**26.** a) $\widehat{Lift} = 573.257 - 15491.5(1/Class)$
  b) This model predicts 391.0 kg for Asanidze.
  c) He lifted 8.5 kg less than the model predicted.
27. The relationship cannot be made straight by the methods of this chapter.
28. $\dfrac{\widehat{1}}{aveWeight} = 1.603 + 0.00112(Oranges)$
29. a) $\widehat{\sqrt{Left}} = 8.465 - 0.06926(Age)$ b) 52.10 years
  c) No; the residuals plot still shows a pattern.
30. a) The linear model:
  $\widehat{Diameter} = 1.97 + 0.463 \times Age$ gives $R^2 = 98\%$, but there is a clear pattern in the residuals. The pattern oscillates, so it cannot be made straight with the methods of this chapter.
 b) Less strong, since averages are less variable than individuals.

PART II REVIEW

1. % over 50, 0.69.
  % under 20, -0.71.
  % Graduating on time, -0.51.
  % Full-time Faculty, 0.09
2. a) If no meals are eaten together, the *GPA* is 2.73.
  b) For each increase of 1 meal, the *GPA* increases by 0.11.
  c) 3.15
  d) The predicted GPA was higher than the observed value.
  e) There is no indication that there is a causal relation. There may be lurking variables.
3. a) There does not appear to be a linear relationship.
  b) Nothing, there is no reason to believe that the results for the Finger Lakes region are representative of the vineyards of the world.
  c) $\widehat{CasePrice} = 92.77 + 0.567 \times Years$.
  d) Only 2.7 % of the variation in case price is accounted for by the ages of vineyards. Most of that is due to two outliers. We are better off using the mean price rather than this model.
4. a) No, no apparent relationship.
  b) There is one very large leverage point.
  c) Weaker. The cross point is extreme in the x direction and low on the y-axis, causing the correlation to weaken (become closer to 0).
  d) The slope of the line would increase.
5. a) $\widehat{TwinBirths} = -5119590 + 2618.25 \times Year$.
  b) Each year, the number of twins born in a year increases, on average, by approximately 2618.25.
  c) 143,092.5 births. The scatterplot appears to be somewhat linear, but there is some curvature in the pattern. There is no reason to believe that the increase will continue to be linear 5 years beyond the data.
  d) The residuals plot shows a definite curved pattern, so the relation is not linear.
6. a) 0.914
  b) $\widehat{DJIA} = -2294.01 + 341.095(Year-1970)$.
  c) Each year, the *DJIA* increased by about 341 points, on average.
  d) The relationship does not appear to be linear, as the residuals have a definite pattern.
7. a) -0.520
  b) Negative, not strong, somewhat linear, but with more variation as pH increases.
  c) The BCI would also be average.
  d) The predicted *BCI* will be 1.56 SDs of *BCI* below the mean *BCI*.
8. a) Number of motorboat registrations.
  b) The association is fairly strong, linear, and has a positive direction.
  c) 0.946
  d) 89.5% of the variation in manatees killed is explained by the variation in powerboat registrations.
  e) No, there is no proof of causality.

