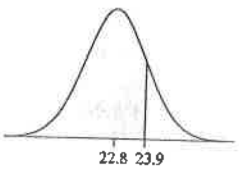
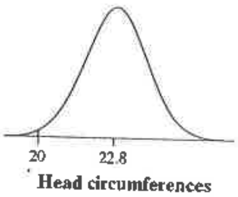
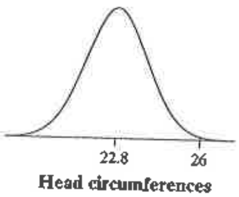
T2.11 (a) Jane's performance was better. She did more curl-ups than 85% of girls her age. This means that she qualified for both the Presidential award and for the National award. Matt did more curl-ups than 50% of boys his age. This means that less than 50% of the boys his age did better than he did, whereas less than 15% of the girls her age did better than Jane. Matt qualified for the National award, but did not qualify for the Presidential award. (b) Since Jane's position in her distribution is so much higher than Matt's position in his distribution, Jane's z-score is likely to be bigger than Matt's z-score.

T2.12 (a) The z-value that corresponds to this soldier's head circumference is $z = \dfrac{23.9 - 22.8}{1.1} = 1$. So the proportion of observations lower than this is 0.8413 (using Table A). This means that this soldier's head circumference is in approximately the 84th percentile.



(b) Standardizing the left endpoint we get $z = \dfrac{20 - 22.8}{1.1} = -2.55$. Using Table A, the area below −2.55 is 0.0054. Standardizing the right endpoint we get $z = \dfrac{26 - 22.8}{1.1} = 2.91$. The area below 2.91 (using Table A) is 0.9982, so the area above 2.91 is $1 - 0.9982 = 0.0018$. This means that the area in both tails is $0.0054 + 0.0018 = 0.0072$. So approximately 0.7% of soldiers require custom helmets.





(c) The quartiles of a standard Normal distribution are −0.67 and 0.67. To find the quartiles of the head circumference distribution, we solve the following equations for x.

$$-0.67 = \frac{x - 22.8}{1.1} \Rightarrow x = -0.67(1.1) + 22.8 = 22.063$$

$$0.67 = \frac{x - 22.8}{1.1} \Rightarrow x = 0.67(1.1) + 22.8 = 23.537$$

This means that $Q_1 = 22.063$ and $Q_3 = 23.537$. So $IQR = Q_3 - Q_1 = 23.537 - 22.063 = 1.474$ inches.

T2.13 No, these data do not seem to follow a Normal distribution. First, there is a large difference between the mean and the median. The mean is 48.25 and the median is 37.80. The Normal distribution is symmetric so the mean and median should be quite close in a Normally distributed data set. This data set appears to be highly skewed to the right. This can be seen by the fact that the mean is so much larger than the median. It can also be seen by the fact that the distance between the minimum and the median is $37.80 - 2 = 35.80$, but the distance between the median and the maximum is $204.90 - 37.80 = 167.10$.

CHAPTER 3

Section 3.1

Answers to Check Your Understanding

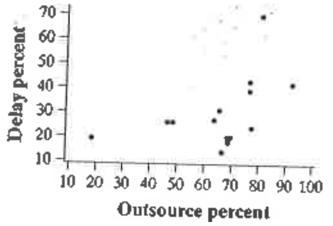
Page 144 1. Explanatory variable: number of cans of beer. Response variable: blood alcohol level. 2. There are two explanatory variables: amount of debt and income. Response variable: stress caused by college debt. ***Page 149*** 1. The relationship is positive. The longer the duration of the eruption, the longer the wait between eruptions. One reason for this may be that if the geyser erupted for longer, it expended more energy and it will take longer to build up the energy needed to erupt again. 2. The form is roughly linear with two clusters. The clusters indicate that in general there are two types of eruptions: one shorter, the other somewhat longer. 3. The relationship is fairly strong. Two points define a line, and in this case we could think of each cluster as a point, so the two clusters seem to define a line. 4. There are a few outliers around the clusters, but not many and not very distant from the main grouping of points. 5. The Starnes family needs to know how long the last eruption lasted in order to predict how long until the next one. ***Page 154*** 1. Estimates of r will vary. (a) The correlation is about 0.9. There is a strong, positive linear relationship between the number of boats registered in Florida and the number of manatees killed. (b) The correlation is about 0.5. There is a moderate, positive linear relationship between the number of named storms predicted and the actual number of named storms. (c) The correlation is about 0.3. There is a weak, positive linear relationship between the healing rate of the two front limbs of the newts. (d) The correlation is about −0.1. There is a weak, negative linear relationship between last year's percent return and this year's percent return in the stock market. 2. The correlation would decrease. This point has the effect of strengthening the observed linear relationship that we see.

