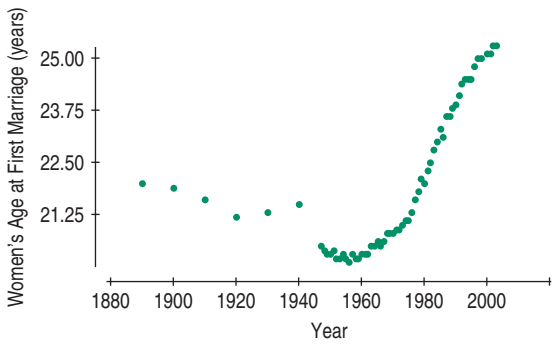
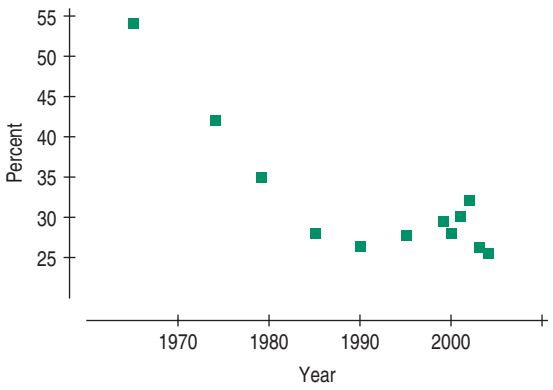
EXERCISES

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).



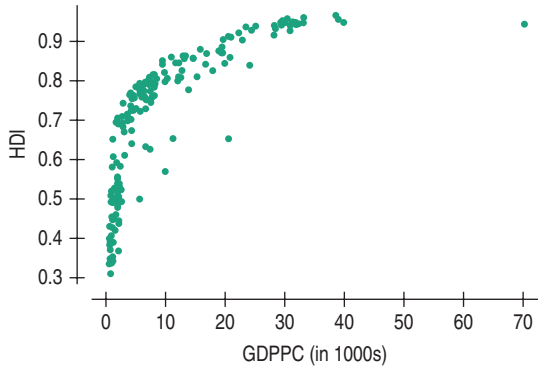
a) Is there a clear pattern? Describe the trend.
b) Is the association strong?
c) Is the correlation high? Explain.
d) Is a linear model appropriate? Explain.

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/).



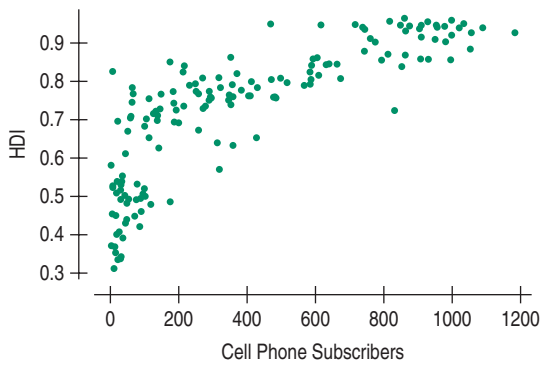
a) Is there a clear pattern? Describe the trend.
b) Is the association strong?
c) Is a linear model appropriate? Explain.

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*.



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:

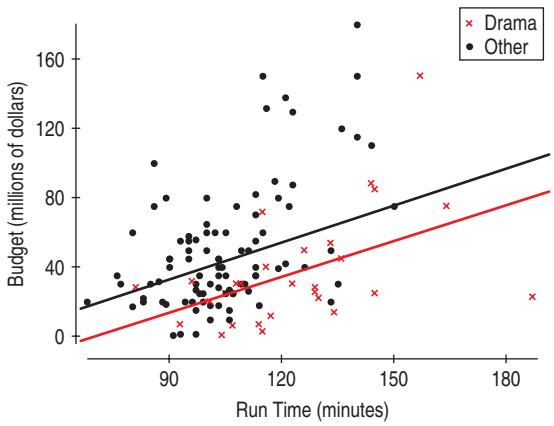


a) Explain why fitting a linear model to these data might be misleading.
b) If you fit a linear model to the data, what do you think a scatterplot of residuals versus predicted *HDI* will look like?

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.

