# Chapter 14

# Simple Linear Regression

**Learning Objectives**

1. Understand how regression analysis can be used to develop an equation that estimates mathematically how two variables are related.

2. Understand the differences between the regression model, the regression equation, and the estimated regression equation.

3. Know how to fit an estimated regression equation to a set of sample data based upon the least-squares method.

4. Be able to determine how good a fit is provided by the estimated regression equation and compute the sample correlation coefficient from the regression analysis output.

5. Understand the assumptions necessary for statistical inference and be able to test for a significant relationship.

6. Know how to develop confidence interval estimates of *y* given a specific value of *x* in both the case of a mean value of *y* and an individual value of *y*.

7. Learn how to use a residual plot to make a judgement as to the validity of the regression assumptions.

8. Know the definition of the following terms:

independent and dependent variable

simple linear regression

regression model

regression equation and estimated regression equation

scatter diagram

coefficient of determination

standard error of the estimate

confidence interval

prediction interval

residual plot

**Solutions:**

1 a.

b. There appears to be a positive linear relationship between *x* andy.

c. Many different straight lines can be drawn to provide a linear approximation of the relationship between *x and* y; in part (d) we will determine the equation of a straight line that “best” represents the relationship according to the least squares criterion.

d. 









e. 

2. a.

b. There appears to be a negative linear relationship between *x* and *y*.

c. Many different straight lines can be drawn to provide a linear approximation of the relationship between *x and y*; in part (d) we will determine the equation of a straight line that “best” represents the relationship according to the least squares criterion.

d. 









e. 

3. a.

b. 







c. 

4. a.

b. There appears to be a positive linear relationship between the percentage of women working in the five companies (*x*) and the percentage of management jobs held by women in that company (*y*)

c. Many different straight lines can be drawn to provide a linear approximation of the relationship between *x* and *y*; in part (d) we will determine the equation of a straight line that “best” represents the relationship according to the least squares criterion.

d. 











e. 

5. a.

b. There appears to be a positive relationship between price and rating. The sign that says “Quality: You Get What You Pay For” does fairly reflect the price-quality relationship for ellipticals.

c. Let *x* = price ($) and *y* = rating.











d. or approximately 71

6. a.

b. The scatter diagram indicates a positive linear relationship between *x* = average number of passing yards per attempt and *y* = the percentage of games won by the team.

c. 









d. The slope of the estimated regression line is approximately 17.2. So, for every increase of one yard in the average number of passes per attempt, the percentage of games won by the team increases by 17.2%.

e. With an average number of passing yards per attempt of 6.2, the predicted percentage of games won is = -70.391 + 17.175(6.2) = 36%. With a record of 7 wins and 9 loses, the percentage of wins that the Kansas City Chiefs won is 43.8 or approximately 44%. Considering the small data size, the prediction made using the estimated regression equation is not too bad.

7. a.

b. Let *x* = years of experience and *y* = annual sales ($1000s)











c. or $116,000

8. a.

b. The scatter diagram indicates a positive linear relationship between *x* = speed of execution rating and *y* = overall satisfaction rating for electronic trades.

c. 









1. The slope of the estimated regression line is approximately .9077. So, a one unit increase in the speed of execution rating will increase the overall satisfaction rating by approximately .9 points.
2. The average speed of execution rating for the other brokerage firms is 3.4. Using this as the new value of *x* for Zecco.com, we can use the estimated regression equation developed in part (c) to estimate the overall satisfaction rating corresponding to *x* = 3.4.



Thus, an estimate of the overall satisfaction rating when *x* = 3.4 is approximately 3.3.

9. a.

b. The scatter diagram indicates a positive linear relationship between *x* = price ($) and *y* = overall rating.

c. 









d. We can use the estimated regression equation developed in part (c) to estimate the overall satisfaction rating corresponding to *x* = 200.



Thus, an estimate of the overall rating when *x* = $200 is approximately 70.

10. a.

b. The scatter diagram indicates a positive linear relationship between *x* = percentage increase in the stock price and *y* = percentage gain in options value. In other words, options values increase as stock prices increase.

c. 





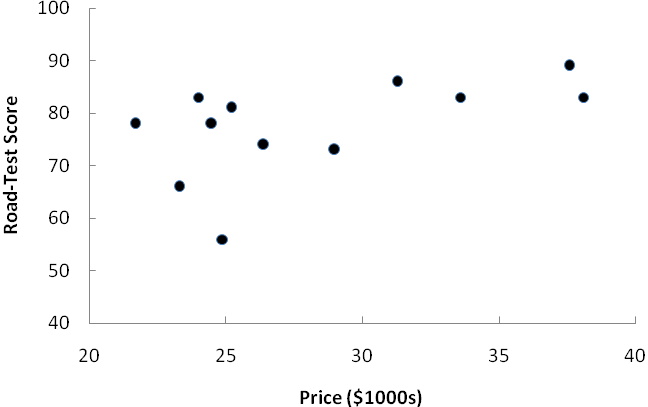




d. The slope of the estimated regression line is approximately 2.7. So, for every percentage increase in the price of the stock the options value increases by 2.7%.

e. The rewards for the CEO do appear to be based upon performance increases in the stock value. While the rewards may seem excessive, the executive is being rewarded for his/her role in increasing the value of the company. This is why such compensation schemes are devised for CEOs by boards of directors. A compensation scheme where an executive got a big salary increase when the company stock went down would be bad. And, if the stock price for a company had gone down during the periods in question, the value of the CEOs options would also go down.

11. a.



b. There appears to be a positive linear relationship between *x* = priceand *y* = road-test score.

c. 









d. The slope is .8947. A sporty car that has a ten thousand dollar higher price can be expected to have a 10(.8947) = 8.947, or approximately a 9 point higher road-test score.

e. 

12. a.

b. The scatter diagram indicates a positive linear relationship between *x* = hotel room rate and the amount spent on entertainment.

c. 









d. With a value of *x* = $128, the predicted value of *y* for Chicago is



Note: In The Wall Street Journal article the entertainment expense for Chicago was $146. Thus, the estimated regression equation provided a good estimate of entertainment expenses for Chicago.

