Simple Linear Regression

11.1 Creating the Least Squares Equation

To complete this section of homework watch Chapter Eleven, Lecture Examples: 160, 161, 162, 163, 161tech, and 162tech.

1. An educator wants to see if there is a relationship between the number of absences for a student and the student's final grade. Use the data below to find the least squares prediction line and to answer questions 3 and 4: VS

Absences	10	12	2	0	8	5
Final Grade	70	65	96	94	75	82

2. Use the output provided from Minitab below to form a least squares model to predict the average final grade based on the number of absences, then use the model to determine the average grade for students who missed 6 days of class?

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	96.784	1.994	48.54	<0.0001
Absences	-2.6677	0.2661	-10.03	0.0006

3. What does x = 0 represent in the model?

4. What average grade does the model predict for students who have perfect attendance?

5. The following set of data is randomly selected from a STA 2122 class of mine from spring 2010. The list includes clicker points earned in class (clickers allow students to answer questions in class and to have their answers scored) and their final averages. Use the data to find the least squares prediction line: ($\sum x = 171, \sum x^2 = 4,737, \sum y = 648, \sum y^2 = 53,856, \sum xy = 15,013$) VS

Clicker points	32	11	34	41	16	15	7	15
Class Average	99	70	91	101	79	72	68	68

6. I plugged an entire class of 200 students into SPSS in order to calculate the least squares line to predict class average using the number of clicker points earned over the term. The results were as follows: $\hat{y} = 0.512x + 70.196$. VS

: *indicates the exercise has a video devoted to it in the corresponding section of STATSprofessor.com*

a. What does x = 0 represent here?
b. What is the expected grade for students who do not earn any clicker points?
c. What is the average grade for a student who has 20 clicker points?
d. I pulled the grade and clicker points for a randomly chosen student from a different class of statistics from the spring term 2010. That student had 35 clicker points and a 94% in the class. Plug 35 points into our model and determine the prediction error for this case.

7. The following table compares age at death and systolic blood pressure. Use the data to find the least squares prediction line and to answer questions 8 and 9:
$(\sum x = 2678, \sum x^2 = 365,446, \sum y = 1166, \sum y^2 = 69,346, \sum xy = 153,860)$

BP	age at death		BP	age at death
158	46		134	59
157	46		157	59
157	49		150	62
160	49		117	62
131	50		126	64
138	51		109	65
160	53		120	68
122	54		111	69
123	57		107	71
122	58		119	74

8. What if any interpretation do we have for x = 0 in the model below?

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	102.546	9.569	10.72	<0.0001
BloodPressure	-0.33044	0.07079	-4.67	0.0002

Regression Equation

AgeAtDeath = 102.546 − 0.33044 BloodPressure

9. What is the expected age at death for people with a systolic blood pressure of 140?

10. What is the expected age at death for people with a systolic blood pressure of 159?

11. Use the data below to create the least squares prediction line and to predict the average weight for supermodels that are 69 inches tall.
$\sum x = 632, \sum x^2 = 44,399.5, \sum y = 1089, \sum y^2 = 132,223, \sum xy = 76,546$

Height	71	70.5	71	72	70	70	66.5	70	71
Weight	125	119	128	128	119	127	105	123	115

12. The following computer output from Minitab is for a least squares prediction line, which is used to predict the average weight in pounds of supermodels given their height in inches. Interpret the slope for this linear model.

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	-151.70	78.50	-1.93	0.0946
Height	3.883	1.118	3.47	0.0103

13. Use the data below to create the least squares prediction line for predicting the best finishing time of the New York City marathon given the temperature.
$\sum x = 478, \sum x^2 = 29,070, \sum y = 1,176.617, \sum y^2 = 173,068.7, \sum xy = 70,318.99$ VS

Temp	55	61	49	62	70	73	51	57
Time	145.283	148.717	148.3	148.1	147.617	146.4	144.667	147.533

14. The model below is the least squares prediction line for predicting the best finishing time of the New York City marathon given the temperature. Find the prediction error for the actual best finish time in 1990 which was 150.75 minutes with a temperature of 73 degrees.

Regression Equation

Time = 145.186 + 0.03165 Temperature

15. True or False: The least squares line has the smallest sum of squared errors when compared to all other linear models.

16. Why do we say that the least squares line provides the “best fit” of any linear model?

17. True or False: The sum of errors for the least squares line is always positive.

18. What is the sum of all the errors, $\sum(y-\hat{y})$, made by any least squares line?

19. True or False: The strength of the linear association between two variables is not important when using a least squares prediction model.

20. A researcher collected data involving the frequency of chirps made by a ground cricket, at various ground temperatures. He showed that there was a significant linear relationship between temperature and the frequency of chirps. The data he used is below. Would it be a good idea to use the model to estimate the average chirp frequency when the temperature was 51 degrees? Why or why not? VS

Temp	88.6	71.6	93.3	84.3	80.6	75.2	69.7	82	69.4	83.3	78.6	82.6	80.6	83.5	76.3
Chirps	20	16	19.8	18.4	17.1	15.5	14.7	17.1	15.4	16.2	15	17.2	16	17	14.1

11.1 Answers

1. $\hat{y} = -2.67x + 96.78$

2. solution:
$$\hat{y} = -2.6677x + 96.784$$
$$\hat{y} = -2.6677(6) + 96.784 = 80.8$$
$$\hat{y}(6) = 80.8$$

3. It represents a student who missed zero class days = perfect attendance.

4. A 96.78%.

5. $\hat{y} = 1.07x + 58.04$

6. a. no clicker points b. 70.196 c. 80.436 d. 5.884

7. $\hat{y} = -0.330x + 102.546$

8. There is no practical interpretation for x = 0, since that would imply the age at death for a person with zero blood pressure (which would basically mean your heart stopped working).

9. About 56 years old.

10. About 50 years old.

11. $\hat{y} = 3.88x - 152$; 116 lbs

12. The slope indicates that for a unit increase in height, there is a corresponding average increase in weight of 3.883 pounds in supermodels.

13. $\hat{y} = 0.032x + 145.2$

14. 150.75 - 147.536 = 3.214 minutes. This may not seem too bad, but this is a lot of error since the standard deviation for the eight finishing times is only 1.47. The reason for the poor performance is that there is not a strong linear relationship between the two variables. That is the next phase of our work in this section: to measure the strength of the linear relationship.