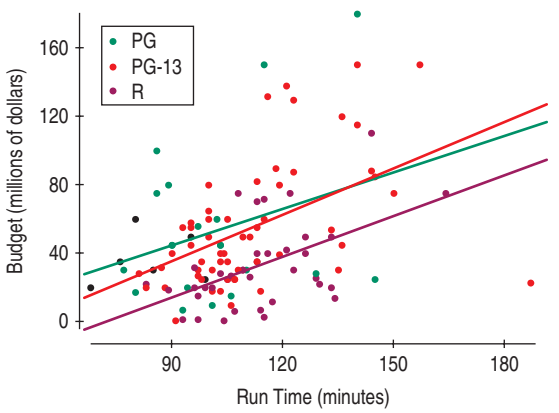
6. **Bad model?** A student who has created a linear model is disappointed to find that her R^2 value is a very low 13%.
   a) Does this mean that a linear model is not appropriate? Explain.
   b) Does this model allow the student to make accurate predictions? Explain.

T 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:



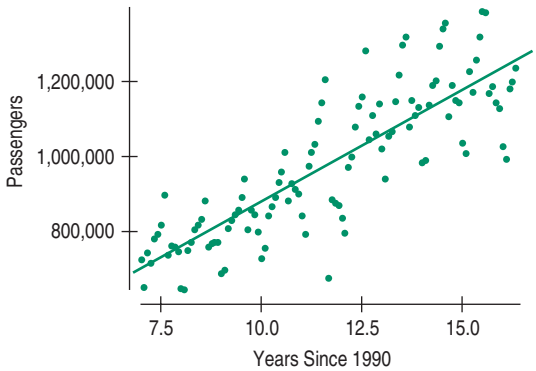
   a) What are the units for the slopes of these lines?
   b) In what way are dramas and other movies similar with respect to this relationship?
   c) In what way are dramas different from other genres of movies with respect to this relationship?

T 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) 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?

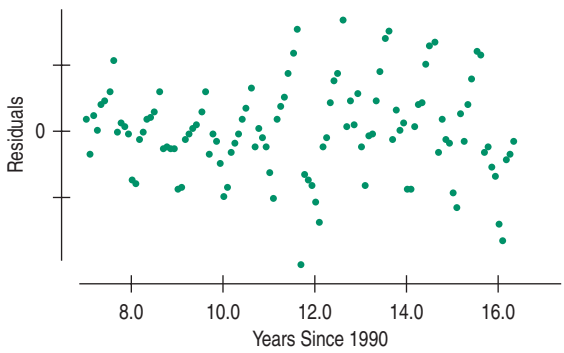
T 9. **Oakland passengers.** The scatterplot below shows the number of passengers departing from Oakland (CA) airport month by month since the start of 1997. Time is shown as years since 1990, with fractional years used to represent each month. (Thus, June of 1997 is 7.5—halfway through the 7th year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



Here's a regression and the residuals plot:

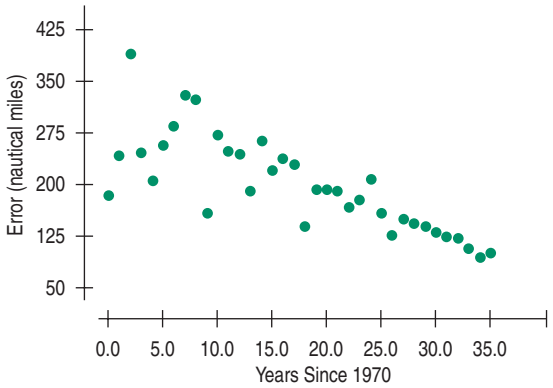
Dependent variable is: Passengers
R-squared = 71.1 % s = 104330

Variable	Coefficient
Constant	282584
Year-1990	59704.4



   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 ($YearsSince1990 = 20$)? Explain.
   e) There's a point near the middle of this time span with a large negative residual. Can you explain this outlier?

T 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).

