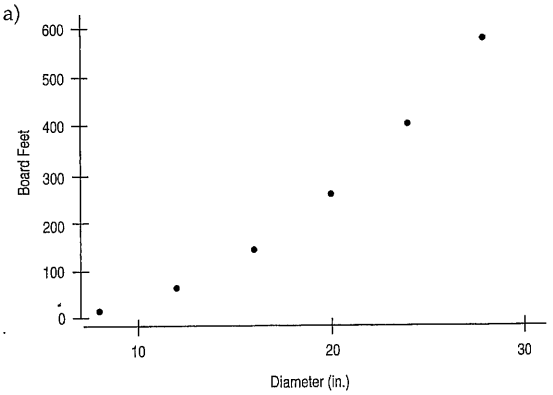
23. a)



$\widehat{\sqrt{bdft}} = -4 + diam$

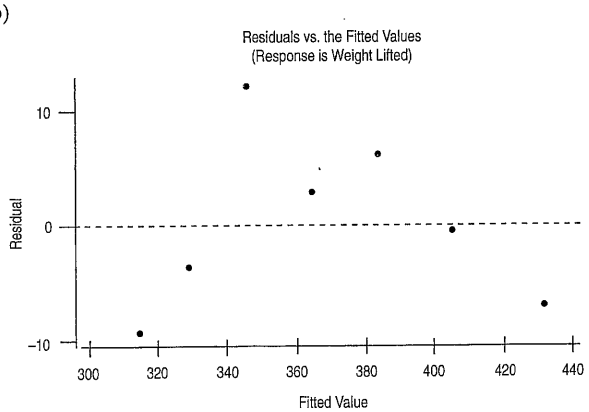
The model is exact.

b) 36 board feet.

c) 1024 board feet.

24. a) $\widehat{lift} = 179.92 + 2.4(class)$.

b)



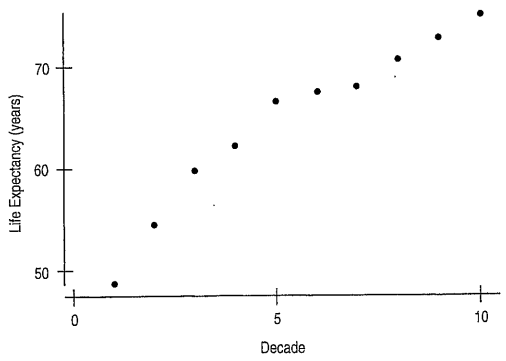
Residuals show a curved pattern; need a different model.

c) $\widehat{lift} = 746.508 - 3295.99 \times \frac{1}{\sqrt{class}}$

d) Less pattern in the residuals plot

e) Boevski's large positive residual indicates he lifted much more than expected.

25.



$\log \widehat{life} = 3.879 + 0.18497 \log decade$

26. a) $\widehat{lift} = 751.069 - 3347.4 \times \frac{1}{\sqrt{class}}$

b)

c)

27. Th
ch

28. a)

29. a)

b)

30. No
lin

31. a) l

b)

c)

i.

32. a) The linear model:
Predicted Diameter = 1.97 + 0.463 × *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 predicted GPA is 2.73.
b) For each increase of 1 meal, the predicted 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) *Predicted CasePrice* = 92.77 + 0.567 × *Years*.
d) Only 2.7% of the variation in case price is accounted for by the ages of vineyards. Most of that is caused by 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) *Predicted TwinBirths* = −4316980 + 2214.19 × *Year*.
b) For each 1-year increase in time, the number of twins born in a year increases by approximately 2214.2.
c) 115,835.2 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 out from the data.
d) The residuals plot shows a definite curved pattern, so the relation is not linear.

6. a) 0.811
b) *Predicted DJIA* = −603335 + 305.471 × *Year*.

c) For each 1-year increase, the DOW increased by about 305.5 points, on average.
d) The relationship does not appear to be linear, as the residuals have a definite pattern. The errors do not appear to be independent.

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) 0.924
d) 85.4% of the variation in manatees killed is explained by the variation in powerboat registrations.
e) No, there is no reason to assign causality.

9. a) *Predicted Manatee Deaths* = −45.9 + 0.132 × *Powerboat Registrations* (in 1000s).
b) For each increase of 10,000 motorboat registrations, the predicted number of manatees killed increases by 1.32.
c) If there were 0 motorboat registrations, the number of manatee deaths would be −45.9. This is obviously a silly extrapolation.
d) The predicted number is 67.4 deaths. The actual number of deaths was 81. The residual is 81 − 67.4 = 13.6. The model underestimated the number of deaths by 13.6.
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.

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 × 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.

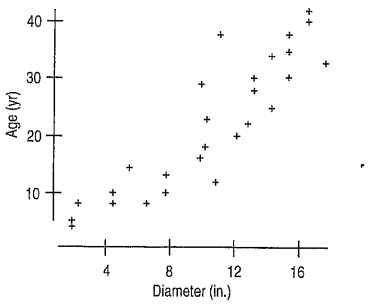
15. a) *Predicted Horsepower* = 3.50 + 34.314 × *Weight*.
b) Thousands. For the equation to have predicted values between 60 and 160, the X values would have to be in thousands of pounds.
c) Yes. The residual plot does not show any pattern.
d) 115.0 horsepower.

