Continued

Then, they set about recruiting the players within their budget that helped them achieve their goal. The A's are still considered one of the most efficient teams when it comes to recruiting. The Oakland A's went to the playoffs in 2000–2003, 2006, 2012–2014, and 2018–2020. The improvement of the Oakland A's performance is a case where empiricism triumphed over belief, conjecture, and speculation.

Interpretation of b_0 and b_1

The intercept coefficient, b_0, is the average value of the dependent variable, y, when the independent variable, x, is equal to zero.

The slope coefficient, b_1, is the average change in the dependent variable, y, for a one unit change in the independent variable, x.

DEFINITION

5.2 Exercises

Basic Concepts

1. What is the value of a model?
2. What is regression analysis?
3. What is the difference between a dependent and an independent variable?
4. What is $\hat{y}$? How does this differ from y?
5. What is the technique used to estimate the linear regression coefficients?
6. What is the relationship between scatterplots and linear regression?
7. Why is it often difficult to accurately describe real world situations using a linear regression equation?
8. What is the sum of squared errors and what does it measure?
9. Explain why the best line is referred to as the least squares line.
10. What measure should be minimized in order to find the least squares line?
11. What is the equation for finding the slope of the least squares line?
12. What is the equation for finding the intercept of the least squares line?
13. Interpret the intercept coefficient, b_0.
14. Interpret the slope coefficient, b_1.

Exercises

15. 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.

$$\textit{Estimated Sales Volume} = \hat{y} = 49.25 + 0.51\ \textit{Advertising Expenditures}$$

   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?

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

$$\textit{Estimated Salary} = \hat{y} = 25689.10 + 2148.35\ \textit{Years of Experience}$$

a. What is the dependent variable? What is the independent variable?

b. What is the value that estimates b_0 in this particular equation?

c. What is the value that estimates b_1 in this particular equation?

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

17. Plot the following lines.

a. $y = 2 + 3x$

b. $y = 4 + 8x$

c. $y = 9 - 2x$

d. $y = x$

18. Plot the following lines.

a. $y = 100 + 50x$

b. $y = 0.5 + 0.7x$

c. $y = 20 - 5x$

19. Consider the following estimated regression equation.

$$\hat{y} = 10x - 5$$

a. Complete the following table.

Predicted Values					
x	2	5	7	9	10
y					

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.

20. Consider the following data.

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

a. Draw a scatterplot of the data.

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

c. Suppose that $\hat{y} = 3 + 0.8x$ 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 this data?

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > Mount Pleasant Real Estate Data

21. Using the Mount Pleasant Real Estate data set from the companion website, consider the following:

a. Suppose we want to predict *List Price* based on *Square Footage*. Write the estimated regression equation in terms of *List Price* and *Square Footage*. (Assume the parameters of this model have not been estimated.)

b. Create a scatterplot using the *List Price* and *Square Footage* variables and draw a least squares regression line.

c. Suppose we determine that an equation that fits the data reasonably well is $\textit{Estimated List Price} = \hat{y} = -144{,}193 + 256.3753\ \textit{Square Footage}$. Using this equation, complete the following table. Round values to the nearest whole number.

Housing Prices and Square Footage				
Observed Selling Price	**Observed Square Footage**	**Predicted Selling Price (Thousands of Dollars)**	**Error**	**Squared Error**
$375,000	1797			
$423,600	2135			
$448,315	1895			
$515,250	2423			
$556,400	2800			
$600,500	3045			
$583,620	3115			
$683,025	3210			
$635,250	3143			
$615,300	2730			
$731,410	3340			
$860,750	3521			
$835,000	4236			
$815,500	3841			

d. Compute the sum of squared errors for the data in the table in part **c.** Round your answer to the nearest whole number.

e. Is there a variable other than *Square Footage* that might predict *List Price* more accurately? Create a scatterplot using that variable and *List Price*, and compare it to the plot in part **b.** Is the grouping tighter or more dispersed?

22. In the story of Moneyball, Billy Beane and Paul DePodesta determined that 95 wins out of 162 games was necessary in order for the Oakland A's to make it to the MLB playoffs. They also determined that there was a strong linear relationship between season run differential and the number of wins in a season. Using the Moneyball data set from the companion website, subset the data to only include data prior to the year 2002 (1962-2001), and answer the following questions.