15. True, the SSE for the least squares line is a minimum.

16. Because the least square line has the minimum SSE of any linear model for a given set of data.

17. False, the SE is always zero for the least squares line.

18. The SE is always equal to zero for the least squares line.

19. False, the strength of the linear relationship is very important. If the linear relationship is very weak, it is usually not wise to use a linear model for prediction.

20. It is not a good idea since 51 degrees is far outside of the data range we used to create the model.

Need more exercises?

11.2 Finding S for the Random Error Terms

To complete this section of homework watch Chapter Eleven, Lecture Examples: 164, 165*, and *165tech.

21. An educator wants to see if there is a relationship between the number of absences for a student and the student's final grade. Use the data below to find the S^2 for the random error of the least squares prediction line: VS

Absences	10	12	2	0	8	5
Final Grade	70	65	96	94	75	82

22. Study the Minitab output below, which provides the least squares line for data relating the number of absences for a student and the student's final grade. Find the S^2 for the random error of the least squares prediction line.

Model Summary

S	R-sq
2.93598	95.63%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	97.771	1.743	56.08	<0.0001	
Absences	-2.7887	0.2108	-13.23	<0.0001	1.00

Regression Equation

Grade = 97.771 − 2.7887 Absences

23. The following set of data is randomly selected from a STA 2122 class of mine from spring 2010. The list includes clicker points earned in class and their final averages. Use the data below to find the S^2 for the random error of the least squares prediction line: VS

$$\left(\sum x = 171, \sum x^2 = 4,737, \sum y = 648, \sum y^2 = 53,856, \sum xy = 15,013\right)$$

Clicker points	32	11	34	41	16	15	7	15
Class Average	99	70	91	101	79	72	68	68

24. The following ANOVA table was created from the age at death and systolic blood pressure data we looked at in the last section. Can you use the ANOVA table to find the S^2 for the random error of the least squares prediction line? Note: $S^2 = \dfrac{SSE}{n-2}$

Analysis of Variance

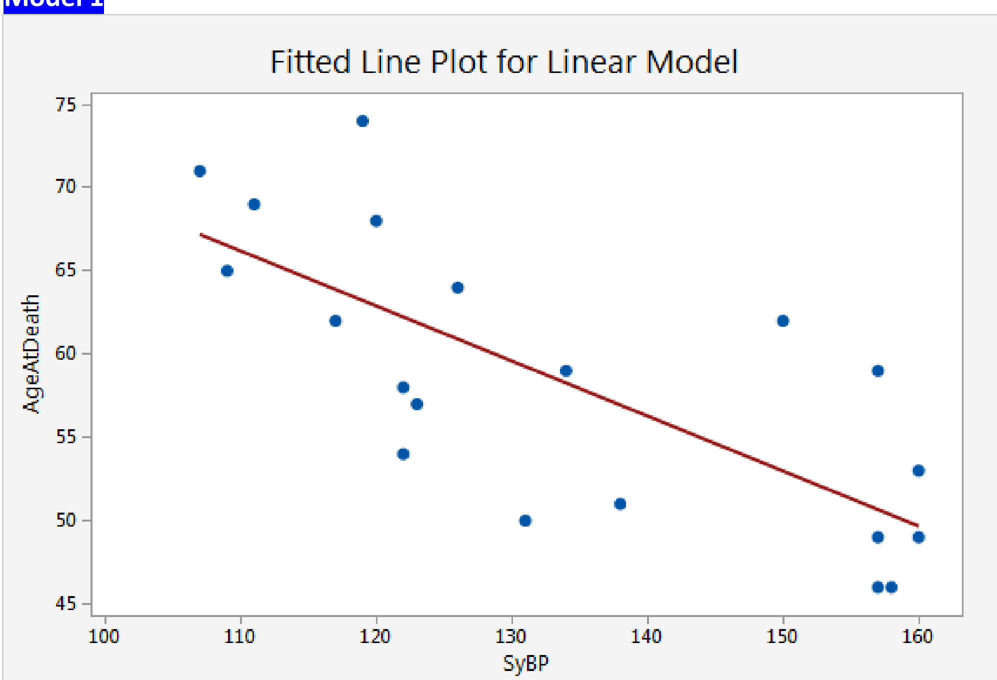
Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	749.24	749.235	21.79	0.0002
SyBP	1	749.24	749.235	21.79	0.0002
Error	18	618.96	34.387		
Total	19	1368.20			

25. The following table compares age at death and systolic blood pressure. Use the data below to find the S^2 for the random error of the least squares prediction line: (n = 20 $\sum x = 2678, \sum x^2 = 365,446, \sum y = 1166, \sum y^2 = 69,346$, $\sum xy = 153,860$)

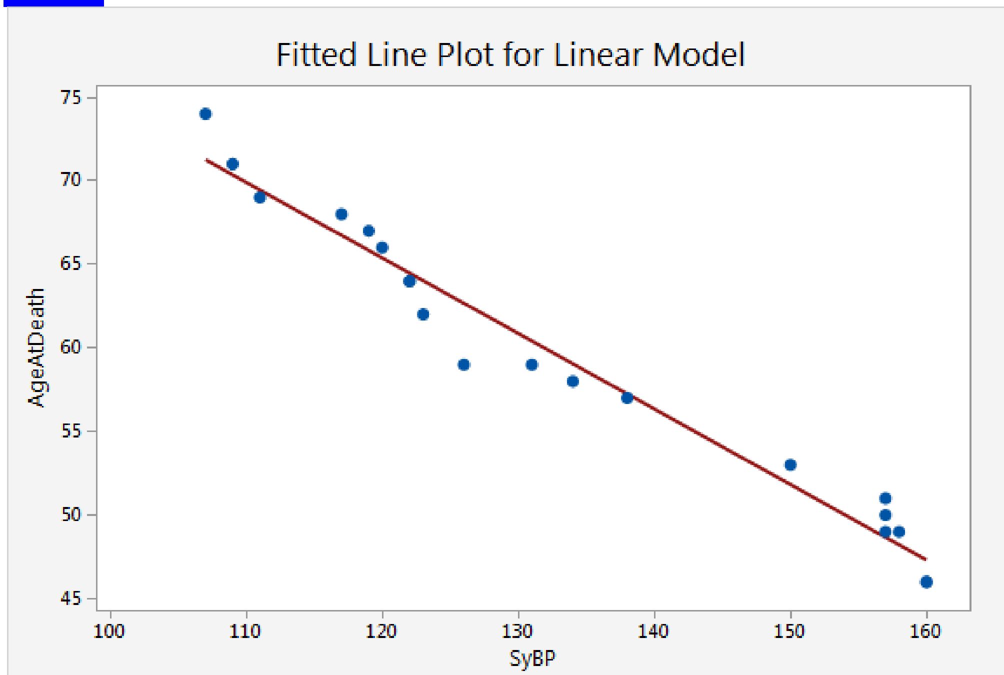
BP	age at death		BP	age at death
158	46		134	59
157	46		157	59
157	49		150	62
160	49		117	62
131	50		126	64
138	51		109	65
160	53		120	68
122	54		111	69
123	57		107	71
122	58		119	74

26. Consider the following set of graphs (model 1 and 2). For one model, the $S^2 = 2.53$, and for the other $S^2 = 34.387$. Can you tell which model has the smaller S^2 value?