13. a.

b. Let *x* = adjusted gross income and *y* = reasonable amount of itemized deductions







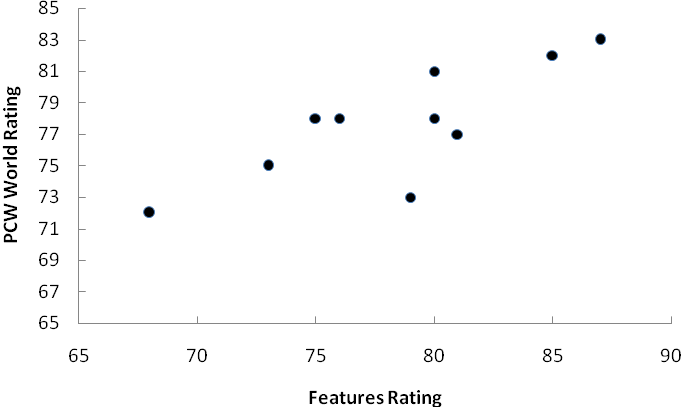




c. or approximately $13,080.

The agent's request for an audit appears to be justified.

14. a.



b. There appears to be a positive linear relationship between *x* = features ratingand *y* = PCW World Rating.

c. 









d. or 73

15. a. The estimated regression equation and the mean for the dependent variable are:



The sum of squares due to error and the total sum of squares are



Thus, SSR = SST - SSE = 80 - 12.4 = 67.6

b. *r*2 = SSR/SST = 67.6/80 = .845

The least squares line provided a very good fit; 84.5% of the variability in *y* has been explained by the least squares line.

c. 

16. a. The estimated regression equation and the mean for the dependent variable are:



The sum of squares due to error and the total sum of squares are



Thus, SSR = SST - SSE = 1850 - 230 = 1620

b. *r*2 = SSR/SST = 1620/1850 = .876

The least squares line provided an excellent fit; 87.6% of the variability in *y* has been explained by the estimated regression equation.

c. 

Note: the sign for *r* is negative because the slope of the estimated regression equation is negative.

(*b*1 = -3)

17. The estimated regression equation and the mean for the dependent variable are:



The sum of squares due to error and the total sum of squares are



Thus, SSR = SST - SSE = 281.2 – 127.3 = 153.9

*r*2 = SSR/SST = 153.9/281.2 = .547

We see that 54.7% of the variability in *y* has been explained by the least squares line.



18. a. 



SSR = SST – SSR = 1800 – 287.624 = 1512.376

b. 

c. 

19. a. The estimated regression equation and the mean for the dependent variable are:

= 80 + 4*x* = 108

The sum of squares due to error and the total sum of squares are



Thus, SSR = SST - SSE = 2442 - 170 = 2272

b. *r*2 = SSR/SST = 2272/2442 = .93

We see that 93% of the variability in *y* has been explained by the least squares line.

c. 

20. a. 









b. SST = 52,120,800 SSE = 7,102,922.54

SSR = SST – SSR = 52,120,800 - 7,102,922.54 = 45,017,877

= SSR/SST = 45,017,877/52,120,800 = .864

The estimated regression equation provided a very good fit.

c. 

Thus, an estimate of the price for a bike that weighs 15 pounds is $6989.

21. a. 









b. $7.60

c. The sum of squares due to error and the total sum of squares are:



Thus, SSR = SST - SSE = 5,648,333.33 - 233,333.33 = 5,415,000

*r*2 = SSR/SST = 5,415,000/5,648,333.33 = .9587

We see that 95.87% of the variability in *y* has been explained by the estimated regression equation.

d. 

22. a. = 74 SSE = 173.88

The total sum of squares is



Thus, SSR = SST - SSE = 756 – 173.88 = 582.12

*r*2 = SSR/SST = 582.12/756 = .77

b. The estimated regression equation provided a good fit because 77% of the variability in *y* has been explained by the least squares line.

c. 

This reflects a strong positive linear relationship between price and rating.

23. a. *s*2 = MSE = SSE / (*n* - 2) = 12.4 / 3 = 4.133

b. 

c. 



d. 

Using *t* table (3 degrees of freedom), area in tail is between .01 and .025

*p*-value is between .02 and .05

Using Excel or Minitab, the *p*-value corresponding to *t* = 4.04 is .0272.

Because *p*-value, we reject *H*0: 1 = 0

e. MSR = SSR / 1 = 67.6

*F* = MSR / MSE = 67.6 / 4.133 = 16.36

Using *F* table (1 degree of freedom numerator and 3 denominator), *p*-value is between .025 and .05

Using Excel or Minitab, the *p*-value corresponding to *F* = 16.36 is .0272.

Because *p*-value, we reject *H*0: 1 = 0

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source  of Variation | Sum  of Squares | Degrees  of Freedom | Mean  Square | *F* | *p*-value |
| Regression | 67.6 | 1 | 67.6 | 16.36 | .0272 |
| Error | 12.4 | 3 | 4.133 |  |  |
| Total | 80.0 | 4 |  |  |  |

24. a. *s*2 = MSE = SSE/(*n* - 2) = 230/3 = 76.6667

b. 

c. 



d. 

Using *t* table (3 degrees of freedom), area in tail is less than .01; *p*-value is less than .02

Using Excel or Minitab, the *p*-value corresponding to *t* = -4.59 is .0193.

Because *p*-value, we reject *H*0: **1 = 0

e. MSR = SSR/1 = 1620

*F* = MSR/MSE = 1620/76.6667 = 21.13

Using *F* table (1 degree of freedom numerator and 3 denominator), *p*-value is less than .025

Using Excel or Minitab, the *p*-value corresponding to *F* = 21.13 is .0193.

Because *p*-value, we reject *H*0: **1 = 0

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source  of Variation | Sum  of Squares | Degrees  of Freedom | Mean  Square | *F* | *p*-value |
| Regression | 1620 | 1 | 1620 | 21.13 | .0193 |
| Error | 230 | 3 | 76.6667 |  |  |
| Total | 1850 | 4 |  |  |  |

25. a. *s*2 = MSE = SSE/(*n* - 2) = 127.3/3 = 42.4333



b. 





Using *t* table (3 degrees of freedom), area in tail is between .05 and .10

*p*-value is between .10 and .20

Using Excel or Minitab, the *p*-value corresponding to *t* = 1.90 is .1530.

Because *p*-value >, we cannot reject *H*0: **1 = 0; *x* and *y* do not appear to be related.

c. MSR = SSR/1 = 153.9 /1 = 153.9

*F* = MSR/MSE = 153.9/42.4333 = 3.63

Using *F* table (1 degree of freedom numerator and 3 denominator), *p*-value is greater than .10

Using Excel or Minitab, the *p*-value corresponding to *F* = 3.63 is .1530.

Because *p*-value >, we cannot reject *H*0: **1 = 0; *x* and *y* do not appear to be related.