Answers to Odd-Numbered Section 3.1 Exercises

1. Explanatory variable: water temperature. Response variable: weight change (growth). Both are quantitative.

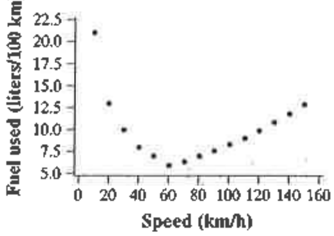
3. (a) Students with higher IQs tend to have higher GPAs, and those with lower IQs generally have lower GPAs. The plot does show a positive association. (b) Roughly linear, because a line through the scatterplot of points would provide a good summary. The positive association is moderately strong because most of the points would be close to the line. (c) An IQ of about 103, a GPA of about 0.4.

5. Here is a scatterplot.



7. (a) Positive, somewhat curved, moderately weak association. (b) The outlier is Hawaiian Airlines. Without this outlier, the relationship is more linear but still not very strong.

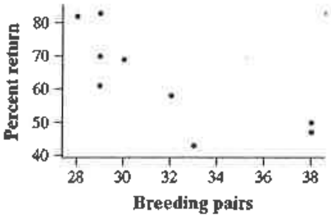
9. (a) Here is a scatterplot.



(b) The relationship is curved. High amounts of fuel were used for low and high values of speed, and low amounts of fuel were used for moderate speeds. That's because the best fuel efficiency is obtained by driving at moderate speeds. (c) Both are present. At low speeds the graph shows a negative association. At higher speeds the association is positive. (d) Very strong, with little deviation from a curve that can be drawn through the points.

11. (a) Several southern states lie at the lower edges of their clusters. Students in the southern states do not do as well as their counterparts in other portions of the country. (b) It has a much lower mean SAT Math score than the other states with a similar percent of students taking the exam.

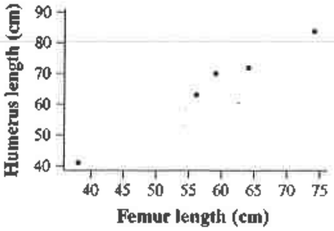
13. State: Is the relationship between the number of breeding pairs of merlins and the percent of males who return the next season negative? Plan: Begin with a scatterplot, and compute the correlation if appropriate. Do: A scatterplot of the percent returning against the number of breeding pairs shows the expected negative association. Though slightly curved, it is reasonable to compute $r = -0.7943$ as a measure of the strength of the linear association. Conclude: This supports the theory: a smaller percent of birds survive following a successful breeding season.



15. (a) $r = 0.9$. (b) $r = 0$. (c) $r = 0.7$. (d) $r = -0.3$. (e) $r = -0.9$.

17. (a) Gender is a categorical variable. (b) r is at most 1. (c) r has no units.

19. (a) The scatterplot shows a strong positive linear relationship between the two measurements. Thus, all five specimens appear to be from the same species.

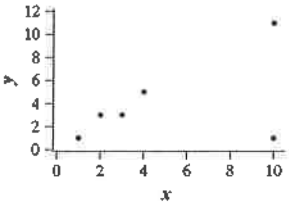


(b) The femur measurements have mean 58.2 and standard deviation 13.2. The humerus measurements have mean 66 and standard deviation 15.89. The sum of the products of the standardized values is 3.97659, so the correlation coefficient is $r = 0.9941$.

21. (a) There is a strong, positive linear association between the salt content and calories of hot dogs. (b) It would tend to decrease the strength of the linear relationship, and thus, the correlation.

23. (a) The correlation would not change. It does not have units associated with it, so a change in units for either variable (or both) will not change the correlation. (b) The correlation would not change. The correlation does not distinguish between the explanatory and response variables.

25. Here is a scatterplot. The one unusual point (10, 1) is responsible for reducing the correlation.

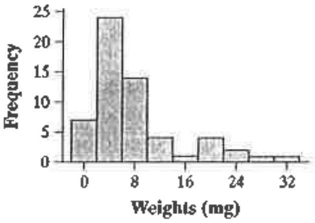


27. a

29. d

31. c

33. A histogram is shown below. The distribution is sharply right-skewed, with several possible high outliers. The five-number summary is 0.1, 3.5, 5.4, 9, 33.8.



