PART

4

Analyzing Association and Extended Statistical Methods

11.1 Practicing the Basics

11.1 Gender gap in politics? In the United States, is there a gender gap in political beliefs? That is, do women and men tend to differ in their political thinking and voting behavior? The table taken from the 2012 GSS relates gender and political party identification. Subjects indicated whether they identified more strongly with the Democratic or Republican party or as Independents.

	Political Party Identification			
Gender	**Democrat**	**Independent**	**Republican**	**Total**
Female	421	398	244	1063
Male	278	367	198	843

Source: Data from CSM, UC Berkeley.

a. Identify the response variable and the explanatory variable.

b. Construct a table that shows the conditional distributions of the response variable. Interpret.

c. Give a hypothetical example of population conditional distributions for which these variables would be independent.

d. Sketch bar graphs to portray the distributions in part b and in part c.

11.2 Beliefs of new employees Every year, a large-scale poll of new employees conducted by the human resources management department at a consulting firm asks their opinions on a variety of issues. In 2015, although women were more likely to rate their time management skills as "above average," they were also twice as likely as men to indicate that they frequently felt overwhelmed by all they have to do (38.4% versus 19.3%).

a. If results for the population of new employees were similar to these, would gender and feelings of being overwhelmed be independent or dependent?

b. Give an example of hypothetical population percentages for which these variables would be independent.

11.3 Williams College admission Data from 2013 posted on the Williams College website shows that of all 3,195 males applying, 18.2% were admitted, and of all 3,658 females applying, 16.9% were admitted. Let X denote gender of applicant and Y denote whether admitted.

a. Which conditional distributions do these percentages refer to, those of Y at given categories of X, or those of X at given categories of Y? Set up a table showing the two conditional distributions.

b. Are X and Y independent or dependent? Explain. (*Hint:* These percentages refer to the population of all applicants in 2013.)

11.4 Happiness and gender The contingency table shown relates happiness and gender for the 2012 GSS.

	Happiness			
Gender	**Not Too Happy**	**Pretty Happy**	**Very Happy**	**Total**
Female	154	592	336	1082
Male	123	502	257	882

Source: Data from CSM, UC Berkeley.

a. Identify the response variable and the explanatory variable.

b. Construct a table or graph showing the conditional distributions. Interpret.

c. Give an example of population conditional distributions (i.e., proportions or percentages) that would seem to be consistent with this sample and for which happiness and gender would be independent.

11.5 Marital happiness and income In the GSS, subjects who were married were asked about the happiness of their marriage, the variable coded as HAPMAR.

a. Go to the GSS website sda.berkeley.edu/GSS/, click GSS with no weight as the default, and construct a contingency table for 2012 relating family income (measured as in Table 11.1) to marital happiness: Enter FINRELA(r:4;3;2) as the row variable (where 4 = "above average," 3 = "average," and 2 = "below average") and HAPMAR(r:3;2;1) as the column variable (3 = not too, 2 = pretty, and 1 = very happy). As the selection filter, enter YEAR(2012). Under Output Options, put a check in the row box (instead of in the column box) for the Percentaging option and put a check in the Summary Statistics box further below. Click on *Run the Table*.

b. Construct a table or graph that shows the conditional distributions of marital happiness, given family income. How would you describe the association?

c. Compare the conditional distributions to those in Table 11.2. For a given family income, what tends to be higher, general happiness or marital happiness for those who are married? Explain.

11.6 What is independent of happiness? Which one of the following variables would you think most likely to be independent of happiness: belief in an afterlife, family income, quality of health, region of the country in which you live, satisfaction with job? Explain the basis of your reasoning.

11.7 Sample evidence about independence Refer to the previous exercise. Go to the GSS website and construct a table relating happiness (HAPPY) to the variable you chose (AFTERLIF, FINRELA, HEALTH, REGION, or JOBSAT). Inspect the conditional distributions and indicate whether independence seems plausible, with the sample conditional distributions all being quite similar.

11.8 Lung cancer and smoking In a study conducted by a pharmaceutical company, 605 out of 790 smokers and 122 out of 434 nonsmokers were diagnosed with lung cancer.

a. Construct a 2×2 contingency table relating smoking (SMOKING, categories smoker and nonsmoker) as the rows to lung cancer (LUNGCANCER, categories present and absent) as the columns.

b. Find the four expected cell counts when assuming independence. Compare them to the observed cell counts, identifying cells having more observations than expected.

c. For this data, $X^2 = 272.89$. Verify this value by plugging into the formula for X^2 and computing the sum.

11.9 Happiness and gender For the 2×3 table on gender and happiness in Exercise 11.4 (shown below), software tells us that $X^2 = 1.04$ and the P-value $= 0.59$.

	Happiness		
Gender	**Not**	**Pretty**	**Very**
Female	154	592	336
Male	123	502	257

a. State the null and alternative hypothesis, in context, to which these results apply.

b. Interpret the P-value.

11.10 What gives a P-value = 0.01? How large an X^2 test statistic value provides a P-value of 0.01 for testing independence for the following table dimensions?

a. 2×2

b. 2×3

c. 3×5

d. 4×5

e. 5×9

11.11 Marital happiness and income In Exercise 11.5 when you used the GSS to download a 3×3 table for family income and marital happiness in 2012, you should have obtained results similar to the following table.

	Marital Happiness		
Income	**Not**	**Pretty**	**Very**
Above	6	62	139
Average	7	125	283
Below	6	69	115

a. State the null and alternative hypotheses for the test.

b. What is the number of degrees of freedom for the chi-squared test?

c. The chi-squared statistic for the table equals $X^2 = 4.58$. (i) What value do you expect for X^2 if the null hypothesis were true? (ii) How many standard deviations is 4.58 from this expected value? (*Hint:* The standard deviation of the chi-squared distribution equals $\sqrt{2 \times df}$). (iii) Is $X^2 = 4.58$ an extreme value? Explain.

d. How large an X^2 value would give a P-value of exactly 0.05?

e. Find (at least approximately, using Table C in the appendix) the P-value and give a conclusion for the test in context.

f. Verify that the expected cell count in the first cell equals 4.84. Could this be a problem? Explain.

11.12 First and second free throw independent? In pro basketball games during 1980–1982, when Larry Bird of the Boston Celtics made his first free throw, 251 out of 285 times he made the second one, and when he missed his first free throw, 48 out of 53 times he made the second one.

a. Form a 2×2 contingency table that cross-tabulates the outcome of the first free throw (with categories made and missed) and the outcome of the second free throw (made and missed).

b. When we use MINITAB to analyze the contingency table, we get the result

```
Pearson Chi-Square = 0.273, DF = 1,
P-Value = 0.602
```

Does it seem as if his success on the second shot depends on whether he made the first? Explain how to interpret the result of the chi-squared test. (Here, for the chi-squared test to apply, we assume that each pair of free throws attempted is independent of any other pair during the same or other games. This is reasonable in professional sports.)

11.13 Cigarettes and marijuana The table on the following page refers to a survey[2] in which senior high school students in Dayton, Ohio, were randomly sampled. It cross-tabulates whether a student had ever smoked cigarettes and whether a student had ever used marijuana. Analyze these data by

[2] *Source*: Data from personal communication from Harry Khamis, Wright State University.

(a) finding and interpreting conditional distributions with marijuana use as the response variable and (b) reporting all five steps of the chi-squared test of independence.

	Marijuana	
Cigarettes	**Yes**	**No**
Yes	914	581
No	46	735

11.14 Smoking and alcohol Refer to the previous exercise. A similar table relates cigarette use to alcohol use. The MINITAB output for the chi-squared test follows.

a. True or false: If we use cigarette use as the column variable and alcohol use as the row variable, then we will get different values for the chi-squared statistic and the P-value shown in the output.

b. Explain what value you would get for the z statistic and P-value if you conducted a significance test of H_0: $p_1 = p_2$ against H_a: $p_1 \neq p_2$, where p_1 is the population proportion of non-cigarette users who have drunk alcohol and p_2 is the population proportion of cigarette users who have drunk alcohol.

Dayton student survey

```
Row: cigarette  Columns: alcohol
            no          yes
no         281          500
yes         46         1449

Pearson Chi-Square = 451.404,
DF = 1, P-Value = 0.000
```

11.15 Help the environment In 2010 the GSS asked whether a subject was willing to accept cuts in the standard of living to help the environment (GRNSOL), with categories (vw = very willing, fw = fairly willing, nwu = neither willing nor unwilling, nvw = not very willing, nw = not at all willing). When this was cross-tabulated with whether the respondent is currently in school or has retired (WORKSTAT), as shown below, $X^2 = 9.56$.

a. What are the hypotheses for the test to which X^2 refers?

b. Report r and c and the df value on which X^2 is based.

c. Is the P-value less than (i) 0.05? (ii) 0.025? Explain.

d. What conclusion would you make, using a significance level of (i) 0.05 and (ii) 0.025? State your conclusion in the context of this study.

Help the Environment					
Status	**vw**	**fw**	**nwu**	**nvw**	**nw**
in school	14	47	41	43	27
retired	34	189	177	169	189

11.16 Primary food choice of alligators For alligators caught in two Florida lakes, the following table shows their primary food choice. The four food categories refer to fish, invertebrates (such as snails, insects, or crayfish), birds and reptiles (such as egrets or turtles), and others, including mammals or plants. Is there evidence that primary food choice differs between the two lakes?

	Primary Food				
Lake	**Fish**	**Invertebrates**	**Birds & Reptiles**	**Others**	***n***
Hancock	30	4	8	13	55
Trafford	13	18	12	10	53

a. Find the conditional sample distributions of primary food choice in lakes Hancock and Trafford.

b. Set up the hypotheses of interest.

c. The X^2 value for this table equals 16.79. Based on the df for the corresponding chi-squared distribution, can this be considered large? Why?

d. The P-value for the chi-squared test is less than 0.001. Write the conclusion of the test in context.

11.17 Cognitive behavioral therapy and anxiety A study used 1496 patients suffering from low levels of anxiety. The study randomly assigned each subject to a cognitive behavioral therapy (CBT) treatment or a placebo treatment. In this study, increased anxiety levels were observed for 45 of the 729 subjects taking a placebo and for 29 of the 767 subjects taking CBT.

a. Report the data in the form of a 2×2 contingency table.

b. Show how to carry out all five steps of the null hypothesis that having an anxiety attack is not associated with whether one is taking a placebo or CBT. (You should get a chi-squared statistic equal to 4.5.) Interpret.

11.18 z test for anxiety study Refer to the previous exercise. The printout from MINITAB reports

```
Test for difference = 0 (vs not = 0):
      Z = 2.12 P-Value = 0.033
```

a. Define population proportions p_1 and p_2 and state the hypotheses for that test.

b. Explain how the result of the chi-squared test in part b of the previous exercise corresponds to this z test result.

11.19 Severity of fever after flu shot The study mentioned in Example 5 also looked at the severity of fever (rated as mild, moderate or severe) for all subjects who developed one after receiving a flu shot. The following table shows counts for subjects randomized to the group that received the active ingredient of the flu shot and the placebo group that received a sugar injection.

	Severity		
Group	**Mild**	**Moderate**	**Severe**
Active	39	12	12
Placebo	19	11	4

a. Researchers want to know whether the distribution of the severity of fever is the same in both groups. Formulate appropriate null and alternative hypotheses.

b. The X^2 value for these data equals 2.49. Based on the df for the chi-squared distribution, argue that this value is not extreme.

c. The P-value for the chi-squared test equals 0.287. Write the conclusion of the test in context.

11.20 What is independent of happiness? Refer to Exercises 11.6 and 11.7. For the variables that you thought might be independent,

a. At the GSS website, conduct all five steps of the chi-squared test.

b. Based on part a, which inference is most appropriate? (i) We accept the hypothesis that the variables are independent; (ii) the variables may be independent; (iii) the variables are associated.

11.21 Testing a genetic theory In an experiment on chlorophyll inheritance in corn, for 1,103 seedlings of self-fertilized heterozygous green plants, 854 seedlings were green and 249 were yellow. Theory predicts that 75% of the seedlings would be green.

a. Specify a null hypothesis for testing the theory.

b. Find the value of the chi-squared goodness-of-fit statistic and report its *df*.

c. Report the P-value and interpret.

11.22 Footfall by quarters Based on a random sample of 1098 customers at a grocery store, the table shows how many arrived in the first, second, third, and fourth quarter of the year. Is there evidence that the probabilities of arrival of customers in a given quarter are not equal?

Footfall				
Jan-Mar	**Apr-Jun**	**Jul-Sep**	**Oct-Dec**	***n***
198	340	318	242	1098

a. Formulate the null and alternative hypotheses.

b. Find the expected values and compute the chi-squared statistic, either by hand or using software.

c. How many degrees of freedom is the test based on?

d. Find the P-value and write a conclusion in context.

11.23 Checking a roulette wheel Karl Pearson devised the chi-squared goodness-of-fit test partly to analyze data from an experiment to analyze whether a particular roulette wheel in Monte Carlo was fair, in the sense that each outcome was equally likely in a spin of the wheel. For a given European roulette wheel with 37 pockets (with numbers 0, 1, 2, …, 36), consider the null hypothesis that the wheel is fair.

a. For the null hypothesis, what is the probability for each pocket?

b. For an experiment with 3,700 spins of the roulette wheel, find the expected number of times each pocket is selected.

c. In the experiment, the 0 pocket occurred 110 times. Show the contribution to the X^2 statistic of the results for this pocket.

d. Comparing the observed and expected counts for all 37 pockets, we get $X^2 = 34.4$ Specify the *df* value and indicate whether there is strong evidence that the roulette wheel is not balanced. (*Hint:* Recall that the *df* value is the mean of the distribution.)

11.3 Practicing the Basics

11.24 Democrat, race, and gender The two tables show 2012 GSS data on whether someone is identified as Democrat, by race and by gender.

	Democrat		
race	Yes	No	Total
black	212	88	300
white	422	1046	1468

	Democrat		
gender	Yes	No	Total
female	421	660	1081
male	278	601	879

a. Find the difference of proportions between blacks and whites and between females and males. Interpret each. Which variable has a stronger association with whether someone identifies as Democrat, race or gender? Explain.

b. Find the ratio of proportions between blacks and whites and between females and males. Interpret each. Which variable has a stronger association with whether someone identifies as Democrat, race or gender? Explain.

c. Find the odds of identifying as Democrat for blacks and whites and interpret each. Then find the odds ratio and interpret.

11.25 Death penalty associations Table 11.10, summarized again here, showed the associations between death penalty opinion and gender or race.

	Opinion	
race	Favor	Oppose
black	70%	30%
white	48%	52%

	Opinion	
gender	Favor	Oppose
female	69%	31%
male	61%	39%

a. True or false: The table with the larger X^2 statistic necessarily has the stronger association. Explain.

b. To make an inference about the strength of association in the population, you can construct confidence intervals around the sample differences of proportions. The 95% confidence intervals are (0.15, 0.28) comparing whites and blacks and (0.04, 0.12) comparing males and females. In the *population,* can you make a conclusion about which variable is more strongly associated with the death penalty opinion? Explain.

11.26 Smoking and alcohol The table refers to a survey of senior high school students in Dayton, Ohio. It cross-tabulates

whether a student had ever smoked cigarettes and whether a student had ever drunk alcohol and shows counts and the conditional distributions of alcohol use.

	Alcohol		
Cigarettes	**Yes**	**No**	**Total**
Yes	1449 (97%)	46 (3%)	1495 (100%)
No	500 (64%)	281 (36%)	781 (100%)

a. Describe the strength of association by using the difference between users and nonusers of cigarettes in the proportions who have used alcohol. Interpret.

b. Describe the strength of association by using the relative risk of using alcohol, comparing those who have used or not used cigarettes. Interpret.

c. Find the odds of having used alcohol for users and nonusers of cigarettes. Interpret each. Then, describe the strength of association, using the odds ratio.

11.27 Gender and dominant hand usage The following table cross-tabulates dominant hand usage by gender for 200 individuals. Find and interpret a measure of association, treating hand usage as the response variable.

	Dominant Hand	
Gender	**Right handed**	**Left handed**
Male	86	18
Female	88	8

11.28 Smelling and mortality A recent study (Pinto et al., Olfactory Dysfunction Predicts 5-Year Mortality in Older Adults. PLoS ONE 9(10):e107541, 2014) mentions that anosmic (those with almost no sense of smell) older adults had more than three times the odds of death over a 5-year span compared to normosmic (those with normal smell) individuals. Does this imply that anosmic older adults were more than three times as likely to die over the next 5 years than normosmic ones? Explain.

11.29 Vioxx In September 2004, the pharmaceutical company Merck withdrew its blockbuster drug rofecoxib (a painkiller, better known under its brand name, Vioxx) from the worldwide market amid concerns about its safety. By that time, millions of people had used the drug. In a 2000 study comparing rofecoxib to a control group (naproxen), it was mentioned that "Myocardial infarctions were less common in the naproxen group than in the rofecoxib group (0.1 percent vs. 0.4 percent; 95 percent confidence interval for the difference, 0.1 to 0.6 percent; relative risk, 0.2; 95 percent confidence interval, 0.1 to 0.7)". (*Source:* Bombadier et al., *New England Journal of Medicine*, vol. 343, 2000, pp. 1520-8.)

a. Find and interpret the difference of proportions between the naproxen and rofecoxib groups.

b. Interpret the stated relative risk.

c. In this study, myocardial infarctions were how much more likely to occur in the rofecoxib group than in the naproxen group?

11.30 Egg and cell derived vaccine When comparing the cell-derived flu vaccine mentioned in Example 9 to a more traditionally manufactured egg-derived vaccine, the following data were obtained.

	Developed Flu		
Group	**Yes**	**No**	**Total**
Cell-derived	26	3874	3900
Egg-derived	24	3876	3900

a. Find the relative risk of developing the flu and interpret.

b. Find the odds ratio and interpret.

c. Based on either measure, is it likely that the probability of developing the flu is similar in the two groups? Explain.