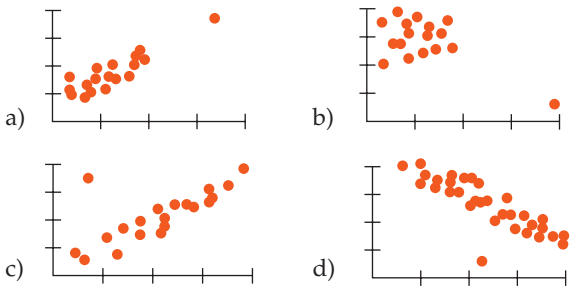


Dependent variable is: Error
R-squared = 63.0 % s = 42.87

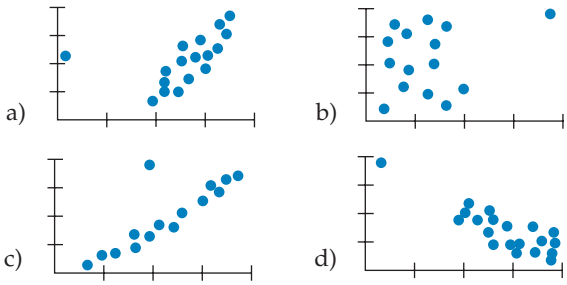
Variable	Coefficient
Intercept	292.089
Years-1970	−5.22924

a) Interpret the slope and intercept of the model.
b) Interpret s_e 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?

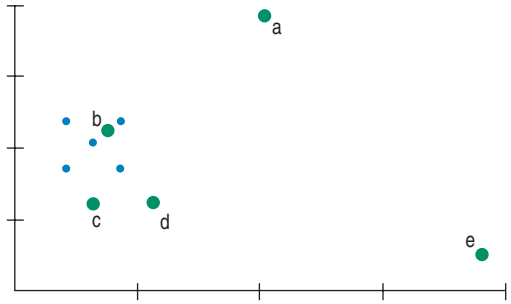
11. Unusual points. Each of the four scatterplots that follow shows a cluster of points and one "stray" point. For each, answer these questions:
1) In what way is the point unusual? Does it have high leverage, a large residual, or both?
2) Do you think that point is an influential point?
3) If that point 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
2) −0.40
3) 0.00
4) 0.05
5) 0.75



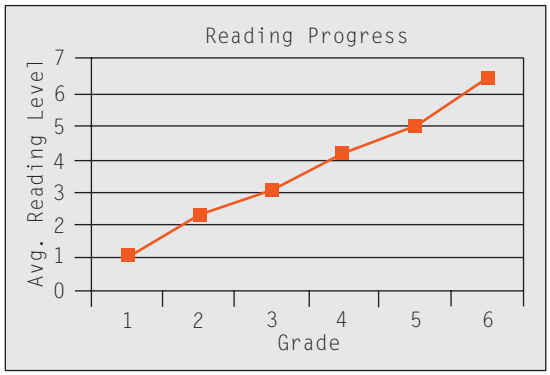
14. The extra point revisited. The original five points in Exercise 13 produce a regression line with slope 0. Match each of the green points (a–e) with the slope of the line after that one point is added:
1) −0.45
2) −0.30
3) 0.00
4) 0.05
5) 0.85

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.

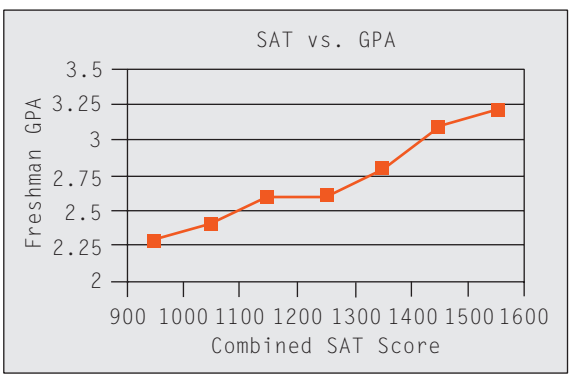
17. Reading. To measure progress in reading ability, students at an elementary school take a reading comprehension test every year. Scores are measured in "grade-level" units; that is, a score of 4.2 means that a student is reading at slightly above the expected level for a fourth grader. The school 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.

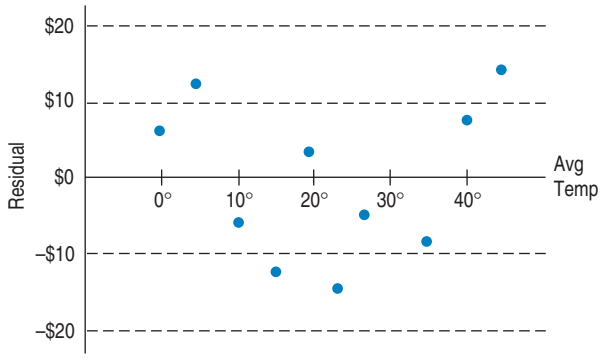


a) Does this graph indicate that students are making satisfactory progress in reading? Explain.
b) What would you estimate the correlation between *Grade* and *Average Reading Level* to be?
c) If, instead of this plot showing average reading levels, the principal had produced a scatterplot of the reading 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.