Model 1



Model 2



27. Use the data below to find the S^2 for the random error of the least squares prediction line used to predict the average weight for super models given their height: VS
($\sum x = 632, \sum x^2 = 44{,}399.5, \sum y = 1089, \sum y^2 = 132{,}223, \sum xy = 76{,}546$)

Height	71	70.5	71	72	70	70	66.5	70	71
Weight	125	119	128	128	119	127	105	123	115

28. In the problem involving data on absenteeism and grades, we found that $S^2 = 7.7037$. Find S and then state the largest deviation we would expect between any of the actual data points and our least squares line.

29. Use the data below to find the S^2 for the random error of the least squares prediction line: The data includes the finishing time of the New York City marathon and the temperature on the day of the race. ($\sum x = 478, \sum x^2 = 29{,}070, \sum y = 1{,}176.617, \sum y^2 = 173{,}068.7$, $\sum xy = 70{,}318.99$)

Temp	55	61	49	62	70	73	51	57
Time	145.283	148.717	148.3	148.1	147.617	146.4	144.667	147.533

30. We found $S^2 = 23.804$ for the data used to predict the average weight for super models given their height. Find S and then state the largest deviation we would expect between any of the actual data points and our least squares line. VS

: indicates the exercise has a video devoted to it in the corresponding section of STATSprofessor.com

11.2 Answers

21. $S^2 = 7.7037$

SSXX	108.8333	B1	-2.6677
SSXY	-290.333	SSE	30.8147
SSYY	805.3333		

22. $S^2 = 2.93598^2 = 8.61998$

23. $S^2 = 19.990$

SSXX	1081.875	B1	1.074061
SSXY	1162	SSE	119.9408
SSYY	1368		

24. Since $S^2 = \dfrac{SSE}{n-2}$ it is the same as the sum of square for error divided by its degrees of freedom, or, in other words, it is the same as the MSE from our output. $S^2 = 34.387$

25. $S^2 = 34.387$

SSXX	6861.8	B1	-0.33044
SSXY	-2267.4	SSE	618.9647
SSYY	1368.2		

26. You can see that the values hug the line more closely in model 2, so that is the model with the smaller s value.

27. $S^2 = 23.804$

SSXX	19.05556	B1	3.883382
SSXY	74	SSE	166.6297
SSYY	454		

28. We would expect 95% of all of our observed values to lie within 5.551 of the least squares line.

29. $SS_{xy} = 16.12425, SS_{xx} = 509.5, SS_{yy} = 15.21626$, SSE = 14.70597, $S^2 = 2.45099$

30. We would expect 95% of all of our observed values to lie within 9.758 of the least squares line.

Need more exercises?

: *indicates the exercise has a video devoted to it in the corresponding section of STATSprofessor.com*

11.3 Finding the Standard Error of the Slope Estimator

To complete this section of homework watch Chapter Eleven, Lecture Example 165.5 and 165.5tech.

31. Use the data below to find $S_{\hat{\beta}_1}$ the standard error of the slope estimator. VS

X	1	2	3	4	5
Y	2	2	4	3	6

32. In the previous sections, we looked at data relating age at death and systolic blood pressure. Can you use the Minitab output below to identify, $S_{\hat{\beta}_1}$, the standard error of the slope estimator?

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	102.546	9.569	10.72	<0.0001
SyBP	-0.33044	0.07079	-4.67	0.0002

11.3 Answers

31. $n = 5, SS_{xx} = 10, SS_{xy} = 9, SS_{yy} = 11.2$, $SSE = 11.2 - \frac{9}{10} * 9 = 3.1, S^2 = \frac{3.1}{3} = 1.033$,

$S = 1.01653, S_{\hat{\beta}_1} = \frac{1.01653}{\sqrt{10}} = 0.3215$

32. The standard error for the terms in the model can be found under the SE Coef column of the output. The slope estimator is found in the SyBP row for this output. $S_{\hat{\beta}_1} = 0.07079$

Need more exercises?

11.4 Hypothesis Tests about the Slope β1

To complete this section of homework watch Chapter Eleven, Lecture Examples 166, 167, 166tech, and 167tech.

33. Pierce (1949) measured the frequency (the number of wing vibrations per second) of chirps made by a ground cricket, at various ground temperatures. Since crickets are ectotherms (cold-blooded), the rate of their physiological processes and their overall metabolism are influenced by temperature. Consequently, there is reason to believe that temperature would have a profound effect on aspects of their behavior, such as chirp frequency. Use a 1% significance level to test the claim that there is a positive linear relationship between temperature and the frequency of chirps. ($SS_{xx} = 631.64933, SS_{xy} = 136.23333, SS_{yy} = 41.993333$)

Temp	88.6	71.6	93.3	84.3	80.6	75.2	69.7	82	69.4	83.3	78.6	82.6	80.6	83.5	76.3
Chirps	20	16	19.8	18.4	17.1	15.5	14.7	17.1	15.4	16.2	15	17.2	16	17	14.1

: indicates the exercise has a video devoted to it in the corresponding section of STATSprofessor.com

34. A study looking at weight and blood fat content was conducted on a random selection of 20 participants. The Minitab output for the data is provided below. Use the results and a 10% significance level to test the claim that there is a positive linear relationship between weight and the amount of fat content in the blood. Do these variables appear to be strongly associated?

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	195.67	95.57	2.05	0.0555
Weight	0.8323	0.6155	1.35	0.1931

Regression Equation

BloodFatContent = 195.67 + 0.8323 Weight

35. We tend to think that our height has something to do with the heights of our parents. Perhaps there is a linear relationship between a father's height and the height of his son. Heights of fathers and their sons are included below. Use a 5% significance level to test the claim that there is a positive relationship between the heights.