26. a. In the statement of exercise 18, = 23.194 + .318*x*

In solving exercise 18, we found SSE = 287.624











Using *t* table (4 degrees of freedom), area in tail is between .005 and .01

*p*-value is between .01 and .02

Using Excel, the *p*-value corresponding to *t* = 4.58 is .010.

Because *p*-value, we reject *H*0: = 0; there is a significant relationship between price and overall score

b. In exercise 18 we found SSR = 1512.376

MSR = SSR/1 = 1512.376/1 = 1512.376

*F* = MSR/MSE = 1512.376/71.906 = 21.03

Using *F* table (1 degree of freedom numerator and 4 denominator), *p*-value is between .025 and .01

Using Excel, the *p*-value corresponding to *F* = 11.74 is .010.

Because *p*-value, we reject *H*0: = 0

c.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source  of Variation | Sum  of Squares | Degrees  of Freedom | Mean  Square | *F* | *p*-value |
| Regression | 1512.376 | 1 | 1512.376 | 21.03 | .010 |
| Error | 287.624 | 4 | 71.906 |  |  |
| Total | 1800 | 5 |  |  |  |

27. a. Let *x* = number of megapixels and *y* = price ($)









b. SSE =  SST == 103,690

Thus, SSR = SST - SSE = 103,690 – 20,730.27 = 82,959.73

MSR = SSR/1 = 82,959.73

MSE = SSE/(*n* - 2) = 20,730.27/8 = 2591.28

*F* = MSR / MSE = 82,959.73/2591.28 = 32.015

Using *F* table (1 degree of freedom numerator and 8 denominator), *p*-value is less than .01

Using Excel, the *p*-value corresponding to *F* = 32.015 is .000.

Because *p*-value, we reject *H*0: **1 = 0

Number of megapixels and price are related.

c. *r*2 = SSR/SST = 82,959.73/103,690= .80

The estimated regression equation provided a good fit; we should feel comfortable using the estimated regression equation to estimate the price given the number of megapixels.

d. or approximately $238

28. The sum of squares due to error and the total sum of squares are



Thus, SSR = SST - SSE = 3.5800 – 1.4379 = 2.1421

*s*2 = MSE = SSE / (*n* - 2) = 1.4379 / 9 = .1598



We can use either the *t* test or *F* test to determine whether speed of execution and overall satisfaction are related.

We will first illustrate the use of the *t* test.







Using *t* table (9 degrees of freedom), area in tail is less than .005; *p*-value is less than .01

Using Excel or Minitab, the *p*-value corresponding to *t* = 3.66 is .000.

Because *p*-value, we reject *H*0: = 0

Because we can reject *H*0: = 0 we conclude that speed of execution and overall satisfaction are related.

Next we illustrate the use of the *F* test.

MSR = SSR / 1 = 2.1421

*F* = MSR / MSE = 2.1421 / .1598 = 13.4

Using *F* table (1 degree of freedom numerator and 9 denominator), *p*-value is less than .01

Using Excel or Minitab, the *p*-value corresponding to *F* = 13.4 is .000.

Because *p*-value, we reject *H*0: = 0

Because we can reject *H*0: = 0 we conclude that speed of execution and overall satisfaction are related.

The ANOVA table is shown below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source  of Variation | Sum  of Squares | Degrees  of Freedom | Mean  Square | *F* | *p*-value |
| Regression | 2.1421 | 1 | 2.1421 | 13.4 | .000 |
| Error | 1.4379 | 9 | .1598 |  |  |
| Total | 3.5800 | 10 |  |  |  |

29. SSE =233,333.33 SST == 5,648,333.33

Thus, SSR = SST – SSE = 5,648,333.33 –233,333.33 = 5,415,000

MSE = SSE/(*n* - 2) = 233,333.33/(6 - 2) = 58,333.33

MSR = SSR/1 = 5,415,000

*F* = MSR / MSE = 5,415,000 / 58,333.25 = 92.83

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source of Variation | Sum  of Squares | Degrees of Freedom | Mean Square | *F* | *p*-value |
| Regression | 5,415,000.00 | 1 | 5,415,000 | 92.83 | .0006 |
| Error | 233,333.33 | 4 | 58,333.33 |  |  |
| Total | 5,648,333.33 | 5 |  |  |  |

Using *F* table (1 degree of freedom numerator and 4 denominator), *p*-value is less than .01

Using Excel or Minitab, the *p*-value corresponding to *F* = 92.83 is .0006.

Because *p*-value, we reject *H*0: **1 = 0. Production volume and total cost are related.

30. SSE =173.88 SST == 756

Thus, SSR = SST – SSE = 756 – 173.88 = 582.12

*s*2 = MSE = SSE/(*n*-2) = 173.88/6 = 28.98



 = 8,155,000





Using *t* table (1 degree of freedom numerator and 8 denominator), area in tail is less than .005

*p*-value is less than .01

Using Excel or Minitab, the *p*-value corresponding to *t* = 4.48 is .0042.

Because *p*-value, we reject *H*0: **1 = 0

There is a significant relationship between price and rating.

31. SST = 52,120,800 SSE = 7,102,922.54

SSR = SST – SSR = 52,120,800 - 7,102,922.54 = 45,017,877

MSR = SSR/1 = 45,017,877

MSE = SSE/(*n* - 2) = 7,102,922.54/8 = 887,865.3

*F* = MSR / MSE = 45,017,877/887,865.3 = 50.7

Using *F* table (1 degree of freedom numerator and 8 denominator), *p*-value is less than .01

Using Excel, the *p*-value corresponding to *F* = 32.015 is .000.

Because *p*-value, we reject *H*0: **1 = 0

Weight and price are related.

32. a. *s* = 2.033





b. = .2 + 2.6= .2 + 2.6(4) = 10.6



10.6  3.182 (1.11) = 10.6  3.53

or 7.07 to 14.13

c. 

d. 

10.6  3.182 (2.32) = 10.6  7.38

or 3.22 to 17.98

33. a.     *s*  =  8.7560

b.  







44  3.182 (4.3780) = 44  13.93

or 30.07 to 57.93

c. 

d. 

44  3.182(9.7895) = 44  31.15

or 12.85 to 75.15

34. *s* = 6.5141









18.40  3.182(3.0627) = 18.40  9.75

or 8.65 to 28.15





18.40  3.182(7.1982) = 18.40  22.90

or -4.50 to 41.30

The two intervals are different because there is more variability associated with predicting an individual value than there is a mean value.

