In addition to the formal assumptions previously stated, a linear model should only be used to fit data that appear to be reasonably linear. Because of the wide availability of computer programs that calculate least squares estimates, you will not need to manually calculate estimates very often.

13.1 Exercises

Basic Concepts

1. What is regression analysis?
2. Give two examples of why businesses might be interested in studying the relationship between two variables.
3. What is the difference between a dependent and an independent variable?
4. What is a simple linear regression model? Give the equation that describes a simple linear regression model and define all terms in the equation.
5. What is the estimated simple linear regression equation and how is it used?
6. What is $\hat{y}$? How does this differ from y?
7. What is the technique used to estimate the simple linear regression coefficients?
8. What is the relationship between scatterplots and simple linear regression?
9. Why is it often difficult to accurately describe real world situations using a simple linear regression equation?
10. What is the correlation coefficient? Why is the correlation coefficient insufficient when describing an exact linear relationship between x and y?
11. What is the residual of a model?
12. What is the sum of squared errors and what does it measure?
13. Explain why the best line is referred to as the least squares line.
14. What measure should be minimized in order to find the least squares line?
15. What is the equation for finding the slope of the least squares line?
16. What is the equation for finding the intercept of the least squares line?
17. When finding the least squares line manually, which must be calculated first: the slope or the y-intercept?
18. Interpret the intercept coefficient, b_0.
19. Interpret the slope coefficient, b_1.
20. Why is the magnitude of the prediction errors important when estimating a regression model?
21. What is the mean error for a least squares model?
22. Describe what the magnitude of the variation in the error terms tells us about the reliability of the regression model.
23. What is mean square error?
24. How many degrees of freedom are associated with the error term in a simple linear regression model?
25. What is the square root of the mean square error known as?
26. Describe where the summary statistics for the standard error and mean square error are found in a standard regression summary output in Microsoft Excel.

©HAWKES LEARNING

27. Is there a universal rule on how large is *large* with regard to standard error in a model?

28. What is estimated by the mean square error and what is estimated by the standard error?

29. Why is there an error term incorporated in the simple linear model?

30. What does the error term represent?

31. List the four assumptions about the error term in the simple linear model.

32. List the parameters of the simple linear regression model, and identify their estimates.

Exercises

33. Consider the following simple linear regression model. Write the estimated simple linear regression equation that corresponds to this model.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

34. Consider the following estimated simple linear regression equation.

$$\hat{y}_i = b_0 + b_1 x_i$$

a. What population parameter does b_0 estimate?

b. What population parameter does b_1 estimate?

c. Is error incorporated into the estimated model? Explain.

35. Suppose that a company wishes to predict sales volume based on the amount of advertising expenditures. The sales manager thinks that sales volume and advertising expenditures are modeled according to the following linear equation. Both sales volume and advertising expenditures are in thousands of dollars.

$$\text{Estimated Sales Volume} = 49.25 + 0.51\left(\text{Advertising Expenditures}\right)$$

a. What is the dependent variable in this model? Explain.

b. What is the independent variable in this model? Explain.

c. What is the estimated sales volume for this company when the marketing department spends \$40,000 on advertising?

d. If the company had a target sales volume of \$100,000, how much should the sales manager allocate for advertising in the budget?

e. What is the sales manager forgetting to account for when using this linear equation to determine sales volume? What kinds of problems could this cause for the company?

36. Suppose the following estimated regression equation was determined to predict salary based on years of experience.

$$\text{Estimated Salary} = 25689.10 + 2148.35\left(\text{Years of Experience}\right)$$

a. What is the dependent variable?

b. What is the independent variable?

c. What is the value that estimates β_0 in this particular equation?

d. What is the value that estimates β_1 in this particular equation?

e. What is the estimated salary for an employee with 15 years of experience?

37. Plot the following lines.

a. $y = 2 + 3x$

b. $y = 4 + 8x$

c. $y = 9 - 2x$

d. $y = x$

©HAWKES LEARNING

38. Plot the following lines.

a. $y = 100 + 50x$

b. $y = 0.5 + 0.7x$

c. $y = 20 - 5x$

39. Consider the following estimated regression equation.

$$\hat{y}_i = 10x_i - 5$$

a. Complete the following table.

Predicted Values	
x	$\hat{y}$
2	
5	
7	
9	
10	

b. Do these two variables appear to have a positive or negative relationship?

c. For these two variables, what sign would you expect the correlation coefficient to have? Explain.

