1. **Marriage age.** We've looked at the ages of men at first marriage. How about women? 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.

   ![Scatterplot of Women's Age at First Marriage](image)

   a) Do you think there is a clear pattern? Describe the trend.
   b) Is the association strong?
   c) Is the correlation high? Explain.
   d) Do you think a linear model is appropriate for these data? Explain.

2. **Ages of couples.** The graph shows the ages of both men and women at first marriage.

   ![Graph of Men's and Women's Ages at First Marriage](image)

   Clearly, the pattern for men is very 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.

   ![Timeplot of Age Difference at First Marriage](image)

   a) What is the correlation between age difference and year?
   b) Interpret the slope of this line.
   c) Predict the average age difference in 2010.
   d) Describe reasons why you might not place much faith in that prediction.

3. **Marriage age 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 2005.

<table>
<thead>
<tr>
<th>1900-1940 (10-yr intervals):</th>
<th>1955-1995 (5-yr intervals):</th>
</tr>
</thead>
<tbody>
<tr>
<td>21.9,</td>
<td>20.2,</td>
</tr>
<tr>
<td>21.6,</td>
<td>20.2,</td>
</tr>
<tr>
<td>21.2,</td>
<td>20.6,</td>
</tr>
<tr>
<td>21.3,</td>
<td>20.8,</td>
</tr>
<tr>
<td>21.5,</td>
<td>21.1,</td>
</tr>
<tr>
<td>20.3,</td>
<td>22.0,</td>
</tr>
<tr>
<td></td>
<td>23.3,</td>
</tr>
<tr>
<td></td>
<td>23.9,</td>
</tr>
<tr>
<td></td>
<td>24.5</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.

4. **Ages of couples, again.** Is the trend of decreasing difference in age at first marriage seen in Exercise 2 stronger recently? Here are the scatterplot and residual plot for all the data, and a regression analysis using only the data from 1975 through 1998.
Dependent variable is: Men – Women
Cases selected according to post75
R-squared = 46.3%

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>SE(Coeff)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>49.9021</td>
<td>10.93</td>
</tr>
<tr>
<td>Year</td>
<td>-0.023957</td>
<td>0.0055</td>
</tr>
</tbody>
</table>

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

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 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 for all the individual students, would you expect the correlation to be the same, higher, or lower? Explain.
d) While 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.

8. 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?

9. Heating. After keeping track of his heating expenses for several winters, a homeowner believes he can estimate the monthly cost $C$ from the average daily Fahrenheit temperature using the model

$$\hat{C} = 133 - 2.13 \text{temp}.$$  

The residuals plot for his data is shown.
10. Speed. How does the speed at which you drive impact 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 \( mpg = 32 - 0.1 \text{ mph} \) and created this residuals 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 gas mileage, what can you say about those predictions?
d) What gas mileage does the model predict when the car is driven at 50 mph?
e) What was the actual gas mileage when the car was driven at 45 mph?
f) Do you think there appears to be a strong association between speed and fuel economy? Explain.
g) Do you think this is the appropriate model for that association? Explain.

12. More unusual points. Each of the scatterplots below shows a cluster of points and one "stray" point. For each, answer these questions:
1) Do you think that point is an outlier?
2) Do you think that point is an influential point?
3) If that point were removed from the data, would the correlation become stronger or weaker? Explain.
4) If that point were removed from the data, 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. 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 gestation period is related to the animal's lifespan? The first scatterplot shows gestation period (in days) vs. lifespan (in years) for 18 species of mammals. The highlighted point at the far right represents humans.

![First Scatterplot]

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.

![Second Scatterplot]

Dependent variable is: Gestation
R-squared = 72.9%
Variable Coefficient
Constant -39.5172
LifExp 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.

16. Elephants and hippos. We removed humans from the scatterplot in Exercise 15 because our species was an outlier in life expectancy. The resulting scatterplot 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.

17. Law enforcement. Federal employees with the authority to carry firearms and make arrests are sometimes assaulted, injured, or killed. The table below summarizes the rates of assault and injury (or death) for these employees for 5 years, 1995-1999. Can the assault rate be used to predict injuries or deaths?