35. a. 

b. *s* = 145.89







3833.8  2.776 (68.54) = 3833.8  190.27

or $3643.53 to $4024.07

c. 



3833.8  2.776 (161.19) = 3833.8  447.46

or $3386.34 to $4281.26

d. As expected, the prediction interval is much wider than the confidence interval. This is due to the fact that it is more difficult to predict the starting salary for one new student with a GPA of 3.0 than it is to estimate the mean for all students with a GPA of 3.0.

36. a. 





116  2.306(1.6503) = 116  3.8056

or 112.19 to 119.81 ($112,190 to $119,810)

b. 



116  2.306(4.8963) = 116  11.2909

or 104.71 to 127.29 ($104,710 to $127,290)

c. As expected, the prediction interval is much wider than the confidence interval. This is due to the fact that it is more difficult to predict annual sales for one new salesperson with 9 years of experience than it is to estimate the mean annual sales for all salespersons with 9 years of experience.

37. a. 

*s*2 = 1.88 *s* = 1.37





= 4.68 + 0.16= 4.68 + 0.16(52.5) = 13.08

13.08  2.571 (.52) = 13.08  1.34

or 11.74 to 14.42 or $11,740 to $14,420

b. = 1.47

13.08  2.571 (1.47) = 13.08  3.78

or 9.30 to 16.86 or $9,300 to $16,860

c. Yes, $20,400 is much larger than anticipated.

d. Any deductions exceeding the $16,860 upper limit could suggest an audit.

38. a. = 1246.67 + 7.6(500) = $5046.67

b. 

*s2* = MSE = 58,333.33 *s* = 241.52





5046.67  4.604 (267.50) = 5046.67  1231.57

or $3815.10 to $6278.24

c. Based on one month, $6000 is not out of line since $3815.10 to $6278.24 is the prediction interval. However, a sequence of five to seven months with consistently high costs should cause concern.

39. a. Let *x* = miles of track and *y* = weekday ridership in thousands.











b. SST =3620.9 SSE = 1038.7 SSR = 2582.1

*r*2 = SSR/SST = 2582.1/3620.9 = .713

The estimated regression equation explained 71.3% of the variability in *y*; a good fit.

c. *s*2 = MSE = 1038.7/5 = 207.7







45.9  2.571(5.47) = 45.9  14.1

or 31.8 to 60

d. 



45.9  2.571(15.41) = 45.9  39.6

or 6.3 to 85.5

The prediction interval is so wide that it would not be of much value in the planning process. A larger data set would be beneficial.

40. a. 9

b. = 20.0 + 7.21*x*

c. 1.3626

d. SSE = SST - SSR = 51,984.1 - 41,587.3 = 10,396.8

MSE = 10,396.8/7 = 1,485.3

*F* = MSR / MSE = 41,587.3 /1,485.3 = 28.00

Using *F* table (1 degree of freedom numerator and 7 denominator), *p*-value is less than .01

Using Excel or Minitab, the *p*-value corresponding to *F* = 28.00 is .0011.

Because *p*-value= .05, we reject H0: *B*1 = 0.

Selling price is related to annual gross rents.

e. = 20.0 + 7.21(50) = 380.5 or $380,500

41. a. = 6.1092 + .8951*x*

b. 

Using the *t* table (8 degrees of freedom), area in tail is less than .005

*p*-value is less than .01

Using Excel or Minitab, the *p*-value corresponding to *t* = 6.01 is .0003.

Because *p*-value= .05, we reject H0: *B*1 = 0

Maintenance expense is related to usage.

c. = 6.1092 + .8951(25) = 28.49 or $28.49 per month

42 a. = 80.0 + 50.0*x*

b. 30

c. *F* = MSR / MSE = 6828.6/82.1 = 83.17

Using *F* table (1 degree of freedom numerator and 28 denominator), *p*-value is less than .01

Using Excel or Minitab, the *p*-value corresponding to *F* = 83.17 is .000.

Because *p*-value < = .05, we reject H0: *B*1 = 0.

Annual sales is related to the number of salespersons.

d. = 80 + 50 (12) = 680 or $680,000

43. a.

b. There appears to be a positive relationship between the two variables. Students that graduate from the schools with higher tuition and fees tend to receive a higher starting salary and bonus.

The Minitab output is shown below:

The regression equation is

Salary & Bonus ($1000s) = 33.8 + 1.92 Tuition & Fees ($1000s)

Predictor Coef SE Coef T P

Constant 33.788 9.340 3.62 0.002

Tuition & Fees ($1000s) 1.9154 0.2689 7.12 0.000

S = 7.60875 R-Sq = 73.8% R-Sq(adj) = 72.4%

Analysis of Variance

Source DF SS MS F P

Regression 1 2937.1 2937.1 50.73 0.000

Residual Error 18 1042.1 57.9

Total 19 3979.2

d. The *p*-value = .000 < = .05 (*t* or *F*); significant relationship

e. *r*2 = .738. The least squares line provided a good fit; approximately 74% of the variability in salary and bonus can be explained by the linear relationship with tuition and fees.

f. = 33.788 + 1.9154(43) = 116.15 or approximately $116,000.

Note to Instructor: The average starting salary and bonus reported by U.S. News & World Report for the University of Virginia was $121,000.

44. a. Scatter diagram:

b. There appears to be a negative linear relationship between the two variables. The heavier helmets tend to be less expensive.

c. The Minitab output is shown below:

The regression equation is

Price = 2044 - 28.3 Weight

Predictor Coef SE Coef T P

Constant 2044.4 226.4 9.03 0.000

Weight -28.350 3.826 -7.41 0.000

S = 91.8098 R-Sq = 77.4% R-Sq(adj) = 76.0%

Analysis of Variance

Source DF SS MS F P

Regression 1 462761 462761 54.90 0.000

Residual Error 16 134865 8429

Total 17 597626

= 2044.4 – 28.35 Weight

d. Significant relationship: *p*-value = .000 < α = .05

e. *r*2 = 0.774; A good fit

45. a. 









b. The residuals are 3.48, -2.47, -4.83, -1.6, and 5.22

c.

With only 5 observations it is difficult to determine if the assumptions are satisfied. However, the plot does suggest curvature in the residuals that would indicate that the error term assumptions are not satisfied. The scatter diagram for these data also indicates that the underlying relationship between *x* and *y* may be curvilinear.

d. 



The standardized residuals are 1.32, -.59, -1.11, -.40, 1.49.

e. The standardized residual plot has the same shape as the original residual plot. The curvature observed indicates that the assumptions regarding the error term may not be satisfied.