16. Gender and colorblindness are both categorical variables, not quantitative. Correlation is meaningless for them, so we say 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) *Predicted Interval* = 33.967 + 10.358 × *Duration*.
c) As the duration of the previous eruption increases by 1 minute, the time between eruptions increases by about 10.4 minutes on average.
d) Since 77% of the variation in interval is accounted for by duration, and the error standard deviation is 6.16 minutes, the prediction will be relatively accurate.
e) 75.4 minutes.
f) A residual is the observed value minus the predicted value. So the residual = 79 − 75.4 = 3.6 minutes, indicating that the model underestimated the interval in this case.

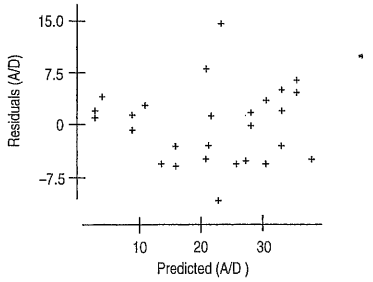
18. a) Yes, the R^2 values indicate that 97.2% of the Indian crocodile length and 98% of the Australian crocodile length is explained by the head size.
b) The slopes of the regression equations are similar, as are the R^2 values.
c) The two models have different y-intercepts. It means that the Indian 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) *Predicted Age* = −0.97 + 2.21 × *Diameter*.

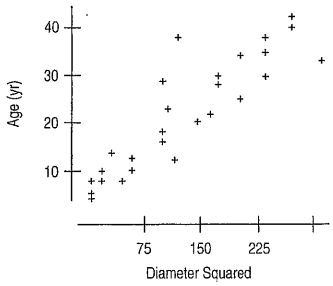
d)



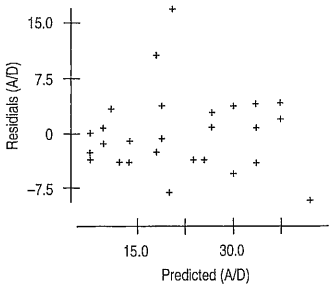
The residuals show a curved pattern (and two outliers).

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) *Predicted Age* = 7.24 + 0.11 × *Diameter Squared.*

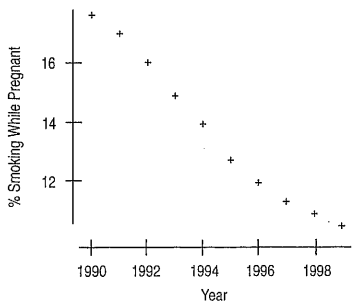


c) The residuals appear to be more randomly scattered and have less of a pattern. But there are still some points with large residuals.

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, somewhat linear, and negative. However, there is some curvature to the scatterplot.

b) −0.991.

c) Averaging increases the strength of the correlation, making it more negative.

d) *Predicted Percentage* = 1745.84 − 0.868 × *Year.* Each year, the average percent of women smoking while pregnant has decreased by about 0.87%.

23. a) The model predicts % smoking from year, not the other way around.

b) *Predicted Year* = 2009.9 − 1.130 × *% Smoking.*

c) The smallest % smoking given is 10.4, and an extrapolation for $x = 0$ 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) *Predicted January Temperature* = 108.80 − 2.111 × *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 0.92 SDs above the mean long jump on average.

30. a) *Predicted High Jump* = −17.93 + 0.321 × *Long Jump.*

b) As the long jump increases by 1 inch, the high jump increases by 0.32 inches on average.

c) 91.21 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) *Predicted Long Jump* = 96.88 + 2.616 × *High Jump.*

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 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. *Predicted Calories* = 560.7 − 3.08 × *Time.*
Each minute extra at the table is associated with 3.08 fewer calories being consumed on average. Perhaps the hungry children eat fast and eat more.

34. a) *Predicted Gallons* = 12451.2 − 6.2 × *Year.*
b) The model is designed to predict gallons based on the year. That question asks to predict the year based on the gallons.
c) *Predicted Year* = 2008.2 − 0.16 × *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, and so the correlation appears to be higher than individuals would be. Any conclusions about individuals would be suspect.

36. a) *Predicted Weight* = −1971.26 + 1.091 × *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) *Predicted Weight* = −1121.66 + 0.673 × *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 original data.
c) Predicted Log(*Profit*) = −0.11 + 0.648 × 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:
$$1/\sqrt{drain\ time} = 0.0024 + 0.219\ diameter.$$

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.
log Predicted *Cost per Chip* = 2.67 − 0.502 × log *Chips Produced.*

43. The predicted values are (12, 774), (24, 738), (36, 702), (48, 666).
The residuals are (12, 26), (24, –58), (36, 38), (48, –6).
The squared residuals are (12, 676), (24, 3364), (36, 1444), (48, 36).
The "least squares" regression equation minimizes the sum of the squares of the residuals.

CHAPTER 11

1. Yes. You cannot predict the outcome beforehand.
2. The outcomes cannot be predicted beforehand. Each of the individual outcomes, 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. 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) Suppose a hand has 4 aces. This might be represented by 1, 1, 1, 1, and any other number. The likelihood for the first ace in the hand is not the same as for the second or third or fourth. But with this simulation, the likelihood is the same for each.
6. a) The numbers would represent the sums, but the sums are not all equally likely. The simulation assumes they are equally likely.
b) The number of boys in a family of 5 children is not equally likely; for example, having a total of 5 boys is less likely than having 3 boys out of 5 children. The simulation assigns the same likelihood to each event.
c) The likelihoods for out, single, double, triple, and home run are not the same, but the simulation assumes they are.