1. **Marriage age 2003.** Is there evidence that the age at which women get married has changed over the past 100 years? The scatterplot shows the trend in age at first marriage for American women (www.census.gov).

   ![Graph](chart.png)

   a) Is there a clear pattern? Describe the trend.
   b) Is the association strong?
   c) Is the correlation high? Explain.
   d) Is a linear model appropriate? Explain.

2. **Smoking 2004.** The Centers for Disease Control and Prevention track cigarette smoking in the United States. How has the percentage of people who smoke changed since the danger became clear during the last half of the 20th century? The scatterplot shows percentages of smokers among men 18–24 years of age, as estimated by surveys, from 1965 through 2004 (www.cdc.gov/nchs/).

   ![Graph](chart2.png)

   a) Is there a clear pattern? Describe the trend.
   b) Is the association strong?
   c) Is a linear model appropriate? Explain.

3. **Human Development Index.** The United Nations Development Programme (UNDP) uses the Human Development Index (HDI) in an attempt to summarize in one number the progress in health, education, and economics of a country. In 2006, the HDI was as high as 0.965 for Norway and as low as 0.331 for Niger. The gross domestic product per capita (GDPPC), by contrast, is often used to summarize the overall economic strength of a country. Is the HDI related to the GDPPC? Here is a scatterplot of HDI against GDPPC.

   ![Graph](chart3.png)

   a) Explain why fitting a linear model to these data might be misleading.
   b) If you fit a linear model to the data, what do you think a scatterplot of residuals versus predicted HDI will look like?
   c) There is an outlier (Luxembourg) with a GDPPC of around $70,000. Will setting this point aside improve the model substantially? Explain.

4. **HDI Revisited.** The United Nations Development Programme (UNDP) uses the Human Development Index (HDI) in an attempt to summarize in one number the progress in health, education, and economics of a country. The number of cell phone subscribers per 1000 people is positively associated with economic progress in a country. Can the number of cell phone subscribers be used to predict the HDI? Here is a scatterplot of HDI against cell phone subscribers:

   ![Graph](chart4.png)

   a) Explain why fitting a linear model to these data might be misleading.
   b) If you fit a linear model to the data, what do you think a scatterplot of residuals versus predicted HDI will look like?

5. **Good model?** In justifying his choice of a model, a student wrote, “I know this is the correct model because $R^2 = 99.4%$.”

   a) Is this reasoning correct? Explain.
   b) Does this model allow the student to make accurate predictions? Explain.
6. **Bad model?** A student who has created a linear model is disappointed to find that her $R^2$ value is a very low 13%.
   a) Does this mean that a linear model is not appropriate? Explain.
   b) Does this model allow the student to make accurate predictions? Explain.

7. **Movie Dramas.** Here’s a scatterplot of the production budgets (in millions of dollars) vs. the running time (in minutes) for major release movies in 2005. Dramas are plotted in red and all other genres are plotted in black. A separate least squares regression line has been fitted to each group. For the following questions, just examine the plot:
   a) What are the units for the slopes of these lines?
   b) In what way are dramas and other movies similar with respect to this relationship?
   c) In what way are dramas different from other genres of movies with respect to this relationship?

8. **Movie Ratings.** Does the cost of making a movie depend on its audience? Here’s a scatterplot of the same data we examined in Exercise 7. Movies with an R rating are colored purple, those with a PG-13 rating are red, and those with a PG rating are green. Regression lines have been found for each group. (The black points are G-rated, but there were too few to fit a line reliably.)
   a) Interpret the slope and intercept of the model.
   b) What does the value of $R^2$ say about the model?
   c) Interpret $s_e$ in this context.
   d) Would you use this model to predict the numbers of passengers in 2010? Explain.
   e) There’s a point near the middle of this time span with a large negative residual. Can you explain this outlier?

9. **Oakland passengers.** The scatterplot below shows the number of passengers departing from Oakland (CA) airport month by month since the start of 1997. Time is shown as years since 1990, with fractional years used to represent each month. (Thus, June of 1997 is 7.5—halfway through the 7th year after 1990.)

   a) In what ways is the relationship between run times and budgets similar for the three ratings groups?
   b) How do the costs of R-rated movies differ from those of PG-13 and PG rated movies? Discuss both the slopes and the intercepts.
   c) The film *King Kong*, with a run time of 187 minutes, is the red point sitting at the lower right. If it were omitted from this analysis, how might that change your conclusions about PG-13 movies?