Section 3.2

Answers to Check Your Understanding

Page 167 1. The slope is 40. We predict that a rat will gain 40 grams of weight per week. 2. The y intercept is 100. This suggests that we expect a rat at birth to be 100 grams. 3. We predict the rat's weight to be 740 grams. 4. The time is measured in weeks for this equation, so 2 years becomes 104 weeks. We then predict the rat's weight to be 4260 grams, which is equivalent to 9.4 pounds (about the weight of a large newborn human). This is unreasonable and is the result of extrapolation. ***Page 176*** 1. We predict the fat gain for this person to be 1.3722 kg. So the residual is $2.3 - 1.3722 = 0.9278$. 2. The residual says that this person gained 0.9278 kg more than we would have predicted using the least-squares line as a model. 3. The line overpredicted the fat gain the most for the person who had an NEA change of 580 and a fat gain of 0.4. This person's predicted fat gain is 1.51, and so the residual is -1.11. Based on the list of residuals given in the previous example, this is the largest negative residual and therefore the point for which the fat gain was most overpredicted. ***Page 179*** 1. There is a moderate positive linear relationship with one outlier in the bottom-right corner of the plot. 2. The average error (residual) in predicting the backpack weight is 2.27 lb using the least-squares regression line. ***Page 181*** 1. c 2. d

Answers to Odd-Numbered Section 3.2 Exercises

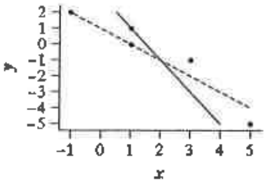
35. The equation is $\hat{y} = 80 - 6x$ where $\hat{y}$ = the estimated weight of the soap and x = the number of days since the bar was new.

37. (a) The slope is 1.109. We predict highway mileage will increase by 1.109 mpg for each 1 mpg increase in city mileage. (b) The intercept is 4.62 mpg. This is not statistically meaningful, because this would represent the highway mileage for a car that gets 0 mpg in the city. (c) With city mpg of 16, the predicted highway mpg is 4.62 + 1.109(16) = 22.36 mpg. With city mpg of 28, the predicted highway mpg is 4.62 + 1.109(28) = 35.67 mpg.

39. (a) The slope is −0.0053; the pH decreased by 0.0053 units per week on average. (b) The y intercept is 5.43, and it provides an estimate for the pH level at the beginning of the study. (c) The pH is predicted to be 4.635 at the end of the study.

41. No. The data was collected weekly for 150 weeks. 1000 months corresponds to roughly 4000 weeks, which is well outside the observed time period. This constitutes extrapolation.

43. The dotted line in the scatterplot is the line $\hat{y} = 1 - x$ and the solid line is the line $\hat{y} = 3 - 2x$. The dotted line comes closer to all the data points. Thus, the line $\hat{y} = 1 - x$ fits the data better.



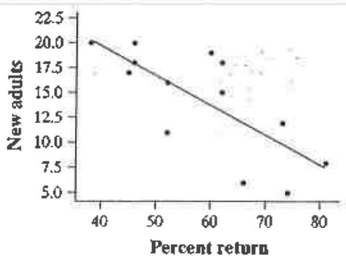
45. The residual is −0.085. The line predicted a pH value for that week that was 0.085 too large.

47. (a) The equation for predicting y = husband's height from x = wife's height is $\hat{y} = 33.67 + 0.54x$. (b) The predicted height is 69.85 inches. 67 inches is one standard deviation above the mean for women. So the predicted value for husband's height would be $\bar{y} + rs_y = 69.85$.

49. (a) $r^2 = 0.25$. Thus, the straight-line relationship explains 25% of the variation in husbands' heights. (b) The average error (residual) when using the line for prediction is 1.2 inches.

51. (a) The regression line is $\hat{y} = -3.5519 + 0.101x$. (b) $r^2 = 0.4016$. Thus, 40.16% of the variation in GPA is accounted for by the linear relationship with IQ. (c) The predicted GPA for this student is $\hat{y} = 6.8511$ and the residual is −6.3211. The student had a GPA that was 6.3211 points worse than expected for someone with an IQ of 103.

53. (a) Here is a scatterplot.