9. a) $\widehat{Manatee\,Deaths} = -45.67 \times 0.1315\ Powerboat\,Registrations$ (in 1000s).
  b) According to the model, for each increase of 10,000 motorboat registrations, the number of manatees killed increases by approximately 1.315.
  c) If there were 0 motorboat registrations, the number of manatee deaths would be -45.67. This is obviously a silly extrapolation.
  d) The predicted number is 82.41 deaths. The actual number of deaths was 79. The residual is $79 - 82.41 = -3.41$. The model overestimated the number of deaths by 3.41.
  e) Negative residuals would suggest that the actual number of deaths was lower than the predicted number.
  f) Over time, the number of motorboat registrations has increased and the number of manatee kills has increased. The trend may continue. Extrapolation is risky, however, because the government may enact legislation to protect the manatee.
10. a) 73 points   b) 7 points   c) $r = 0.75$ d) 100 points
  e) The regression equation is designed to predict final exam scores based on midterm exam scores. You would need to find the regression equation to predict midterm scores based on final exam scores.
  f) -85 points
  g) Increase. The point is unusual and has a high negative residual that would decrease the correlation; removing it would increase the correlation and the R^2 value.
  h) Slope will increase.
11. a) -0.984 b) 96.9%   c) 32.95 mph   d) 1.66 mph
  e) Slope will increase.
  f) Correlation will weaken (become less negative).
  g) Correlation is the same, regardless of units.
12. a) 0.473   b) A weak linear association in a positive direction.
  c) The actual score was higher than the predicted value for Monday.
  d) The predicted Monday score would be (0.473) SDs, or 2.37 points below the mean for Monday, or 34.9 points.
  e) Predicted Monday Score $= 14.6 + 0.54 \times$ Friday Score.
  f) 36.0 points
13. a) Weight (but unable to verify linearity).
  b) As weight increases, mileage decreases.
  c) *Weight* accounts for 81.5% of the variation in *Fuel Efficiency*.
14. a) *Displacement* and *Weight* (but unable to verify linearity).
  b) No. Large engines (higher displacement) weigh more, so the engine itself may influence the weight somewhat. More likely, heavy cars are equipped with larger engines because their added weight needs a larger engine (higher displacement) to drive.
  c) The change in units will not affect the correlation.
  d) The predicted fuel economy will be 0.786 SDs below the mean fuel economy.
15. a) $\widehat{Horsepower} = 3.50 + 34.314 \times Weight$.
  b) Thousands. For the equation to have predicted values between 0 and 160, the *X* values would have to be in thousands of pounds.
  c) Yes. The residual plot does not show any pattern.
  d) 115.0 horsepower.
16. *Sex* and *Colorblindness* are both categorical variables, not quantitative. Correlation is meaningless for them, but we can say that the variables are associated.
17. a) The scatterplot shows a fairly strong linear relation in a positive direction. There seem to be two distinct clusters of data.
  b) $\widehat{Interval} = 33.967 \div 10.358 \times Duration$.
  c) The time between eruptions increases by about 10.4 minutes per minute of *Duration* on average.
  d) Since 77% of the variation in *Interval* is accounted for by *Duration* and the error standard deviation is 6.16 minutes, prediction will be relatively accurate.
  e) 75.4 minutes.
  f) A residual is the observed value minus the predicted value, the residual $= 79 - 75.4 = 3.6$ minutes, indicating the model underestimated the interval in this case.

18. a) Yes, the R^2 values indicate that 97.2% of the Indian crocodile length and 98% of the Australian crocodile length are explained by the head size.
b) The slopes of the regression equations are similar, as are the R^2 values.
c) The two models have different y-intercepts. It means that the Indian crocodile is smaller.
d) Predicted body length for the Indian crocodile is 389.4 cm but is 458.2 cm for the Australian croc. The skeleton was probably from an Indian crocodile.

19. a) $r = 0.888$. Although r is high, you must look at the scatterplot and verify that the relation is linear in form.
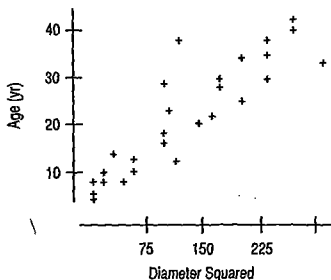b)
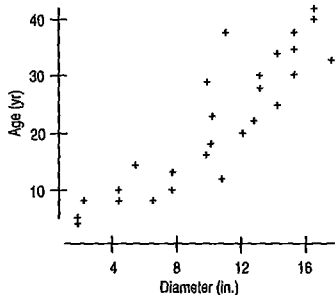
The association between diameter and age appears to be strong, somewhat linear, and positive.
c) $\widehat{Age} = -0.97 + 2.21 \times Diameter$.
d)
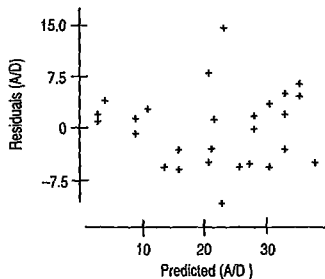
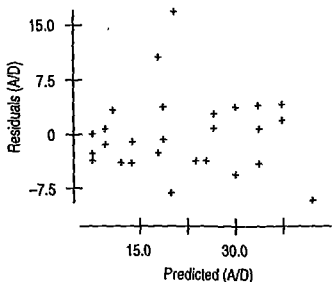The residuals show a curved pattern (and two outliers).
e) The residuals for five of the seven largest trees (15 in. or larger) are positive, indicating that the predicted values underestimate the age.

20. a) Yes.

b) $\widehat{Age} = 7.24 + 0.11 \times Diameter^2$.

c) The residuals appear to be more randomly scattered and have less of a pattern. But there are still some points with large residuals.
d) 43.9 years old.

21. Most houses have areas between 1000 and 5000 square feet. Increasing 1000 square feet would result in either $1000(.008) = 8$ thousand dollars, $1000(.08) = 80$ thousand dollars, $1000(.8) = 800$ thousand dollars, or $1000(8) = 8000$ thousand dollars. Only $80,000 is reasonable, so the slope must be 0.08.