Data

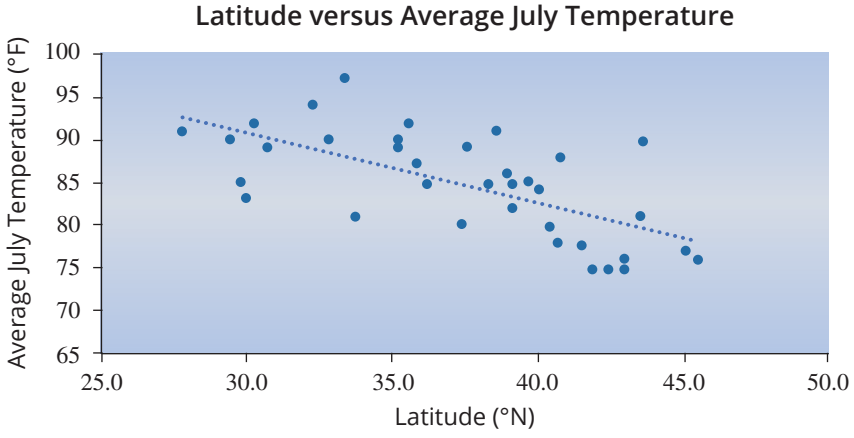
stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > Moneyball

a. What is the regression equation for predicting wins (W) using run differential (RD)?

b. What is the run differential necessary in order to win 95 games? Use the regression equation from part **a.** to calculate your answer. Do not round any values until the final answer. Round your final answer to the nearest whole number.

23. The summary statistics for the latitudes and average July temperature of 36 cities in the United States are shown in the table. The sample data is also displayed in a scatterplot with the regression line included.

	Mean	Standard Deviation
Latitude (°N)	37.4	4.89
Average July Temperature (°F)	85	6.04
Correlation	−0.66	



a. What is indicated by the negative value of the correlation coefficient?

b. Use the summary statistics to determine the equation of the regression line that predicts the average July temperature when given the latitude of a certain U. S. city.

c. St Louis, Missouri is located at 38.6°N latitude. What is the average July temperature that is predicted by the regression model?

24. 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.

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

$$\textit{Estimated Current Value of Home} = \hat{y} = 72{,}331 + 3.14\ \textit{Annual Salary}$$

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 $52,000?

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.

26. 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.

$$\textit{Estimated Points Scored} = \hat{y} = -97.2 + 0.645\ \textit{Minutes Played}$$

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.

27. 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 **a.** 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 **a.**?

d. Use the model created in part **c.** 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?

28. There were ten homes sold in the Seaside Subdivision in April 2021. The data shown below shows the square footage and sale price for each of the houses. A table of summary statistics for each variable and the correlation coefficient is shown as well.

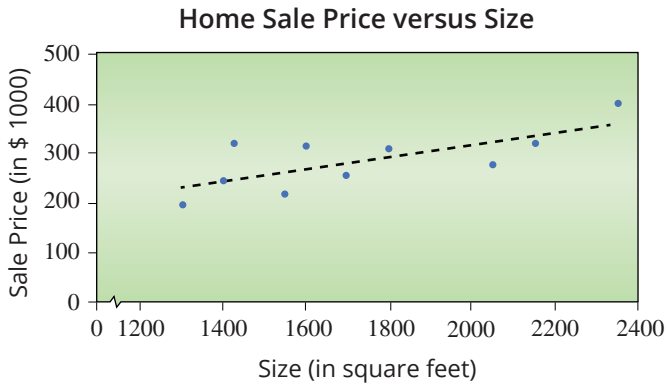
a. Determine the equation of the least squares regression line.

b. Interpret the value of the slope coefficient in the context of the data.

c. Calculate the error for the house with 1550 square feet.

Size (sq. ft.)	1400	1600	2050	1800	1300	1550	2350	2150	1425	1700
Sale Price (in $1000)	245	312	279	308	199	219	405	324	319	255

Variable	Mean	Standard Deviation
Explanatory - Size	1732.5	350.4
Response - Price	286.5	60.2
Correlation Coefficient	0.735	



5.3 Evaluating the Fit of a Linear Model

The Importance of Errors

The usefulness of the regression model depends on the magnitude of the prediction errors you expect the model to produce. The Jeep Cherokee Limited model is

$$Asking\ Price = \hat{y} = b_0 + b_1\ Age + \text{error}.$$

Yet we ignored the error component in the previous section when we predicted the prices of the Jeep Cherokee Limited for different ages (in Table 5.2.3). For instance, when we predicted the price of a two-year-old Jeep Cherokee Limited, we found

$$Estimated\ Asking\ Price = \hat{y} = \$47{,}030.83 - \$3{,}846.09 \cdot (2) = \$39{,}338.65.$$