(b) The least-squares regression line is $\hat{y} = 31.9 - 0.304x$. Minitab output is shown at top right.

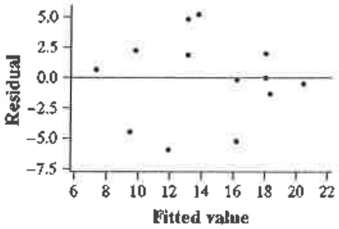
```
The regression equation is
newadults = 31.9 - 0.304 %returning

Predictor   Coef      SE Coef      T      P
Constant    31.934    4.838      6.60  0.000
%returning  -0.30402  0.08122   -3.74  0.003

S = 3.66689   R-Sq = 56.0% R-Sq(adj) = 52.0%
```

(c) The slope tells us that as the percent of returning birds increases by 1, we predict the number of new birds will decrease by 0.304. The y intercept provides a prediction that we will see 31.9 new adults in a colony when the percent of returning birds is 0. This is extrapolation. (d) The predicted value for the number of new adults is 13.66, or about 14.

55., (a) A residual plot suggests that the line is a decent fit. The points are all scattered around a residual value of 0.



(b) The point with the largest residual has a residual of about −6. This means that the line overpredicted the number of new adults by 6.

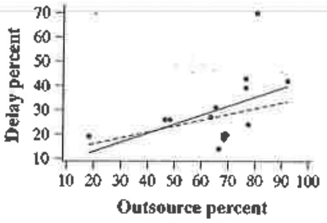
57. 56% of the variation in the number of new adult birds is explained by the straight-line relationship. The typical error when using the line for prediction is 3.67 new adults.

59. (a) There is a positive linear association between the two variables. There is more variation in the field measurements for larger laboratory measurements. (b) The points for the larger depths fall systematically below the line $y = x$, which means that the field measurements are too small compared with the laboratory measurements. (c) The slope would decrease and the intercept would increase.

61. No; the data show a clearly curved pattern in the residual plot.

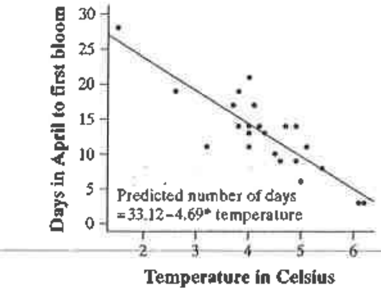
63. (a) The regression line is $\hat{y} = 157.68 - 2.99x$. Following a season with 30 breeding pairs, we predict that about 68% of males will return. (b) The linear relationship explains 63.1% of the variation in the percent of returning males.(c) $r = -0.79$; the sign is negative because it has the same sign as the slope coefficient. (d) Since $s = 9.46$, the typical error when using the line to predict the return rate of males is about 9.46%.

65. (a) Here is a scatterplot (with regression lines).

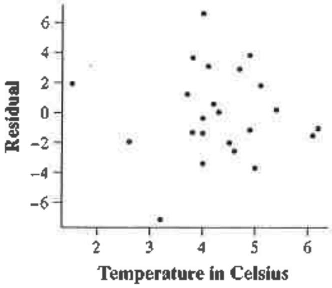


(b) The correlation is $r = 0.4765$ with all points. It rises slightly to 0.4838 when the outlier is removed; this is too small a change to consider the outlier influential for correlation. (c) With all points, $\hat{y} = 4.73 + 0.3868x$ (the solid line), and the prediction for $x = 76$ is

R3.5 (a) The association is negative, linear, and fairly strong.



(b) The least-squares regression equation is $\hat{y} = 33.12 - 4.69x$ where y represents the number of days and x represents the temperature. For every 1 degree increase in average March temperature, we predict the number of days in April until first bloom to decrease by 4.69. The y intercept is outside the range of data and therefore has no meaningful interpretation. (c) Predicted number of days until first bloom is 16.7. We predict the first cherry blossom to appear on April 17. (d) Predicted number of days until first bloom is $\hat{y} = 12.015$. The observed value was 10. The residual is then -2.015. (e) The plot is the residuals versus the explanatory variable. There is no discernible pattern in the residuals. They are clustered about 0 in a random fashion.



(f) $r^2 = 0.72$ and $s = 3.02$. 72% of the variation in the number of days in April until the first cherry blossom appears is explained by the linear relationship with the average temperature (in Celsius) in March. We expect a typical prediction error of 3.02 days.