$SS_{xx} = 160.9333333, SS_{xy} = 41.17333333, SS_{yy} = 124.1493333$ VS

Father H	70	69	64	71	68	66	74	73	62	69	67	68	72	66	72
Son's H	62.5	64.6	69.1	73.9	67.1	64.4	71.1	71	67.4	69.3	64.9	68.1	66.5	67.5	66.5

36. In 46 states, data on liquor consumption per capita and liver disease death rates was collected. Minitab output has been provided below for the data. Use the results and a 5% significance level to test the claim that there is a positive linear relationship between the number of liquor drinks per capita consumed and the death rate from liver disease. Does the results imply that drinking too much causes liver disease?

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	21.965	7.185	3.06	0.0038
Drinks	0.7222	0.1168	6.19	<0.0001

37. Is age a predictor of male waist sizes? As men age their metabolism slows, so it makes sense that as men age, their waistlines expand. Use the data below and a 1% significance level to test for a positive linear relationship between age and waist circumference (in cm).

$SS_{xx} = 3811.214286, SS_{xy} = 1104.007143, SS_{yy} = 942.0035714$ VS

Age	58	22	32	31	28	46	41	56	20	54	17	73	52	25
Waist	90.6	78.1	96.5	87.7	87.1	92.4	78.8	103.3	89.1	82.5	86.7	103.3	91.8	75.6

: indicates the exercise has a video devoted to it in the corresponding section of STATSprofessor.com

38. A real-estate agent looked at the sales price of 28 randomly selected homes and the age of the homes to determine if there was a negative linear relationship between the age of a home and its selling price. The results of her analysis are provided below. Use the results and a 10% significance level to test the claim that there is a negative linear relationship between the age of a home and its selling price.

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	49.296	7.434	6.63	<0.0001
Age	-0.3067	0.1917	-1.60	0.1218

39. Abdominal fat is dangerous for women (and men). It has been linked to heart disease. The data below lists waist measurements for women and their overall cholesterol level. At the 2% significance level, test for a linear relationship between waist circumference in women and their cholesterol level. ($SS_{xx} = 1860.944, SS_{xy} = 8698.371, SS_{yy} = 221,713.4$) VS

Waist	67.2	82.5	66.7	93	82.6	75.4	73.6	81.4	99.4	67.7	100.7	99.3	85.7	85.7
Cholest.	264	181	267	384	98	62	126	89	531	130	175	280	149	112

40. True or False: If we are unable to reject the null hypothesis when testing if the slope parameter is equal to zero, this indicates that there is no relationship between the two variables in the model.

41. Is there a relationship between being overweight and an early death? Use the data below and a 2.5% significance level to test for a negative linear relationship between the number of pounds overweight and the age at death. ($SS_{xx} = 2933.8, SS_{xy} = -1577.4, SS_{yy} = 1368.2$)

Pound over weight	Age at Death	X2	Y2	XY
40	46	1600	2116	1840
29	46	841	2116	1334
30	49	900	2401	1470
6	49	36	2401	294
18	50	324	2500	900
16	51	256	2601	816
0	53	0	2809	0
13	54	169	2916	702
1	57	1	3249	57
12	58	144	3364	696
14	59	196	3481	826
1	59	1	3481	59
2	62	4	3844	124
3	62	9	3844	186
1	64	1	4096	64
0	65	0	4225	0
1	68	1	4624	68
-3	69	9	4761	-207
-5	71	25	5041	-355
-1	74	1	5476	-74

11.4 Answers

33. $Claim: \beta_1 > 0$

$H_0: \beta_1 \leq 0$

$H_A: \beta_1 > 0$

SSXX	631.6493333		B1	0.215679		Sb1	0.039189
SSXY	136.2333333		SSE	12.6107		test stat	5.503607
SSYY	41.99333333		s2	0.970054			

$Critical Value: 2.650$

$Initial Conclusion:$ Reject the null, support the alternative.

There is sufficient evidence to support the claim that there is a positive linear relationship.

34.

$Claim: \beta_1 > 0$

$H_0: \beta_1 \leq 0$

$H_A: \beta_1 > 0$

p-value = 0.1931

Since the p-value is larger than 0.10, we do not reject the null hypothesis. This indicates that we cannot support the notion of a positive linear relationship/association for these two variables.

35. $Claim: \beta_1 > 0$

$H_0: \beta_1 \leq 0$

$H_A: \beta_1 > 0$

SSXX	160.9333		B1	0.255841		Sb1	0.233037
SSXY	41.17333		SSE	113.6155		test stat	1.097857
SSYY	124.1493		s2	8.739655			

$Critical Value: 1.771$

$Initial Conclusion:$ Do not reject the null, do not support the alternative.

There is not sufficient evidence to support the claim that there is a positive linear relationship.

36. Solution:

$Claim: \beta_1 > 0$

$H_0: \beta_1 \leq 0$

$H_A: \beta_1 > 0$

p-value = less than 0.0001
Since the p-value is less than 0.0001, we reject the null hypothesis. This indicates that we support the notion of a positive linear relationship/association for these two variables. The results indicate that drinking and liver disease appear together, but it does not indicate that drinking causes liver disease. That cannot be proven with this set of statistical tools.

37. $Claim: \beta_1 > 0$

$H_0: \beta_1 \leq 0$

$H_A: \beta_1 > 0$

SSXX	3811.214286		B1	0.289673		Sb1	0.116639
SSXY	1104.007143		SSE	622.2021		test stat	2.483505
SSYY	942.0035714		s2	51.85018			

$CriticalValue: 2.681$

$InitialConclusion:$ Do not reject the null, do not support the alternative.

There is not sufficient evidence to support the claim that there is a positive linear relationship.

38. Solution:

$Claim: \beta_1 < 0$

$H_0: \beta_1 \geq 0$

$H_A: \beta_1 < 0$

p-value = 0.1218
Since the p-value is larger than 0.10, we do not reject the null hypothesis. This indicates that we cannot support the notion of a negative linear relationship/association for these two variables.

39. $Claim: \beta_1 \neq 0$

$H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$

SSXX	1860.944		B1	4.674173		Sb1	2.847403
SSXY	8698.371		SSE	181055.7		test stat	1.641556
SSYY	221713.4		s2	15087.98			

$CriticalValues: \pm 2.681$

$InitialConclusion:$ Do not reject the null, do not support the alternative.

There is not sufficient evidence to support the claim that there is a linear relationship.

40. False: If we are unable to reject the null hypothesis when testing if the slope parameter is equal to zero, this indicates that there is no **linear** relationship between the two variables in the model.