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?

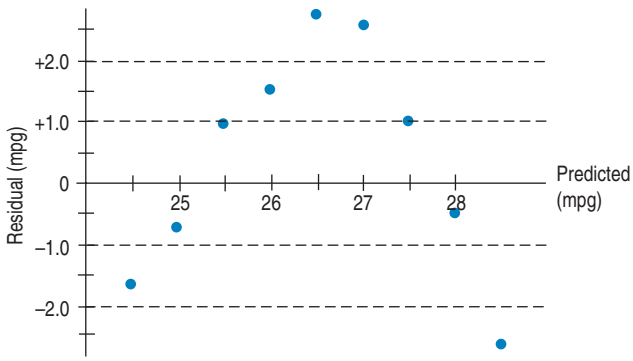


19. Heating. After keeping track of his heating expenses for several winters, a homeowner believes he can estimate the monthly cost from the average daily Fahrenheit temperature by using the model $\widehat{Cost} = 133 - 2.13\,Temp$. Here is the residuals plot for his data:



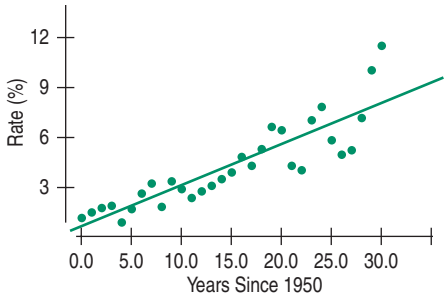
a) Interpret the slope of the line in this context.
b) Interpret the *y*-intercept of the line in this context.
c) During months when the temperature stays around freezing, would you expect cost predictions based on this model to be accurate, too low, or too high? Explain.
d) What heating cost does the 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.

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 $\widehat{Fuel\,Efficiency} = 32 - 0.1\,Speed$ and created this residual plot:



a) Interpret the slope of this line in context.
b) Explain why it's silly to attach any meaning to the *y*-intercept.
c) When this model predicts high *Fuel Efficiency*, what can you say about those predictions?
d) What *Fuel Efficiency* does the model predict when the car is driven at 50 mph?
e) What was the actual *Fuel Efficiency* when the car was driven at 45 mph?
f) Do you think there appears to be a strong association between *Speed* and *Fuel Efficiency*? Explain.
g) Do you think this is the appropriate model for that association? 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/).

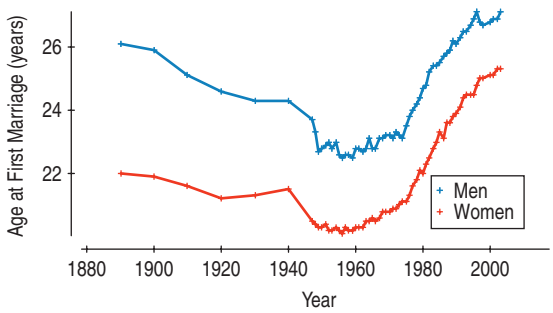


Dependent variable is: Rate
R-squared = 77.4 % s = 1.239

Variable	Coefficient
Intercept	0.640282
Year − 1950	0.247637

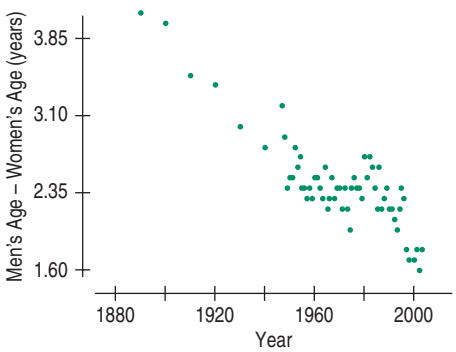
a) What is the correlation between *Rate* and *Year*?
b) Interpret the slope and intercept.
c) What does this model predict for the interest rate in the year 2000?
d) Would you expect this prediction to have been accurate? Explain.