46. a. 

b.

The assumption that the variance is the same for all values of *x* is questionable. The variance appears to increase for larger values of *x*.

47. a. Let *x* = advertising expenditures and *y* = revenue



b. SST = 1002 SSE = 310.28 SSR = 691.72

MSR = SSR / 1 = 691.72

MSE = SSE / (*n* - 2) = 310.28/ 5 = 62.0554

*F* = MSR / MSE = 691.72/ 62.0554= 11.15

Using *F* table (1 degree of freedom numerator and 5 denominator), *p*-value is between .01 and .025

Using Excel or Minitab, the *p*-value corresponding to *F* = 11.15 is .0206.

Because *p*-value = .05, we conclude that the two variables are related.

c.

d. The residual plot leads us to question the assumption of a linear relationship between *x* and *y*. Even though the relationship is significant at the .05 level of significance, it would be extremely dangerous to extrapolate beyond the range of the data.

48. a. 

b. The assumptions concerning the error term appear reasonable.

49. a. The Minitab output follows:

The regression equation is

Price ($) = 22636 + 59.0 Square Footage

Predictor Coef SE Coef T P

Constant 22636 20460 1.11 0.283

Square Footage 58.96 12.08 4.88 0.000

S = 19166.0 R-Sq = 57.0% R-Sq(adj) = 54.6%

Analysis of Variance

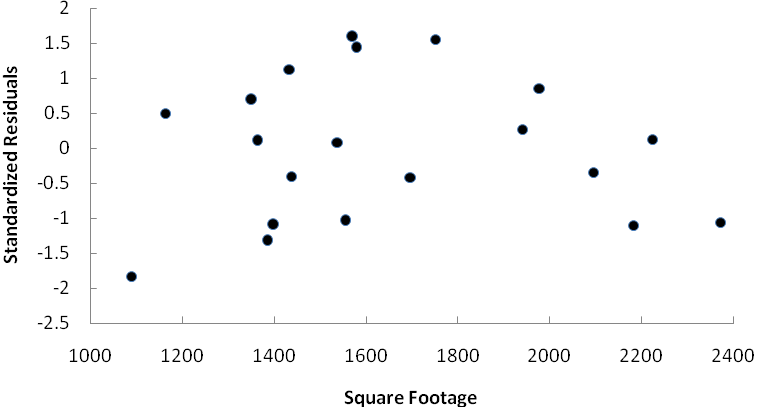
Source DF SS MS F P

Regression 1 8748562231 8748562231 23.82 0.000

Residual Error 18 6612039769 367335543

Total 19 15360602000

b.



c. The residual plot leads us to question the assumption of a linear relationship between square footage and price. Therefore, even though the relationship is very significant (*p*-value = .000), using the estimated regression equation make predictions of the price for a house with square footage beyond the range of the data is not recommended.

50. a. The Minitab output follows:

The regression equation is

Y = 66.1 + 0.402 X

Predictor Coef SE Coef T p

Constant 66.10 32.06 2.06 0.094

X 0.4023 0.2276 1.77 0.137

S = 12.62 R-sq = 38.5% R-sq(adj) = 26.1%

Analysis of Variance

SOURCE DF SS MS F p

Regression 1 497.2 497.2 3.12 0.137

Residual Error 5 795.7 159.1

Total 6 1292.9

Unusual Observations

Obs. X Y Fit SEFit Residual St.Resid

1 135 145.00 120.42 4.87 24.58 2.11R

R denotes an observation with a large standardized residual.

The standardized residuals are: 2.11, -1.08, .14, -.38, -.78, -.04, -.41

The first observation appears to be an outlier since it has a large standardized residual.

b.



The standardized residual plot indicates that the observation *x* = 135, *y* = 145 may be an outlier; note that this observation has a standardized residual of 2.11.

c. The scatter diagram is shown below

The scatter diagram also indicates that the observation *x* = 135, *y* = 145 may be an outlier; the implication is that for simple linear regression an outlier can be identified by looking at the scatter diagram.

51. a. The Minitab output is shown below:

The regression equation is

Y = 13.0 + 0.425 X

Predictor Coef SE Coef T p

Constant 13.002 2.396 5.43 0.002

X 0.4248 0.2116 2.01 0.091

S = 3.181 R-sq = 40.2% R-sq(adj) = 30.2%

Analysis of Variance

SOURCE DF SS MS F p

Regression 1 40.78 40.78 4.03 0.091

Residual Error 6 60.72 10.12

Total 7 101.50

Unusual Observations

Obs. X Y Fit Stdev.Fit Residual St.Resid

7 12.0 24.00 18.10 1.20 5.90 2.00R

8 22.0 19.00 22.35 2.78 -3.35 -2.16RX

R denotes an observation with a large standardized residual.

X denotes an observation whose X value gives it large influence.

The standardized residuals are: -1.00, -.41, .01, -.48, .25, .65, -2.00, -2.16

The last two observations in the data set appear to be outliers since the standardized residuals for these observations are 2.00 and -2.16, respectively.

b. Using Minitab, we obtained the following leverage values:

.28, .24, .16, .14, .13, .14, .14, .76

MINITAB identifies an observation as having high leverage if *hi* > 6/*n*; for these data, 6/*n* = 6/8 = .75. Since the leverage for the observation *x* = 22, *y* = 19 is .76, Minitab would identify observation 8 as a high leverage point. Thus, we conclude that observation 8 is an influential observation.

c.

The scatter diagram indicates that the observation *x* = 22, *y* = 19 is an influential observation.

52. a.

The scatter diagram does indicate potential influential observations. For example, the 22.2% fundraising expense for the American Cancer Society and the 16.9% fundraising expense for the St. Jude Children’s Research Hospital look like they may each have a large influence on the slope of the estimated regression line. And, with a fundraising expense of on 2.6%, the percentage spend on programs and services by the Smithsonian Institution (73.7%) seems to be somewhat lower than would be expected; thus, this observeraton may need to be considered as a possible outlier

b. A portion of the Minitab output follows:

The regression equation is

Program Expenses (%) = 91.0 - 0.917 Fundraising Expenses (%)

Predictor Coef SE Coef T P

Constant 90.981 3.177 28.64 0.000

Fundraising Expenses (%) -0.9172 0.3392 -2.70 0.027

S = 7.47387 R-Sq = 47.7% R-Sq(adj) = 41.2%

Analysis of Variance

Source DF SS MS F P

Regression 1 408.35 408.35 7.31 0.027

Residual Error 8 446.87 55.86

Total 9 855.22

Unusual Observations

Program

Fundraising Expenses

Obs Expenses (%) (%) Fit SE Fit Residual St Resid

3 2.6 73.70 88.60 2.67 -14.90 -2.13R

5 22.2 71.60 70.62 5.90 0.98 0.21 X

R denotes an observation with a large standardized residual.