11.31 Risk of dying for teenagers According to summarized data from 1999 to 2006 accessed from the Centers of Disease Control and Prevention, the annual probability that a male teenager at age 19 is likely to die is about 0.00135 and 0.00046 for females age 19. (www.cdc.gov)

a. Compare these rates using the difference of proportions, and interpret.

b. Compare these rates using the relative risk, and interpret.

c. Which of the two measures seems more useful when both proportions are very close to 0? Explain.

11.32 Recreation and happiness The table shows data on indulgence in recreational activities and happiness for 398 individuals in a city.

	Happiness		
Recreation	**Not Too Happy**	**Happy**	**Very Happy**
Seldom	37	12	8
Sometimes	25	85	44
Often	9	42	136

a. The chi-squared test of independence has $X^2 = 168.78$. What conclusion would you make using a significance level of 0.05? Interpret.

b. Does this large chi-squared value imply there is a strong association between recreation and happiness? Explain.

c. Find the difference in the proportion of being not too happy between those who seldom indulge into recreation and those who often indulge into recreation. Interpret that difference.

d. Find and interpret the relative risk of being not too happy, comparing the lowest and highest recreation group. Interpret.

11.33 Party ID and gender The table shows the 2012 GSS data on gender and political party identification from Exercise 11.1. (The row totals are slightly different from the second table in Exercise 11.24 because selecting Independent is ignored.) The chi-squared test of independence has $X^2 = 10.04$ with a P-value of 0.0066,

indicating a significant association. Let's describe this association:

	Political Party Identification			
Gender	**Democrat**	**Independent**	**Republican**	**Total**
Female	421	398	244	1063
Male	278	367	198	843

Source: Data from CSM, UC Berkeley.

a. Estimate the difference between females and males in the proportion who identify themselves as Republicans. Interpret.

b. Estimate the difference between females and males in the proportion who identify themselves as Democrat. Interpret.

c. Estimate the ratio between females and males in the proportion who identify themselves as Republican. Interpret.

d. Estimate the ratio between females and males in the proportion who identify themselves as Democrat. Interpret.

e. What can you say about the strength of the association between gender and whether identifying as Republican? What about gender and whether identifying as Democrat?

11.34 Chi-squared versus measuring association For the table on recreation and happiness in Exercise 11.32, the chi-squared statistic equals 168.78 ($df = 4$, P-value < 0.0001). Explain the difference between the purposes of the chi-squared test in part a and the descriptive analysis in parts c and d in that exercise, which compares conditional distributions by using measures of association. (*Hint:* Is a chi-squared test a descriptive or inferential analysis?)

11.35 Standardized residuals for happiness and income

TRY The table displays the observed and expected cell counts and the standardized residuals for testing independence of happiness and family income, for GSS data.

Rows: Income	Columns: Happiness			
	not	pretty	very	All
above	29	178	135	342
	42.6	194.6	104.8	
	-2.490	-2.021	3.955	
average	83	494	277	854
	106.4	485.9	261.7	
	-3.410	0.787	1.598	
below	104	314	119	537
	66.9	305.5	164.5	
	5.830	0.889	-5.131	
All	216	986	531	1733

a. How would you interpret the standardized residual of −2.49?

b. Interpret the standardized residuals highlighted in green.

c. Interpret the standardized residuals highlighted in red.

11.36 Happiness and religious attendance The table shows MINITAB output for data from the 2008 GSS on happiness and frequency of attending religious services (1 = at most several times a year, 2 = once a month to nearly every week, 3 = every week to several times a week).

```
Rows: religion      Columns:       happiness
Not too happy     Pretty happy     Very happy

1     201             609             268
    4.057           1.731          -5.107
2      46             224             132
   -2.566           0.456           1.541
3      66             265             196
   -2.263          -2.376           4.385
Cell Contents:      Count
                    Adjusted residual
Pearson Chi-Square = 36.445, DF = 4,
P-Value = 0.000
```

a. Based on the chi-squared statistic and P-value, give a conclusion about the association between the variables.

b. The numbers below the counts in the table are standardized residuals. Which cells have strong evidence that in the population there are more subjects than if the variables were independent?

c. Which cells have strong evidence that in the population there are fewer subjects than if the variables were independent?

11.37 Recreation and happiness Exercise 11.32 showed the association between recreation and happiness. The table shown here gives the standardized residuals for those data in parentheses.

a. Explain what a relatively small standardized residual such as −0.5 in the second cell represents.

b. Identify the cells in which you would infer that the population has more cases than would occur if recreation and happiness were independent. Pick one of these cells and explain the association relative to independence.

	Happiness		
Recreation	**Not Too Happy**	**Happy**	**Very Happy**
Seldom	37 (8.4)	12 (−1.8)	8 (−3.6)
Sometimes	25 (−0.5)	85 (4.3)	44 (−3.4)
Often	9 (−4.2)	42 (−2.9)	136 (5.1)

11.38 Happiness and marital status The screen shot from the GSS website shows standardized residuals (called *Z*-statistics) and cell counts for 2012 GSS data on happiness (the column) and marital status (the row). The color coding is based on the magnitude of the *Z*-statistic (= standardized residual) and is explained in the legend. Summarize the association by indicating which marital statuses have strong evidence of (i) more and (ii) fewer people in the population in the Very Happy category than if the variables were independent.

Frequency Distribution

Cells contain: -Z-statistic -N of cases		HAPPY 1 VERY HAPPY	HAPPY 2 PRETTY HAPPY	HAPPY 3 NOT TOO HAPPY	ROW TOTAL
MARITAL	1: MARRIED	10.3 375	-3.8 458	-8.1 64	--- 897
	2: WIDOWED	-2.4 35	1.2 97	1.5 29	--- 161
	3: DIVORCED	-4.2 64	1.9 191	2.8 60	--- 315
	4: SEPARATED	-2.7 10	-1.2 32	5.3 24	--- 66
	5: NEVER MARRIED	-5.5 109	2.4 316	3.8 100	--- 525
	COL TOTAL	--- 593	--- 1,094	--- 277	--- 1,964

Color coding:	<-2.0	<-1.0	<0.0	>0.0	>1.0	>2.0	Z
N in each cell:	Smaller than expected			Larger than expected			

11.39 Gender gap? The table in Exercise 11.1 on gender and party identification is shown again. The largest standardized residuals in absolute value were +2.98 for females who identified as Democrats and −2.98 for males who identified as Democrats. Interpret.

	Political Party Identification			
Gender	**Democrat**	**Independent**	**Republican**	**Total**
Female	421	398	244	1063
Male	278	367	198	843

Source: Data from CSM, UC Berkeley.

11.40 Ideology and political party Go to the GSS website sda.berkeley.edu/GSS/. Construct the 7 × 8 contingency table relating political ideology (POLVIEWS) and party identification (PARTYID) for the year 2012. (Enter YEAR (2012) in the Selection filter.) Select Summary Statistics and *Z*-statistic under Output Options to get the value of the chi-squared statistic and standardized residuals.

a. Summarize the results of carrying out the chi-squared test.

b. What do you learn from the residuals that you did not learn from the chi-squared test?

11.5 Practicing the Basics

11.41 Keeping old dogs mentally sharp In an experiment with beagles ages 7–11, the dogs attempted to learn how to find a treat under a certain black-colored block and then relearn that task with a white-colored block. The control group of dogs received standard care and diet. The diet and exercise group were given dog food fortified with vegetables and citrus pulp and vitamin E and C supplements plus extra exercise and social play. All 12 dogs in the diet and exercise group were able to solve the entire task, but only 2 of the 8 dogs in the control group could do so. (Background material from N. W. Milgram et al., *Neurobiology of Aging,* vol. 26, 2005, pp. 77–90.)

a. Show how to summarize the results in a contingency table.

b. Conduct all steps of Fisher's exact test of the hypothesis that whether a dog can solve the task is independent of the treatment group. Use the two-sided alternative hypothesis. Interpret.

c. Why is it improper to conduct the chi-squared test for these data?

11.42 Tea-tasting results Consider the tea-tasting experiment of Example 11 and Table 11.16. Consider the possible sample table in which all four of her predictions about the cups that had milk poured first are correct. Using software, find the P-value for the one-sided alternative. Interpret the P-value.

11.43 Claritin and nervousness An advertisement by Schering Corporation for the allergy drug Claritin mentioned that in a pediatric randomized clinical trial, symptoms of nervousness were shown by 4 of 188 patients on Claritin and 2 of 262 patients taking placebo. Denote the population proportion who would show such symptoms by p_1 for Claritin and by p_2 for placebo. The computer printout shows results of significance tests for H_0: $p_1 = p_2$.

a. Report the P-value for the small-sample test with H_a: $p_1 \neq p_2$. Interpret in the context of this study.

b. Is it appropriate to conduct the chi-squared test for these data? Why or why not?

Analyses of Claritin data

```
Rows: treatment  Columns: nervousness

              yes      no

Claritin        4     184
Placebo         2     260

Statistic                          P-Value
Fisher's Exact Test (2-Tail)          0.24
Chi-squared = 1.55                    0.21
```

11.44 AIDS and condom use Chatterjee et al. (1995, p. 132) described a study about the effect of condoms in reducing the spread of AIDS. This two-year Italian study followed heterosexual couples where one partner was infected with the HIV virus. Of 171 couples who always used condoms, 3 partners became infected with HIV, whereas of 55 couples who did not always use condoms, 8 partners became infected. Test whether the rates are significantly different.

a. Define p_1 and p_2 in this context and specify the null and two-sided alternative hypotheses.

b. For these data, software reports

```
Pearson Chi-Square = 14.704, DF = 1,
P-Value = 0.0001
*Note* 1 cell with expected count
less than 5 Fisher's exact test:
P-Value = 0.0007
```

Report the result of the test that you feel is most appropriate for these data. For the test you chose, report the P-value and interpret in the context of this study.

11.45 Fitness workshop worthwhile? TECH During a fitness workshop, participants were told about the importance of proper meal timings in staying fit. All 5 female participants changed their meal times as suggested during the workshop. Out of 12 male participants, 6 shifted to proper meal times and the rest didn't. Is there evidence that female participants had a larger probability of shifting to proper meal timings?

a. Write these results as a 2×2 contingency table.

b. The sampling distribution (derived from all possible 6188 permutations) for the possible counts in the first cell of the contingency table is

cell count:	0	1	2	3	4	5
#perms:	6	165	1100	1980	462	6188

Find the proportion of tables that have a cell count as large or larger than the one observed.

c. Find the permutation P-value and write a conclusion for the hypothesis when using a 5% significance level.

d. Check your results by entering the contingency table from part a into the Fisher's Exact Test web app accessible from the book's website.

11.46 Proper meal timings enhance fitness? TRY TECH Refer to the previous exercise. Two months after the workshop, participants were asked whether shifting to proper meal timings was beneficial or not, with the possible answers (i) Better, I am fitter than before and (ii) Same, I feel the same. Results of that survey are shown in the following table. Is there evidence that shifting to proper meal timings was beneficial for fitness?

	Fitness status after 2 months	
Shifted to proper meal timings	**Better**	**Same**
Yes	9	2
No	2	4

a. Formulate the null and alternative hypotheses of interest.

b. Suggest a test statistic to use for this test.

c. If all expected cell counts were at least 5, what distribution would the test statistic from part b follow?

d. Of 10,000 random permutations, 475 resulted in a X^2 statistic as large or larger than the observed one of 3.2. Approximate the permutation P-value and write a conclusion for the test.

e. Use the Permutation Test of Independence web app accessible from the book's website to enter the contingency table and replicate the results from part d.

12.1 Practicing the Basics

12.1 Car mileage and weight The Car Weight and Mileage data file on the book's website shows the weight (in pounds) and mileage (miles per gallon) of 25 different model autos.

TRY

a. Identify the natural response variable and explanatory variable.

b. The regression of mileage on weight has MINITAB regression output

```
Term          Coef     SE Coef   T-Value P-Value
Constant     45.65      2.60     17.54   0.000
weight    −0.005222  0.000627    −8.33   0.000
```

State the prediction equation and report the y-intercept and slope.

c. Interpret the slope in terms of a 1,000-pound increase in the vehicle weight.

d. Does the y-intercept have any contextual meaning for these data? (*Hint:* The weight values range between 2,460 and 6,400 pounds.)

12.2 Predicting car mileage Refer to the previous exercise.

a. Find the predicted mileage for the Toyota Corolla, which weighs 2,590 pounds.

b. Find the residual for the Toyota Corolla, which has observed mileage of 38.

c. Sketch a graphical representation of the residual in part b.

12.3 Predicting maximum bench strength in males For the Male Athlete Strength data file on the book's website, the prediction equation relating y = maximum bench press (maxBP) in kilograms to x = repetitions to fatigue bench press (repBP) is $\hat{y} = 117.5 + 5.86\,x$.

a. Find the predicted maxBP for a male athlete with a repBP of 35, which was one of the highest repBP values.

b. Find the predicted maxBP for a male athlete with a repBP of 0, which was the lowest repBP value.

c. Interpret the y-intercept. Use the slope to describe how predicted maxBP changes as repBP increases from 0 to 35.

12.4 Higher income with experience TRY Suppose the regression line $\mu_y = -10{,}000 + 9500x$ models the relationship for the population of working adults in a country between x = experience (in years) and the mean of y = annual income (in US dollars). The conditional distribution of y at each value of x is modeled as normal with $\sigma = 6500$. Use this regression model to describe the mean and the variability around the mean for the conditional distribution at an experience of (a) 5 years and (b) 10 years.

12.5 Ensuring linear relationship In a linear regression model, how does one ensure that the relationship between the dependent variable and the independent variable is linear? Explain.

12.6 Fast food and indigestion Let y = number of times fast food was eaten in the past month and x = number of times indigestion happened in the past month, measured for all students at your school. Explain the mean and variability aspects of the regression model $\mu_y = \alpha + \beta x$ in the context of these variables. In your answer, explain why (a) it is more sensible to use a straight line to model the *means* of the conditional distributions rather than individual observations and (b) the model needs to allow variation around the mean.

12.7 Study time and college GPA TECH Exercise 3.39 in Chapter 3 showed data collected at the end of an introductory statistics course to investigate the relationship between x = study time per week (average number of hours) and y = college GPA. The table here shows the data for the eight males in the class on these variables and on the number of class lectures for the course that the student reported skipping during the term.

Student	Study Time	GPA	Skipped
1	14	2.8	9
2	25	3.6	0
3	15	3.4	2
4	5	3.0	5
5	10	3.1	3
6	12	3.3	2
7	5	2.7	12
8	21	3.8	1

a. Create a data file and use it to construct a scatterplot between x and y. Interpret.

b. Find the prediction equation and interpret the slope.

c. Find the predicted GPA for a student who studies 25 hours per week.

d. Find and interpret the residual for Student 2, who reported $x = 25$.

12.8 GPA and skipping class TECH Refer to the previous exercise. Now let x = number of classes skipped and y = college GPA.

a. Construct a scatterplot. Does the association seem to be positive or negative?

b. Find the prediction equation and interpret the y-intercept and slope.

c. Find the predicted GPA and residual for Student 1.

12.9 Cell phone specs TRY TECH Refer to the cell phone data set available on the book's website, which shows various specs of a random sample of cell phones. Engineers would like to analyze how the weight (measured in grams) of a phone depends on the size of the battery, the heaviest component of a cell phone. Here, the size is measured by the capacity of the battery, which is the amount of energy it can supply on a full charge (measured in milliampere-hours, mAh).

a. Identify the response and explanatory variables and then construct and interpret the scatterplot. Mention any outliers you see.

b. What effect will the outlier have on the fitted regression equation? Will its residual be positive or negative, and will it be small or large in absolute value?

c. Remove the outlier and find the prediction equation. Predict the weight of a cell phone when the battery capacity is (i) 1,000 mAh and (ii) 1,500 mAh.

d. Interpret the slope in context.

12.10 Exercise and watching TV TECH For the Georgia Student Survey file on the book's website, let y = exercise and x = watch TV (minutes per day).

a. Construct a scatterplot. Identify an outlier that could have an impact on the fit of the regression model. What would you expect its effect to be on the slope?

b. Fit the model with and without that observation. Summarize its impact.

12.11 *t*-score? A regression analysis is conducted with 32 observations.

a. What is the *df* value for inference about the slope β?

b. Which two *t* test statistic values would give a P-value of 0.10 for testing H_0: $\beta = 0$ against H_a: $\beta \neq 0$?

c. What is the value of the *t*-score that you multiply the standard error with to find the margin of error for a 90% confidence interval for β?

12.12 Predicting house prices For the House Selling Prices FL data file on the book's website, MINITAB results of a regression analysis are shown for 100 homes relating y = selling price (in dollars) to x = the size of the house (in square feet).

a. Using this output, go through all steps of a significance test of independence (testing whether the population slope equals 0) by (i) mentioning the necessary assumptions for inferences to be valid, (ii) stating the hypotheses, (iii) giving the value of the test statistic, (iv) stating the P-value, and (v) writing a conclusion.

b. Show that a 95% confidence interval for the population slope is (64, 90). (*Hint:* The *t*-score with $df = 98$ equals 1.984.)

c. A builder had claimed that the selling price increases $100, on average, for every extra square foot. Based on part b, what would you conclude about this claim?

House selling prices and size of home

Term	Coef	SE Coef	T-Value	P-Value
Constant	9161	10760	0.85	0.397
size	77.01	6.63	11.62	0.000

12.13 Confidence interval for slope Refer to the previous exercise, which mentioned a confidence interval of (64, 90) for the slope. The 100 houses included in the data set had sizes ranging from 370 square feet to 4,050 square feet.

a. Interpret what the confidence interval implies for a one-unit increase in the size of the house.

b. Repeat part a, now using a more meaningful increase of 100 square feet in the size of a house.

