than $\hat{\sigma}^2_{Full}$ and C_p tends to be large. For any models where $\hat{\sigma}^2$ is similar to $\hat{\sigma}^2_{Full}$, C_p provides a penalty, $2p-n$, to favor models with a smaller number of terms. For a fixed number of terms, minimizing C_p is equivalent to minimizing SSE, which is also equivalent to maximizing R^2.

The C_p statistic assumes that $\hat{\sigma}^2_{Full}$ is an unbiased estimate of the overall residual variability, σ^2. If the full model has several terms that are not useful in predicting the response (i.e., several coefficients are essentially zero), then $\hat{\sigma}^2_{Full}$ will overestimate σ^2 and C_p will be small.*

When the current model explains the data as well as the full model, $\hat{\sigma}^2/\hat{\sigma}^2_{Full} = 1$. Then $C_p = (n-p)(1)+2p-n = p$ Thus, the objective is often to find the smallest C_p value that is close to p.

Akaike's Information Criterion (AIC) and Bayes' Information Criterion (BIC)

Two additional model selection criteria are the Akaike information criterion (AIC) and the Bayesian information criterion (BIC).† Both of these criteria are popular because they are also applicable to regression models fit by maximum likelihood techniques (such as logistic regression), whereas R^2 and C_p are appropriate only for least squares regression models.

Calculations for these two criteria are provided below. These statistics also include a measure of the variability of the residuals plus a penalty term:

$$AIC = n[\log(\hat{\sigma}^2)] + 2p$$

$$BIC = n[\log(\hat{\sigma}^2)] + p[\log(n)]$$

where n is the sample size, p is the number of coefficients in the model, and $\hat{\sigma}^2$ estimates the variance of the residuals in the model.

AIC and BIC are identical except for their penalties, $2p$ and $p[\log(n)]$, respectively. Thus, AIC and BIC will tend to select slightly different models based on n. AIC and BIC both select models that correspond to the smallest value.

KEY CONCEPT
No individual criterion (R^2_{adj}, C_p, AIC, or BIC) is universally better than the other selection criteria. While these tools are helpful in selecting models, they do not produce a model that is necessarily "best."

Model Validation

Often our goal is not just to describe the sample data, but to generalize to the entire population from which the sample was drawn. Even if a regression model is developed that fits the existing sample data very well and satisfies the model assumptions, there is no guarantee that the model will accurately predict new observations.

Variable selection techniques choose variables that account for the variation in the response. When there are many explanatory variables, it is likely that at least some of the terms selected don't explain patterns seen in the entire population; they are included simply because of chance variability seen in the sample.

To validate that a regression model is useful for predicting observations that were not used to develop the model, do the following:

- Collect new data from the same population as the original data. Use the new data to determine the predictive ability of the original regression model.
- Split the original data. For example, randomly select 75% of the observations from the original data set, develop a model, and check the appropriate model assumptions. Test the predictive ability of the model on the remaining 25% of the data set. This is often called **cross-validation**.
- When the data set is not large enough to split, try **jackknife validation**. Hold out one observation at a time, develop a model using the $n - 1$ remaining observations, and then estimate the value of the observation that was held out. Repeat this process for each of the n observations and calculate a predictive R^2 value to evaluate the predictive ability of the model.
- Use theory and prior experience to compare your regression model with other models developed with similar data.

Chapter Summary

In this chapter, we discussed techniques to describe, predict, and test hypotheses about the relationship between a quantitative response variable and multiple explanatory variables. The goals of a regression model will influence which techniques should be used and which conclusions can be drawn. The Cars activities in this chapter focused on developing a model that could describe the relationship between the explanatory variables and response variable as well as predict the value of the response based on a function of the explanatory variables.

Iterative techniques such as best subsets regression are often very useful in identifying which terms should be included in a model. The process of selecting explanatory variables to include in a model often involves iterative techniques, in which numerous models are created and compared at each step in the process. Iterative techniques should not be used when the goal of multiple regression is to test hypotheses. Stepwise regression is used to find the model with the highest R2 value; however, it does not provide much useful information about the model. For example, important variables (such as Liter in our investigation) are often not included in the stepwise regression models. Best subsets regression is a more useful iterative technique because it allows the researcher to better identify important explanatory variables, even if multicollinearity (highly correlated explanatory variables) exists. Table 3.3 lists the key measures used in variable selection.

No individual criterion (\$*{R^2}*{adj}, C_p, AIC, or BIC) is universally better than the other selection criteria. These tools are helpful in selecting models, but they do not produce a model that is necessarily "best."