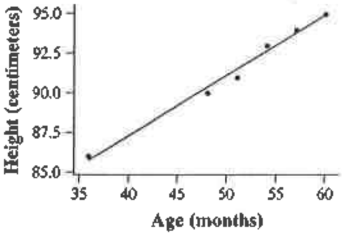
R3.6 (a) The regression equation is $\hat{y} = 0.16x + 30.2$, where $x =$ pre-exam total and $y =$ final exam score. For every extra point earned on the pre-exam total, we predict that the score on the final exam will increase by about 0.16. (b) Julie's predicted final-exam score is $\hat{y} = 78.2$. (c) $r^2 = 0.36$, so only 36% of the variability in the final-exam scores is accounted for by the linear relationship with pre-exam totals. Julie has a good reason to think that this is not a good estimate.

R3.7 (a) The correlation would decrease because Hawaii is far above average for both the maximum 24-hour precipitation and the maximum annual precipitation. (b) The blue line is the line calculated with all 50 states. Hawaii's point is influential and pulls the line up toward it. The other line is the one with all states except Hawaii. (c) The correlation will not change since it does not have units; s will decrease since it would now be measured in feet as well; the slope of the regression line would not change since both x and y are measured in the same units, leading to a slope without units, but the y intercept would decrease since it changes units from inches to feet. If we switch the explanatory and response variables, the correlation will not change, but the standard error and the least-squares line will.

Answers to Chapter 3 AP Statistics Practice Test

T3.1 d
T3.2 e
T3.3 c
T3.4 a
T3.5 a
T3.6 c
T3.7 b
T3.8 e
T3.9 b
T3.10 c

T3.11 (a) Here is a scatterplot, with regression line added.



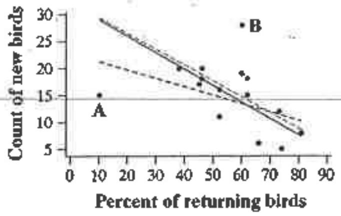
(b) The regression line for predicting $y =$ height from $x =$ age is $\hat{y} = 71.95 + 0.3833x$. (c) At age 480 months, we would predict Sarah's height to be $\hat{y} = 71.95 + 0.3833(480) = 255.934$ cm. Her height in inches would be $\frac{255.934}{2.54} = 100.76$ in. (d) This height is impossibly large (about 8 feet, 4 inches) because we used extrapolation. Obviously, the linear trend does not continue all the way out to 40 years. Our data were based only on the first 5 years of life.

T3.12 (a) The unusual point is the one in the upper-right corner with isotope value about -19.3 and silicon value about 345. This point is unusual in that it has such a high silicon value for the given isotope value. (b) (i) If the point were removed, the correlation would increase because this point does not follow the linear pattern of the other points. (ii) And since this point has a higher silicon value, if it were removed, the slope of the regression line would decrease and the y intercept would increase.

T3.13 (a) The regression equation is $\hat{y} = 92.29 - 0.05762x$. The variable y represents the percent of the grass burned, and x represents the number of wildebeest. (b) The slope of the regression line suggests that for every increase of 1000 wildebeest (this is a 1 unit increase in x since x is measured in terms of 1000s of wildebeest), we predict that the percent of grass area burned will decrease by about 0.058. (c) $r = -0.804$. There is a moderately strong, negative linear relationship between wildebeest abundance and percent of grass area burned. (d) The linear model is appropriate for describing the relationship between wildebeest abundance and percent of grass area burned. The residual plot shows a fairly "random" scatter of points around the "residual = 0" line. There is one large positive residual at 1249 thousand wildebeest. Since $r^2 = 0.646$, 64.6% of the variation in percent of grass area burned is explained by the least-squares regression of percent of grass area burned on wildebeest abundance. That leaves 35.4% of the variation in percent of grass area burned unexplained by the linear relationship.

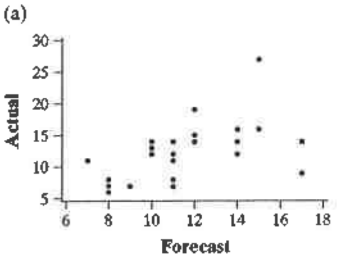
34.13%. With Hawaiian Airlines removed, $\hat{y} = 10.878 + 0.2495x$ (the dotted line), and the prediction is 29.84%. This difference in prediction indicates that the outlier is influential for regression.