22.
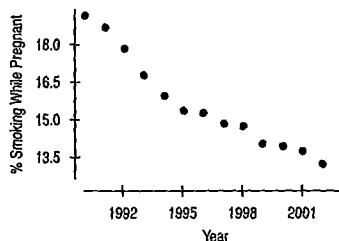
a) The association is very strong, straight, and negative. However, there is some curvature to the scatterplot.
b) -0.972
c) Averaging increases the strength of the correlation, making it more negative.
d) $\widehat{Percentage} = 946.92 - 0.4662 \times Year$. The model suggests that each year the average percent of women smoking while pregnant has decreased by about 0.47%. (Up-and-down curvature can't be straightened.)

23. a) The model predicts % smoking from year, not the other way around.
b) $\widehat{Year} = 2027.91 - 202.74 \times \% \, Smoking$.
c) The smallest % smoking given is 12.7, and an extrapolation to $x = 0$ is probably too far from the given data. The prediction is not very reliable in spite of the strong correlation.

24. a) There is a very weak linear relationship between quality of service and tip size.
b) $R^2 = 1.21\%$, indicating that the variation in quality of service accounts for only about 1% of the variation in tip percentages.

25. The relation shows a negative direction, with a somewhat linear form, but perhaps with some slight curvature. There are several model outliers.

26. a) Latitude, since the correlation -0.848 is stronger (but unable to verify linearity).
b) The same.
c) The same. Changes in units do not change the shape of the association nor correlation of the data.
d) The predicted average January temperature would be 0.738 SDs below the mean average January temperature.

27. a) 71.9%
b) As latitude increases, the January temperature decreases.
c) $\widehat{January \, Temperature} = 108.80 - 2.111 \times Latitude$.
d) As the latitude increases by 1 degree, the average January temperature drops by about 2.11 degrees, on average.
e) The y-intercept would indicate that the average January temperature is 108.8 when the latitude is 0. However, this is extrapolation and may not be meaningful.
f) 24.4 degrees.
g) The equation underestimates the average January temperature.

28. No. First, the R^2 indicates that only 4.6% of the variation in *Depression* is accounted for by *Internet Use*. That is a very low percentage. Even if it were higher, there may be lurking variables that explain the increase in both.

29. a) The scatterplot shows a strong, linear, positive association.
b) There is an association, but it is likely that training and technique have increased over time and affected both jump performances.

c) Neither; the change in units does not affect the correlation.

d) The long-jumper would jump 0.925 SDs above the mean long jump, on average.

30. a) $\widehat{High\ Jump} = -20.76 + 0.331 \times Long\ Jump$.

b) As the long jump increases by 1 inch, the high jump increases by 0.331 inches, on average, according to the model.

c) 95.09 inches.

d) This equation is designed to predict high jump based on long jump. To predict the long jump, you would need the equation predicting long jump based on high jump.

e) $\widehat{Long\ Jump} = 99.26 + 2.5853 \times High\ Jump$. (99.055 + 2.588 High Jump with technology)

31. a) No relation; the correlation would probably be close to 0.

b) The relation would have a positive direction and the correlation would be strong, assuming that students were studying French in each grade level. Otherwise, no correlation.

c) No relation; correlation close to 0.

d) The relation would have a positive direction and the correlation would be strong, since vocabulary would increase with each grade level.

32. a) There is a strong, fairly linear, positive trend for the data; as the year of birth increases, the preterm birth rate increases.

b) The highest preterm birth rate is for mothers receiving adequate prenatal care utilization, and the lowest preterm birth rate is for mothers receiving inadequate prenatal care utilization. The slope is about the same for these relations. Mothers receiving intensive prenatal care utilization had more preterm births than mothers receiving inadequate prenatal care and fewer preterm births than mothers receiving adequate prenatal care; however, their rate of increase is higher.

c) No, there is undoubtedly an underlying variable explaining these differences. Perhaps the level of prenatal care was determined by the complications the mothers experienced during the pregnancy.

33. $\widehat{Calories} = 560.7 - 3.08 \times Time$.

Each minute extra at the table results in 3.08 fewer calories being consumed, on average. Perhaps the hungry children eat fast and eat more.

34. a) $\widehat{Gallons} = 12451.2 - 6.2 \times Year$.

b) The model is designed to predict gallons based on the year. That question asks for a prediction of the year based on the gallons.

c) $\widehat{Year} = 2008.2 - 0.16 \times Gallons$ (in 1000s). Prediction is 2008.

d) The prediction in the wrong direction is close because the relationship is so strong. The line that minimizes the y-residuals also does a pretty good job at minimizing the x-residuals.