<table>
<thead>
<tr>
<th>Agency</th>
<th>Assaults (per 1000)</th>
<th>Killed or Injured (per 1000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bureau of Alcohol, Tobacco, and Firearms (BATF)</td>
<td>31.1</td>
<td>2.2</td>
</tr>
<tr>
<td>Capitol Police</td>
<td>5.0</td>
<td>3.6</td>
</tr>
<tr>
<td>Customs Service</td>
<td>9.7</td>
<td>5.1</td>
</tr>
<tr>
<td>Drug Enforcement Agency (DEA)</td>
<td>17.9</td>
<td>1.1</td>
</tr>
<tr>
<td>Federal Bureau of Investigation (FBI)</td>
<td>3.9</td>
<td>1.2</td>
</tr>
<tr>
<td>Immigration and Naturalization Service (INS)</td>
<td>14.1</td>
<td>2.5</td>
</tr>
<tr>
<td>Internal Revenue Service (IRS)</td>
<td>1.7</td>
<td>0.2</td>
</tr>
<tr>
<td>U.S. Marshal Service</td>
<td>9.7</td>
<td>3.0</td>
</tr>
<tr>
<td>National Park Service</td>
<td>38.7</td>
<td>15.0</td>
</tr>
<tr>
<td>Postal Service</td>
<td>5.7</td>
<td>2.9</td>
</tr>
<tr>
<td>Secret Service</td>
<td>9.7</td>
<td>3.0</td>
</tr>
</tbody>
</table>

18. Smoking. 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 table below shows percentages of smokers among men 18-24 years of age, as estimated by surveys. Create a model to describe the changes in smoking rate. Justify decisions you make in using these data to construct your model.

<table>
<thead>
<tr>
<th>Year</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1965</td>
<td>54.1</td>
</tr>
<tr>
<td>1974</td>
<td>42.1</td>
</tr>
<tr>
<td>1979</td>
<td>35.0</td>
</tr>
<tr>
<td>1983</td>
<td>32.9</td>
</tr>
<tr>
<td>1985</td>
<td>28.0</td>
</tr>
<tr>
<td>1990</td>
<td>26.6</td>
</tr>
</tbody>
</table>
19. Illegitimate 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.

<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>

20. Life expectancy. 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.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentina</td>
<td>2.5</td>
<td>77</td>
<td>Guatemala</td>
<td>4.7</td>
<td>68</td>
</tr>
<tr>
<td>Bahamas</td>
<td>2.2</td>
<td>77</td>
<td>Honduras</td>
<td>4.0</td>
<td>72</td>
</tr>
<tr>
<td>Barbados</td>
<td>1.8</td>
<td>78</td>
<td>Jamaica</td>
<td>2.5</td>
<td>77</td>
</tr>
<tr>
<td>Belize</td>
<td>3.5</td>
<td>74</td>
<td>Mexico</td>
<td>2.8</td>
<td>75</td>
</tr>
<tr>
<td>Bolivia</td>
<td>4.0</td>
<td>64</td>
<td>Nicaragua</td>
<td>3.6</td>
<td>71</td>
</tr>
<tr>
<td>Brazil</td>
<td>2.2</td>
<td>70</td>
<td>Panama</td>
<td>2.5</td>
<td>76</td>
</tr>
<tr>
<td>Canada</td>
<td>1.5</td>
<td>82</td>
<td>Paraguay</td>
<td>4.0</td>
<td>72</td>
</tr>
<tr>
<td>Chile</td>
<td>2.2</td>
<td>79</td>
<td>Peru</td>
<td>3.1</td>
<td>71</td>
</tr>
<tr>
<td>Colombia</td>
<td>2.7</td>
<td>74</td>
<td>Puerto Rico</td>
<td>3.1</td>
<td>71</td>
</tr>
<tr>
<td>Costa Rica</td>
<td>25.0</td>
<td>79</td>
<td>U. S.</td>
<td>2.1</td>
<td>80</td>
</tr>
<tr>
<td>Dom. Rep.</td>
<td>2.8</td>
<td>73</td>
<td>Uruguay</td>
<td>2.3</td>
<td>78</td>
</tr>
<tr>
<td>Ecuador</td>
<td>3.1</td>
<td>71</td>
<td>Venezuela</td>
<td>2.9</td>
<td>76</td>
</tr>
<tr>
<td>El Salvador</td>
<td>3.2</td>
<td>72</td>
<td>Virgin Is.</td>
<td>2.4</td>
<td>79</td>
</tr>
</tbody>
</table>

21. Inflation. The Consumer Price Index (CPI) tracks the prices of consumer goods in the United States, as shown in the table. It indicates, for example, that the average item costing $17.90 in 1928 cost $168.80 in the year 2000.