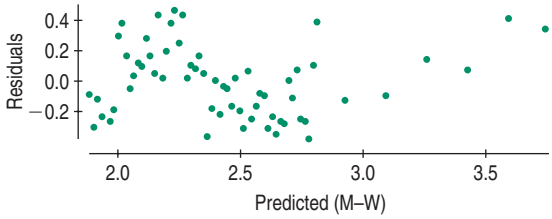
22. Ages of couples 2003. The graph shows the ages of both men and women at first marriage (www.census.gov).



Clearly, the pattern for men is similar to the pattern for women. But are the two lines getting closer together?

Here's a timeplot showing the *difference* in average age (men's age − women's age) at first marriage, the regression analysis, and the associated residuals plot.



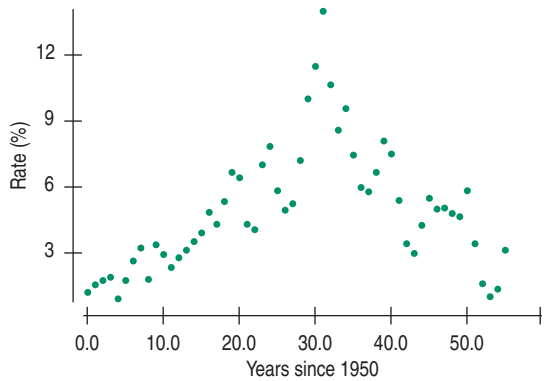


Dependent variable is: Age Dif ference
R-squared = 75.1 % s = 0.2333

Variable	Coefficient
Constant	35.0617
Year	−0.016565

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.



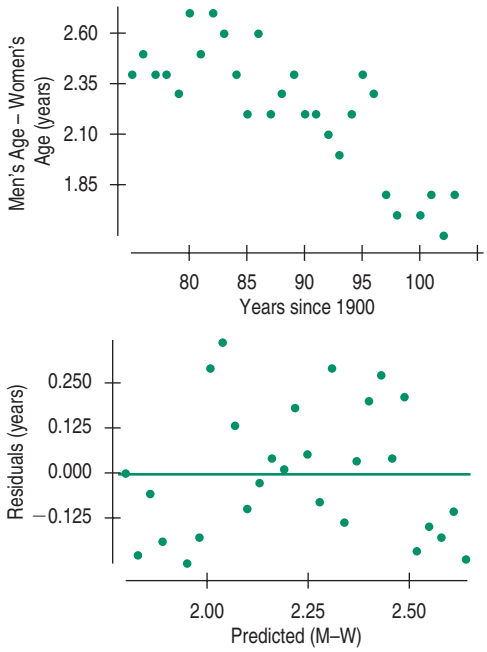
Here's a regression model for the data since 1980.

Dependent variable is: Rate
R-squared = 74.5 % s = 1.630

Variable	Coefficient
Intercept	21.0688
Year − 1950	−0.356578

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.

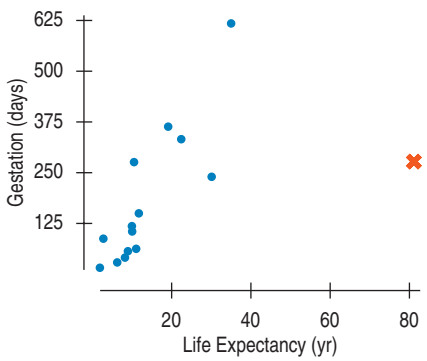


Dependent variable is: Men − Women
R-Squared = 65.6 % s = 0.1869

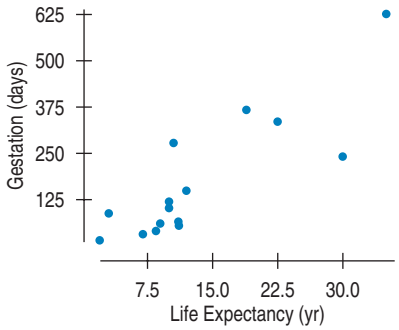
Variable	Coefficient
Intercept	4.88424
Year	−0.029959

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.



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.
c) Here are the scatterplot and regression analysis for the 17 nonhuman species. Comment on the strength of the association.