40. Consider the following data.

Observed Values	
x	y
0	2
1	4
5	9
6	7
8	8

a. Draw a scatterplot of the data.

b. Draw a line which you believe fits the data.

c. Suppose that $\hat{y}_i = 3 + 0.8x_i$ is a line that fits the data reasonably well. Complete the following table.

Observed and Predicted Values				
Observed x	Observed y	Predicted y	Error	Squared Error
0	2			
1	4			
5	9			
6	7			
8	8			

d. What is the sum of squared errors for these data?

©HAWKES LEARNING

41. Consider the following data regarding home sale prices and square footage.

Housing Prices and Square Footage	
Selling Price (Thousands of Dollars)	Square Footage
199.9	1065
228.0	1254
235.0	1300
285.0	1577
239.0	1600
293.0	1750
285.0	1800
365.0	1870
295.0	1935
290.0	1948
385.0	2254
505.0	2600
425.0	2800
415.0	3000

a. Suppose we want to predict selling price based on square footage. Write the estimated regression equation in terms of selling price and square footage. (Assume the parameters of this model have not been estimated.)

b. Create a scatterplot of the data and draw a line of best fit.

c. Suppose we determine that an equation that fits the data reasonably well is

$$\text{Estimated Selling Price} = 52.35 + 0.14(\text{Square Footage}).$$

Complete the following table.

Housing Prices and Square Footage				
Observed Selling Price (Thousands of Dollars)	Observed Square Footage	Predicted Selling Price (Thousands of Dollars)	Error	Squared Error
199.9	1065			
228.0	1254			
235.0	1300			
285.0	1577			
239.0	1600			
293.0	1750			
285.0	1800			
365.0	1870			
295.0	1935			
290.0	1948			
385.0	2254			
505.0	2600			
425.0	2800			
415.0	3000			

d. Compute the sum of squared errors for these data.

©HAWKES LEARNING

42. Consider the following data.

x	1	2	3	4	5
y	1	3	4	4	6

a. Plot the data points on a scatterplot.

b. Determine the least squares line. Use x as the independent variable.

c. Plot the least squares line on the scatterplot.

d. Use the model to compute the error for each data point.

43. Consider the following data.

x	–2	–1	0	3	5
y	1	3	5	4	8

a. Plot the data points on a scatterplot.

b. Determine the least squares line. Use x as the independent variable.

c. Plot the least squares line on the scatterplot.

d. Use the model to compute the error for each data point.

44. Comparing the least squares lines in Exercises 42 and 43, which line fits the data better? Explain your answer.

45. Suppose a linear regression analysis produced the following equation relating an individual's salary to the current value of his or her home.

$$\text{Estimated Current Value of Home} = 12331 + 3.14\left(\text{Annual Salary}\right)$$

a. Which of the variables in the model is the dependent variable?

b. Which of the variables in the model is the independent variable?

c. What would be the predicted current value of home for someone earning a salary of $32,000?

d. If a person earned $5000 additional income, how much of an increase in home value would be predicted?

e. In terms of the problem, interpret the estimate of the slope in the model.

f. In terms of the problem, interpret the estimate of the intercept in the model.

g. Do you believe annual salary is a causal factor in explaining the price of someone's home? Explain.

46. Suppose a linear regression analysis produced the following equation relating a basketball player's total points scored to the number of minutes played in a season.

$$\text{Estimated Points Scored} = -97.2 + 0.645\left(\text{Minutes Played}\right)$$

a. Which of the variables in the model is the dependent variable?

b. Which of the variables in the model is the independent variable?

c. What would be the predicted value of total points scored for a basketball player who plays 500 minutes in a season?

d. If a basketball player played an additional 100 minutes, how much of an increase in total points scored would be predicted?

e. In the model, which of the coefficients is the slope?

f. In the model, which of the coefficients is the intercept?

g. Do you believe the number of minutes played is a causal factor in explaining the total points scored? Explain.

©HAWKES LEARNING

47. Suppose you were studying the educational level of husbands and wives (measured in number of years of education). You have randomly selected 10 couples and have obtained the data in the following table.

Education Level										
Husband	12	16	16	18	20	17	23	14	12	16
Wife	14	16	14	16	16	18	18	12	16	20

a. Suppose you wanted to predict the husband's years of education based on the wife's. Use the data to estimate the appropriate model.

b. Use the model in part **b.** to predict the husband's educational level if married to a woman with 16 years of education.

c. Suppose you wanted to predict the years of education for the wife based on the husband's years of education. Use the data to create the appropriate model. Did you get the same model as in part **b.**?

d. Use the model created in part **d.** to predict the wife's educational level if married to a husband with 16 years of education.

e. Do you believe there is a causal relationship between the two variables? If so, which direction is the causality? Does the husband's education cause the wife to have more or less education, or vice versa?