X denotes an observation whose X value gives it large leverage.

c. The slope of the estimtaed regression equation is -0.917. Thus, for every 1% increase in the amount spent on fundraising the percentage spent on program expresses will decrease by .917%; in other words, just a little under 1%. The negative slope and value seem to make sense in the context of this problem situation.

d. The Minitab output in part (b) indicates that there are two unusual observations:

* Observation 3 (Smithsonian Institution) is an outlier because it has a large standardized residual.
* Observation 5 (American Cancer Society) is an influential observation becasuse has high leverage.

Although fundraising expenses for the Smithsonian Institution are on the low side as compared to most of the other super-sized charities, the percentage spent on program expenses appears to be much lower than one would expect. It appears that the Smithsonian’s administrative expenses are too high. But, thinking about the expenses of running a large museum like the Smithsonian, the percetage spent on administrative expenses may not be unreasonable and is just due to the fact that operating costs for a museum are in general higher than for some other types of organizations. The very large value of fundraising expenses for the American Cancer Society suggests that this obervation has a large influence on the estiamted regresion equation. The following Minitab output shows the results if this observatoin is deleted from the original data.

The regression equation is

Program Expenses (%) = 91.3 - 1.00 Fundraising Expenses (%)

Predictor Coef SE Coef T P

Constant 91.256 3.654 24.98 0.000

Fundraising Expenses (%) -1.0026 0.5590 -1.79 0.116

S = 7.96708 R-Sq = 31.5% R-Sq(adj) = 21.7%

The y-intercept has changed slightly, but the slope has changed from -.917 to -1.00.

53. a.

b. There appears to be a positive relationship between the two variables. But, observation 9 (U.S.) appears to be an observation with high leverage and may be very influential in terms of fitting a linear model to the data.

c. The Minitab output follows.

The regression equation is

Debt = 49.1 + 0.123 Gold Value

Predictor Coef SE Coef T P

Constant 49.08 15.12 3.25 0.014

Gold Value 0.12299 0.07847 1.57 0.161

S = 32.0394 R-Sq = 26.0% R-Sq(adj) = 15.4%

Analysis of Variance

Source DF SS MS F P

Regression 1 2522 2522 2.46 0.161

Residual Error 7 7186 1027

Total 8 9708

Unusual Observations

Gold

Obs Value Debt Fit SE Fit Residual St Resid

9 487 93.2 109.0 29.5 -15.8 -1.27 X

X denotes an observation whose X value gives it large leverage.

d. The Minitab output identifies observation 9 as an observation whose x value gives it large leverage.

e. Looking at the scatter diagram in part (a) it looks like observation 9 will have a lot of influence on the estimated regression equation. To investigate this we can simply drop the observation from the data set and fit a new estimated regression equation. The Minitab output we obtained follows.

The regression equation is

Debt = 30.8 + 0.342 Gold Value

Predictor Coef SE Coef T P

Constant 30.77 19.85 1.55 0.172

Gold Value 0.3422 0.1804 1.90 0.107

S = 30.3907 R-Sq = 37.5% R-Sq(adj) = 27.1%

Analysis of Variance

Source DF SS MS F P

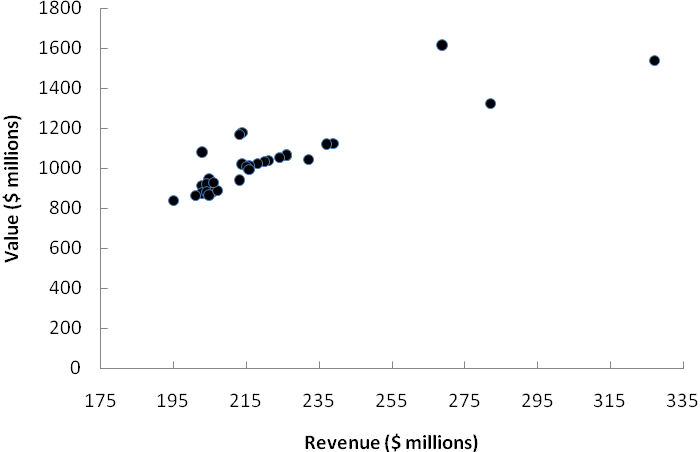
Regression 1 3324.2 3324.2 3.60 0.107

Residual Error 6 5541.6 923.6

Total 7 8865.7

Note that the slope of the estimated regression equation is now .342 as compared to a value of .123 when this observation is included. Thus, we see that this observation has a big impact on the value of the slope of the fitted line and hence we would say that it is an influential observation.

54. a.



The scatter diagram does indicate potential outliers and/or influential observations. For example, the data for the Washington Redskins, New England Patriots, and the Dallas Cowboys not only have the three highest revenues, they also have the highest team values.

b. A portion of the Minitab output follows:

The regression equation is

Value = - 252 + 5.83 Revenue

Predictor Coef SE Coef T P

Constant -252.1 130.8 -1.93 0.064

Revenue 5.8317 0.5863 9.95 0.000

S = 87.2441 R-Sq = 76.7% R-Sq(adj) = 76.0%

Analysis of Variance

Source DF SS MS F P

Regression 1 753008 753008 98.93 0.000

Residual Error 30 228346 7612

Total 31 981354

Unusual Observations

Obs Revenue Value Fit SE Fit Residual St Resid

9 269 1612.0 1316.6 31.8 295.4 3.64R

19 282 1324.0 1392.5 38.6 -68.5 -0.88 X

21 214 1178.0 995.9 16.0 182.1 2.12R

22 213 1170.0 990.1 16.2 179.9 2.10R

32 327 1538.0 1654.9 63.7 -116.9 -1.96 X

R denotes an observation with a large standardized residual.

X denotes an observation whose X value gives it large leverage.

c. The Minitab output indicates that there are five unusual observations:

* Observation 9 (Dallas Cowboys) is an outlier because it has a large standardized residual.
* Observation 19 (New England Patriots) is an influential observation becasuse has high leverage.
* Observation 21 (New York Giants) is an outlier because it has a large standardized residual.
* Observation 22 (New York Jets) is an outlier because it has a large standardized residual.
* Observation 32 (Washington Redskins) is an influential observation becasuse has high leverage.