Dependent variable is: Gestation
R-Squared = 72.2 %

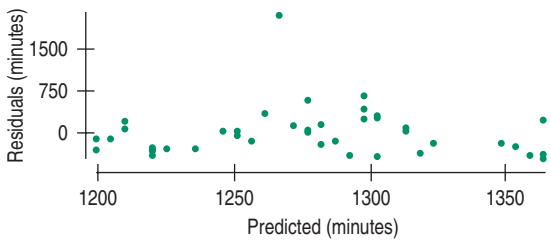
Variable	Coefficient
Constant	−39.5172
Lif Exp	15.4980

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.

T **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:

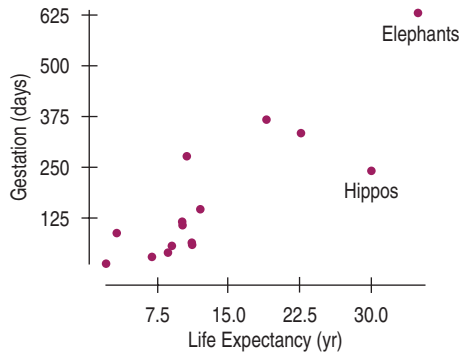
Dependent variable is: T ime
R-Squared = 1.3 % s = 443.8

Variable	Coefficient
Intercept	−8950 .40
Year	5.14171



a) What does the R^2 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.



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.

T **28. Another swim 2006.** In Exercise 26 we saw that Vicki Keith's round-trip swim of Lake Ontario was an obvious outlier among the other one-way times. Here is the new regression after this unusual point is removed:

Dependent variable is: T ime
R-Squared = 4.1 % s = 292.6

Variable	Coefficient
Intercept	−11048.7
Year	6.17091

a) In this new model, the value of s_e is much smaller. Explain what that means in this context.
b) Now would you be willing to say that the Lake Ontario swimmers are getting faster (or slower)?

T **29. Marriage age 2003 revisited.** Suppose you wanted to predict the trend in marriage age for American women into the early part of this century.
a) How could you use the data graphed in Exercise 1 to get a good prediction? Marriage ages in selected years starting in 1900 are listed below. Use all or part of these data to create an appropriate model for predicting the average age at which women will first marry in 2010.

1900–1950 (10-yr intervals): 21.9, 21.6, 21.2, 21.3, 21.5, 20.3
1955–2000 (5-yr intervals): 20.2, 20.2, 20.6, 20.8, 21.1, 22.0, 23.3, 23.9, 24.5, 25.1

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.

30. Unwed births. The National Center for Health Statistics reported the data below, showing the percentage of all births that are to unmarried women for selected years between 1980 and 1998. Create a model that describes this trend. Justify decisions you make about how to best use these data.

Year	1980	1985	1990	1991	1992	1993	1994	1995	1996	1997	1998
%	18.4	22.0	28.0	29.5	30.1	31.0	32.6	32.2	32.4	32.4	32.8

T **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/).

Country	Births/ Woman	Life Exp.	Country	Births/ Woman	Life Exp.
Argentina	2.3	74.6	Guatemala	4.4	67.6
Bahamas	2.3	70.5	Honduras	3.6	68.2
Barbados	1.7	75.4	Jamaica	2.4	70.8
Belize	3.0	71.9	Mexico	2.2	75.1
Bolivia	3.7	64.5	Nicaragua	3.2	70.1
Brazil	2.3	70.9	Panama	2.6	75.1
Canada	1.5	79.8	Paraguay	3.7	71.2
Chile	2.0	78.0	Peru	2.8	70.4
Colombia	2.4	72.6	Puerto Rico	1.9	77.5
Costa Rica	24.9	78.7	United States	2.0	77.4
Dominican Republic	2.8	67.8	Uruguay	2.1	75.2
Ecuador	2.7	74.5	Venezuela	2.7	73.7
El Salvador	2.8	71.1	Virgin Islands	2.2	78.6

a) Create a scatterplot relating these two variables, and describe the association.
b) Are there any countries that do not seem to fit the overall pattern?
c) Find the correlation, and interpret the value of R^2.
d) Find the equation of the regression line.
e) Is the line an appropriate model? Describe what you see in the residuals plot.
f) Interpret the slope and the y-intercept of the line.
g) If government leaders wanted to increase life expectancy in their country, should they encourage women to have fewer children? Explain.

T **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.

T **33. Inflation 2006.** The Consumer Price Index (CPI) tracks the prices of consumer goods in the United States, as shown in the table on the next page (ftp://ftp.bis.gov). It

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

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

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.

T **34. 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?

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