12.14 House prices in bad part of town Refer to the previous exercise. Of the 100 homes, 25 were in a part of town considered less desirable. For a regression analysis using y = selling price and x = size of house for these 25 homes,

a. You plan to test H_0: $\beta = 0$ against H_a: $\beta > 0$. Explain what H_0 means and why a data analyst might choose a one-sided H_a for this test.

b. For this one-sided alternative hypothesis, how large would the *t* test statistic need to be to get a P-value equal to (i) 0.05 and (ii) 0.01?

12.15 Strength through leg press The high school female athlete strength study also considered prediction of y = maximum leg press (maxLP) using x = number of 200-pound leg presses (LP200). MINITAB results of a regression analysis are shown.

Regression Equation maxLP = 233.9 + 5.271 LP200

Term	Coef	SE Coef	T-Value	P-Value
Constant	233.9	13.1	17.90	0.000
LP200	5.271	0.547	9.64	0.000

a. Show all steps of a two-sided significance test of the hypothesis of independence.

b. Use the quoted *se* to find and then interpret a 95% confidence interval for the true slope. (*Hint:* $t_{0.025} = 2.00$.) What do you learn from the interval that you cannot learn from the significance test in part a?

12.16 More boys are bad? A study of 375 women who lived in pre-industrial Finland (by S. Helle et al., *Science*, vol. 296, p. 1085, 2002), using Finnish church records from 1640 to 1870, found that there was roughly a linear relationship between y = life length (in years) and x = number of sons the woman had, with a slope estimate of −0.65 ($se = 0.29$).

a. Interpret the sign of the slope. Is the effect of having more boys good or bad?

b. Show all steps of the test of the hypothesis that life length is independent of number of sons for the two-sided alternative hypothesis. Interpret the P-value.

c. Construct a 95% confidence interval for the true slope. Interpret. Is it plausible that the effect is relatively weak, with true slope near 0?

12.17 More girls are good? Repeat the previous exercise using x = number of daughters the woman had, for which the slope estimate was 0.44 ($se = 0.29$).

12.18 CI and two-sided tests correspond Refer to the previous two exercises. Using significance level 0.05, what decision would you make? Explain how that decision is in agreement with whether 0 falls in the confidence interval. Do this for the data for both the boys and the girls.

12.19 Investment and rate of interest. A market research company wants to study the relationship between y = investment (in pounds) and x = rate of interest (in percentage), for a British commercial bank. For the last four months, the observations are as shown in the table. The correlation equals 0.857.

Investment	Rate of interest
4000	4.0
7000	5.0
8000	6.0
9000	9.0

a. Find the mean and standard deviation for each variable.

b. Using the formulas for the slope and the *y*-intercept or software, find the regression line.

c. The *se* of the slope estimate is 364.21. Test the null hypothesis that these variables are independent, using a significance level of 0.05.

12.20 GPA and study time—revisited Refer to the association you investigated in Exercise 12.7 between study time and college GPA. Using software with the data file you constructed, conduct a significance test of the hypothesis of independence for the one-sided alternative of a positive population slope. Report the hypotheses, appropriate assumptions, sample slope, its standard error, the test statistic, and the P-value and interpret.

12.21 GPA and skipping class—revisited Refer to the association you investigated in Exercise 12.8 between skipping class and college GPA. Using software with the data file you constructed, construct a 90% confidence interval for the slope in the population. Interpret.

12.22 Battery capacity Refer to the cell phone data set from Exercise 12.9 about various specs of cell phones. Treat the weight of the phone as the response and the capacity of the battery as the explanatory variable. Remove the outlier (phone no. 70).

a. Is there evidence for an association between these two variables? Show all steps of a relevant significance test with significance level 0.05 and interpret.

b. Confirm, using the output of the software, that the 95% confidence interval for the population slope equals (0.028, 0.060). Interpret the interval and explain the correspondence with the result of the significance test in part a.

12.23 Euros and thousands of euros If a slope is 1.63 when x = investment in thousands of euros, then what is the slope when x = investment in euros? (*Hint*: A €1 change has only 1/1000 of the impact of a €1000 change.)

12.24 When can you compare slopes? Although the slope does not measure association, it *is* useful for comparing effects for two variables that have the *same* units. Let x = GDP (thousands of pounds per capita). For predicting y = consumer expenditure, the prediction equation is $\hat{y} = 3034.89 + 0.52x$. For predicting y = investment expenditure, the prediction equation is $\hat{y} = 2037.73 + 0.27x$.

a. Explain how to interpret the two slopes.

b. Explain why a one-unit increase in GDP has a slightly greater impact on consumer expenditure than on investment expenditure.

12.25 Sketch scatterplot Sketch a scatterplot, identifying quadrants relative to the sample means as in Figure 12.2, for which (a) the slope and correlation would be negative and (b) the slope and correlation would be approximately zero.

12.26 Sit-ups and the 40-yard dash Is there a relationship between x = how many sit-ups you can do and y = how fast you can run 40 yards (in seconds)? The MINITAB output of a regression analysis based on the female athlete strength study is shown here.

```
Coefficients
Term          Coef    SE Coef  T-Value  P-Value
Constant     6.707      0.178    37.70    0.000
SIT-UP    −0.02435    0.00635   −3.83    0.000
Model Summary
       S       R-sq
0.327208     21.10%
```

a. Find the predicted time in the 40-yard dash for a subject who can do (i) 10 sit-ups and (ii) 40 sit-ups. (The minimum and maximum in the study were 10 and 39.) Relate the difference in predicted times to the slope.

b. The correlation between these two variables equals −0.46 (the negative square root of the R-Sq value shown, because the estimated slope has a negative sign). Predict by how many standard deviations the time of the 40-yard dash decreases for a one standard deviation increase in the number of sit-ups the athlete can do.

c. The number of sit-ups had mean 27.175 and standard deviation 6.887. The time in the 40-yard dash had mean 6.045 and standard deviation 0.365. Show how the correlation relates to the slope and verify the value given in part b.

d. The time difference for the 40-yard run between two athletes that are one standard deviation apart in the number of sit-ups they can do is predicted to be how much? First, answer in terms of standard deviations of the time for the 40-yard dash. Then, using the standard deviation given in part c, answer in absolute terms (i.e., in seconds).

12.27 Body fat For the Male Athlete Strength data file on the book's website, the correlation between weight (pounds) and percent body fat (BF%) equals 0.883.

a. Interpret the sign and the strength of the correlation.

b. Find and interpret r^2.

c. If weight were measured instead with metric units, would any results differ in parts a and b? Explain.

12.28 Verbal and math GRE scores All graduate students who attend an Irish university must submit their math and verbal GRE scores. Both the scores have a mean of 150 and a standard deviation of 6.5. The regression equation relating y = verbal GRE score and x = math GRE score is $\hat{y} = 30 + 0.80x$.

a. Find the predicted verbal GRE score for a student who has the mean math GRE score of 150. (*Note*: At the x value equal to $\bar{x}$, the predicted value of y equals $\bar{y}$.)

b. Show how to find the correlation. Interpret its value as a standardized slope. (*Hint:* Both standard deviations are equal.)

c. Find r^2 and interpret its value.

12.29 GRE score regression toward mean Refer to the previous exercise.

a. Predict the verbal GRE score for a student whose math GRE score = 170.

b. The correlation is 0.8. Interpret the prediction in part a in terms of regression toward the mean.

12.30 GPA and TV watching For the Georgia Student Survey data file on the book's website, the correlation between daily time spent watching TV and college GPA is −0.35.

a. Interpret r and r^2. Use the interpretation of r^2 that (i) refers to the prediction error and (ii) the percent of variability explained.

b. One student is 2 standard deviations above the mean on time watching TV. (i) Would you expect that student to be above or below the mean on college GPA? Explain (ii) How many standard deviations would you expect that student to be away from the mean on college GPA? Use your answer to explain "regression toward the mean."

12.31 GPA and study time Refer to the association you investigated in Exercise 12.7 between study time and college GPA. Using software or a calculator with the data file you constructed for that exercise,

a. Find and interpret the correlation.

b. Find and interpret r^2. Use the interpretation of r^2 that (i) refers to the prediction error and (ii) the percent of variability explained.

12.32 Placebo helps cholesterol? A clinical trial admits subjects suffering from high cholesterol, who are then randomly assigned to take a drug or a placebo for a 12-week study. For the population, without taking any drug, the correlation between the cholesterol readings at times 12 weeks apart is 0.70. The mean cholesterol reading at any given

time is 200, with the same standard deviation at each time. Consider all the subjects with a cholesterol level of 300 at the start of the study, who take placebo.

a. What would you predict for their mean cholesterol level at the end of the study?

b. Does this suggest that placebo is effective for treating high cholesterol? Explain.

12.33 Was the advertising strategy helpful? Among the 100 different varieties of bread made by a bakery, the marketing manager selected the 10 worst-selling bread types and promoted them through a special advertising strategy. Both the mid-month and the end-month sales had an average of 70 packets with a standard deviation of 10 packets considering all the different bread types, and the correlation was 0.50 between the two types of sales. The average for the specially promoted bread types increased from 50 packets in the middle of the month to 60 packets at the end of the month. Can we conclude that the advertising strategy was successful? Explain by identifying the response and explanatory variables and the role of regression toward the mean.

12.34 What's wrong with your stock fund? Last year you looked at all the financial firms that had stock growth funds. You picked the growth fund that had the best performance last year (ranking at the 99th percentile on performance) and invested all your money in it this year. This year, with its new investments, it ranked only at the 65th percentile on performance. Your friend suggests that its stock picker became complacent or was burned out. Can you give another explanation?

12.35 Golf regression In the first round of a golf tournament, five players tied for the lowest round, at 65. The mean score of all players was 75. If the mean score of all players is also 75 in the second round, what does regression toward the mean suggest about how well we can expect the five leaders to do, on the average, in the second round? (*Hint:* Suppose the standard deviation is also the same in each round.)

12.36 Car weight and mileage The Car Weight and Mileage data file on the book's website shows the weight and the mileage per gallon of gas of 25 cars of various models. The regression of mileage on weight has $r^2 = 0.75$. Explain how to interpret this in terms of how well you can predict a car's mileage if you know its weight.

12.37 Food and drink sales The owner of Bertha's Restaurant is interested in whether an association exists between the amount spent on food and the amount spent on drinks for the restaurant's customers. She decides to measure each variable for every customer in the next month. Each day she also summarizes the mean amount spent on food and the mean amount spent on drinks. Which correlation between amounts spent on food and drink do you think would be higher, the one computed for the 2500 customers in the next month, or the one computed using the means for the 30 days of the month? Why? Sketch a sample scatterplot showing what you expect for each case as part of your answer.

12.38 Yale and UConn For which student body do you think the correlation between high school GPA and college GPA would be higher: Yale University or the University of Connecticut? Explain why.

12.39 Violent crime and single-parent families Use software to analyze the U.S. Statewide Crime data file on the book's website on y = violent crime rate and x = percentage of single parent families.

a. Construct a scatterplot. What does it show?

b. One point is quite far removed from the others, having a much higher value on both variables than the rest of the sample, but it fits in well with the linear trend exhibited by the rest of the points. Show that the correlation changes from 0.77 to 0.59 when you delete this observation. Why does it drop so dramatically?

12.40 Correlations for the strong and for the weak Refer to the High School Female Athlete and Male Athlete Strength data files on the book's website.

a. Find the correlation between number of 60-pound bench presses before fatigue and bench press maximum for females and between bench presses before fatigue and bench press maximum for males. Interpret.

b. Find the correlation using only the x values (i) below the median of 10 for females and below the median of 17 for males and (ii) above the median of 10 for females and above the median of 17 for males. Compare to the correlation in part a. Why are they so different?

12.4 Practicing the Basics

12.41 Poor predicted sales The MINITAB output shows the large standardized residuals for studying sales in thousands of pounds as a response using marketing in thousands of pounds as the explanatory variable.

Large standardized residuals for sales study:

Month	Marketing Spend	Fit	Resid	Std Resid
11	157.00	1895.20	29.10	2.56
25	159.00	1298.88	18.62	2.14
38	75.00	1586.12	−19.12	−2.23

a. Explain how to interpret all the entries in the row of the output for month 11, where Marketing Spend = 157.00 (in thousands of pounds).

b. Out of 57 observations, is it surprising that 3 observations would have standardized residuals with absolute value above 2.0? Explain.

12.42 Loves TV and exercise For the Georgia Student Survey file on the book's website, let y = time exercising and x = time watching TV. One student reported watching TV an average of 180 minutes a day and exercising 60 minutes a day. This person's residual was 48.8 and the standardized residual was 6.41.

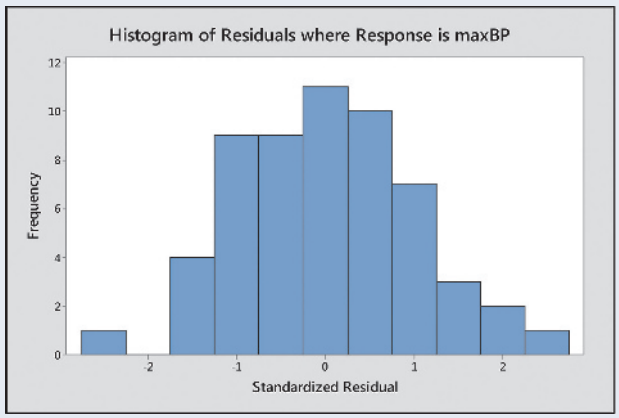
a. Interpret the residual, and use it to find the predicted value of exercise.

b. Interpret the standardized residual.

12.43 Bench press residuals The figure is a histogram of the standardized residuals for the regression of maximum bench press on number of 60-pound bench presses for the high school female athletes.

a. Which distribution does this figure provide information about?

b. What would you conclude based on this figure?



12.44 Predicting house prices The House Selling Prices FL data file on the book's website has several predictors of house selling prices. The table here shows the ANOVA table for a regression analysis of y = the selling price (in thousands of dollars) and x = the size of house (in thousands of square feet). The prediction equation is $\hat{y} = 9.2 + 77x$.

ANOVA table for selling price and size of house:

Source	DF	SS	MS	F	P
Regression	1	182220	182220	135.07	0.000
Error	98	132213	1349		
Total	99	314433			

a. What was the sample size? (*Hint:* You can figure it out from the residual *df*.)

b. The sample mean house size was 1.53 thousand square feet. What was the sample mean selling price? (*Hint:* What does $\hat{y}$ equal when $x = \bar{x}$?)

c. Estimate the standard deviation of the selling prices for homes that have $x = 1.53$. Interpret.

d. Report an approximate prediction interval within which you would expect about 95% of the selling prices to fall for homes of size $x = 1.53$.

12.45 Predicting annual salary For a random sample of residents from a district in South Carolina, a regression analysis is conducted of y = salary in thousands of dollars and x = years of education. MINITAB reports the tabulated results for observations at $x = 15$.

Variable	Setting		
Years of Education	15		
Fit	SE Fit	95% CI	95% PI
148.0	10.6	(129, 162)	(70, 210)

a. Interpret the value listed under "Fit."

b. Interpret the interval listed under "95% CI."

c. Interpret the interval listed under "95% PI."

12.46 CI versus PI Using the context of the previous exercise, explain the difference between the purpose of a 95% prediction interval (PI) for an observation and a 95% confidence interval (CI) for the mean of y at a given value of x. Why would you expect the PI to be wider than the CI?

12.47 ANOVA table for leg press Exercise 12.15 referred to an analysis of leg strength for 57 female athletes, with y = maximum leg press and x = number of 200-pound leg presses until fatigue, for which $\hat{y} = 233.89 + 5.27x$. The table shows ANOVA results from SPSS for the regression analysis.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	121082.4	1	121082.400	92.875	.000[a]
	Residual	71704.442	55	1303.717		
	Total	192786.8	56			

a. Show that the residual standard deviation is 36.1. Interpret it.

b. For this sample, $\bar{x} = 22.2$. For female athletes with $x = 22$, what would you estimate the variability to be of their maximum leg press values? If the y values are approximately normal, find an interval within which about 95% of them would fall.

12.48 Predicting leg press Refer to the previous exercise. MINITAB reports the tabulated results for observations at $x = 25$.

Variable	Setting		
LegPress	25		
Fit	SE Fit	95% CI	95% PI
365.7	5.02	(355.6, 375.7)	(292.6, 438.7)

a. Show how MINITAB got the "Fit" of 365.7.

b. Using the predicted value and *se* value, explain how MINITAB got the interval listed under "95% CI." Interpret this interval.

c. Interpret the interval listed under "95% PI."

12.49 Variability and *F* Refer to the previous two exercises.

a. In the ANOVA table, show how the Total SS breaks into two parts and explain what each part represents.

b. From the ANOVA table, explain why the overall sample standard deviation of y values is $s_y = \sqrt{192787/56} = 58.7$. Explain the difference between the interpretation of this standard deviation and the residual standard deviation s of 36.1.

c. Exercise 12.15 reported a t statistic of 9.64 for testing independence of these variables. Report the F test statistic from the table here and explain how it relates to that t statistic.

12.50 Assumption violated For prediction intervals, an important inference assumption is a constant standard deviation σ of y values at different x values. In practice, the standard deviation often tends to be larger when μ_y is larger.

a. Sketch a hypothetical scatterplot for which this happens, using observations on x = family income and y = amount donated to charity.

b. Explain why a 95% prediction interval would not work well at very small or at very large x values.

12.51 Understanding an ANOVA table For a random sample of Indian states, the ANOVA table shown refers to hypothetical data on x = tax revenue in Indian rupees and y = agricultural subsidies in Indian rupees.

a. Fill in the blanks in the table.

b. For what hypotheses can the F test statistic be used?