55. No. Regression or correlation analysis can never prove that two variables are causally related.

56. The estimate of a mean value is an estimate of the average of all *y* values associated with the same *x*. The estimate of an individual *y* value is an estimate of only one of the *y* values associated with a particular *x*.

57. The purpose of testing whetheris to determine whether or not there is a significant relationship between *x* and *y*. However, rejectingdoes not necessarily imply a good fit. For example, if is rejected and *r*2 is low, there is a statistically significant relationship between *x* and *y* but the fit is not very good.

58. a.

b. A portion of the Minitab output is shown below:

The regression equation is

S&P = - 669 + 0.157 DJIA

Predictor Coef SE Coef T P

Constant -669.0 130.7 -5.12 0.000

DJIA 0.15727 0.01015 15.49 0.000

S = 9.60811 R-Sq = 94.9% R-Sq(adj) = 94.5%

Analysis of Variance

Source DF SS MS F P

Regression 1 22146 22146 239.89 0.000

Residual Error 13 1200 92

Total 14 23346

c. Using the *F* test, the *p*-value corresponding to *F* = 239.89 is .000. Because the *p*-value =.05, we reject ; there is a significant relationship.

d. With R-Sq = 94.9%, the estimated regression equation provided an excellent fit.

e. 

f. The DJIA is not that far beyond the range of the data. With the excellent fit provided by the estimated regression equation, we should not be too concerned about using the estimated regression equation to predict the S&P500.

59. a. The Minitab output is shown below:

The regression equation is

Share Price ($) = - 2.99 + 0.911 Fair Value ($)

Predictor Coef SE Coef T P

Constant -2.987 5.791 -0.52 0.610

Fair Value ($) 0.91128 0.09783 9.31 0.000

S = 12.0064 R-Sq = 76.9% R-Sq(adj) = 76.1%

Analysis of Variance

Source DF SS MS F P

Regression 1 12507 12507 86.76 0.000

Residual Error 26 3748 144

Total 27 16255

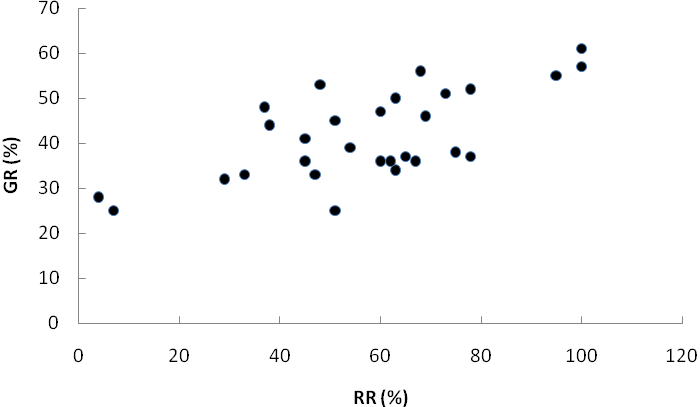
= -2.987 + .91128 Fair Value ($)

b. Significant relationship: *p*-value = .000 < *α* = .05

c. = -2.987 + .91128 Fair Value ($) = -2.987 + .91128(50) = 42.577 or approximately $42.58

d. The estimated regression equation should provide a good estimate because *r*2 = 0.769

60. a.



The scatter diagram indicates a positive linear relationship between the two variables. Online universities with higher retention rates tend to have higher graduation rates.

b. The Minitab output follows:

The regression equation is

GR(%) = 25.4 + 0.285 RR(%)

Predictor Coef SE Coef T P

Constant 25.423 3.746 6.79 0.000

RR(%) 0.28453 0.06063 4.69 0.000

S = 7.45610 R-Sq = 44.9% R-Sq(adj) = 42.9%

Analysis of Variance

Source DF SS MS F P

Regression 1 1224.3 1224.3 22.02 0.000

Residual Error 27 1501.0 55.6

Total 28 2725.3

Unusual Observations

Obs RR(%) GR(%) Fit SE Fit Residual St Resid

2 51 25.00 39.93 1.44 -14.93 -2.04R

3 4 28.00 26.56 3.52 1.44 0.22 X

R denotes an observation with a large standardized residual.

X denotes an observation whose X value gives it large leverage.

c. Because the *p*-value = .000 < *α* =.05, the relationship is significant.

d. The estimated regression equation is able to explain 44.9% of the variability in the graduation rate based upon the linear relationship with the retention rate. It is not a great fit, but given the type of data, the fit is reasonably good.

e. In the Minitab output in part (b), South University is identified as an observation with a large standardized residual. With a retention rate of 51% it does appear that the graduation rate of 25% is low as compared to the results for other online universities. The president of South University should be concerned after looking at the data. Using the estimated regression equation, we estimate that the gradation rate at South University should be 25.4 + .285(51) = 40%.

f. In the Minitab output in part (b), the University of Phoenix is identified as an observation whose x value gives it large influence. With a retention rate of only 4%, the president of the University of Phoenix should be concerned after looking at the data.

61. The Minitab output is shown below:

The regression equation is

Expense = 10.5 + 0.953 Usage

Predictor Coef SE Coef T p

Constant 10.528 3.745 2.81 0.023

X 0.9534 0.1382 6.90 0.000

S = 4.250 R-sq = 85.6% R-sq(adj) = 83.8%

Analysis of Variance

SOURCE DF SS MS F p

Regression 1 860.05 860.05 47.62 0.000

Residual Error 8 144.47 18.06

Total 9 1004.53

Fit Stdev.Fit 95% C.I. 95% P.I.

39.13 1.49 ( 35.69, 42.57) ( 28.74, 49.52)

a. = 10.528 + .9534 Usage

b. Since the *p*-value corresponding to *F* = 47.62 = .000 < ** = .05, we reject H0: *β*1 = 0.

c. The 95% prediction interval is 28.74 to 49.52 or $2874 to $4952

d. Yes, since the expected expense is = 10.528 + .9534(30) = 39.13 or $3913.

62. a. The Minitab output is shown below:

The regression equation is

Defects = 22.2 - 0.148 Speed

Predictor Coef SE Coef T P

Constant 22.174 1.653 13.42 0.000

Speed -0.14783 0.04391 -3.37 0.028

S = 1.489 R-Sq = 73.9% R-Sq(adj) = 67.4%

Analysis of Variance

Source DF SS MS F P

Regression 1 25.130 25.130 11.33 0.028

Residual Error 4 8.870 2.217

Total 5 34.000

Predicted Values for New Observations

New Obs Fit SE Fit 95.0% CI 95.0% PI

1 14.783 0.896 ( 12.294, 17.271) ( 9.957, 19.608)

b. Since the *p*-value corresponding to *F* = 11.33 = .028 < *α* = .05, the relationship is significant.

c.  = .739; a good fit. The least squares line explained 73.9% of the variability in the number of defects.

d. Using the Minitab output in part (a), the 95% confidence interval is 12.294 to 17.271.