10. **Tracking hurricanes.** In a previous chapter, we saw data on the errors (in nautical miles) made by the National Hurricane Center in predicting the path of hurricanes. The scatterplot on the next page shows the trend in the 24-hour tracking errors since 1970 (www.nhc.noaa.gov).
Dependent variable is: Error
R-squared = 63.0%  s = 42.87

Variable Coefficient
Intercept 292.089
Years-1970 5.22924

a) Interpret the slope and intercept of the model.
b) Interpret in this context.
c) The Center had a stated goal of achieving an average tracking error of 125 nautical miles in 2009. Will they make it? Why do you think so?
d) What if their goal were an average tracking error of 90 nautical miles?
e) What cautions would you state about your conclusion?

11. Unusual points. Each of the four scatterplots that follow shows a cluster of points and one “stray” point. For each, answer these questions:
1) In what way is the point unusual? Does it have high leverage, a large residual, or both?
2) Do you think that point is an influential point?
3) If that point were removed, would the correlation become stronger or weaker? Explain.
4) If that point were removed, would the slope of the regression line increase or decrease? Explain.

12. More unusual points. Each of the following scatterplots shows a cluster of points and one “stray” point. For each, answer these questions:
1) In what way is the point unusual? Does it have high leverage, a large residual, or both?
2) Do you think that point is an influential point?
3) If that point were removed, would the correlation become stronger or weaker? Explain.
4) If that point were removed, would the slope of the regression line increase or decrease? Explain.

13. The extra point. The scatterplot shows five blue data points at the left. Not surprisingly, the correlation for these points is \( r = 0 \). Suppose one additional data point is added at one of the five positions suggested below in green. Match each point (a–e) with the correct new correlation from the list given.
1) -0.90  4) 0.05
2) -0.40  5) 0.75
3) 0.00

14. The extra point revisited. The original five points in Exercise 13 produce a regression line with slope 0. Match each of the green points (a–e) with the slope of the line after that one point is added:
1) -0.45  4) 0.05
2) -0.30  5) 0.85
3) 0.00

15. What’s the cause? Suppose a researcher studying health issues measures blood pressure and the percentage of body fat for several adult males and finds a strong positive association. Describe three different possible cause-and-effect relationships that might be present.

16. What’s the effect? A researcher studying violent behavior in elementary school children asks the children’s parents how much time each child spends playing computer games and has their teachers rate each child on the level of aggressiveness they display while playing with other children. Suppose that the researcher finds a moderately strong positive correlation. Describe three different possible cause-and-effect explanations for this relationship.

17. Reading. To measure progress in reading ability, students at an elementary school take a reading comprehension test every year. Scores are measured in “grade-level” units; that is, a score of 4.2 means that a student is reading at slightly above the expected level for a fourth grader. The school principal prepares a report to parents that includes a graph showing the mean reading score for
each grade. In his comments he points out that the strong positive trend demonstrates the success of the school’s reading program.

A college admissions officer, defending the college’s use of SAT scores in the admissions process, produced the graph below. It shows the mean GPAs for last year’s freshmen, grouped by SAT scores. How strong is the evidence that SAT Score is a good predictor of GPA? What concerns you about the graph, the statistical methodology or the conclusions reached?

19. Heating. After keeping track of his heating expenses for several winters, a homeowner believes he can estimate the monthly cost from the average daily Fahrenheit temperature by using the model \( \text{Cost} = 133 - 2.13 \times \text{Temp} \). Here is the residuals plot for his data:

a) Does this graph indicate that students are making satisfactory progress in reading? Explain.
b) What would you estimate the correlation between Grade and Average Reading Level to be?
c) If, instead of this plot showing average reading levels, the principal had produced a scatterplot of the reading levels of all the individual students, would you expect the correlation to be the same, higher, or lower? Explain.
d) Although the principal did not do a regression analysis, someone as statistically astute as you might do that. (But don't bother.) What value of the slope of that line would you view as demonstrating acceptable progress in reading comprehension? Explain.

d. What heating cost does the model predict for a month that averages ?

e. During one of the months on which the model was based, the temperature did average 10°. What were the actual heating costs for that month?
f. Should the homeowner use this model? Explain.
g. Would this model be more successful if the temperature were expressed in degrees Celsius? Explain.

18. Grades. A college admissions officer, defending the college’s use of SAT scores in the admissions process, produced the graph below. It shows the mean GPAs for last year’s freshmen, grouped by SAT scores. How strong is the evidence that SAT Score is a good predictor of GPA? What concerns you about the graph, the statistical methodology or the conclusions reached?