Source	DF	SS	MS	F
Regression	1	500000	_____	_____
Error	35	300000	_____	
Total	36	800000		

12.52 Predicting cell phone weight Refer to the cell phone data file on the book's website. Regress y = weight on x = capacity of battery, excluding the outlier (phone no. 70).

a. Stating the necessary assumptions, find a 95% confidence interval for the mean weight of cell phones with a battery capacity of 1500mAh. Interpret the interval.

b. Find a 95% prediction interval for the weight of a cell phone with a battery capacity of 1500mAh. Interpret.

c. Explain the difference between the purposes of the intervals in part a and part b.

12.53 Cell phone ANOVA Report the ANOVA table for the previous exercise.

a. Verify that total SS = residual SS + regression SS. Explain what each of the three sums of squares represent.

b. Find the estimated residual standard deviation of y. Interpret it.

c. Find the sample standard deviation s_y of y values. Explain the difference between the interpretation of this standard deviation and the residual standard deviation in part b.

12.5 Practicing the Basics

12.54 Savings grow exponentially You invest $100 in a savings account with interest compounded annually at 10%.

TRY

a. How much money does the account have after one year?

b. How much money does the account have after five years?

c. How much money does the account have after x years?

d. How many years does it take until your savings more than double in size?

12.55 Growth by year versus decade It is expected that the female population in a city will double in two decades.

a. Explain why this is possible for a growth rate of 3.6% a year. (*Hint:* What does $(1.036)^{20}$ equal?)

b. You might think that a growth rate of 5% a year would result in 100% growth (i.e. the female population doubles) over two decades. Explain why a growth rate of 5% a year would actually cause the female population to multiply by 2.65 over two decades.

12.56 Moore's law today The following data show the number of components (per square inch, in millions) being packed on a Pentium-type chip, for years 1994 to 2015. Let x be the number of years since 1994 (e.g., $x = 0$ for 1994, $x = 3$ for 1997, ..., $x = 21$ for 2015) and let y be the number of components on a chip. The prediction equation when fitting an exponential regression model to these data equals $\hat{y} = 151.61 \times 1.191^x$.

TRY

Chip	Year	Count	Chip	Year	Count
Pentium	1994	14	PentiumE	2008	1794
PentiumII	1997	24	PentiumG	2010	3043
PentiumIII	1999	48	PentiumG	2011	2482
Pentium4	2000	125	PentiumG	2013	5645
PentiumD	2005	720	PentiumG	2015	5103
PentiumE	2007	971			

a. What is the predicted number of components per square inch in 2015?

b. By how much does the predicted number of components increase per year over the time range that these data cover?

c. The correlation between the logarithm of the count and x equals 0.985, and the scatterplot with the log counts shows a linear trend. What does this suggest about whether the exponential regression model is appropriate for these data?

12.57 U.S. population growth The table shows the approximate U.S. population size (in millions) at 10-year intervals beginning in 1900. Let x denote the number of decades since 1900. That is, 1900 is $x = 0$, 1910 is $x = 1$, and so forth. The exponential regression model fitted to $y =$ population size and x gives $\hat{y} = 81.14 \times 1.1339^x$.

U.S. population sizes (in millions) from 1900 to 2010

Year	Population Size	Year	Population Size	Year	Population Size
1900	76.2	1940	132.1	1980	226.5
1910	92.2	1950	150.7	1990	248.7
1920	106.0	1960	179.3	2000	281.4
1930	122.8	1970	203.3	2010	308.7

Source: U.S. Bureau of the Census.

a. Show that the predicted population sizes are 81.14 million in 1900 and 323.3 million in 2010.

b. Explain how to interpret the value 1.1339 in the prediction equation.

c. The correlation equals 0.98 between the log of the population size and the year number. What does this suggest about whether the exponential regression model is appropriate for these data?

12.58 Future shock Refer to the previous exercise, for which predicted population growth was 14.18% per decade. Suppose the growth rate is now 15% per decade. Explain why the population size will (a) double after five decades, (b) quadruple after 100 years (10 decades), and (c) be 16 times its original size after 200 years. (The exponentially increasing function has the property that its doubling time is constant. This is a fast increase, even though the annual rate of growth seems small.)

12.59 Age and death rate Let x denote a person's age and let y be the death rate, measured as the number of deaths per thousand individuals of a fixed age within a period of a year. For women in a European country, these variables follow approximately the equation $\hat{y} = 0.34(1.081)^x$.

a. Interpret 0.34 and 1.081 in this equation.

b. Find the predicted death rate when age is (i) 25, (ii) 55, and (iii) 80.

c. After how many years does the death rate double? (*Hint:* What is x such that $(1.081)^x = 2$?)

12.60 Leaf litter decay Ecologists believe that organic material decays over time according to an *exponential decay* model. This is the case $0 < \beta < 1$ in the exponential regression model, for which μ_y decreases over time. The rate of decay is determined by a number of factors, including composition of material, temperature, and humidity. In an experiment carried out by researchers at the University of Georgia Ecology Institute, leaf litter was allowed to sit for a 20-week period in a bag in a moderately forested area. Initially, the total weight of the organic mass in the bag was 75.0 kg. Each week, the remaining amount (y) was measured. The table shows the weight y by $x =$ number of weeks of time that have passed.

x	y	x	y	x	y	x	y	x	y	x	y
0	75.0	1	60.9	2	51.8	3	45.2	4	34.7	5	34.6
6	26.2	7	20.4	8	14.0	9	12.3	11	8.2	15	3.1
20	1.4										

a. Construct a scatterplot. Why is a straight-line model inappropriate?

b. Show that the ordinary regression model gives the fit $\hat{y} = 54.98 - 3.59x$. Find the predicted weight after $x = 20$ weeks. Does this prediction make sense? Explain.

c. Plot the log of y against x. Does a straight-line model now seem appropriate?

d. The exponential regression model has prediction equation $\hat{y} = 80.6(0.813)^x$. Find the predicted weight (i) initially and (ii) after 20 weeks.

e. Interpret the coefficient 0.813 in the prediction equation.

12.61 More leaf litter Refer to the previous exercise.

a. The correlation equals -0.890 between x and y and -0.997 between x and $\log(y)$. What does this tell you about which model is more appropriate?

b. The half-life is the time for the weight remaining to be one-half of the original weight. Use the equation $\hat{y} = 80.6(0.813)^x$ to predict the half-life of this organic material. (*Hint:* By trial and error, find the value of x for which $(0.813)^x$ is about $1/2$.)

13.1 Practicing the Basics

13.1 Predicting weight For a study of female college athletes, the prediction equation relating y = total body weight (in pounds) to x_1 = height (in inches) and x_2 = percent body fat is $\hat{y} = -121 + 3.50x_1 + 1.35x_2$.

TRY

a. Find the predicted total body weight for a female athlete at the mean values of 66 and 18 for x_1 and x_2.

b. An athlete with $x_1 = 66$ and $x_2 = 18$ has actual weight $y = 115$ pounds. Find the residual and interpret it.

13.2 Does study help GPA? For the Georgia Student Survey file on the book's website, the prediction equation relating y = college GPA to x_1 = high school GPA and x_2 = study time (hours per day), is $\hat{y} = 1.13 + 0.643x_1 + 0.0078x_2$.

a. Find the predicted college GPA of a student who has a high school GPA of 3.5 and who studies three hours a day.

b. For students with fixed study time, what is the change in predicted college GPA when high school GPA increases from 3.0 to 4.0?

13.3 Predicting visitor satisfaction For all the restaurants in a city, the prediction equation for y = average monthly visitor satisfaction rating (range 0–4.0 where 0 = very poor and 4 = very good) and x_1 = the monthly food quality score given by the food inspection authority (range 0–4.0 where 0 = very poor and 4 = very good) and x_2 = the number of visitors in a month is $\hat{y} = 0.35 + 0.55x_1 + 0.0015x_2$.

a. Find the predicted average monthly visitor satisfaction rating for a restaurant having (i) a monthly food quality score of 4.0 and 800 visitors in a month and (ii) a monthly food quality score of 2.0 and 200 visitors in a month.

b. For restaurants with $x_2 = 500$, show that $\hat{y} = 1.10 + 0.55x_1$.

c. For restaurants with $x_2 = 600$, show that $\hat{y} = 1.25 + 0.55x_1$. Thus, compared to part b, the slope for x_1 is still 0.55, and increasing x_2 by 100 (from 500 to 600) shifts the intercept upward by $100 \times$ (slope for x_2) $= 100(0.0015) = 0.15$ units.

13.4 Interpreting slopes on average monthly visitor satisfaction Refer to the previous exercise.

a. Explain why setting x_2 at a variety of values yields a collection of parallel lines relating $\hat{y}$ to x_1. What is the value of the slope for those parallel lines?

b. Since the slope 0.55 for x_1 is larger than the slope 0.0015 for x_2, does this imply that x_1 has a larger effect than x_2 on y in this sample? Explain.

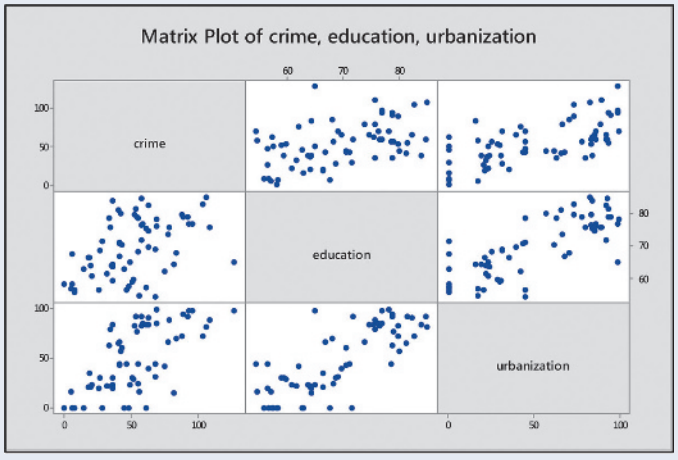
13.5 Does more education cause more crime? The FL Crime data file on the book's website has data for the 67 counties in Florida on

y = crime rate: Annual number of crimes in county per 1000 population

x_1 = education: Percentage of adults in county with at least a high school education

x_2 = urbanization: Percentage in county living in an urban environment.

The figure shows a scatterplot matrix. MINITAB multiple regression results are also displayed.



Scatterplot matrix for crime rate, education, and urbanization.

Multiple regression for y = crime rate, x_1 = education, and x_2 = urbanization.

Term	Coef	SE Coef	T-Value	P-Value
Constant	59.1	28.4	2.08	0.041
education	−0.583	0.472	−1.23	0.221
urbanization	0.683	0.123	5.54	0.000

a. Find the predicted crime rate for a county that has 0% in an urban environment and (i) 70% high school graduation rate and (ii) 80% high school graduation rate.

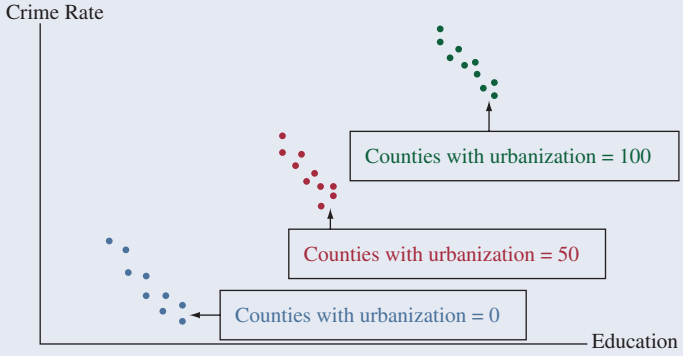
b. Use results from part a to explain how education affects the crime rate, controlling for urbanization, interpreting the slope coefficient −0.58 of education.

c. Using the prediction equation, show that the equation relating crime rate and education when urbanization is fixed at (i) 0, (ii) 50, and (iii) 100, is as follows:

x_2	$\hat{y} = 59.1 - 0.58x_1 + 0.68x_2$
0	$\hat{y} = 59.1 - 0.58x_1$
50	$\hat{y} = 93.2 - 0.58x_1$
100	$\hat{y} = 127.4 - 0.58x_1$

Sketch a plot with these lines and use it to interpret the effect of education on crime rate, controlling for urbanization.

d. The scatterplot matrix shows that education has a *positive* association with crime rate, but the multiple regression equation shows that the association is *negative* when we keep x_2 = urbanization fixed. Consider the hypothetical figure that follows. Sketch lines that represent (i) the prediction equation from a simple regression model using only education and ignoring the information on urbanization and (ii) the prediction equation from the multiple regression model for counties having urbanization = 50. Use these lines to explain the difference in the interpretation of the slope for education in simple and multiple regression models with regard to ignoring or controlling for urbanization. (*Note:* The reversal in the association between crime rate and education is an example of **Simpson's paradox**; see Example 16 in Sec. 3.4 and Example 18 in Sec. 10.5).



Hypothetical scatterplot for crime rate and education, labeling by urbanization.

13.6 Crime rate and income Refer to the previous exercise. MINITAB reports the following results for the multiple regression of y = crime rate on x_1 = median income (in thousands of dollars) and x_2 = urbanization.

Results of regression analysis

Term	Coef	SE Coef	T-Value	P-Value
Constant	40.0	16.4	2.44	0.017
income	−0.791	0.805	−0.98	0.330
urbanization	0.642	0.111	5.78	0.000

a. Report the prediction equations relating crime rate to income at urbanization levels of (i) 0 and (ii) 100. Interpret.

b. For the simple regression model relating y = crime rate to x = income, MINITAB reports

crime = −11.6 + 2.61 income

Interpret the effect of income, according to the sign of its slope. How does this effect differ from the effect of income in the multiple regression equation?

c. Use the estimated slope for income in the simple and multiple regression model to explain the difference in the interpretation of the slope when (i) ignoring urbanization (ii) controlling urbanization. (Note: The reversal in the association between income and education is an example of **Simpson's paradox**.)

13.7 The economics of golf The earnings of a PGA Tour golfer are determined by performance in tournaments. A study analyzed tour data to determine the financial return for certain skills of professional golfers. The sample consisted of 393 golfers competing in one or both of the 2002 and 2008 seasons. The most significant factors

that contribute to earnings were the percent of attempts a player was able to hit the green in regulation (GIR), the number of times that a golfer made par or better after hitting a bunker divided by the number of bunkers that were hit (SS), the average of putts after reaching the green (AvePutt), and the number of PGA events entered (Events). The resulting coefficients from multiple regression to predict yearly earnings (in $) are:

Predictor	Coefficient
Constant	26,417,000
GIR	168,300
SS	33,859
AvePutt	−19,784,000
Events	−44,725

Source: Some data from K. Rinehart, *Major Themes in Economics*, 2009.

a. State the prediction equation for a PGA Tour golfer's yearly earnings.

b. Explain how to interpret the coefficient for AvePutt.

c. Find the predicted earnings for a golfer who had a GIR score of 60, SS score of 50, AvePutt of 1.5 and participated in 20 events.

13.8 Comparable number of bedrooms and house size effects In Example 2, the prediction equation between y = selling price and x_1 = house size and x_2 = number of bedrooms was $\hat{y} = 60{,}102 + 63.0x_1 + 15{,}170x_2$.

a. For fixed number of bedrooms, how much is the house selling price predicted to increase for each square foot increase in house size? Why?

b. For a fixed house size of 2000 square feet, how does the predicted selling price change for two, three, and four bedrooms?

13.9 Controlling has an effect The slope of x_1 is not the same for multiple linear regression of y on x_1 and x_2 as compared to simple linear regression of y on x_1, where x_1 is the only predictor. Explain why you would expect this to be true. Does the statement change when x_1 and x_2 are uncorrelated?

13.10 House selling prices Using software with the House Selling Prices OR data file on the book's website, analyze y = selling price, x_1 = house size, and x_2 = lot size.

a. Construct box plots for each variable and a scatterplot matrix or scatter plots between y and each of x_1 and x_2. Interpret.

b. Find the multiple regression prediction equation.

c. If house size remains constant, what, if any, is the effect of an increase in lot size? Why do you think this is?

13.11 Used cars The following data (also available from the book's website) is from a random sample of campus newspaper ads on used cars for sale. Consider the age and horsepower (HP) of a car to predict its selling price. (The variable Type stands for whether the car is from the United States, coded as 1, or a foreign car, coded as 0. This variable will be considered in Exercise 13.48.)

a. Construct a scatterplot matrix (or separate scatterplots) to investigate the relationship among price, age, and horsepower and interpret.

b. Find the multiple regression prediction equation for the selling price in terms of age and horsepower of a car. What is the predicted price for a car that has a horsepower of 80 and (i) is 8 years old, (ii) 10 years old, rounded to the nearest hundred?

c. Based on this multiple regression, can you predict the price difference between a car with 60 HP and a car with 80 HP without knowing the ages of the two cars? Explain.

Car	Price	Age	HP	Type	Car	Price	Age	HP	Type
1	8700	9	55	1	11	5050	10	44	1
2	11200	7	75	1	12	14800	7	75	0
3	9000	9	69	0	13	1800	12	55	1
4	10300	10	76	0	14	7200	10	82	1
5	10500	8	76	0	15	8600	9	73	1
6	5250	12	49	1	16	13000	9	130	0
7	12000	5	120	0	17	11790	8	95	0
8	2500	11	79	1	18	12350	7	124	0
9	8300	8	90	0	19	6100	10	74	1
10	9300	8	38	0					

13.2 Practicing the Basics

13.12 Predicting average monthly visitor satisfaction
TRY Refer to Exercise 13.3 about the multiple linear regression of a restaurant's average monthly visitor satisfaction rating (y) on the monthly food quality score (x_1) and the number of visitors in a month (x_2). Following is the ANOVA table of the same multiple linear regression:

ANOVA table for y = average monthly visitor satisfaction rating

Source	DF	SS	MS	F-Value	P-Value
Regression	2	7.12	3.560	131.85	0.00
Error	110	2.97	0.027		
Total	112	10.09			

a. Show how R^2 is calculated from the SS values and report its value.

b. Interpret the R^2 value. Does the multiple regression equation help us predict the average monthly visitor satisfaction rating much better than we could without knowing that equation?

c. Find the multiple correlation. Interpret.