48. Consider the following summary output.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.911653228
R Square	0.831111609
Adjusted R Square	0.79733393
Standard Error	0.253142413
Observations	7

ANOVA

	df	*SS*	*MS*	*F*
Regression	1	1.576737452	1.576737	24.60535
Residual	5	0.320405405	0.064081	
Total	6	1.897142857		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	4.021621622	0.181401491	22.16973	3.47E-06
X Variable 1	-0.22297297	0.044950802	-4.96038	0.004247

a. What is the mean square error for these data?

b. What is the standard error of the model?

©HAWKES LEARNING

49. Consider the following data.

Observed Values	
x	*y*
15	110
18	135
25	150
24	149
26	158
40	169

a. Suppose that, using statistical software, we determine that $b_0 = 93.2922$ and $b_1 = 2.1030$. Complete the following table.

Observed versus Predicted Values				
Observed *x*	Observed *y*	Predicted *y*	Error	Squared Error
15	110			
18	135			
25	150			
24	149			
26	158			
40	169			

b. Compute the sum of squared errors.

c. Compute the mean square error.

d. Compute the standard error of the model.

e. Do you believe these estimates of b_0 and b_1 provide a reliable estimated regression equation for these data? Explain.

50. Consider the following data regarding students' college GPAs and high school GPAs.

GPAs	
College GPA	High School GPA
2.80	3.42
3.54	3.56
2.88	3.13
2.15	3.27
2.22	3.38
3.31	4.13
2.13	3.95
2.39	3.81
3.01	4.33
2.68	2.85

a. Suppose we want to predict college GPA based on high school GPA. Write the estimated regression equation in terms of college GPA and high school GPA. (Assume the parameters of the model have not been estimated.)

b. Suppose we determine, using statistical software, that the estimated regression equation is

$$\text{Estimated College GPA} = 1.88 + 0.2319(\text{High School GPA}).$$

Complete the following table.

©HAWKES LEARNING

GPAs				
Observed College GPA	Observed High School GPA	Predicted College GPA	Error	Squared Error
2.80	3.42			
3.54	3.56			
2.88	3.13			
2.15	3.27			
2.22	3.38			
3.31	4.13			
2.13	3.95			
2.39	3.81			
3.01	4.33			
2.68	2.85			

c. Compute the sum of squared errors for the model.

d. Compute the standard error of the model.

51. The regression equation that relates the delivery time with number of pizzas and distance is given by $\widehat{\text{Delivery Time}} = 1.79 + 1.95(\text{Number of Pizzas}) + 1.57(\text{Distance})$.

a. Estimate the delivery time to deliver 5 pizzas at a distance of 2 miles.

b. The observed data shows that the time taken to deliver 5 pizzas at a distance of 2 miles is 16 minutes. Find the residual.

c. Interpret the meaning of the residual in the context of the problem.

13.2 Residual Analysis

When performing regression analysis, residual analysis is a useful technique to help us determine if the model we are using is appropriate. By studying the estimated errors (i.e., the residuals), we can check the underlying assumptions of the regression model. Before one can adequately make predictions with the estimated regression model, the analyst should ensure that the assumptions of the model are valid. Residual analysis is the method used to validate those assumptions.

All estimates, intervals, and hypothesis tests in regression analysis are based on assuming that the model is correct. If the model is not correct (i.e., at least one of the assumptions is not valid), the formulas and methods will also be incorrect.

Validating the assumptions of the simple linear regression model revolve around the error term (ε). You may recall that the assumptions of the simple linear regression model are:

1. The average response at each value of the independent variable is a linear function. That is, there is a linear relationship between x and y.
2. The errors, ε_i, are assumed to be independent of each other.
3. The errors, ε_i, at each value of x_i are normally distributed.
4. The errors, ε_i, at each value of x_i have equal variances, σ_ε^2.

One method of validating the assumptions is by performing a graphical analysis of the residuals. The most frequently used graph is that of the residuals (e_i) vs. the fitted values ($\hat{y}_i$). It is a scatter plot with the residuals on the vertical axis and the fitted values on the horizontal axis. This plot is used to detect non-linearity, unequal variances, and outliers.

In Figure 13.2.1, (a) is a scatterplot of the raw data, y vs. x, with a simple linear regression line fit through the data. Note that the scatterplot is somewhat curvilinear. When we plot the

©HAWKES LEARNING