<table>
<thead>
<tr>
<th>Year</th>
<th>CPI</th>
<th>Year</th>
<th>CPI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1916</td>
<td>10.4</td>
<td>1960</td>
<td>29.3</td>
</tr>
<tr>
<td>1920</td>
<td>16.5</td>
<td>1964</td>
<td>30.9</td>
</tr>
<tr>
<td>1924</td>
<td>17.3</td>
<td>1968</td>
<td>34.1</td>
</tr>
<tr>
<td>1928</td>
<td>17.9</td>
<td>1972</td>
<td>41.1</td>
</tr>
<tr>
<td>1932</td>
<td>12.9</td>
<td>1976</td>
<td>55.6</td>
</tr>
<tr>
<td>1936</td>
<td>14.2</td>
<td>1980</td>
<td>77.8</td>
</tr>
<tr>
<td>1940</td>
<td>13.9</td>
<td>1984</td>
<td>101.9</td>
</tr>
<tr>
<td>1944</td>
<td>17.4</td>
<td>1988</td>
<td>115.7</td>
</tr>
<tr>
<td>1948</td>
<td>23.7</td>
<td>1992</td>
<td>138.1</td>
</tr>
<tr>
<td>1952</td>
<td>26.5</td>
<td>1996</td>
<td>154.4</td>
</tr>
<tr>
<td>1956</td>
<td>26.8</td>
<td>2000</td>
<td>168.8</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.
c) Sometimes re-expressing the data can reveal interesting information concealed there. Create a new variable containing the logarithm of the CPI. Plot log(CPI) against year. What historical events had a clear impact on the cost of living? (The next chapter explores more advantages that may arise by re-expressing data.)
Answers

1. 
   a) The trend appears to be somewhat linear up to about 1940, but from 1940 to about 1970 the trend appears to be nonlinear. From 1975 or so, to present, the trend appears to be linear.
   
   b) Relatively strong for certain time periods.
   
   c) No, as a whole, the graph is clearly nonlinear. Within certain time periods the correlation is high (1975 to the present).
   
   d) Overall, no. You could fit a linear model to the time period 1975 to 1995, but why? You don't need to interpolate, since every year is reported and extrapolation seems very dangerous.

2. 
   a) \(-0.846\)
   
   b) For each increase of 10 years, the difference in average age at first marriage is decreasing by about 0.16 years.
   
   c) About 1.81 years.
   
   d) The latest data point is before the year 2000. Extrapolating for 2010 is risky because it depends on the assumption that the trend will continue in the same manner.

3. 
   a) Answers may vary. Using the data for 1955-1995, the scatterplot is relatively linear with some curvature. The residual plot shows a definite trend, indicating that the data are not linear. If you used the line, for 2005, the predicted age is 25.3 years.
   
   b) Not much since, the data are not truly linear, and 2005 is 10 years from the last data point (extrapolating is risky).
   
   c) No, that extrapolation of more than 50 years would be absurd. There's no reason to believe the trend from 1955 to 1995 will continue.

4. 
   a) The data from the late 1800s to 1950 are high leverage points. Since they follow generally the same linear trend as the 1975-1998 data, those data points increase the correlation and the \(R^2\) value.
   
   b) Yes, although there are a couple of unusual residual values in the bottom left corner of the residual graph.
   
   c) For each increase of 10 years, the average age difference drops about 0.24 years.
   
   d) These data begin with \(X = 1975\). The y-intercept is the predicted difference in age at the year AC = 0, quite an extrapolation.

5. 
   a) No. We need to see the scatterplot first to see if the conditions are satisfied.
   
   b) No, the linear model might not fit the data at all.

6. 
   a) No, it just means that the model doesn't explain more than 13% of the variation of the response variable.
   
   b) She should look at \(s\) to see accuracy of predictions. \(R^2\) is not a direct measure of accuracy.

7. 
   a) The graph shows that on average, students progress at about one reading level per year. This graph shows averages for each grade. The linear trend has been increased by using averages.
   
   b) Very close to 1.
   
   c) The individual data points would show much more scatter, and the correlation would be lower.
   
   d) A slope of 1 would indicate that for each 1-year grade level increase, the average reading level is increasing by 1 year.

8. 
   a) As the temperature increases by 1°F, the cost decreases by $2.13. So warmer temperatures indicate lower costs.
   
   b) For a temperature of 0°F, the cost is $133.
   
   c) Too high; the residuals (observed - predicted) around 32 degrees are negative, showing that the model overestimates the costs.
   
   d) $111.70
   
   e) About $105.70
   
   f) No, the residuals show a definite curved pattern. The data are probably not linear.
   
   g) No, there would be no difference. The relationship does not depend on the units.

9. 
   a) The slope indicates that as the speed increases by 1 mph, the fuel economy decreases by 0.1 mpg; for every 10 mph of increase in speed, the fuel economy decreases by 1 mpg.
   
   b) The intercept is the mileage at 0 mph, which doesn't make any sense.
   
   c) The residuals are negative, so the model is overestimating the mpg.
   
   d) 27 mpg.
   
   e) 28.5 mpg.
   
   f) The association is strong, but not linear.
   
   g) No, the residuals indicate a clear nonlinear relationship.