13.13 Predicting weight Let's use multiple regression to predict total body weight (TBW, in pounds) using data from a study of female college athletes. Possible predictors are HGT = height (in inches), %BF = percent body fat, and age. The display shows the correlation matrix for these variables.

	TBW	HGT	%BF	AGE
TBW	-	0.74	0.39	−0.19
HGT	0.74	-	0.10	−0.12
%BF	0.39	0.10	-	0.02
AGE	−0.19	−0.12	0.02	-

a. Which explanatory variable gives by itself the best predictions of weight? Explain.

b. With height as the sole predictor, $\hat{y} = -106 + 3.65$ (HGT) and $r^2 = 0.55$. If you add %BF as a predictor, you know that R^2 will be at least 0.55. Explain why.

c. When you add % body fat to the model, $\hat{y} = -121 + 3.50(\text{HGT}) + 1.35(\%\text{BF})$ and $R^2 = 0.66$. When you add age to the model, $\hat{y} = -97.7 + 3.43(\text{HGT}) + 1.36(\%\text{BF}) - 0.960(\text{AGE})$ and $R^2 = 0.67$. Once you know height and % body fat, does age seem to help you in predicting weight? Explain, based on comparing the R^2 values.

13.14 When does controlling have little effect? Refer to the previous exercise. Height has a similar estimated slope for each of the three models. Why do you think that controlling for % body fat and then age does not change the effect of height much? (*Hint:* How strongly is height correlated with the other two variables?)

13.15 Price of used cars For the 19 used cars listed in the Used Cars data file on the book's website (see also Exercise 13.11), modeling the mean of y = used car price in terms of x_1 = age results in $r^2 = 0.66$.

Adding x_2 = horsepower (HP) to the model yields the results in the following display.

```
Regression Equation

Price = 19349 − 1406 Age + 25.5 HP

Coefficients

Term          Coef     SE Coef    T-Value    P-Value
Constant     19349        4053       4.77      0.000
Age          −1406         320      −4.40      0.000
HP            25.5        22.3       1.14      0.270


Model Summary

      S      R-sq
2084.69    68.69%

Analysis of Variance

Source       DF     SS         MS     F-value  P-value
Regression    2  152567500  76283750   17.55    0.000
Error        16   69534753   4345922
Total        18  222102253
```

a. Report R^2 and show how it is determined by SS values in the ANOVA table.

b. Interpret its value as a proportional reduction in prediction error.

c. Interpret its value as the percentage of the variability in the response variable that can be explained.

13.16 Price, age, and horsepower In the previous exercise, $r^2 = 0.66$ when age is the predictor and $R^2 = 0.69$ when both age and HP are predictors. Why do you think that the predictions of price don't improve much when HP is added to the model? (The correlation between HP and price is $r = 0.56$, and the correlation between HP and age is $r = -0.51$.)

13.17 Softball data For the Softball data set on the book's website, for each game, the variables are a team's number of runs scored (RUNS), number of hits (HIT), number of errors (ERR), and the difference (DIFF) between the number of runs scored by that team and by the other team, which is the response variable. MINITAB reports

```
Difference = -4.03 + 0.0260 Hits
+ 1.04 Run - 1.22 Errors
```

a. If you know the team's number of runs and number of errors in a game, explain why it does not help much to know how many hits the team has.

b. Explain why the result in part a is also suggested by knowing that $R^2 = 0.7594$ for this model, whereas $R^2 = 0.7593$ when only runs and errors are the explanatory variables in the model.

13.18 Slopes, correlations, and units In Example 2 on y = house selling price, x_1 = house size, and x_2 = number of bedrooms, $\hat{y} = 60{,}102 + 63.0x_1 + 15{,}170x_2$, and $R = 0.72$.

a. Interpret the value of the multiple correlation.

b. Suppose house selling prices are changed from dollars to *thousands* of dollars. Explain why if each price in Table 13.1 is divided by 1000, the prediction equation changes to $\hat{y} = 60.102 + 0.063x_1 + 15.170x_2$.

c. In part b, does the multiple correlation change to 0.00072? Justify your answer.

13.19 Predicting college GPA Using software with the Georgia Student Survey data file from the book's website, find and interpret the multiple correlation and R^2 for the relationship between y = college GPA, x_1 = high school GPA, and x_2 = study time. Use both interpretations of R^2 as the reduction in prediction error and the percentage of the variability explained.

13.3 Practicing the Basics

13.20 Predicting CPI For a random sample of 100 students in a German university, the result of regressing the college cumulative performance index (CPI) on the high school grade (HSG), the average monthly attendance percentage (AMAP) and the average daily study time (ADST) follows.

```
Regression Equation
CPI = 1.1362 + 0.6615 HSG + 0.2301 AMAP + 0.0075 ADST
Coefficients
Term          Coef     SE Coef    T-Value    P-Value
Constant    1.1362      0.5731       1.98      0.050
HSG         0.6615      0.1578       4.19      0.000
AMAP        0.3301      0.1473       2.24      0.027
ADST        0.0075      0.0151       0.50      0.621
Model Summary
       S       R-sq
0.218624     75.82%
```

a. Explain in nontechnical terms what it means if the population slope coefficient for high school grade (HSG) equals 0.

b. Show all the steps for testing the hypothesis that this slope equals 0.

13.21 Study time helps CPI? Refer to the previous exercise.

a. Report and interpret the P-value for testing the hypothesis that the population slope coefficient for study time equals 0.

b. Find a 95% confidence interval for the true slope for average daily study time (ADST). Explain how the result is in accord with the result of the test in part a.

c. Does the result in part a imply that in the corresponding population, the average daily study time (ADST) has no association with college CPI? Explain. (*Hint:* What is the impact of also having HSG and AMAP in the model?)

13.22 Variability in college CPI Refer to the previous two exercises.

a. Report the residual standard deviation. What does this describe?

b. Interpret the residual standard deviation by predicting where approximately 95% of the college CPI fall when high school grade (HSG) = 3.80, average monthly attendance percentage (AMAP) = 0.90 and average daily study time (ADST) = 5.0 hours per day, which are the sample means.

13.23 Does leg press help predict body strength? Chapter 12 analyzed strength data for 57 female high school athletes. Upper body strength was summarized by the maximum number of pounds the athlete could bench press (denoted maxBP). This was predicted well by the number of times she could do a 60-pound bench press (denoted BP60). Can we predict maxBP even better if we also know how many times an athlete can perform a 200-pound leg press? The table shows results after adding this second predictor (denoted LP200) to the model.

Term	Coef	SE Coef	T-Value	P-Value
Constant	60.60	2.87	21.10	0.000
BP60	1.332	0.188	7.10	0.000
LP200	0.211	0.152	1.39	0.171

```
Analysis of Variance
Source       DF  Adj SS   Adj MS   F-Value   P-Value
Regression    2  6473.3  3236.65    51.39     0.000
Error        54  3401.3    62.99
Total        56  9874.6
```

a. Does LP200 have a significant effect on maxBP if BP60 is also in the model? Show all steps of a significance test to answer this.

b. Show that the 95% confidence interval for the slope for LP200 equals 0.21 ± 0.30, roughly $(-0.1, 0.5)$. Based on this interval, does LP200 seem to have a strong impact, or a weak impact, on predicting maxBP if BP60 is also in the model?

c. Given that LP200 is in the model, provide evidence from a significance test that shows why it *does* help to add BP60 to the model.

13.24 Leg press uncorrelated with strength? The P-value of 0.17 in part a of the previous exercise suggests that LP200 plausibly had no effect on maxBP once BP60 is in the model. Yet when LP200 is the sole predictor of BP, the correlation is 0.58 and the significance test for its effect has a P-value of 0.000, suggesting very strong evidence of an effect. Explain why this is not a contradiction.

13.25 Interpret strength variability Refer to the previous two exercises. The sample standard deviation of maxBP was 13.3. The residual standard deviation of maxBP when BP60 and LP200 are predictors in a multiple regression model is 7.9.

a. Explain the difference between the interpretations of these two standard deviations.

b. If the conditional distributions of maxBP are approximately bell shaped, explain why most maximum bench press values fall within about 16 pounds of the regression equation when the predictors BP60 and LP200 are near their sample mean values.

c. At BP60 $= 11$ and LP200 $= 22$, which are close to the sample mean values, software reports $\hat{y} = 80$ and a 95% prediction interval of 80 ± 16, or (64, 96). Is this interval an inference about where the population maxBP values fall or where the population *mean* of the maxBP values fall (for subjects having BP60 $= 11$ and LP200 $= 22$)? Explain.

d. Refer to part c. Would it be unusual for a female athlete with these predictor values to be able to bench press more than 100 pounds? Why?

13.26 Any predictive power? Refer to the previous three exercises.

TRY

a. State and interpret the null hypothesis tested with the F statistic in the ANOVA table given in Exercise 13.23.

b. From the F table (Table D), which F statistic value would have a P-value of 0.05 for these data?

c. Report the observed F test statistic and its P-value. Interpret the P-value, and make a decision for a 0.05 significance level. Explain in nontechnical terms what the result of the test means.

13.27 Predicting restaurant revenue An Italian restaurant keeps monthly records of its total revenue, expenditure on advertising, prices of its own menu items, and the prices of its competitors' menu items.

a. Specify notation and formulate a multiple regression equation for predicting the monthly revenue using the available data. Explain how to interpret the parameters in the equation.

b. State the null hypothesis that you would test if you want to analyze whether advertising is helpful, for the given prices of items in the restaurant's own menu and the prices of its competitors' menu items.

c. State the null hypothesis that you would test if you want to analyze whether *at least one* of the predictors has some effect on monthly revenue.

13.28 Regression for human development A study investigated an index of human development in a South American country, which had $\bar{y} = 27.3$ and $s = 5.5$. Two explanatory variables were x_1 = literacy rate in percentage (mean = 44.4, $s = 22.6$) and x_2 = daily per capita income in dollars (mean = 56.6, $s = 25.3$). Based on a random sample of 50 cities in the country, some regression results are also shown as follows:

---

y = index of human development, x_1 = life literacy rate and x_2 = daily per capita income

```
Coefficients
Term                 Coef     SE Coef   T-Value   P-Value
Constant            28.23        2.17     12.98     0.000
literacy rate      0.1033      0.0325      3.18     0.003
daily per capita  −0.0975      0.0291     −3.35     0.002
income
Model Summary
      S      R-sq
4.55644    33.92%
Analysis of Variance
Source       DF     SS      MS     F-Value   P-Value
Regression    2   394.2  197.12      9.49     0.000
Error        37   768.2   20.76
Total        39  1162.4
```

a. Find the 95% confidence interval for β_1.

b. Explain why the interval in part a means that an increase of 10 units in literacy rate corresponds to anywhere from a 0.4- to 1.7-unit increase in mean human development, controlling for daily per capita income. (This lack of precision reflects the small sample size.)

13.29 Gain in human development Refer to the previous exercise.

a. Report the test statistic and P-value for testing H_0: $\beta_1 = \beta_2 = 0$.

b. State the alternative hypothesis that is supported by the result in part a.

c. Does the result in part a imply that necessarily *both* literacy rate and daily per capita income are needed in the model? Explain.

13.30 More predictors for selling price The MINITAB results are shown for predicting selling price using x_1 = size of home, x_2 = number of bedrooms, and x_3 = age.

Regression Equation

Price = 80489 + 62.65 House Size + 13543 Bedrooms − 418 Age

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	80489	21810	3.69	0.000
House Size	62.65	4.73	13.24	0.000
Bedrooms	13543	5380	2.52	0.013
Age	−418	236	−1.77	0.078

a. State the null hypothesis for an F test, in the context of these variables.

b. The F statistic equals 74.23, with P-value $= 0.000$. Interpret.

c. Explain in nontechnical terms what you learn from the results of the t tests reported in the table for the three explanatory variables.

13.31 House prices Use software to do further analyses with the multiple regression model of $y =$ selling price of home in thousands, $x_1 =$ size of home, and $x_2 =$ number of bedrooms, considered in Section 13.1. The data file House Selling Prices OR is on the book's website.

a. Report the F statistic and state the hypotheses to which it refers. Report its P-value and interpret. Why is it not surprising to get a small P-value for this test?

b. Report and interpret the t statistic and P-value for testing H_0: $\beta_2 = 0$ against H_a: $\beta_2 > 0$.

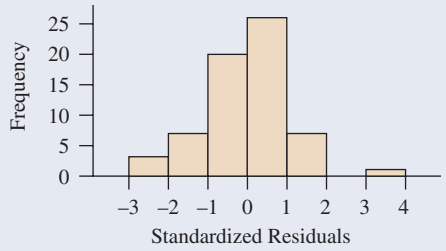
c. Construct a 95% confidence interval for β_2 and interpret. This inference is more informative than the test in part b. Explain why.

13.32 Body weight residuals Examples 4–7 used multiple regression to predict total body weight of college athletes in terms of height, percent body fat, and age. The following figure shows a histogram of the standardized residuals resulting from fitting this model.

TRY

a. About which distribution do these give you information—the overall distribution of weight or the conditional distribution of weight at fixed values of the predictors?

b. What does the histogram suggest about the likely shape of this distribution? Why?



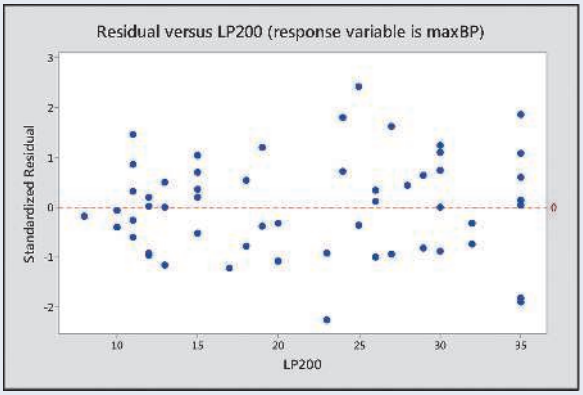
13.33 Strength residuals In Chapter 12, we analyzed strength data for a sample of female high school athletes. The following figure is a residual plot for the multiple regression model relating the maximum number of pounds the athlete could bench press (maxBP) to the number of 60-pound bench presses (BP60) and the number of 200-pound leg presses (LP200). It plots the standardized residuals against the values of LP200.

TRY

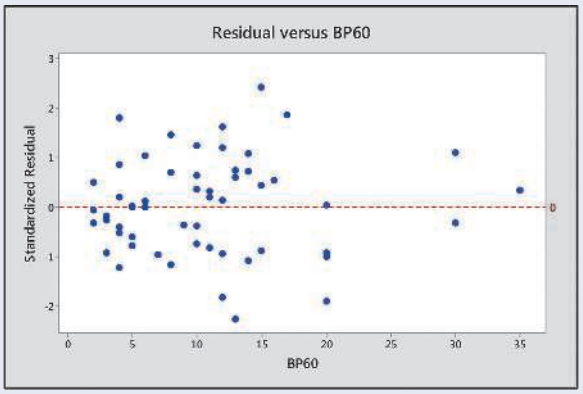
a. You don't see BP60 on the plot, so how do its values affect the analysis?

b. Explain how the plot might suggest less variability at the lower values of LP200.

c. Suppose you remove the three points with standardized residuals around −2. Then is the evidence about variability in part b so clear? What does this suggest about cautions in looking at residual plots?



13.34 More residuals for strength Refer to the previous exercise. The following figure is a residual plot for the model relating maximum bench press to LP200 and BP60. It plots the standardized residuals against the values of BP60. Does this plot suggest any irregularities with the model? Explain.



13.35 Nonlinear effects of age Suppose you fit a straight-line regression model to y = number of hours worked (excluding time spent on household chores) and x = age of the subject. Values of y in the sample tend to be quite large for young adults and for elderly people, and they tend to be lower for other people. Sketch what you would expect to observe for (a) the scatterplot of x and y and (b) a plot of the residuals against the values of age.

13.36 Population growth with time Suppose you fit a straight-line regression model to x = time and y = population. Sketch what you would expect to observe for (a) the scatterplot of x and y and (b) a plot of the residuals against the values of time.

13.37 Why inspect residuals? When we use multiple regression, what is the purpose of performing a residual analysis? Why is it better to work with standardized residuals than unstandardized residuals to detect outliers?

13.38 College athletes The College Athletes data set on the book's website comes from a study of University of Georgia female athletes. Using the column names from the data set, the response variable 1RM = maximum bench press has explanatory variables LBM = lean body mass (which is weight times 1 minus the proportion of body fat) and REPS70 = number of repetitions before fatigue with a 70-pound bench press. Let's look at all the steps of a regression analysis for these data.

a. The first figure shows a scatterplot matrix. Which two plots in the figure describe the associations with 1RM as a response variable? Describe those associations.

b. Results of a multiple regression analysis are shown. Write down the prediction equation and interpret the coefficient of REPS70.

c. Report R^2 and interpret its value in the context of these variables.

d. Based on the value of R^2, report and interpret the multiple correlation.

e. Interpret results of the F test that 1RM is independent of these two predictors. Show how to obtain the F statistic from the mean squares in the ANOVA table.