20. Speed. How does the speed at which you drive affect your fuel economy? To find out, researchers drove a compact car for 200 miles at speeds ranging from 35 to 75 miles per hour. From their data, they created the model \( \text{Fuel Efficiency} = 32 - 0.1 \times \text{Speed} \) and created this residual plot:

a) Interpret the slope of this line in context.
b) Interpret the y-intercept of the line in this context.
c) During months when the temperature stays around freezing, would you expect cost predictions based on this model to be accurate, too low, or too high? Explain.
d) What heating cost does the model predict for a month that averages 10°?
e) During one of the months on which the model was based, the temperature did average 10°. What were the actual heating costs for that month?
f) Should the homeowner use this model? Explain.
g) Would this model be more successful if the temperature were expressed in degrees Celsius? Explain.

a. Does this graph indicate that students are making satisfactory progress in reading? Explain.
b. What would you estimate the correlation between Grade and Average Reading Level to be?
c. If, instead of this plot showing average reading levels, the principal had produced a scatterplot of the reading levels of all the individual students, would you expect the correlation to be the same, higher, or lower? Explain.
d. Although the principal did not do a regression analysis, someone as statistically astute as you might do that. (But don't bother.) What value of the slope of that line would you view as demonstrating acceptable progress in reading comprehension? Explain.

d. What heating cost does the model predict for a month that averages ?

e. During one of the months on which the model was based, the temperature did average 10°. What were the actual heating costs for that month?
f. Should the homeowner use this model? Explain.
g. Would this model be more successful if the temperature were expressed in degrees Celsius? Explain.

18. Grades. A college admissions officer, defending the college’s use of SAT scores in the admissions process, produced the graph below. It shows the mean GPAs for last year’s freshmen, grouped by SAT scores. How strong is the evidence that SAT Score is a good predictor of GPA? What concerns you about the graph, the statistical methodology or the conclusions reached?

20. Speed. How does the speed at which you drive affect your fuel economy? To find out, researchers drove a compact car for 200 miles at speeds ranging from 35 to 75 miles per hour. From their data, they created the model \( \text{Fuel Efficiency} = 32 - 0.1 \times \text{Speed} \) and created this residual plot:

a) Interpret the slope of this line in context.
b) Interpret the y-intercept of the line in this context.
c) During months when the temperature stays around freezing, would you expect cost predictions based on this model to be accurate, too low, or too high? Explain.
d) What heating cost does the model predict for a month that averages 10°?
e) During one of the months on which the model was based, the temperature did average 10°. What were the actual heating costs for that month?
f) Should the homeowner use this model? Explain.
g) Would this model be more successful if the temperature were expressed in degrees Celsius? Explain.

a. Does this graph indicate that students are making satisfactory progress in reading? Explain.
b. What would you estimate the correlation between Grade and Average Reading Level to be?
c. If, instead of this plot showing average reading levels, the principal had produced a scatterplot of the reading levels of all the individual students, would you expect the correlation to be the same, higher, or lower? Explain.
d. Although the principal did not do a regression analysis, someone as statistically astute as you might do that. (But don't bother.) What value of the slope of that line would you view as demonstrating acceptable progress in reading comprehension? Explain.

d. What heating cost does the model predict for a month that averages ?

e. During one of the months on which the model was based, the temperature did average 10°. What were the actual heating costs for that month?
f. Should the homeowner use this model? Explain.
g. Would this model be more successful if the temperature were expressed in degrees Celsius? Explain.

18. Grades. A college admissions officer, defending the college’s use of SAT scores in the admissions process, produced the graph below. It shows the mean GPAs for last year’s freshmen, grouped by SAT scores. How strong is the evidence that SAT Score is a good predictor of GPA? What concerns you about the graph, the statistical methodology or the conclusions reached?

20. Speed. How does the speed at which you drive affect your fuel economy? To find out, researchers drove a compact car for 200 miles at speeds ranging from 35 to 75 miles per hour. From their data, they created the model \( \text{Fuel Efficiency} = 32 - 0.1 \times \text{Speed} \) and created this residual plot:

a) Interpret the slope of this line in context.
b) Interpret the y-intercept of the line in this context.
c) During months when the temperature stays around freezing, would you expect cost predictions based on this model to be accurate, too low, or too high? Explain.
d) What heating cost does the model predict for a month that averages 10°?
e) During one of the months on which the model was based, the temperature did average 10°. What were the actual heating costs for that month?
f) Should the homeowner use this model? Explain.
g) Would this model be more successful if the temperature were expressed in degrees Celsius? Explain.
21. **Interest rates.** Here’s a plot showing the federal rate on 3-month Treasury bills from 1950 to 1980, and a regression model fit to the relationship between the *Rate* (in %) and *Years since 1950* (www.gpoaccess.gov/eop/).