41. $Claim: \beta_1 < 0$

$H_0: \beta_1 \geq 0$

$H_A: \beta_1 < 0$

SSXX	2933.8		B1	-0.53766		Sb1	0.09924
SSXY	-1577.4		SSE	520.0881		test stat	-5.41782
SSYY	1368.2		s2	28.89378			

$Critical Value: -2.101$

$Initial Conclusion:$ Reject the null, support the alternative.

There is sufficient evidence to support the claim that there is a negative linear relationship.

Need more exercises?

11.5 Confidence Interval for the Slope β1

To complete this section of homework watch Chapter Eleven, Lecture Example 168.

42. A study looking at weight and blood fat content was conducted on a random selection of 20 participants. The Minitab output for the data is provided below. Use the results to find the 95% confidence interval estimate for the true slope β_1 for the linear relationship between weight and the amount of fat content in the blood. What does this interval indicate?

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	195.67	95.57	2.05	0.0555
Weight	0.8323	0.6155	1.35	0.1931

43. Use the cricket data from above to find the 98% confidence interval estimate for the true slope β_1 for the linear relationship between temperature and the frequency of chirps. What does this interval indicate? (n = 15, $SS_{xx} = 631.64933, SS_{xy} = 136.23333, SS_{yy} = 41.993333$) VS

44. In 46 states, data on liquor consumption per capita and liver disease death rates was collected. Minitab output has been provided below for the data. Use the results to find the 98% confidence interval estimate for the true slope β_1 for the linear relationship between the number of liquor drinks per capita consumed and the death rate from liver disease. What does this interval indicate?

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	21.965	7.185	3.06	0.0038
Drinks	0.7222	0.1168	6.19	<0.0001

45. Use the Father/son data above to form a 90% confidence interval for the true slope β_1 for the linear relationship between a father's height and the height of his son. What does the interval indicate? (n = 15, $SS_{xx} = 160.9333333, SS_{xy} = 41.17333333, SS_{yy} = 124.1493333$) VS

46. A real-estate agent looked at the sales price of 28 randomly selected homes and the age of the homes to determine if there was a negative linear relationship between the age of a home and its selling price. The results of her analysis are provided below. Use the results to find the 90% confidence interval estimate for the true slope β_1 for the linear relationship between the age of a home and its selling price. What does this interval indicate?

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	49.296	7.434	6.63	<0.0001
Age	-0.3067	0.1917	-1.60	0.1218

47. Use the male age/waist circumference data above to form a 98% confidence interval for the true slope β_1 for the linear relationship between age and waist circumference (in cm). (n = 14, $SS_{xx} = 3811.214286, SS_{xy} = 1104.007143, SS_{yy} = 942.0035714$) VS

48. True or False: If the confidence interval for β_1 contains zero, it is possible that there is no linear relationship between the two variables in the model.

49. Use the pounds/age of death data above to form a 95% confidence interval for the true slope β_1 for the linear relationship between the number of pounds overweight and the age at death. (n = 20, $SS_{xx} = 2933.8, SS_{xy} = -1577.4, SS_{yy} = 1368.2$)

11.5 Answers

42. $E = 2.101 * 0.6155 = 1.2931655$ CI: $[-0.4609, 2.1255]$ This indicates that the slope could be zero, which would mean no linear relationship.
43. $[0.1118, 0.3195]$ This indicates a positive relationship since both limits are > 0.
44. $E = 2.423 * 0.1168 = 0.2830064$ CI: $[0.439, 1.01]$ This indicates a positive relationship since both limits are > 0.

: *indicates the exercise has a video devoted to it in the corresponding section of STATSprofessor.com*

45. $[-0.1569, 0.6685]$ This indicates that the slope could be zero, which would mean no linear relationship.
46. $E = 1.706 * 0.1917 = 0.3270402 \quad \text{CI:}\left[-0.6337, 0.0203\right]$ This indicates that the slope could be zero, which would mean no linear relationship.
47. $[-0.0230, 0.6024]$ This indicates that the slope could be zero, which would mean no linear relationship.
48. **True:** If the confidence interval for β_1 contains zero, it is possible that there is no linear relationship between the two variables in the model.
49. $[-0.7462, -0.3292]$ This indicates a negative relationship since both limits are < 0.

Need more exercises?

11.6 Finding r the Coefficient of Correlation

To complete this section of homework watch Chapter Eleven, Lecture Example 169.

50. A real-estate agent looked at the sales price of 28 randomly selected homes and the age of the homes to determine if there was a negative linear relationship between the age of a home and its selling price. The results of her analysis are provided below. Can you tell from the output if the correlation coefficient will be positive or negative? Try to find the correlation coefficient r using the provided output.

Model Summary

S	R-sq
13.7651	8.96%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	49.296	7.434	6.63	<0.0001
Age	-0.3067	0.1917	-1.60	0.1218

51. An educator wants to see if there is a relationship between the number of absences for a student and the student's final grade. Use the data below to find the correlation coefficient r.

Absences	10	12	2	0	8	5
Final Grade	70	65	96	94	75	82

52. Abdominal fat is dangerous for women (and men). It has been linked to heart disease. The data below lists waist measurements for women and their overall cholesterol level. Use the data below to find the correlation coefficient r for a linear relationship between waist circumference in women and their cholesterol level.

$n = 14, SS_{xx} = 1860.943571, SS_{xy} = 8698.371429, SS_{yy} = 221713.4286$

Waist	67.2	82.5	66.7	93	82.6	75.4	73.6	81.4	99.4	67.7	100.7	99.3	85.7	85.7
Cholest.	264	181	267	384	98	62	126	89	531	130	175	280	149	112

53. Use the data below to find the correlation coefficient r. The data includes the finishing time of the New York City marathon and the temperature on the day of the race.