67. (a) Here is the scatterplot with two new points. Point A is a horizontal outlier. Point B is a vertical outlier.

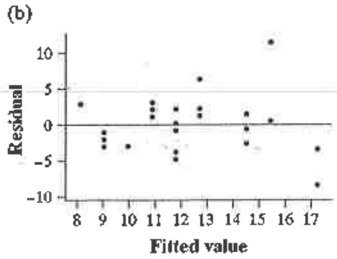


(b) The three regression formulas are $\hat{y} = 31.9 - 0.304x$ (the original data, solid line); $\hat{y} = 22.8 - 0.156x$ (with Point A, dashed line); $\hat{y} = 32.3 - 0.293x$ (with Point B, gray dashed line). Adding Point B has little impact. Point A is influential; it pulls the line down.

69. State: How accurate are Dr. Gray's forecasts? Plan: Construct a scatterplot with Gray's forecast as the explanatory variable and, if appropriate, find the regression equation. Then make a residual plot and calculate r^2 and s. Do: The scatterplot shows a moderate positive association; the regression line is $\hat{y} = 1.688 + 0.9154x$ with $r^2 = 0.30$ and $s = 4.0$. The relationship is strengthened by the large number of storms in the 2005 season, but it is weakened by 2006 and 2007, when Gray's forecasts were the highest, but the actual numbers of storms were unremarkable. As an indication of the influence of the 2005 season, we might find the regression line without that point; it is $\hat{y} = 3.977 + 0.6699x$, with $r^2 = 0.265$ and $s = 3.14$.



Finally, the residual plot does not indicate any problems with fitting the linear equation.



Conclude: If Gray forecasts $x = 16$ tropical storms, we expect 16.33 storms in that year. However, we do not have very much confidence in this estimate, because the regression line explains only 30% of the variation in tropical storms and the typical error we should expect when using this line for prediction is 4 storms. (If we exclude 2005, the prediction is 14.7 storms, but this estimate is less reliable than the first.)

71. b

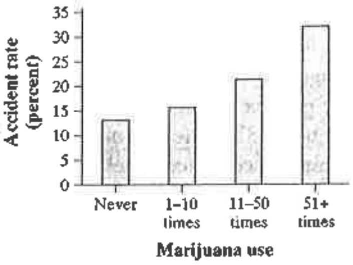
73. b

75. b

77. d

79. About 92.92%

81. (a) There is evidence of an association between accident rate and marijuana use. Those people who use marijuana more are more likely to have caused accidents.



(b) This was an observational study. If we wanted to see whether using marijuana *caused* more accidents, then we would have to set up an experiment where we randomly assigned people to use more or less marijuana.

Answers to Chapter 3 Review Exercises

R3.1 (a) Explanatory variable: weight of a person. Response variable: mortality rate. (b) No. Since obese people tend to be poor, their higher mortality rate may be due to low-quality medical care and not to their weight.

R3.2 (a) The direction of the scatterplot is positive but appears curved, not linear. The strength of the association is moderate. (b) The hippopotamus is unusual because its lifespan is longer than would be expected given its gestation period. The Asian elephant is unusual because it has the second-longest gestation time and has a longer lifetime than expected. The observation for the giraffe tends to follow the curvilinear shape, with possibly a little shorter lifespan than expected based on the pattern in the remaining data.

R3.3 (a) The slope is 0.0138 minutes per meter. We predict that if the depth of the dive is increased by one meter, it will add 0.0138 minutes (about 0.83 seconds) to the time spent underwater. (b) When Depth = 200, the regression line estimates DiveDuration to be $\hat{y} = 5.45$ minutes. (c) The intercept suggests that a dive of no depth would last an average of 2.69 minutes; this does not make any sense.

R3.4 (a)Theleast-squaresregressionlineis $\hat{y} = 7288.54 + 11630.6x$ where y represents the mileage of the cars and x represents the age. (b) The residual for this car is -12072.14. (c) The slope of the line is 11,630.6. We expect that cars will be driven 11,630.6 miles per year on average. (d) Since the slope is positive, the correlation $r = 0.906$. This shows that there is a strong, linear relationship between the age of cars and their mileage. (e) The line fits reasonably well. The residual plot shows no large pattern. The typical error (residual) when using this line to predict mileage is 19,280. This suggests that although the line fits reasonably well, it would not be very useful in practice since our predictions would be off by an average of approximately 20,000 miles.