While iterative techniques are useful in reducing a large number of explanatory variables to a more manageable set, a researcher should ask the following questions to evaluate the resulting model:

- Were the techniques used to create the model appropriate based on the goals of the regression model?
- Do the coefficients make sense? Are the magnitudes of the coefficients reasonable? If the coefficients have the opposite sign than expected, multicollinearity may be present, the range of the explanatory variables may be too small, or important explanatory variables may be missing from the model.
- Do the residual plots identify any outliers or patterns that indicate unexplained structure that should be included in the model?

If the goal is to use hypothesis testing to determine how each of the explanatory variables impacts the response, iterative techniques are not appropriate. In addition, hypothesis tests about specific explanatory variables are not reliable when multicollinearity or lack of normality exists.

Model assumptions need to be met if the goal is to test hypotheses. While least squares regression models can be calculated without checking model assumptions, identifying patterns in residual plots that may indicate **heteroskedasticity**, **autocorrelation**, **outliers**, or **lack of normality** is important to creating a good model. If a pattern exists in any of the residual plots, it is likely that another model exists that better explains the response variable. Researchers need to be somewhat creative in deciding which graphs to create and how to adapt a model based on what they see.

[[[The table is not displaying. But it works when I knit on separate pdf]]]

Table 3.3: (#tab:tab3.3)Table 3.3 Variable selection criteria.

Statistic	Selection Criteria
$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \overline{y})^2}$	Larger is better
$R^2_{\text{adj}} = 1 - \left(\frac{n-1}{n-p}\right) \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \overline{y})^2}$	Larger is better
$C_p = (n-p)\left(\frac{\hat{\sigma}^2}{\hat{\sigma}^2_{\text{Full}}}\right) + (2p - n)$	Close to p is better
$\text{AIC} = n[\log(\hat{\sigma}^2)] + 2p$	Smaller is better
$\text{BIC} = n[\log(\hat{\sigma}^2)] + p[\log(n)]$	Smaller is better

where

n is the sample size \ p is the number of coefficients in the model (including β_0)

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-p}$$

estimates the variance of the residuals in the current model tested. \ $\hat{\sigma}^2_{\text{Full}}$ estimates the variance of the residuals in the full model (the model with all explanatory variables).

where

n is the sample size, p is the number of coefficients in the model (including β_0),

$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-p}$ estimates the variance of the residuals in the current model tested. $\hat{\sigma}^2_{\text{Full}}$ estimates the variance of the residuals in the full model (the model with all explanatory variables).

Histograms or normal probability plots of residuals are used to determine if the residuals follow the normal distribution. Autocorrelation may exist when patterns appear in the ordered residual plots, indicating that each observation is not independent of the prior observation. Heteroskedasticity occurs when the variance of the residuals is not equal. If the variance increases as the expected value increases, a variance stabilizing transformation, such as the natural log or square root of the response, may reduce heteroskedasticity.

Tests of regression coefficients (as in Question 2) and **extra sum of squares tests** can be used for exploratory purposes or to test a theory. The individual t-test for coefficients can be unreliable if the explanatory variables are correlated. In addition, p-values for this t-test become less reliable as more tests are conducted.

The following hypothesis test can be analyzed with Table 3.2:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$$
$$H_a : \text{at least one of the coefficients in the null hypothesis is not } 0$$

The extra sum of squares test shown in Table 3.4 is used to compare a full model (with k terms) to a reduced model (with p terms), where $k > p$. Every term in the reduced model must also be in the full model.

$$H_0 : \beta_p = \beta_{p+1} = \cdots = \beta_{k-1} = 0$$
$$H_a : \text{at least one of the coefficients in the null hypothesis is not } 0$$

Interaction terms and squared (or cubed) terms can be tested with the extra sum of squares test to determine if the additional terms improve the regression model. Testing models with all possible interaction and squared terms can become complex very quickly, so these terms are not typically tested unless there is some reason to include them.

Table 3.4: (#tab:tab3.4)Table 3.4 The extra sum of squares F-statistic is the ratio of the mean square for the extra $k-p$ terms to the MSE_{Full} with $k-p$ and $n-k$ degrees of freedom.

Source	df	SS	MS	F-Statistic
Reduced model	$p-1$	$\text{SSR}_{\text{Reduced}}$	$\frac{\text{SSR}_{\text{Reduced}}}{df_{\text{Reduced}}}$	$\frac{MS_{\text{Reduced}}}{\text{MSE}}$
Extra $k-p$ terms	$k-p$	$\text{SSR}_{\text{Full}} - \text{SSR}_{\text{Reduced}}$	$\frac{\text{SSR}_{\text{Full}} - \text{SSR}_{\text{Reduced}}}{k-p}$	$\frac{MS_{\text{Extra}}}{\text{MSE}}$
Error	$n-k$	SSE_{Full}	$\text{MSE}_{\text{Full}} = \frac{\text{SSE}_{\text{Full}}}{n-k}$	
Total	$n-1$	$\text{SST} = \sum_{i=1}^{n}(y_i - \overline{y})^2$		