$\sum x = 478, \sum x^2 = 29,070, \sum y = 1,176.617, \sum y^2 = 173,068.7$, $\sum xy = 70,318.99$

Temp	55	61	49	62	70	73	51	57
Time	145.283	148.717	148.3	148.1	147.617	146.4	144.667	147.533

54. Pierce (1949) measured the frequency (the number of wing vibrations per second) of chirps made by a ground cricket, at various ground temperatures. Since crickets are ectotherms (cold-blooded), the rate of their physiological processes and their overall metabolism are influenced by temperature. Consequently, there is reason to believe that temperature would have a profound effect on aspects of their behavior, such as chirp frequency. Use the data below to find the correlation coefficient r for the linear relationship between temperature and the frequency of chirps. ($SS_{xx} = 631.64933, SS_{xy} = 136.23333, SS_{yy} = 41.993333$)

Temp	88.6	71.6	93.3	84.3	80.6	75.2	69.7	82	69.4	83.3	78.6	82.6	80.6	83.5	76.3
Chirps	20	16	19.8	18.4	17.1	15.5	14.7	17.1	15.4	16.2	15	17.2	16	17	14.1

55. Use the data below to find the correlation coefficient r for the height/weight relationship for super models. ($\sum x = 632, \sum x^2 = 44,399.5, \sum y = 1089, \sum y^2 = 132,223$, $\sum xy = 76,546$)

Height	71	70.5	71	72	70	70	66.5	70	71
Weight	125	119	128	128	119	127	105	123	115

56. A study looking for a relationship between the area of the skin covered in tattoos and the number of sexual partners finds a correlation coefficient r = 0.987. What does this indicate? VS

57. A study looking at the demand for high speed internet and the price of the service finds that the correlation coefficient for the sample data is r = 0.012. What does this indicate?

11.6 Answers

50. Since the $r^2 = 8.96\% = 0.0896$, we can take the square root to find the absolute value of r. $|r| = \sqrt{0.0896} = 0.2993$ however, to determine the sign of r, we need to look at the slope estimator. It has a negative sign, so our answer should be negative: $r = -0.2993$.

51. r = -0.981 indicating a very strong negative linear relationship.

52. r = 0.428 indicating a moderate to weak linear relationship.

53. r = 0.183 indicating a very weak linear relationship.

54. r = 0.836 indicating a strong positive linear relationship.

55. r = 0.796 indicating a strong positive linear relationship.

56. It indicates that there is a strong positive linear relationship between the two variables. It does not mean that one causes the other. It only means that they tend to appear together in subjects.

57. It indicates very little or no linear relationship; however, it does not indicate that there is no relationship between the variables. For example, there could be a quadratic relationship between the variables.

Need more exercises?

11.7 Finding r-squared the Coefficient of Determination

To complete this section of homework watch Chapter Eleven, Lecture Examples 170, 171, and 171tech.

Use the data below to answer the next six questions:

The following data is from a study of crime in large cities.

City	Population	% earning < $5000	% unemployed	Annual # of Murders per million
1	587000	16.5	6.2	11.2
2	643000	20.5	6.4	13.4
3	635000	26.3	9.3	40.7
4	692000	16.5	5.3	5.3
5	1248000	19.2	7.3	24.8
6	643000	16.5	5.9	12.7
7	1964000	20.2	6.4	20.9
8	1531000	21.3	7.6	35.7
9	713000	17.2	4.9	8.7
10	749000	14.3	6.4	9.6
11	7895000	18.1	6	14.5
12	762000	23.1	7.4	26.9
13	2793000	19.1	5.8	15.7
14	741000	24.7	8.6	36.2
15	625000	18.6	6.5	18.1
16	854000	24.9	8.3	28.9
17	716000	17.9	6.7	14.9
18	921000	22.4	8.6	25.8
19	595000	20.2	8.4	21.7
20	3353000	16.9	6.7	25.7

58. Minitab was used to analyze the linear relationship between the unemployment rate and the percent of the city earning less than $5,000. Use the results to answer the questions that follow:

Model Summary

S	R-sq
1.92828	66.49%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	4.531	2.578	1.76	0.0958
%unemploy	2.1902	0.3665	5.98	<0.0001

a) Is there a significant linear relationship between these variables?
b) Interpret the coefficient of determination, r^2
c) Find and interpret the correlation coefficient, r.

59. Use the crime data above to calculate the correlation coefficient for the variables *"% earning < $5,000"* and "*murder rate*". Is there a linear relationship? If so, is it positive or negative? VS
Note:

SSxx	199.732
SSxy	511.192
SSyy	1855.202

60. Calculate the coefficient of determination, r^2 for the problem above and interpret it. VS

61. Use the crime data above to calculate the correlation coefficient for the variables *"% unemployed"* and *"murder rate".* Is there a linear relationship? If so, is it positive or negative? VS

Note:

SSxx	27.6855
SSxy	196.001
SSyy	1855.202

62. Calculate the coefficient of determination, r^2 for the problem above and interpret it. VS

63. Use the crime data above to calculate the correlation coefficient for the variables population (in millions) and murder rate. Is there a linear relationship? If so, is it positive or negative?
Note:

SSxx	55.15098
SSxy	-21.4627
SSyy	1855.202

64. Calculate the coefficient of determination, r^2 for the problem above and interpret it.

65. In a recent study, researchers formed a least squares model using the variables X: saturated fat calories consumed daily and Y: body fat percentage. The SS_{yy} = 1,286.1, and the SSE = 1037.2. Find and interpret the coefficient of determination, r^2 for the model. VS

66. If we are told the coefficient of determination is r^2 = 0.928, can we know what r is just by taking the square root? If not why? (assume simple linear regression was done with one predictor variable X) VS

11.7 Answers

58. Solution:
    a. The p-value for the test of the slope estimator is very small, so there is a significant linear relationship.
    b. 66.49% of the variation in the percent of a city that earns less than $5,000 can be explained by using the unemployment rate as a predictor.
    c. r = 0.8154, which indicates strong positive linear correlation, as the unemployment rises, more people work for less.
59. $r = \frac{511.192}{\sqrt{199.732 * 1855.202}} \approx 0.8398$; Positive correlation.
60. $r^2 = 0.705$; 70.5% of the variation in the murder rate can be explained by using the percent of the population earning under $5,000 per year as a predictor.
61. $r = \frac{196.001}{\sqrt{27.6855 * 1855.202}} \approx 0.8648$; Positive correlation.
62. $r^2 = 0.7480$; 74.8% of the variation in the murder rate can be explained by the use of the unemployment rate as a predictor.
63. $r = \frac{-21.463}{\sqrt{55.15098 * 1855.202}} \approx -0.0671$; There is not a significant linear relationship.
64. $r^2 = 0.0045$; 0.45% of the variation in the murder rate between two cities can be explained by using population as a predictor.
65. $r^2 = \frac{1286.1 - 1037.2}{1286.1} \approx 0.1935$; Only about 19% of the total variation can be explained by the model.
66. We can only determine the absolute value of r this way, so we would need more information to determine if r was negative or positive. For example if r = - 0.821 r^2 = 0.674, but if r = 0.821 we get the same r^2.