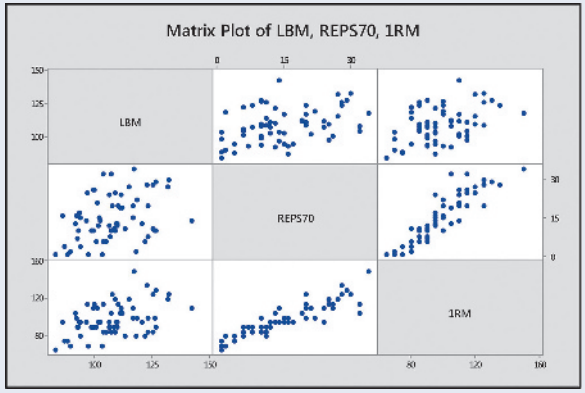
f. Once REPS70 is in the model, does it help to have LBM as a second predictor? Answer by showing all steps of a significance test for a regression parameter.

g. Examine the histogram shown of the residuals for the multiple regression model. What does this describe, and what does it suggest?

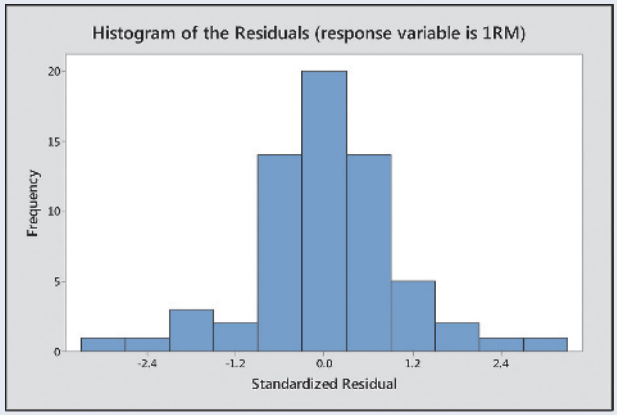
h. Examine the plot shown of the residuals plotted against values of REPS70. What does this describe, and what does it suggest?

i. From the plot in part h, can you identify a subject whose 1RM value was considerably lower than expected based on the predictor values? Identify by indicating the approximate values of REPS70 and the standardized residual for that subject.

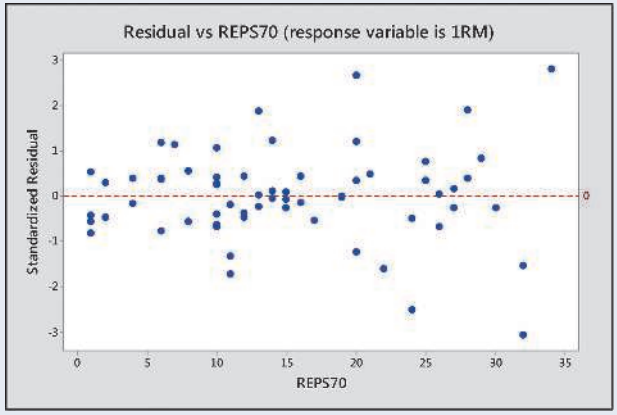
```
Regression Equation
1RM = 55.01 + 0.1668 LBM + 1.658 REPS70
Coefficients
Term        Coef    SE Coef   T-Value   P-Value
Constant   55.01      7.74      7.11     0.000
LBM       0.1668    0.0752      2.22     0.030
REPS70     1.658     0.109     15.21     0.000
Model Summary
      S      R-sq
7.11957    83.17%
Analysis of Variance
Source       DF     SS       MS     F-Value   P-Value
Regression    2   15283   7641.5    150.75     0.000
Error        61    3092     50.7
Total        63   18375
```



Scatterplot matrix for Exercise 13.38.



Residual plot for Exercise 13.38, part g.



Residual plot for Exercise 13.38, part h.

13.39 House prices Use software with the House Selling Prices OR data file on the book's website to do residual analyses with the multiple regression model for y = house selling price (in thousands), x_1 = lot size, and x_2 = number of bathrooms.

a. Find a histogram of the standardized residuals. What assumption does this check? What do you conclude from the plot?

b. Plot the standardized residuals against the lot size. What does this check? What do you conclude from the plot?

13.40 Selling prices level off In the previous exercise, suppose house selling price tends to increase with a straight-line trend for small to medium size lots, but then levels off as lot size gets large, for a fixed value of number of bathrooms. Sketch the pattern you'd expect to get if you plotted the residuals against lot size. What assumption of the multiple regression model is violated?

13.41 U.S. and foreign used cars Refer to the used car data file from Exercise 13.11. The prediction equation relating y = selling price of used car (in \$) as a function of x_1 = age of car and x_2 = type of car (1 = US, 0 = Foreign) is $\hat{y} = 20{,}493 - 1{,}185x_1 - 2{,}379x_2$.

a. Using this equation, find the prediction equation relating selling price and age, separately for U.S. and foreign cars.

b. Predict by how much the price changes for a one year increase in the age of the car. Does this apply for both types of cars? Explain.

c. Find the predicted price of a (i) U.S. and (ii) foreign car that is eight years old. Show how the difference between them relates to a parameter estimate for the model.

13.42 Mountain bike prices The Mountain Bike data file on the book's website shows selling prices for mountains bikes. When y = mountain bike price (\$) is regressed on x_1 = weight of bike (lbs) and x_2 = the type of suspension (0 = full, 1 = front end), $\hat{y} = 2{,}741.62 - 53.752x_1 - 643.595x_2$.

a. Interpret the estimated effect of the weight of the bike.

b. Interpret the estimated effect of the type of suspension on the mountain bike.

13.43 Predict using house size and condition For the House Selling Prices OR data set, when we regress y = selling price (in thousands) on x_1 = house size and x_2 = condition (1 = Good, 0 = Not Good), we get the results shown.

Regression of selling price of house in thousands versus house size and condition

```
Term           Coef      SE Coef   T-Value   P-Value
Constant        96.27      13.46      7.15     0.000
House Size   0.066463   0.004682     14.20     0.000
Condition       12.93      17.20      0.75     0.453
S = 81.7874 R-Sq = 50.6%
```

a. Report the regression equation. Find and interpret the separate lines relating predicted selling price to house size for good condition homes and for homes in not good condition.

b. Sketch how selling price varies as a function of house size for homes in good condition and for homes in not good condition.

c. Estimate the difference between the mean selling price of homes in good and in not good condition, controlling for house size.

13.44 Quality and productivity The table shows data from 27 automotive plants on y = number of assembly defects per 100 cars and x = time (in hours) to assemble each vehicle. The data are in the Quality and Productivity file on the book's website.

Number of defects in assembling 100 cars and time to assemble each vehicle

Plant	Defects	Time	Plant	Defects	Time	Plant	Defects	Time
1	39	27	10	89	17	19	69	54
2	38	23	11	48	20	20	79	18
3	42	15	12	38	26	21	29	31
4	50	17	13	68	21	22	84	28
5	55	12	14	67	26	23	87	44
6	56	16	15	69	30	24	98	23
7	57	18	16	69	32	25	100	25
8	56	26	17	70	31	26	140	21
9	61	20	18	68	37	27	170	28

Source: Data from S. Chatterjee, M. Handcock, and J. Simonoff, *A Casebook for a First Course in Statistics and Data Analysis* (Wiley, 1995); based on graph in *The Machine That Changed the World*, by J. Womack, D. Jones, and D. Roos (Macmillan, 1990).

a. The prediction equation is $\hat{y} = 61.3 + 0.35x$. Find the predicted number of defects for a car having assembly time (i) 12 hours (the minimum) and (ii) 54 hours (the maximum).

b. The first 11 plants were Japanese facilities and the rest were not. Let x_1 = time to assemble vehicle and x_2 = whether facility is Japanese (1 = yes, 0 = no). The fit of the multiple regression model is $\hat{y} = 105.0 - 0.78x_1 - 36.0x_2$. Interpret the coefficients that estimate the effect of x_1 and the effect of x_2.

c. Explain why part a and part b indicate that Simpson's paradox has occurred.

d. Explain *how* Simpson's paradox occurred. To do this, construct a scatterplot between y and x_1 in which points are identified by whether the facility is Japanese. Note that the Japanese facilities tended to have low values for both x_1 and x_2.

13.45 Predicting pizza sales A chain restaurant that specializes in selling pizza wants to analyze how y = sales for a customer (the total amount spent by a customer on food and beverage, in pounds) depends on the location of the restaurant, which is classified as inner city, suburbia, or at an interstate exit.

a. Construct indicator variables x_1 for inner city and x_2 for suburbia so you can include location in a regression equation for predicting the sales.

b. For part a, suppose $\hat{y} = 6.9 + 1.2x_1 + 0.5x_2$. Find the difference between the estimated mean sales at inner-city locations and at interstate exits.

13.46 Houses, size, and garage Use the House Selling Prices OR data file on the book's website to regress selling price in thousands on house size and whether the house has a garage.

a. Report the prediction equation. Find and interpret the equations predicting selling price using house size for homes with and without a garage.

b. How do you interpret the coefficient of the indicator variable for whether the home has a garage?

13.47 House size and garage interact? Refer to the previous exercise.

a. Explain what the no interaction assumption means for this model.

b. Sketch a hypothetical scatter diagram, showing points identified by garage or no garage, suggesting that there is actually a substantial degree of interaction.

13.48 Equal slopes for car prices? Refer to Exercise 13.41, with $\hat{y}$ = predicted selling price of used car and x_1 = age of car. When equations are fitted *separately* for U.S. and foreign cars, we get $\hat{y} = 23{,}417 - 1{,}715x_1$ for U.S. cars and $\hat{y} = 15{,}536 - 557x_1$ for foreign cars.

a. In allowing the lines to have different slopes, we allow for an _____ between age and type of car in their effects on the price. (Fill in the correct word.)

b. Predict by how much the price changes for a one-year increase in the age of the car. Do you need to do this separately for each type of car? Explain.

c. Based on the separate prediction equations, find the predicted price of a (i) U.S. and (ii) foreign car that is eight years old.

13.49 Comparing revenue An entrepreneur owns two filling stations—one at an inner city location and the other at an interstate exit location. He wants to compare the regressions of y = total daily revenue on x = number of customers who visit the filling station, for total revenue listed on a daily basis at the inner city location and at the interstate exit location. Explain how you can do this using regression modeling

a. With a single model, having an indicator variable for location that assumes the slopes are the same for each location.

b. With separate models for each location, permitting the slopes to be different.

13.50 Income and credit cards Example 12 used logistic regression to estimate the probability of having a travel credit card when x = annual income (in thousands of euros). Show that the estimated probability of having a travel credit card at the income level of €35,000 equals 0.54.

13.51 Hall of Fame induction Baseball's highest honor is election to the Hall of Fame. The history of the election process, however, has been filled with controversy and accusations of favoritism. Most recently, there is also the discussion about players who used performance enhancement drugs. The Hall of Fame has failed to define what the criteria for entry should be. Several statistical models have attempted to describe the probability of a player being offered entry into the Hall of Fame. How does hitting 400 or 500 home runs affect a player's chances of being enshrined? What about having a .300 average or 1500 RBI? One factor, the number of home runs, is examined by using logistic regression as the probability of being elected:

$$P(\text{HOF}) = \frac{e^{-6.7+0.0175\text{HR}}}{1 + e^{-6.7+0.0175\text{ HR}}}.$$

a. Compare the probability of election for two players who are 10 home runs apart—say, 369 home runs versus 359 home runs.

b. Compare the probability of election for a player with 475 home runs versus the probability for a player with 465 home runs. (These happen to be the figures for Willie Stargell and Dave Winfield.)

13.52 Cancer prediction A breast cancer study at a city hospital in New York used logistic regression to predict the probability that a female has breast cancer. One explanatory variable was x = radius of the tumor (in cm). The results are as follows:

Term	Coef
Constant	−2.165
radius	2.585

The quartiles for the radius were Q1 = 1.00, Q2 = 1.35, and Q3 = 1.85.

a. Find the probability that a female has breast cancer at Q1 and Q3.

b. Interpret the effect of radius by estimating how much the probability increases over the middle half of the sampled radii, between Q1 and Q3.

13.53 Cancer prediction (continued) Refer to the previous exercise. For what values of the radius do you estimate that a female has a probability of (a) 0.50, (b) greater than 0.50, and (c) less than 0.50, of having breast cancer?

13.54 Voting and income A logistic regression model describes how the probability of voting for the Republican candidate in a presidential election depends on x, the voter's total family income (in thousands of dollars) in the previous year. The prediction equation for a particular sample is

$$\hat{p} = \frac{e^{-1.00+0.02x}}{1 + e^{-1.00+0.02x}}.$$

Find the estimated probability of voting for the Republican candidate when (a) income = \$10,000, (b) income = \$100,000. Describe how the probability seems to depend on income.

13.55 Equally popular candidates Refer to the previous exercise.

a. At which income level is the estimated probability of voting for the Republican candidate equal to 0.50?

b. Over what region of income values is the estimated probability of voting for the Republican candidate (i) greater than 0.50 and (ii) less than 0.50?

c. At the income level for which $\hat{p} = 0.50$, give a linear approximation for the change in the probability for each \$1000 increase in income.

13.56 Many predictors of voting Refer to the previous two exercises. When the explanatory variables are x_1 = family income, x_2 = number of years of education, and x_3 = gender (1 = male, 0 = female), suppose a logistic regression reports

Term	Coef	SE Coef
Constant	−2.40	0.12
income	0.02	0.01
education	0.08	0.05
gender	0.20	0.06

For this sample, x_1 ranges from 6 to 157 with a standard deviation of 25, and x_2 ranges from 7 to 20 with a standard deviation of 3.

a. Interpret the effects by using the sign of the coefficient for each predictor.

b. Illustrate the gender effect by finding and comparing the estimated probability of voting Republican for (i) a man with 16 years of education and income $40,000 and (ii) a woman with 16 years of education and income $40,000.

13.57 Graduation, gender, and race The U.S. Bureau of the Census lists college graduation numbers by race and gender. The table shows the data for graduating 25-year-olds.

College graduation

Group	Sample Size	Graduates
White females	31,249	10,781
White males	39,583	10,727
Black females	13,194	2,309
Black males	17,707	2,054

Source: J. J. McArdle and F. Hamagami, *J. Amer. Statist. Assoc.,* vol. 89 (1994), pp. 1107–1123. Data from U.S. Bureau of the Census, American Community Survey 2005–2007.

a. Identify the response variable.

b. Express the data in the form of a three-variable contingency table that cross-classifies whether graduated (yes, no), race, and gender.

c. When we use indicator variables for race (1 = white, 0 = black) and for gender (1 = female, 0 = male), the coefficients of those predictors in the logistic regression model are 0.975 for race and 0.375 for gender. Based on these estimates, which race and gender combination has the highest estimated probability of graduation? Why?

13.58 Death penalty and race The three-dimensional contingency table shown is from a study of the effects of racial characteristics on whether individuals convicted of homicide receive the death penalty. The subjects classified were defendants in indictments involving cases with multiple murders in Florida between 1976 and 1987.

Death penalty verdict by defendant's race and victims' race

Defendant's Race	Victims' Race	Death Penalty Yes	Death Penalty No	Percent Yes
White	White	53	414	11.3
	Black	0	16	0.0
Black	White	11	37	22.9
	Black	4	139	2.8

Source: Data from M. L. Radelet and G. L. Pierce, *Florida Law Rev.*, vol. 43, 1991, pp. 1–34.

a. Based on the percentages shown, controlling for victims' race, for which defendant's race was the death penalty more likely?

b. Let y = death penalty verdict (1 = yes, 0 = no), let d be an indicator variable for defendant's race (1 = white, 0 = black), and let v be an indicator variable for victims' race (1 = white, 0 = black). The logistic regression prediction equation is

$$\hat{p} = \frac{e^{-3.596 - 0.868d + 2.404v}}{1 + e^{-3.596 - 0.868d + 2.404v}}.$$

According to this equation, for which of the four groups is the death penalty most likely? Explain your answer.

13.59 Death penalty probabilities Refer to the previous exercise.

a. Based on the prediction equation, when the defendant is black and the victims were white, show that the estimated death penalty probability is 0.233.

b. The model-estimated probabilities are 0.011 when the defendant is white and victims were black, 0.113 when the defendant and the victims were white, and 0.027 when the defendant and the victims were black. Construct a table cross-classifying defendant's race by victims' race and show the estimated probability of the death penalty for each cell. Use this to interpret the effect of defendant's race.

c. Collapse the contingency table over victims' race, and show that (ignoring victims' race) white defendants were more likely than black defendants to get the death penalty. Comparing this with what happens when you control for victims' race, explain how Simpson's paradox occurs.

14.1 Practicing the Basics

14.1 Restaurant satisfaction The CEO of a company that owns six restaurants wants to evaluate and compare visitor satisfaction across all six restaurants. The company's research department randomly sampled 150 people who had visited any of the restaurants during the past month and asked them to rate their expectations of the restaurant before their visit and to rate the quality of the actual visit. Both observations used a rating scale of 0–5, with 0 = very poor and 5 = excellent. The researchers compared the restaurants on the gap between prior expectation and actual quality, using the difference score, y = performance gap = (prior expectation score − actual quality score).

a. Identify the response variable, the factor, and the categories that form the groups.

b. State the null and alternative hypotheses for conducting an ANOVA.

c. Explain why the *df* values for this ANOVA are $df_1 = 5$ and $df_2 = 144$.

d. How large an F test statistic is needed to get a P-value = 0.05 in this ANOVA?

14.2 Satisfaction with banking A bank conducts a survey in which it randomly samples 400 of its customers. The survey asks the customers which way they use the bank the most: (1) interacting with a teller at the bank, (2) using ATMs, or (3) using the bank's online banking service. It also asks their level of satisfaction with the service they most often use (on a scale of 0 to 10 with 0 = very poor and 10 = excellent). Does mean satisfaction differ according to how they most use the bank?

a. Identifying notation, state the null and alternative hypotheses for conducting an ANOVA with data from the survey.

b. Report the *df* values for this ANOVA. Above what F test statistic values give a P-value below 0.05?

c. For the data, $F = 0.46$ and the P-value equals 0.63. What can you conclude?

d. What were the assumptions on which the ANOVA was based? Which assumption is the most important?