35. There seems to be a strong, positive, linear relationship with one high-leverage point (Northern Ireland) that makes the overall R^2 quite low. Without that point, the R^2 increases to 61.5%. Of course, these data are averaged across thousands of households, so the correlation appears to be higher than it would be for individuals. Any conclusions about individuals would be suspect.

36. a) $\widehat{Weight} = -1971.26|1.091 \times Year$.

b) The scatterplot appears to have a linear trend in a positive direction; however, since the weights are averages, the linear trend would be stronger than for individual players.

c) 214.75 pounds. Possibly, but it is a 10-year extrapolation from a range of only 20 years.

d) 323.88 pounds. No, this is an extrapolation of 110 years, and the average weight seems absurd.

e) 1306.11 pounds. Absolutely ridiculous to have an average weight of over 1000 pounds. The prediction is an extrapolation of 1010 years.

37. a) 3.842   b) 501.187   c) 4.0

38. a) $\widehat{Weight} = -1121.66 + 0.673 \times Year$.

b) 2032.

c) No. That seems to be too much extrapolation for the model.

39. a) 30,818 pounds.

b) 1302 pounds.

c) 31,187.6 pounds.

d) I would be concerned about using this relation if we needed accuracy closer than 1000 pounds or so, as the residuals are more than ±1000 pounds.

e) Negative residuals will be more of a problem, as the predicted weight would overestimate the weight of the truck; trucking companies might be inclined to take the ticket to court.

40. a) Symmetric distributions of x and y may help ensure that the residuals around a line through the data will be more symmetric and normally distributed. The re-expressed data will not have extreme outliers as well.

b) Yes. The scatterplot shows a linear trend in a positive direction. It has some moderate outliers in the x direction. The residuals plot shows some moderate residuals in the negative direction and some curvature, but the model using re-expressed data is surely better than the model using the original data.

c) $\widehat{\log(Profit)} = -0.11 + 0.648 \times \log(Sales)$.   d) $124.43 million.

41. The original data are nonlinear, with a significant curvature. Using reciprocal square root of diameter gave a scatterplot that is nearly linear:

$$\widehat{1/\sqrt{Drain\ Time}} = 0.0024 + 0.219\ Diameter.$$

42. The scatterplot of Cost per Chip vs. Chips Produced has a significant curvature. Taking the log of Chips Produced and the log of Cost per Chip provided a nearly linear scatterplot.

$$\widehat{\log\ Cost\ per\ Chip} = 2.67 - 0.502 \times \log\ Chips\ Produced.$$

CHAPTER 11

1. Yes. You cannot predict the outcome beforehand.

2. The outcomes cannot be predicted beforehand. Each outcome, numbers 00 through 36, should be equally likely.

3. A machine pops up numbered balls. If it were truly random, the outcome could not be predicted and the outcomes would be equally likely. It is random only if the balls generate numbers in equal frequencies.

4. Answers may vary. **Rolling one die or two dice:** If the dice are fair, then each outcome, 1 through 6, should be equally likely. **Spinning a spinner:** Each outcome should be equally likely, but the spinner might be more likely to land on one outcome than another due to friction or design.
Shuffling cards and dealing a hand: If the cards are shuffled adequately (7 times for riffle shuffling), the cards will be approximately equally likely.

5. Use two-digit numbers 00–99; let 00–02 = defect, 03–99 = no defect

6. Use random digits 0–9; let 0 = colorblind, 1–9 = no defect

7. a) 45, 10   b) 17, 22

8. 43, 9, 50, 13, 27

9. If the lottery is random, it doesn't matter which number you play; all are equally likely to win.

10. If the lottery is random, it doesn't matter which number you play; all are equally likely to win.

11. a) The outcomes are not equally likely; for example, tossing 5 heads does not have the same probability as tossing 0 or 9 heads, but the simulation assumes they are equally likely.

b) The even-odd assignment assumes that the player is equally likely to score or miss the shot. In reality, the likelihood of making the shot depends on the player's skill.

c) The likelihood for the first ace in the hand is not the same for the second or third or fourth. But with this simulation the likelihood is the same for each. (And it allows you 5 aces, which could get you in trouble in a real poker game.)

12. a) The numbers would represent the sums, but the sums are not all equally likely, as this simulation assumes.

b) The numbers of boys in a family of 5 children are not all equally likely; for example, having a total of 5 boys is less likely.