11.8 Using the Model to Create an Estimation Interval

To complete this section of homework watch Chapter Eleven, Lecture Example 171.5.

67. The least squares line for the relationship between the number of absences for a student and the student's final grade from an earlier problem was $\hat{y} = -2.67x + 96.78$. Find the 95% confidence interval for the students' average final grade given that the students were absent 7 times.
$n = 6, \bar{x} = 6.167, SS_{xx} = 108.8333, S = 2.775554$ VS
68. The least squares line for the relationship between the height and weight for super models from an earlier problem was $\hat{y} = 3.88x - 152$. Minitab was used to find the 95% confidence interval for the models' average height given that the models are 70 inches tall. Based on the interval, is it reasonable to believe that the average weight for models who are 70 inches tall is 119 pounds?

Regression Equation

Weight = −151.70 + 3.883 Height

Variable	Setting
Height	70

Fit	SE Fit	95% CI	95% PI
120.137	1.645	(116.247, 124.027)	(107.962, 132.312)

69. The least squares line for the relationship between the number of clicker points earned in class and the student's final average from an earlier problem was $\hat{y} = 1.07x + 58.04$. Find the 98% estimation (confidence) interval for the students' average final grade given that the students had 30 clicker points. ($n = 8, \bar{x} = 21.375, SS_{xx} = 1081.875, S = 4.47102$)
70. The least squares line for the relationship between the number of absences for a student and the student's final grade from an earlier problem was $\hat{y} = -2.67x + 96.78$. Minitab was used to find the 95% confidence interval for the students' average final grade given that the students were absent 9 times. Based on the interval, is it reasonable to believe that the average score for students with 9 absences would be higher than 80%?

Regression Equation

Grade = 96.784 − 2.6677 Absences

Variable	Setting
Absences	9

Fit	SE Fit	95% CI	95% PI
72.775	1.361	(68.996, 76.553)	(64.192, 81.358)

71. The least squares line for the relationship between the height and weight for super models from an earlier problem was $\hat{y} = 3.88x - 152$. Find the 99% confidence interval for the models' average weight given that the models are 73 inches tall.
$n = 9, \bar{x} = 70.2222, SS_{xx} = 19.05556, S = 4.8789$ VS
72. The least squares line for the relationship between the number of pounds overweight and the age at death from an earlier problem was $\hat{y} = -0.538x + 63.085$. Find the 90% confidence interval for the women's average age at death given the women were 20 pounds overweight.
($n = 20, \bar{x} = 8.9, SS_{xx} = 2933.8, S = 5.3753$) VS

11.8 Answers

67. [74.885, 81.295]
68. Since the confidence interval includes all values from 116.2 to 124.0, it is reasonable to believe the average could be 119 pounds.
69. [83.95, 96.33]
70. Since the confidence interval includes only values from 68.996 to 76.553, it is not reasonable to believe the average is greater than 80%.
71. [118.976, 143.504]
72. [49.498, 55.152]

11.9 Using the Model to Create a Prediction Interval

To complete this section of homework watch Chapter Eleven, Lecture Examples 172, 173, and 173tech.

73. The least squares line for the relationship between the number of absences for a student and the student's final grade from an earlier problem was $\hat{y} = -2.67x + 96.78$. Find the 95% prediction interval for a student's final grade given that the student was absent 7 times.
    $n = 6, \bar{x} = 6.167, SS_{xx} = 108.8333, S = 2.775554$ VS
74. The least squares line for the relationship between the height and weight for super models from an earlier problem was $\hat{y} = 3.88x - 152$. Minitab was used to find the 95% prediction interval for a model's height given that the models is 70 inches tall. In a phone interview, a model who is 70 inches tall claims she weighs only 105 pounds. Based on the interval, is it reasonable to believe that this model, who is 70 inches tall, weighs 105 pounds?

    Regression Equation

    Weight = −151.70 + 3.883 Height

Variable	Setting
Height	70

Fit	SE Fit	95% CI	95% PI
120.137	1.645	(116.247, 124.027)	(107.962, 132.312)

75. The least squares line for the relationship between the number of clicker points earned in class and their final averages from an earlier problem was $\hat{y} = 1.07x + 58.04$. Find the 98% prediction interval for a student's final grade given that the student had 30 clicker points.
    $n = 8, \bar{x} = 21.375, SS_{xx} = 1081.875, S = 4.47102$
76. The least squares line for the relationship between the number of absences for a student and the student's final grade from an earlier problem was $\hat{y} = -2.67x + 96.78$. Minitab was used to find

: indicates the exercise has a video devoted to it in the corresponding section of STATSprofessor.com

the 95% prediction interval for a student's final grade given that the student was absent 9 times. Based on the interval, is it reasonable to believe that the average score for this student with 9 absences will be less than 68%?

Regression Equation

Grade = 96.784 − 2.6677 Absences

Variable	Setting
Absences	9

Fit	SE Fit	95% CI	95% PI
72.775	1.361	(68.996, 76.553)	(64.192, 81.358)

77. The least squares line for the relationship between the height and weight for super models from an earlier problem was $\hat{y} = 3.88x - 152$. Find the 99% prediction interval for a model's weight given that the model was 73 inches tall. ($n = 9, \bar{x} = 70.2222, SS_{xx} = 19.05556, S = 4.8789$) VS
78. The least squares line for the relationship between the number of pounds overweight and the age at death from an earlier problem was $\hat{y} = -0.538x + 63.085$. Find the 90% prediction interval for a woman's age at death given that the woman was 20 pounds overweight. ($n = 20, \bar{x} = 8.9, SS_{xx} = 2933.8, S = 5.3753$) VS

11.9 Answers

73. [69.75, 86.43]
74. Since the prediction interval contains only the values from 107.962 to 132.312, it does not seem likely that the models weighs only 105 pounds.
75. [74.786, 105.494]
76. Since the prediction interval contains all of the values from 64.192 to 81.358, it is reasonable to think the student's score will be less than 68%.
77. [110.220, 152.260]
78. [42.585, 62.065]

Take a sample exam for chapter 10 & 11

Chapter 11 Mixed Review

79. The SSyy and SSE for a least squares regression model are 23.334 and 4.554 respectively. Find r^2 for the model.

80. The paired data below consists of weights and bench press (2 rep) maximums for 6 randomly selected adult lifters. The equation of the regression line is $\hat{y} = 100.8 + 0.842x$, SSxx is 619.333, the average weight for the study participants was 158.333, and the standard error of estimate, S is 2.803. Find the 95% prediction interval for the maximum (2 rep) bench press for a person who weighs 160 pounds.