14.3 What's the best way to learn French? The following table shows scores on the first quiz (maximum score

10 points) for eighth-grade students in an introductory level French course. The instructor grouped the students in the course as follows:

Group 1: Never studied foreign language before but have good English skills

Group 2: Never studied foreign language before and have poor English skills

Group 3: Studied at least one other foreign language

French scores on the quiz

	Group 1	Group 2	Group 3
	4	**1**	**9**
	6	**5**	**10**
	8		**5**
Mean	6.0	3.0	8.0
Std. Dev.	2.000	2.828	2.646

Source	DF	SS	MS	F	P
Group	2	30.00	15.00	2.50	0.177
Error	5	30.00	6.00		
Total	7	60.00			

a. Defining notation and using results obtained with software, also shown in the table, report the five steps of the ANOVA test.

b. The sample means are quite different, but the P-value is not small. Name one important reason for this. (*Hint*: For given sample means, how do the results of the test depend on the sample sizes?)

c. Was this an experimental study, or an observational study? Explain how a lurking variable could be responsible for Group 3 having a larger mean than the others. (Thus, even if the P-value were small, it is inappropriate to assume that having studied at least one foreign language causes one to perform better on this quiz.)

14.4 What affects the *F* value? Refer to the previous exercise.

a. Suppose that the first observation in the second group was actually 9, not 1. Then the standard deviations are the same as reported in the table, but the sample means are 6, 7, and 8 rather than 6, 3, and 8. Do you think the *F* test statistic would be larger, the same, or smaller? Explain your reasoning, without doing any calculations.

b. Suppose you had the same means as shown in the table but the sample standard deviations were 1.0, 1.8, and 1.6, instead of 2.0, 2.8, and 2.6. Do you think the *F* test statistic would be larger, the same, or smaller? Explain your reasoning.

c. Suppose you had the same means and standard deviations as shown in the table but the sample sizes were 30, 20, and 30, instead of 3, 2, and 3. Do you think the *F* test statistic would be larger, the same, or smaller? Explain your reasoning.

d. In parts a, b, and c, would the P-value be larger, the same, or smaller? Why?

14.5 Outsourcing Example 1 at the beginning of this chapter mentioned a study to compare customer satisfaction at service centers in San Jose, California; Toronto, Canada; and Bangalore, India. Each center randomly sampled 100 people who called during a two-week period. Callers rated their satisfaction on a scale of 0 to 10, with higher scores representing greater satisfaction. The sample means were 7.6 for San Jose, 7.8 for Toronto, and 7.1 for Bangalore. The table shows the results of conducting an ANOVA.

a. Define notation and specify the null hypothesis tested in this table.

b. Explain how to obtain the *F* test statistic value reported in the table from the MS values shown and report the values of df_1 and df_2 for the *F* distribution.

c. Interpret the P-value reported for this test. What conclusion would you make using a 0.05 significance level?

Customer satisfaction with outsourcing

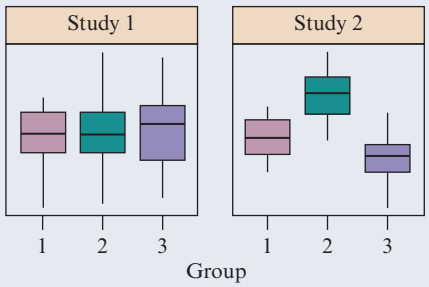
Source	DF	SS	MS	F	P
Group	2	26.00	13.00	27.6	0.000
Error	297	140.00	0.47		
Total	299	60.00			

14.6 ANOVA and box plots For two studies, each comparing three groups, the box plots below show results. (Each box plot is based on a random sample of size 40.)

a. Judging from the box plots, which study will more likely lead to a rejection of the ANOVA null hypothesis of equal population means? Explain.

b. Which study will have the large value for the *F* test statistic? Why?

c. The P-value for the ANOVA *F* test for the second study equals 0.001. Does this necessarily imply that all three population means are different from each other?



14.7 Years of education A recent General Social Survey asked students at an Australian university, "What is the ideal number of years of education for an individual?" Do responses tend to depend on the subjects' area of residence? Results of an ANOVA are shown in the printout, for different residential areas (inner city, suburbia, countryside).

a. Define notation and specify the null hypothesis tested in this printout.

b. Summarize the assumptions made to conduct this test

c. Report the *F* test statistic value and the P-value for this test. Interpret the P-value.

d. Based on part c, can you conclude that each pair of residential area has different population means for an ideal number of years of education? Explain.

Ideal number of years of education by area of residence

Source	DF	SS	MS	F	P
Area of residence	2	9.21	4.61	5.96	0.003
Error	1195	922.82	0.77		
Total	1197	932.03			

14.8 Smoking and personality A study about smoking and personality (by A. Terracciano and P. Costa, *Addiction*, vol. 99, 2004, pp. 472–481) used a sample of 1638 adults in the Baltimore Longitudinal Study on Aging. The subjects formed three groups according to smoking status (never, former, current). Each subject completed a personality questionnaire that provided scores on various personality scales designed to have overall means of about 50 and standard deviations of about 10. The table shows some results for three traits, giving the means with standard deviations in parentheses.

	Never smokers ($n = 828$)	Former smokers ($n = 694$)	Current smokers ($n = 116$)	F
Neuroticism	46.7 (9.6)	48.5 (9.2)	51.9 (9.9)	17.77
Extraversion	50.4 (10.3)	50.2 (10.0)	50.9 (9.4)	0.24
Conscientiousness	51.8 (10.1)	48.9 (9.7)	45.6 (10.3)	29.42

a. For the F test for the extraversion scale, using the 0.05 significance level, what conclusion would you make?

b. Refer to part a. Does this mean that the population means are necessarily equal?

14.9 French cuisine The restaurant guide Zagat compiles customer ratings on the quality of food on a 30-point scale. The data set French Cuisine on the book's website contains a random sample of ratings of French cuisine restaurants in New York, London, and Paris, compiled in June 2015 from the zagat.com website.

a. Using software, construct dot plots or side-by-side box plots that show the data and find the mean and standard deviation of the ratings in each city.

b. Conduct an ANOVA. Report the hypotheses, F test statistic value, P-value, and interpret results. Use a significance level of 0.05.

14.10 Software and French ANOVA Refer to Exercise 14.3. Using software,

a. Create the data file and find the sample means and standard deviations.

b. Find and report the ANOVA table. Interpret the P-value.

c. Change an observation in Group 2 so that the P-value will be smaller. Specify the value you changed and report the resulting F test statistic and the P-value. Explain why the value you changed would have this effect.

14.11 Comparing therapies for anorexia The Anorexia data file on the book's website shows weight change for 72 anorexic teenage girls who were randomly assigned to one of three psychological treatments (cognitive or family therapy and a control treatment). Use software to analyze these data. (The change scores are in the last three columns of the data set. Alternatively, the website also contains a dataset that shows the data in "long" format, with the treatment in one column and the change score in another column.)

a. Construct box plots for the three groups. Use these and sample summary means and standard deviations to describe the three samples.

b. For the one-way ANOVA comparing the three mean weight changes, report the test statistic and P-value. Explain how to interpret.

c. State and check the assumptions for the test in part b.

14.2 Practicing the Basics

14.12 House prices and age For the House Selling Prices OR data file on the book's website, the output shows the result of conducting an ANOVA comparing mean house selling prices (in $1000) by Age Category (New = 0 to 24 years old, Medium = 25 to 50 years old, Old = 51 to 74 years old, Very Old = 75 + years old). It also shows a summary table of means and standard deviations of the selling prices, by age group.

a. Using information given in the tables, show how to construct a 95% confidence interval comparing the population means of new and medium-aged houses.

b. Interpret the confidence interval.

Age	N	Mean	StDev
New	78	305.8	125.9
Medium	72	242.8	79.3
Old	37	217.5	85.4
Very Old	13	316.3	195.4

Source	DF	SS	MS	F	P
Age Condition	3	281852	93951	7.70	0.000
Error	196	2387017	12179		
Total	199	2668870			

14.13 Time on Facebook Do freshmen spent significantly more time on Facebook than other class ranks? A recent study (R. Junco, *Journal of Applied Developmental Psychology*, 2015, vol. 36, p. 18–29) investigated the amount of time per day freshman, sophomores, juniors, and seniors spent on Facebook while doing schoolwork. The students surveyed were U.S. residents from a 4-year, public, primarily residential institution in the Northeastern United States. The data from the survey are available on the book's website, with time measured in minutes. The following computer output shows the mean time each cohort spent on Facebook and an ANOVA table. Construct a 95% confidence interval to compare the population mean time spent per day on Facebook between freshmen and seniors.

Summary statistics for time spent on Facebook by class year and ANOVA table

	FR	SO	JU	SE
n	440	347	403	407
Mean	63.7	56.5	70.4	49.0
s	71.5	67.7	79.0	63.7

Source	DF	SS	MS	F	P
Class	3	102903	34301	6.84	0.000
Error	1593	7988051	5014		
Total	1596	8090954			

14.14 Comparing telephone holding times Examples 2 and 3 analyzed whether telephone callers to an airline would stay on hold different lengths of time, on average, if they heard (a) an advertisement about the airline, (b) Muzak, or (c) classical music. The sample means were 5.4, 2.8, and 10.4, with $n_1 = n_2 = n_3 = 5$. The ANOVA test had $F = 74.6/11.6 = 6.4$ and a P-value of 0.013.

a. A 95% confidence interval comparing the population mean times that callers are willing to remain on hold for classical music and Muzak is (2.9, 12.3). Interpret this interval.

b. The margin of error was 4.7 for this comparison. Without doing a calculation, explain why the margin of error is 4.7 for comparing *each* pair of means.

c. The 95% confidence intervals are (0.3, 9.7) for $\mu_3 - \mu_1$ and $(-2.1, 7.3)$ for $\mu_1 - \mu_2$. Interpret these two confidence intervals. Using these two intervals and the interval from part a, summarize what the airline company learned from this study.

d. The confidence intervals are wide. In the design of this experiment, what could you change to estimate the differences in means more precisely?

14.15 Tukey holding time comparisons Refer to the previous exercise. We could instead use the Tukey method to construct multiple comparison confidence intervals. The Tukey confidence intervals having *overall* confidence level 95% have margins of error of 5.7, compared to 4.7 for the separate 95% confidence intervals in the previous exercise.

a. According to this method, which groups are significantly different?

b. Why are the margins of error larger than with the separate 95% intervals?

14.16 Hamburger sales The market research department of a chain of hamburger restaurants wants to compare the mean monthly sales of hamburgers under three different marketing strategies. It randomly assigns 15 restaurants to the three groups, five per group. The sample means for the three groups were 1800, 1500, and 1200. The table shows the ANOVA table from SPSS.

Hamburger sales

Source	DF	SS	MS	F	P
Group	2	82.00	41.00	1.02	0.389
Error	12	481.00	40.08		
Total	14	563.00			

a. Report and interpret the P-value for the ANOVA F test.

b. For the Tukey 95% multiple comparison confidence intervals comparing each pair of means, calculate the margin of error. Explain why will it be same for all pairs of means.

14.17 Hamburger sales regression Refer to the previous exercise.

a. Set up indicator variables for a regression model so that an F test for the regression parameters is equivalent to the ANOVA test comparing the three means.

b. Express the null hypothesis both in terms of population means and in terms of regression parameters for the model in part a.

c. The prediction equation from fitting the multiple regression model equals $\hat{y} = 1200 + 600x_1 + 300x_2$. Use this equation to find the predicted population means for each group.

14.18 Outsourcing satisfaction Exercise 14.5 showed an ANOVA for comparing mean customer satisfaction scores for three service centers. The sample means on a scale of 0 to 10 were 7.60 in San Jose, 7.80 in Toronto, and 7.10 in Bangalore. Each sample size $= 100$, MS error $= 0.47$, and the F test statistic $= 27.6$ has P-value < 0.001.

a. Explain why the margin of error for separate 95% confidence intervals is the same for comparing the population means for each pair of cities. Show that this margin of error is 0.19.

b. Find the 95% confidence interval for the difference in population means for each pair of service centers. Interpret.

c. The margin of error for Tukey 95% multiple comparison confidence intervals for comparing the service centers is 0.23. Construct the intervals. Interpret.

d. Why are the confidence intervals different in part b and in part c? What is an advantage of using the Tukey intervals?

14.19 Regression for outsourcing Refer to the previous exercise.

a. Set up indicator variables to represent the three service centers.

b. The prediction equation is $\hat{y} = 7.1 + 0.5x_1 + 0.7x_2$, where x_1 is an indicator variable for San Jose and x_2 an indicator variable for Toronto. Find the estimate for the difference in population means (i) between San Jose and Bangalore and (ii) between Toronto and Bangalore.

14.20 Advertising effect on sales Each of 100 restaurants in a fast-food chain is randomly assigned one of four media for an advertising campaign: A = radio, B = TV, C = newspaper, D = mailing. For each restaurant, the observation is the change in sales, defined as the difference between the sales for the month during which the advertising campaign took place and the sales in the same month a year ago (in thousands of dollars).

a. By creating indicator variables, write a regression equation for the analysis to compare mean change in sales for the four media.

b. Explain how you could use the regression model to test the null hypothesis of equal population mean change in sales for the four media.

c. The prediction equation is $\hat{y} = 35 + 5x_1 - 10x_2 + 2x_3$ where x_1, x_2, and x_3 are indicator variables for media A, B, and C, respectively. Estimate the difference in mean change in sales for media (i) A and D, (ii) A and B. (*Hint:* For part (ii), write the prediction equation for the mean for media A, then for media B, and then subtract.)

14.21 French ANOVA Refer to Exercise 14.3 about studying French, with data shown again below. Using software,

a. Compare the three pairs of means with separate 95% confidence intervals. Interpret.

b. Compare the three pairs of means with Tukey 95% multiple comparison confidence intervals. Interpret and explain why the intervals are different than in part a.

Group 1	Group 2	Group 3
4	1	9
6	5	10
8		5

14.22 Multiple comparison for time on Facebook Refer to Exercise 14.13, which investigated the amount of time freshman, sophomores, juniors, and seniors spent on Facebook while doing schoolwork. The data from the study are available on the book's website, where time is measured in minutes.

a. If you want to construct confidence intervals between all possible pairs of means for the four classes, how many intervals do you need to construct?

b. Using software (such as the web app ANOVA accessible from the book's website), find the confidence intervals for all possible pairs of population means such that the overall error rate of all intervals is 0.05 (use Tukey's method). Visualize the confidence intervals in a figure similar to Figure 14.4. (The web app can construct such a plot, which you can download.)

c. Is it true that seniors spent significantly less time than all other classes? Explain.

14.23 Effect of fertilizers An experiment randomly assigns 100 agricultural plots of land to one of four groups of fertilizers: low-dose nitrogen, high-dose nitrogen, low-dose potassium, and high-dose potassium. After three months of using them, the change in harvest is measured (as compared to last year's harvest).

a. Identify the response variable and the two factors.

b. What are the four treatments being compared in this experiment?

c. What comparisons are relevant when we control for dose level?

14.24 Fertilizer main effects For the previous exercise, show a hypothetical set of population means for the four groups that would have

a. A dose effect but no fertilizer effect.

b. A fertilizer effect but no dose effect.

c. A fertilizer effect and a dose effect.

d. No fertilizer effect and no dose effect.

14.25 Political ideology in 2014 The GSS measures political ideology on a seven-point scale, starting with 1 = extremely liberal, to 4 = moderate, to 7 = extremely conservative. The following table shows results from an ANOVA analysis about political ideology, with sex (female, male) and race (black, white) as two factors, using 2014 GSS data.

a. State the null hypothesis to which the F test statistic in the race row refers.

b. Show how to use mean squares to construct the F test statistic for the race main effect, report its P-value, and interpret.

c. Is there evidence that the mean political ideology in the United States differs between females and males, for either blacks or whites? Explain.

Source	DF	SS	MS	F	P
Sex	24	3.0	3.0	1.4	0.229
Race	1	36.5	36.5	17.4	0.000
Error	2203	4524.3	2.1		
Total	2205	4563.8			

14.26 House prices, age, and bedrooms For the House Selling Prices OR data file on the book's website, the output shows the result of conducting a two-way ANOVA of house selling prices (in thousands) by the number of bedrooms in the house and the age (New, Medium, Old, Very Old—see exercise 14.12) of the houses in Corvallis, Oregon.

a. For testing the main effect of age, report the F test statistic value and show how it was formed from other values reported in the ANOVA table.

b. Report the P-value for the main effect test for age and interpret.

Source	DF	SS	MS	F	P
Bedrooms	7	517868	73981	7.325	0.000
Age Condition	3	242042	80681	7.988	0.000
Error	189	1908959	10100		
Total	199	2668870			

14.27 Corn and manure In Example 10, the coefficient of the manure-level indicator variable m is 1.96.

a. Explain why this coefficient is the estimated difference in mean corn yield between the high and low levels of manure, for each level of fertilizer.

b. Explain why the 95% confidence interval for the difference in mean corn yield between the high and low levels of manure is $1.96 \pm 2.11(0.747)$.

14.28 Hang up if recording repeated? Example 2 described an experiment in which telephone callers to an airline were put on hold with an advertisement, Muzak, or classical music in the background. Each caller who was chosen was also randomly assigned to a category of a second factor: whether the recording played was five minutes long or ten minutes long. (In each case, it was repeated at the end.) The table shows the data classified by both factors and the results of a two-way ANOVA.

Telephone holding times by type of recording and repeat time

	Repeat Time	
Recording	**Ten Minutes**	**Five Minutes**
Advertisement	**8, 11, 2**	**5, 1**
Muzak	**1, 4, 3**	**0, 6**
Classical	**13, 8, 15**	**7, 9**

Source	DF	SS	MS	F	P
Recording	2	149.20	74.60	7.09	0.011
Repeat	1	23.51	23.51	2.24	0.163
Error	11	115.69	10.52		
Total	14	288.40			

a. State the null hypothesis to which the F test statistic in the Recording row refers.

b. Show how to use mean squares to construct the F test statistic for the Recording main effect, report its P-value, and interpret.

c. On what assumptions is this analysis based?