10. 
    a) 1) Yes, an outlier.
        
        2) No, not influential for the slope.
        
        3) Correlation would decrease because outlier has large \(z_x\) and \(z_y\), increasing correlation.
        
        4) Slope wouldn't change because the outlier is in line with other points.
    
    b) 1) Yes, an outlier.
        
        2) Yes, influential.
        
        3) Correlation would weaken and become less negative.
        
        4) Slope would increase toward 0, since outlier makes it negative.

    c) 1) Yes, outlier.
        
        2) Yes, somewhat influential.
        
        3) Correlation would increase, since scatter would decrease.
4) Slope would increase slightly.

12. a) 1) Yes, an outlier.
2) Yes, influential.
3) Correlation would become stronger and increase because scatter would decrease.
4) Slope would become stronger and increase because outlier pulls the line toward 0.
b) 1) Yes, an outlier.
2) Yes, influential.
3) Correlation would become weaker because outlier has high $z_x$ and $z_y$ values, increasing correlations.
4) Slope would decrease because outlier pulls the line from nearly flat to a positive slope.
c) 1) Yes, an outlier.
2) No, not influential.
3) Correlation would become stronger because scatter would decrease.
4) About the same; slope may decrease slightly.
d) 1) Yes, an outlier.
2) No, not influential.
3) Correlation would become weaker and become less negative because outlier has large negative $z_x$ $z_y$ value.
4) Slope would stay about the same because outlier is consistent with slope determined by other points.

13. 1) e 2) d 3) c 4) b 5) a
14. 1) d 2) e 3) c 4) b 5) a

15. a) Stronger. Both slope and correlation would increase.
b) Restricting the study to nonhuman animals would justify it.
c) Moderately strong.
d) For every year increase in life expectancy, the gestation period increases by about 15.5 days, on average.
e) About 270.5 days.

16. a) Removing hippos would make the association stronger, since hippos are more of a departure from the pattern.
b) Increase.
c) No, there must be a good reason for removing data points.
d) Yes, removing it lowered the slope from 15.5 to 11.6 days per year.

17. No. There is one high leverage outlier, National Park Service. With that point, R-squared is 46.2%; without that point, R-squared is 0.


19. Answers may vary. There seem to be two time periods. From 1980 to 1993 or so, there was a consistent increase of about 1% a year. A regression of those years shows $\text{Predicted illegitimate percentage} = -1988.94 + 1.014$ year with an $R^2$ of 98.9%. The years from 1994 to 1998 are nearly flat. A model of those years shows: $\text{Predicted illegitimate percentage} = -87.28 + 0.060$ year with an $R^2$ of 17.3%. How to use the data depends on the purpose of the study.

20. a) Except for the outlier, Costa Rica, the data appear to have a linear form in a negative direction.
b) The outlier is Costa Rica, whose data appears to be wrong, with 25 births per woman. That’s impossible.
c) With Costa Rica, $r = 0.081$, $\text{R-squared} = 0.7\%$, indicating that 0.7% of the variation in life expectancy is explained by the variation in births per woman. Without Costa Rica, $r = -0.815$, $\text{R-squared} = 66.4\%$, indicating that 66.4% of the variation in life expectancy is explained by the variation in births/woman.
d) With Costa Rica, $\hat{Y} = 74.2 + 0.08X$; without Costa Rica, $\hat{Y} = 86.8 - 4.4X$.
e) The model with Costa Rica is not appropriate. The residual plot shows a distinct outlier, which is Costa Rica. Removing Costa Rica gives better residuals plots, suggesting that the linear equation is more appropriate.
f) With Costa Rica, the slope is near 0, suggesting that the linear relation is not very useful. The $y$-intercept suggests that with no births, the life expectancy is about 74 years. Without Costa Rica, the slope is $-4.36$, indicating that an average increase of one child per woman predicts a lower life expectancy of 4.36 years, on average. The $y$-intercept indicates that a country with no birth rate would have a life expectancy of 86.8 years. This is extrapolation.
g) No, while there is an association, there is no reason to expect causality. Lurking variables may be involved.

21. a) The scatterplot is clearly nonlinear; however, the last few years, say from 1972 on, do appear to be linear.
   b) Using the data from 1972 to 2000 gives $r = 0.998$ and Predicted CPI = $-9247.62 + 4.710 X$. Predicted CPI in 2010 = 219.72.
   c) The stock market crash and World War II are clearly visible in their impact on CPI.