Weight	169	157	155	145	150	174
Max Bench	240	235	230	225	225	250

81. Find the correlation coefficient, r, for the following data:

Weight	169	157	155	145	150	174
Max Bench	240	235	230	225	225	250

$SS_{xx} = 619.333,\ SS_{xy} = 521.667,\ SS_{yy} = 470.833$

82. The number of alcohol drinks consumed daily and the number of days absent from work have a correlation coefficient of r = 0.896. Does this mean that drinking too much causes work absenteeism?

83. Use the weight and bench press max model to predict the average maximum bench press for lifters who weigh 172 pounds: $\hat{y} = 100.8 + 0.842x$.

84. The paired data below consists of weights and bench press (2 rep) maximums for 6 randomly selected adult lifters. The equation of the regression line is $\hat{y} = 100.8 + 0.842x$, SSxx is 619.333, and the standard error of estimate, S is 2.803. Find the 95% confidence interval for β_1 (the slope) of the regression line that relates weight to maximum bench press.

Weight	169	157	155	145	150	174
Max Bench	240	235	230	225	225	250

85. Suppose you fit a least squares line to 6 data points and the calculated value of SSyy = 470.833, SSxy = 521.667, and $\hat{\beta}_1$ = 0.842
   a. Find s^2, the estimator of σ^2.
   b. What is the largest deviation you might expect between any one of the 6 points and the least squares line (find a deviation that 95% of the observed values will fall within from our line)?

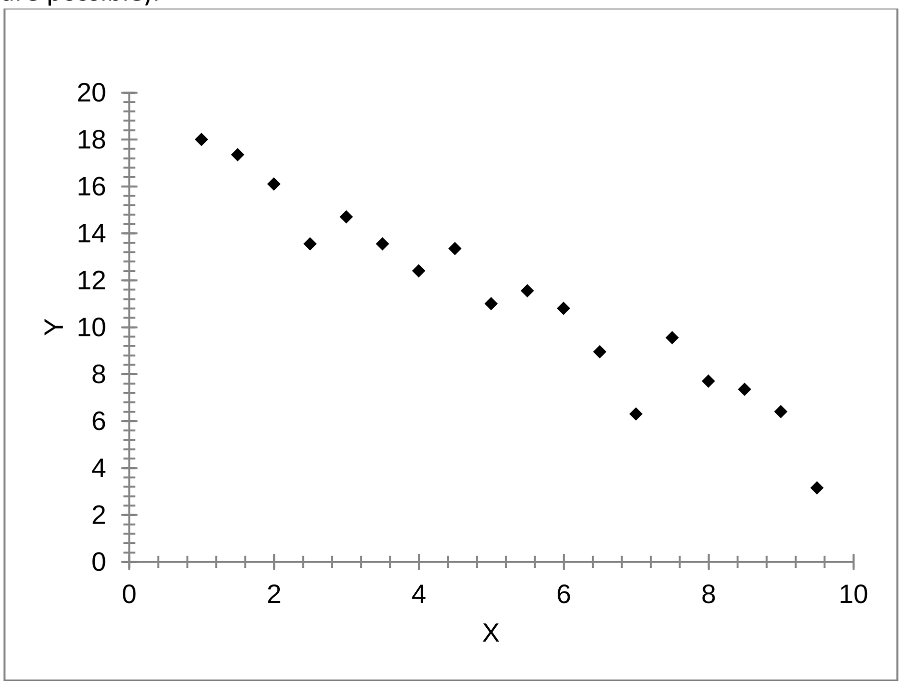
86. The regression equation that follows relates weights and maximum bench press for adult lifters: $\hat{y} = 100.8 + 0.842x$. What is the interpretation of the slope, $\hat{\beta}_1$?

87. The data below include the waist measures of 8 randomly selected men, and the hours they spend exercising per week. Find the equation of the regression line for the given data. What is the interpretation of the y-intercept for this line?

Gym Hours	4	0	3	7	5	5	4.5	3.5
Waist (in.)	31	36	33.5	29.5	30	32.5	31	32

$\sum x = 32,\ \bar{x} = 4,\ \sum x^2 = 156.5,\ SS_{xy} = -27,\ \bar{y} = 31.9375$

88. Determine whether the scatterplot shows little or no association, a negative association, a linear association, a moderately strong association, or a very strong association (multiple associations are possible).



89. The following interval is a 95% prediction interval for the price of a used Honda Civic that has 35,000 miles on it. Interpret the interval: $6,321.00 to $16,112.00.

90. The sample data below were obtained in a study of the relationship between the age of hens and the number of eggs they laid over the past year. At the 5% level of significance, do the data provide sufficient evidence to conclude that the slope of the regression line is not 0 and hence that the age of a hen is useful as a predictor of the number of eggs laid? The regression analysis is given below.

Regression Analysis

r^2	0.968	n	10
Std. Error	9.002	Dep. Var.	**Eggs**

Regression output					confidence interval	
variables	*coefficients*	*std. error*	*t (df=8)*	*p-value*	*95% lower*	*95% upper*
Intercept	266.0500					
Age	-31.5500	2.0130	-15.673	2.74E-07	-36.1920	-26.9080

Chapter 11 Mixed Review Answers:

79. $r^2 = 0.805$
80. 227.10 to 243.94 pounds
81. r = 0.966
82. No, it only implies the two appear together. In other words, we see that drinking more often is related to missing more days of work. One does not need to cause the other. For example, work dissatisfaction could cause a person to miss work and to drink heavily. The same could be said for both chronic pain and marital difficulty. Either of those things could be the true cause of both absenteeism and heavy drinking. Correlation is not causation.
83. 245.624 pounds
84. 0.529 to 1.155
85. a. 7.897 b. 2s = 5.620
86. For every extra pound of body weight, a lifter can expect to increase his maximum bench press by 0.842 pounds.
87. $\hat{y} = 35.727 - 0.947x$
88. There is a strong, negative association or a strong negative linear relationship
89. Based on this regression, the price of a Honda Civic that has 35,000 miles on it should be priced between $6,321.00 and $16,112.00.
90. $Claim: \beta_1 \neq 0$

$H_0 : \beta_1 = 0$

$H_A : \beta_1 \neq 0$

$PValue: 0.000000274$

$InitialConclusion:$ Reject the null, support the alternative.

There is sufficient evidence to support the claim that there is a linear relationship, so the age of a hen is a useful predictor of the number of eggs laid.