14.29 Regression for telephone holding times Refer to the previous exercise. Let $x_1 = 1$ for the advertisement and 0 otherwise, $x_2 = 1$ for Muzak and 0 otherwise, and $x_1 = x_2 = 0$ for classical music. Likewise, let $x_3 = 1$ for repeating in 10-minute cycles and $x_3 = 0$ for repeating in 5-minute cycles. The display shows results of a regression of the telephone holding times on these indicator variables.

Regression for telephone holding times

Term	Coef	SE Coef	T-Value	P-Value
Constant	8.867	1.776	4.99	0.000
x1	−5.000	2.051	−2.44	0.033
x2	−7.600	2.051	−3.71	0.003
x3	2.556	1.709	1.50	0.163

a. Give the corresponding regression model for the population mean and show the equation for the population mean at each setting of the two factor levels. (Create a table similar to Table 14.11.)

b. State the prediction equation. Interpret the parameter estimates.

c. Find the estimated means for the six groups in the two-way cross-classification of type of recording and repeat time.

d. Find the estimated difference between the mean holding times for 10-minute repeats and 5-minute repeats, for a fixed recording type. How can you get this estimate from a coefficient of an indicator variable in the prediction equation?

e. Find a 95% confidence interval for the difference between the mean holding times for 10-minute repeats and 5-minute repeats.

14.30 Wheat crop yield The following table shows the result of fitting the regression model for predicting wheat crop yield with indicator variables for fertilizer level (low: 0, high: 1) and irrigation level (low: 0, high: 1).

Term	Coef	SE Coef	T-Value	P-Value
Constant	10.1500	0.6470	15.69	0.000
Fertilizer	1.8100	0.7271	2.49	0.022
Irrigation	1.1700	0.4471	2.62	0.017

a. Explain why the coefficient of irrigation is the estimated difference in mean wheat crop yield between the high and low levels of irrigation, for each level of fertilizer.

b. Explain why the 95% confidence interval for the difference in mean wheat crop yield between the high and low levels of irrigation is $1.17 \pm 2.09\ (0.4471)$, with $df = 19$ for the MS error.

14.31 Income by gender and degree In 2012, the population mean hourly wage for males was \$17 for high school graduates, \$33 for college graduates and \$43 for males with more advanced degrees. For females, the means were \$14 for high school graduates, \$24 for college graduates, and \$32 for females with more advanced degrees.[3]

a. Identify the response variable and the two factors.

b. Show these means in a two-way classification of the two factors, similar to the table in the margin of Example 9.

c. Compare the differences between males and females for (i) high school graduates and (ii) college graduates. Explain why there is interaction and describe it.

d. Show a set of six population mean wages that would satisfy H_0: no interaction.

14.32 Ideology by gender and race Refer to Example 12, the sample means from which are shown again below (for 2008 data).

Mean political Ideology

	Race	
Gender	**Black**	**White**
Female	4.164	4.2675
Male	3.819	4.4443

[3] *Source*: Data from *The State of Working America*, 12th edition, Economic Policy Institute.

a. Explain how to obtain the following interpretation for the interaction from the sample means: "For females there is no race effect on ideology. For males, whites are more conservative by about half an ideology category, on the average."

b. Suppose that instead of the two-way ANOVA, you performed a one-way ANOVA with gender as the predictor and a separate one-way ANOVA with race as the predictor. Suppose the ANOVA for gender does not show a significant effect. Explain how this could happen, even though the two-way ANOVA showed a gender effect for each race. (*Hint:* Will the overall sample means for females and males be more similar than they are for each race?)

c. Refer to part b. Summarize what you would learn about the gender effect from a two-way ANOVA that you would fail to learn from a one-way ANOVA.

14.33 Attractiveness and getting dates The results in the table are from a study of physical attractiveness and subjective well-being (E. Diener et al., *Journal of Personality and Social Psychology*, vol. 69, 1995, pp. 120–129). A panel rated a sample of college students on their physical attractiveness. The table presents the number of dates in the past three months for students rated in the top or bottom quartile of attractiveness.

a. Identify the response variable and the factors.

b. Do these data appear to show interaction? Explain.

c. Based on the results in the table, specify one of the ANOVA assumptions that these data violate. Is this the most important assumption?

Dates and attractiveness

	Number of DATES, MEN			Number of DATES, WOMEN		
ATTRACTIVENESS	**Mean**	**Std. Dev**	***n***	**Mean**	**Std. Dev**	***n***
More	9.7	10.0	35	17.8	14.2	33
Less	9.9	12.6	36	10.4	16.6	27

14.34 Diet and weight gain A randomized experiment[4] measured weight gain (in grams) of male rats under six diets varying by source of protein (beef, cereal, pork) and level of protein (high, low). Ten rats were assigned to each diet. The data are shown in the table that follows and are also available in the Protein and Weight Gain data file on the book's website.

a. Conduct a two-way ANOVA that assumes a lack of interaction. Report the F test statistic and the P-value for testing the effect of the protein level. Interpret.

b. Now conduct a two-way ANOVA that also considers potential interaction. Report the hypotheses, test statistic, and P-value for a test of no interaction. What do you conclude at the 0.05 significance level? Explain.

c. Refer to part b. Allowing interaction, construct a 95% confidence interval to compare the mean weight gain for the two protein levels, for the beef source of protein.

Weight gain by source of protein and by level of protein

	High Protein	**Low Protein**
Beef	73, 102, 118, 104, 81, 107, 100, 87, 117, 111	90, 76, 90, 64, 86, 51, 72, 90, 95, 78
Cereal	98, 74, 56, 111, 95, 88, 82, 77, 86, 92	107, 95, 97, 80, 98, 74, 74, 67, 89, 58
Pork	94, 79, 96, 98, 102, 102, 108, 91, 120, 105	49, 82, 73, 86, 81, 97, 106, 70, 61, 82

14.35 Regression of weight gain on diet Refer to the previous exercise.

a. Set up indicator variables for protein source and for protein level and specify a regression model with the effects both of protein level and protein source on weight gain.

b. Fit the model in part a and explain how to interpret the parameter estimate for the protein level indicator variable.

c. Show how you could test a hypothesis about beta parameters in the model in part a to analyze the effect of protein source on weight gain.

d. Using the fit of the model, find the estimated mean for each of the six diets. Explain what it means when we say that these estimated means do not allow for interaction between protein level and source in their effects on weight loss.

[4] *Source:* Data from G. Snedecor and W. Cochran, *Statistical Methods*, 6th ed. (Iowa State University Press, 1967), p. 347.

15.1 Practicing the Basics

15.1 Tanning experiment Suppose the tanning experiment described in Examples 1 and 2 used only four participants, two for each treatment.

a. Show the six possible ways the four ranks could be allocated, two to each treatment, with no ties.

b. For each possible sample, find the mean rank for each treatment and the difference between the mean ranks.

c. Presuming H_0 is true of identical treatment effects, construct the sampling distribution of the difference between the sample mean ranks for the two treatments.

15.2 Test for tanning experiment Refer to the previous exercise. For the actual experiment, suppose the participants using the tanning studio got ranks 1 and 2 and the participants using the tanning lotion got ranks 3 and 4.

a. Find and interpret the P-value for the alternative hypothesis that the tanning studio tends to give better tans than the tanning lotion.

b. Find and interpret the P-value for the alternative hypothesis that the treatments have different effects.

c. Explain why it is a waste of time to conduct this experiment if you plan to use a 0.05 significance level to make a decision.

15.3 Comparing clinical therapies A clinical psychologist wants to choose between two therapies for treating severe mental depression. She selects six patients who are similar in their depressive symptoms and overall quality of health. She randomly selects three patients to receive Therapy 1. The other three receive Therapy 2. After one month of treatment, the improvement in each patient is measured by the change in a score for measuring severity of mental depression—the higher the change score, the better. The improvement scores are

Therapy 1: 25, 40, 45

Therapy 2: 10, 20, 30

a. Show all possible ways the ranks from 1 to 6 could be distributed between the two treatments. (*Hint:* There are 20 allocations.)

b. For each possible allocation of ranks, find the mean rank for each treatment and the difference between the mean ranks.

c. Consider the null hypothesis of identical response distributions for the two treatments. Presuming H_0 is true, construct the sampling distribution of the difference between the sample mean ranks for the two treatments.

d. For the actual data shown above, find and interpret the P-value for the alternative hypothesis that the two treatments have different effects. (You can check your answers by entering the data in the Permutation Test for Means web app, selecting Wilcoxon rank sum as the test statistic, and selecting the option for generating all possible permutations.)

15.4 Body mass reduction and smoking Smoking is a known cause of reduced body mass. To validate this, a researcher randomly selects 7 smokers and 7 nonsmokers and records their individual body mass index. The body mass index data in kg/m^2 are:

Smokers: 13.1, 15.2, 16.5, 15.0, 16.4, 19.3, 16.0

Nonsmokers: 19.5, 21.5, 23.5, 18.3, 24.5, 19.6, 18.9

a. State the hypotheses for conducting a one-sided (right tailed) Wilcoxon test.

b. Find the ranks for the two groups and their sum and mean of ranks.

c. Software reports a small-sample one-sided P-value of 0.002. Interpret and explain the significance of the P-value. (You can reproduce these results by entering the data in the Permutation Test for Means web app, selecting Wilcoxon rank sum as the test statistic, and selecting the option for generating all possible permutations.)

15.5 Estimating the effect of smoking Refer to the previous exercise. For these data, MINITAB reports:

```
Point estimate for η1 − η2 is −2.00
95.9 Percent CI for η1 − η2 is (−3.30,−0.70)
W = 31.0
Test of η1 = η2 vs η1 ≠ η2 is significant at 0.006
```

Explain how to interpret the reported (a) point estimate and (b) confidence interval.

15.6 Trading volumes The following data show the number of shares of General Electric stock traded on Mondays and on Fridays from February through April of 2011. The trading volumes (rounded to the nearest million) are as follows:

Mondays: 45, 43, 43, 66, 91, 53, 35, 45, 29, 64, 56

Fridays: 43, 41, 45, 46, 61, 56, 80, 40, 48, 49, 50, 41

Using software,

a. Plot the data. Summarize what the plot shows.

b. State the hypotheses and give the P-value for the Wilcoxon test for comparing the two groups with a two-sided alternative hypothesis. (As one option, you can find an accurate approximation of the exact P-value by using the Permutation Test for Means web app.)

c. A 95.5% confidence interval for comparing the population medians equals (−11, 13). Interpret and explain what (if any) effect the day of the week (Monday versus Friday) has on the median number of shares traded.

d. State the assumptions for the methods in parts b and c.

15.7 Teenage anorexia Previous chapters described a study that used therapy to treat teenage girls who suffered from anorexia. The girls were randomly assigned to the cognitive behavioral treatment (Group 1) or to the control group (Group 2). The study observed the weight change after a period of time. The following output shows results of a nonparametric comparison. (The data set is available online.)

a. Interpret the reported point estimate of the difference between the population medians for the weight changes for the two groups.

b. Interpret the reported confidence interval and summarize the assumptions on which it is based.

c. Report a P-value for testing the null hypothesis of identical population distributions of weight change. Specify the alternative hypothesis, and interpret the P-value.

MINITAB output comparing weight changes

```
                        N     Median
Cognitive_change       29      1.400
Control_change         26     −0.350
Point estimate for η1 − η2 is 3.05
95.0 Percent CI for η1 − η2 is (−0.60,8.10)
W = 907.0
Test of η1 = η2 vs η1 ≠ η2 is significant at 0.1111
```

15.8 How long do you tolerate being put on hold?

TRY

Examples 1–4 and 7 in Chapter 14 referred to the following randomized experiment: An airline analyzed whether telephone callers to their reservations office would remain on hold longer, on average, if they heard (a) an advertisement about the airline, (b) Muzak, or (c) classical music. For 15 callers randomly assigned to these three conditions, the table shows the data. It also shows the ranks for the 15 observations as well as the mean rank for each group and some results from using MINITAB to conduct the Kruskal-Wallis test.

a. State the null and alternative hypotheses for the Kruskal-Wallis test.

b. Identify the value of the test statistic for the Kruskal-Wallis test and state its approximate sampling distribution, presuming H_0 is true.

c. Report and interpret the P-value shown for the Kruskal-Wallis test.

d. To find out which pairs of groups significantly differ, how could you follow up the Kruskal-Wallis test?

Telephone holding times by type of recorded message

Recorded Message	Holding Time Observations	Ranks	Mean Rank
Muzak	0, 1, 3, 4, 6	1, 2.5, 5, 6, 8	4.5
Advertisement	1, 2, 5, 8, 11	2.5, 4, 7, 10.5, 13	7.4
Classical	7, 8, 9, 13, 15	9, 10.5, 12, 14, 15	12.1

Kruskal-Wallis Test: Holding Time Versus Group

Group	N	Median	Ave Rank
Muzak	5	3.000	4.5
advert	5	5.000	7.4
classical	5	9.000	12.1

H = 7.38 DF = 2 P = 0.025 (adjusted for ties)

15.9 What's the best way to learn French? Exercise 14.3 gave the data in the table for scores on the first quiz for ninth-grade students in an introductory-level French course. The instructor grouped the students in the course as follows:

Group 1: Never studied foreign language before, but have good English skills

Group 2: Never studied foreign language before; have poor English skills

Group 3: Studied at least one other foreign language

The table also shows results of using MINITAB to perform the Kruskal-Wallis test.

a. Find the rank associated with each observation and show how to find the mean rank for Group 1.

b. Report and interpret the P-value for the test.

Scores on the quiz

Group 1	Group 2	Group 3
4	1	9
6	5	10
8		5

```
Kruskal-Wallis Test on response
 Group   N    Median    Ave Rank
   1     3     6.000       4.3
   2     2     3.000       2.3
   3     3     9.000       6.2
H = 3.13 DF = 2  P = 0.209  (adjusted for ties)
```

15.10 Tea versus coffee Which of the two beverages do people prefer more—tea or coffee? For the employees surveyed at a company, 24 prefer drinking tea and 30 prefer coffee. Let p denote the corresponding population proportion who prefer tea.

a. Find the test statistic for the sign test of H_0: $p = 0.50$ against H_a: $p \neq 0.50$.

b. Refer to part a. Find and interpret the P-value.

15.11 Cell phones and reaction times Example 13 in Chapter 10 compared reaction times in a simulated driving test for the same students when they were using a cell phone and when they were not. The table shows data for the first four students. For all 32 students, 26 had faster reaction times when not using the cell phone and 6 had faster reaction times when using it.

a. Are the observations for the two treatments independent samples or dependent samples? Explain.

b. Let p denote the population proportion who would have a faster reaction time when not using a cell phone. Estimate p based on this experiment.

c. Using all 32 observations, find the test statistic and the P-value for the sign test of H_0: $p = 1/2$ against H_a: $p > 1/2$. Interpret.

d. What is the parametric method for comparing the scores? What is an advantage of it over the sign test? (*Hint:* Does the sign test use the magnitude of the difference between the two scores or just its direction?)

Reaction times in cell phone study

	Using Cell Phone?		
Student	**No**	**Yes**	**Difference**
1	604	636	32
2	556	623	67
3	540	615	75
4	522	672	150

15.12 Sign test for GRE scores Consider Example 8, for which the changes in the writing portion GRE scores for the first three people who attended a training workshop were 0.5, −0.5, and 1.5. Show how to use the sign test to test that the probability that the difference is positive equals 0.50 against the alternative hypothesis that it is greater than 0.50.

15.13 Does exercise help blood pressure? Exercise 10.50 in Chapter 10 discussed a pilot study of people who suffer from abnormally high blood pressure. A medical researcher decides to test her belief that walking briskly for at least half an hour a day has the effect of lowering blood pressure. She randomly samples three of her patients who have high blood pressure. She measures their systolic blood pressure initially and then again a month later after they participate in her exercise program. The table shows the results. Show how to analyze the data with the sign test. State the hypotheses, find the P-value, and interpret.

Subject	Before	After
1	150	130
2	165	140
3	135	120

15.14 More on blood pressure Refer to the previous exercise. The analysis there did not take into account the *size* of the change in blood pressure. Show how to do this with the Wilcoxon signed-ranks test.

a. State the hypotheses for that test, for the relevant one-sided alternative hypothesis.

b. Construct the sampling distribution for the rank sum of the positive differences when you consider the possible samples that have absolute differences in blood pressure of 20, 25, and 15.

c. Using the sampling distribution from part b, find and interpret the P-value. (When every difference is positive, or when every difference is negative, this test and the sign test give the same P-value for a given alternative hypothesis.)

15.15 More on cell phones Refer to Exercise 15.11. That analysis did not take into account the magnitudes of the differences in reaction times. Show how to do this with the Wilcoxon signed-ranks test, illustrating by using only the four observations shown in the table there.

a. State the hypotheses for the relevant one-sided test.

b. Create the sampling distribution of the sum of ranks for the positive differences.

c. Find the P-value and interpret.

15.16 Use all data on cell phones Refer to the previous exercise. When we use the data for all 32 subjects, MINITAB reports result in the following for the Wilcoxon signed-ranks test.

Wilcoxon signed-ranks test results

```
Test of median = 0.000000 versus median
not = 0.000000
             N for   Wilcoxon           Estimated
       N     Test   Statistic     P      Median
diff  32     32       490.0     0.000     47.25
```

a. State the null and alternative hypotheses for this test.

b. Explain how MINITAB found the value reported for the Wilcoxon test statistic.

c. Report and interpret the P-value.

d. Report and interpret the estimated median.