![Graph showing the federal rate on 3-month Treasury bills from 1950 to 1980, with a regression model fit to the relationship between Rate (in %) and Years since 1950.]

- **Dependent variable is:** Rate
- **R-squared:** 77.4 %  \( s = 0.2333 \)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.640282</td>
</tr>
<tr>
<td>Year - 1950</td>
<td>0.247637</td>
</tr>
</tbody>
</table>

- **a)** What is the correlation between *Rate* and *Year*?
- **b)** Interpret the slope and intercept.
- **c)** What does this model predict for the interest rate in the year 2000?
- **d)** Would you expect this prediction to have been accurate? Explain.

22. **Ages of couples 2003.** The graph shows the ages of both men and women at first marriage (www.census.gov).

![Graph showing ages of couples at first marriage with a regression model for the data since 1980.]

- **Dependent variable is:** Age Difference
- **R-squared:** 77.4 %  \( s = 1.239 \)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>35.0617</td>
</tr>
<tr>
<td>Year</td>
<td>-0.016565</td>
</tr>
</tbody>
</table>

- **a)** What is the correlation between *Age Difference* and *Year*?
- **b)** Interpret the slope of this line.
- **c)** Predict the average age difference in 2015.
- **d)** Describe reasons why you might not place much faith in that prediction.

23. **Interest rates revisited.** In Exercise 21 you investigated the federal rate on 3-month Treasury bills between 1950 and 1980. The scatterplot below shows that the trend changed dramatically after 1980.

![Graph showing the federal rate on 3-month Treasury bills since 1980, with a regression model fit to the relationship between Rate (in %) and Years since 1950.]

- **Dependent variable is:** Rate
- **R-squared:** 75.1 %  \( s = 0.2333 \)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>21.0688</td>
</tr>
<tr>
<td>Year - 1950</td>
<td>-0.016565</td>
</tr>
</tbody>
</table>

- **a)** How does this model compare to the one in Exercise 21?
- **b)** What does this model estimate the interest rate to have been in 2000? How does this compare to the rate you predicted in Exercise 21?
- **c)** Do you trust this newer predicted value? Explain.
- **d)** Given these two models, what would you predict the interest rate on 3-month Treasury bills will be in 2020?

24. **Ages of couples, again.** Has the trend of decreasing difference in age at first marriage seen in Exercise 22 gotten stronger recently? The scatterplot and residual plot for the data from 1975 through 2003, along with a regression for just those years, are on the next page.
21. Dependent variable is: Men - Women

Variable Coefficient
Intercept 4.88424
Year 15.4980

a) Why is $R^2$ higher for the first model (in Exercise 22)?
b) Is this linear model appropriate for the post-1975 data? Explain.
c) What does the slope say about marriage ages?
d) Explain why it’s not reasonable to interpret the y-intercept.

25. Gestation. For women, pregnancy lasts about 9 months. In other species of animals, the length of time from conception to birth varies. Is there any evidence that the gestation period is related to the animal’s lifespan? The first scatterplot shows Gestation Period (in days) vs. Life Expectancy (in years) for 18 species of mammals. The highlighted point at the far right represents humans.

Variable Coefficient
Intercept 4.88424
Year 15.4980

a) For these data, $r = 0.54$, not a very strong relationship. Do you think the association would be stronger or weaker if humans were removed? Explain.
b) Is there reasonable justification for removing humans from the data set? Explain.

d) Interpret the slope of the line.
e) Some species of monkeys have a life expectancy of about 20 years. Estimate the expected gestation period of one of these monkeys.

26. Swim the lake 2006. People swam across Lake Ontario 42 times between 1974 and 2006 (www.soloswims.com). We might be interested in whether they are getting any faster or slower. Here are the regression of the crossing Times (minutes) against the Year of the crossing and the residuals plot:

Variable Coefficient
Intercept -9.5172
Year 15.4980

a) What does the mean for this regression?
b) Are the swimmers getting faster or slower? Explain.
c) The outlier seen in the residuals plot is a crossing by Vicki Keith in 1987 in which she swam a round trip, north to south, and then back again. Clearly, this swim doesn’t belong with the others. Would removing it change the model a lot? Explain.

27. Elephants and hippos. We removed humans from the scatterplot in Exercise 25 because our species was an outlier in life expectancy. The resulting scatterplot (next page) shows two points that now may be of concern. The point in the upper right corner of this scatterplot is for elephants, and the other point at the far right is for hippos.
220 CHAPTER 9 Regression Wisdom

a) By removing one of these points, we could make the association appear to be stronger. Which point? Explain.
b) Would the slope of the line increase or decrease?
c) Should we just keep removing animals to increase the strength of the model? Explain.
d) If we remove elephants from the scatterplot, the slope of the regression line becomes 11.6 days per year. Do you think elephants were an influential point? Explain.