63. a.

There appears to be a negative linear relationship between distance to work and number of days absent.

b. The Minitab output is shown below:

The regression equation is

Days = 8.10 - 0.344 Distance

Predictor Coef SE Coef T p

Constant 8.0978 0.8088 10.01 0.000

X -0.34420 0.07761 -4.43 0.002

S = 1.289 R-sq = 71.1% R-sq(adj) = 67.5%

Analysis of Variance

SOURCE DF SS MS F p

Regression 1 32.699 32.699 19.67 0.002

Residual Error 8 13.301 1.663

Total 9 46.000

Fit Stdev.Fit 95% C.I. 95% P.I.

6.377 0.512 ( 5.195, 7.559) ( 3.176, 9.577)

c. Since the *p*-value corresponding to *F* = 419.67 is .002 <  = .05. We reject H0 : *β*1 = 0.

There is a significant relationship between the number of days absent and the distance to work.

d. *r*2 = .711. The estimated regression equation explained 71.1% of the variability in *y*; this is a reasonably good fit.

e. The 95% confidence interval is 5.195 to 7.559 or approximately 5.2 to 7.6 days.

64. a. The Minitab output is shown below:

The regression equation is

Cost = 220 + 132 Age

Predictor Coef SE Coef T p

Constant 220.00 58.48 3.76 0.006

X 131.67 17.80 7.40 0.000

S = 75.50 R-sq = 87.3% R-sq(adj) = 85.7%

Analysis of Variance

SOURCE DF SS MS F p

Regression 1 312050 312050 54.75 0.000

Residual Error 8 45600 5700

Total 9 357650

Fit Stdev.Fit 95% C.I. 95% P.I.

746.7 29.8 ( 678.0, 815.4) ( 559.5, 933.9)

b. Since the *p*-value corresponding to *F* = 54.75 is .000 < ** = .05, we reject H0: *β*1 = 0.

Maintenance cost and age of bus are related.

c. *r*2 = .873. The least squares line provided a very good fit.

d. The 95% prediction interval is 559.5 to 933.9 or $559.50 to $933.90

65. a. The Minitab output is shown below:

The regression equation is

Points = 5.85 + 0.830 Hours

Predictor Coef SE Coef T p

Constant 5.847 7.972 0.73 0.484

X 0.8295 0.1095 7.58 0.000

S = 7.523 R-sq = 87.8% R-sq(adj) = 86.2%

Analysis of Variance

SOURCE DF SS MS F p

Regression 1 3249.7 3249.7 57.42 0.000

Residual Error 8 452.8 56.6

Total 9 3702.5

Fit Stdev.Fit 95% C.I. 95% P.I.

84.65 3.67 ( 76.19, 93.11) ( 65.35, 103.96)

b. Since the *p*-value corresponding to *F* = 57.42 is .000 < ** = .05, we reject H0: *β*1 = 0.

Total points earned is related to the hours spent studying.

c. 84.65 points

d. The 95% prediction interval is 65.35 to 103.96

66. a. The Minitab output is shown below:

The regression equation is

Horizon = 0.275 + 0.950 S&P 500

Predictor Coef SE Coef T P

Constant 0.2747 0.9004 0.31 0.768

S&P 500 0.9498 0.3569 2.66 0.029

S = 2.664 R-Sq = 47.0% R-Sq(adj) = 40.3%

Analysis of Variance

Source DF SS MS F P

Regression 1 50.255 50.255 7.08 0.029

Residual Error 8 56.781 7.098

Total 9 107.036

The market beta for Horizon is *b*1 = .95

b. Since the *p*-value = 0.029 is less than ** = .05, the relationship is significant.

c. *r*2 = .470. The least squares line does not provide a very good fit.

d. Xerox has higher risk with a market beta of 1.22.

67. a. The Minitab output is shown below:

The regression equation is

Audit% = - 0.471 +0.000039 Income

Predictor Coef SE Coef T P

Constant -0.4710 0.5842 -0.81 0.431

Income 0.00003868 0.00001731 2.23 0.038

S = 0.2088 R-Sq = 21.7% R-Sq(adj) = 17.4%

Analysis of Variance

Source DF SS MS F P

Regression 1 0.21749 0.21749 4.99 0.038

Residual Error 18 0.78451 0.04358

Total 19 1.00200

Predicted Values for New Observations

New Obs Fit SE Fit 95.0% CI 95.0% PI

1 0.8828 0.0523 ( 0.7729, 0.9927) ( 0.4306, 1.3349)

b. Since the *p*-value = 0.038 is less than ** = .05, the relationship is significant.

c. *r*2 = .217. The least squares line does not provide a very good fit.

d. The 95% confidence interval is .7729 to .9927.

68. a.

b. There appears to be a negative relationship between the two variables that can be approximated by a straight line. An argument could also be made that the relationship is perhaps curvilinear because at some point a car has so many miles that its value becomes very small.

c. The Minitab output is shown below.

The regression equation is

Price ($1000s) = 16.5 - 0.0588 Miles (1000s)

Predictor Coef SE Coef T P

Constant 16.4698 0.9488 17.36 0.000

Miles (1000s) -0.05877 0.01319 -4.46 0.000

S = 1.54138 R-Sq = 53.9% R-Sq(adj) = 51.2%

Analysis of Variance

Source DF SS MS F P

Regression 1 47.158 47.158 19.85 0.000

Residual Error 17 40.389 2.376

Total 18 87.547

d. Significant relationship: *p*-value = 0.000 < *α* = .05.

e. = .539; a reasonably good fit considering that the condition of the car is also an important factor in what the price is.

f. The slope of the estimated regression equation is -.0558. Thus, a one-unit increase in the value of x coincides with a decrease in the value of y equal to .0558. Because the data were recorded in thousands, every additional 1000 miles on the car’s odometer will result in a $55.80 decrease in the predicted price.

g. The predicted price for a 2007 Camry with 60,000 miles is = 16.5 -.0588(60) = 12.97 or approximately $13,000. Because of other factors, such as condition and whether the seller is a private party or a dealer, this is probably not the price you would offer for the car. But, it should be a good starting point in figuring out what to offer the seller.