31. Life Expectancy 2004. Data from the World Bank for 26 Western Hemisphere countries can be used to examine the association between female Life Expectancy and the average Number of Children women give birth to (http://devdata.worldbank.org/data-query/).

<table>
<thead>
<tr>
<th>Year</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1980</td>
<td>18.4</td>
</tr>
<tr>
<td>1985</td>
<td>22.0</td>
</tr>
<tr>
<td>1990</td>
<td>28.0</td>
</tr>
<tr>
<td>1991</td>
<td>29.5</td>
</tr>
<tr>
<td>1992</td>
<td>30.1</td>
</tr>
<tr>
<td>1993</td>
<td>31.0</td>
</tr>
<tr>
<td>1994</td>
<td>32.6</td>
</tr>
<tr>
<td>1995</td>
<td>32.2</td>
</tr>
<tr>
<td>1996</td>
<td>32.4</td>
</tr>
<tr>
<td>1997</td>
<td>32.4</td>
</tr>
<tr>
<td>1998</td>
<td>32.8</td>
</tr>
</tbody>
</table>

b) How much faith do you place in this prediction? Explain.
c) Do you think your model would produce an accurate prediction about your grandchildren, say, 50 years from now? Explain.

d) If government leaders wanted to increase life expectancy in their country, should they encourage women to have fewer children? Explain.

32. Tour de France 2007. We met the Tour de France data set in Chapter 2 (in Just Checking). One hundred years ago, the fastest rider finished the course at an average speed of about 25.3 kph (around 15.8 mph). In 2005, Lance Armstrong averaged 41.65 kph (25.88 mph) for the fastest average winning speed in history.

a) Make a scatterplot of Avg Speed against Year. Describe the relationship of Avg Speed by Year, being careful to point out any unusual features in the plot.
b) Find the regression equation of Avg Speed on Year.
c) Are the conditions for regression met? Comment.

d) If we remove elephants from the scatterplot, the slope of the regression line becomes 11.6 days per year. Do you think elephants were an influential point? Explain.

indicates, for example, that the average item costing $17.70 in 1926 cost $201.60 in the year 2006.

### Second stage 2007.
Look once more at the data from the Tour de France. In Exercise 32 we looked at the whole history of the race, but now let’s consider just the post–World War II era.

a) Find the regression of Avg Speed by Year only for years from 1947 to the present. Are the conditions for regression met?
b) Interpret the slope.
c) In 1979 Bernard Hinault averaged 39.8 kph, while in 2005 Lance Armstrong averaged 41.65 kph. Which was the more remarkable performance and why?

<table>
<thead>
<tr>
<th>Year</th>
<th>CPI</th>
<th>Year</th>
<th>CPI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1914</td>
<td>10.0</td>
<td>1962</td>
<td>30.2</td>
</tr>
<tr>
<td>1918</td>
<td>15.1</td>
<td>1966</td>
<td>32.4</td>
</tr>
<tr>
<td>1922</td>
<td>16.8</td>
<td>1970</td>
<td>38.8</td>
</tr>
<tr>
<td>1926</td>
<td>17.7</td>
<td>1974</td>
<td>49.3</td>
</tr>
<tr>
<td>1930</td>
<td>16.7</td>
<td>1978</td>
<td>65.2</td>
</tr>
<tr>
<td>1934</td>
<td>13.4</td>
<td>1982</td>
<td>96.5</td>
</tr>
<tr>
<td>1938</td>
<td>14.1</td>
<td>1986</td>
<td>109.6</td>
</tr>
<tr>
<td>1942</td>
<td>16.3</td>
<td>1990</td>
<td>130.7</td>
</tr>
<tr>
<td>1946</td>
<td>19.5</td>
<td>1994</td>
<td>148.2</td>
</tr>
<tr>
<td>1950</td>
<td>24.1</td>
<td>1998</td>
<td>163.0</td>
</tr>
<tr>
<td>1954</td>
<td>26.9</td>
<td>2002</td>
<td>179.9</td>
</tr>
<tr>
<td>1958</td>
<td>28.9</td>
<td>2006</td>
<td>201.6</td>
</tr>
</tbody>
</table>

a) Make a scatterplot showing the trend in consumer prices. Describe what you see.
b) Be an economic forecaster: Project increases in the cost of living over the next decade. Justify decisions you make in creating your model.

**JUST CHECKING**

### Answers

1. Not high leverage, not influential, large residual
2. High leverage, not influential, small residual
3. High leverage, influential, not large residual