5.3 Exercises

Basic Concepts

1. Why is the magnitude of the prediction errors important when estimating a regression model?
2. What is the mean error for a least squares model?
3. Describe what the magnitude of the variation in the error terms tells us about the reliability of the regression model.
4. What is the variance of the error terms?
5. How many degrees of freedom are associated with the error term in a simple linear regression model?
6. What is the square root of the variance of the error term known as?
7. Describe where the summary statistics for the standard error and mean square error are found in a standard regression summary output in Microsoft Excel.
8. Is there a universal rule on how large is *large* with regard to standard error in a model?
9. What is estimated by the variance of the error term and what is estimated by the standard error?
10. What is the total sum of squares?
11. How are the total sum of squares and the sample variance related?
12. Define error in terms of a regression model.
13. What part of the simple linear regression model captures the unexplained variation?
14. Describe the total sum of squares in terms of explained and unexplained variation.
15. What is the sum of squares of regression?
16. Express SSR in terms of the total sum of squares and the sum of squared errors. Interpret this in terms of model variation.
17. Why will there be errors in virtually all regression models?
18. What is the coefficient of determination? What kinds of values can the coefficient of determination take?
19. Suppose that regression analysis is performed and the resulting model has an R^2 value of 0.856. Interpret this value.
20. How is the coefficient of determination related to the correlation coefficient?

Exercises

21. 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 variance of the error for the data?

b. What is the standard error of the model?

Data

stat.hawkeslearning.com

Discovering Statistics and Data, Fourth Edition > Data Sets > US County Data

22. Using the US County Data from the companion website, use the variables *Diabetes.percent* and *Adult.obesity.percent* to perform the following.

a. Calculate the regression equation to predict *Diabetes.percent* using *Adult.obesity.percent*. Round values to 5 decimal places.

b. What are b_0 and b_1? Round your answers to 5 decimal places.

c. Using the information from parts **a.** and **b.**, complete the following table. Round Predicted *Diabetes.percent* and Error to 3 decimal places, and round Squared Error to 5 decimal places.

Observed versus Predicted Values				
Observed *Adult.obesity.percent*	**Observed *Diabetes.percent***	**Predicted *Diabetes.percent***	**Error**	**Squared Error**
0.408	0.173			
0.275	0.084			
0.375	0.115			
0.349	0.156			
0.312	0.070			
0.382	0.210			

d. Compute the sum of squared errors for the table in part **c.** Round your answer to 5 decimal places.

e. Compute the variance of the error term for the table in part **c.** Round your answer to 5 decimal places.

f. Compute the standard error of the table in part **c.** Round your answer to 5 decimal places.

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

23. In the previous section, we used the Moneyball data set (1962-2001) to determine that a season run differential of about 135 runs was necessary for the Oakland A's to make it to the MLB playoffs. However, Coach Billy Bean and statistician Paul DePodesta needed to figure out how to make that run differential a reality. They found that two of the most statistically significant variables that contributed to the number of runs scored were on-base percentage and slugging percentage. Use the Moneyball data set from the companion website, subsetted to only include the years 1962-2001 since that was the only data available to Beane at the time, and perform the following.

Data

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

a. Calculate the regression equation to predict runs scored (*RS*) using on-base percentage (*OBP*). Round values to 5 decimal places.

b. Calculate the regression equation to predict runs scored (*RS*) using slugging percentage (*SLG*). Round values to 5 decimal places.

c. Calculate the regression equation to predict runs allowed (*RA*) using opponent on-base percentage (*OOBP*). *OOBP* is only measured from 1999 on, so use data for the years 1999-2001 to estimate the equation. Round values to 5 decimal places.

d. Calculate the regression equation to predict runs allowed (*RA*) using opponent slugging percentage (*OSLG*). *OSLG* is only measured from 1999 on, so use data for the years 1999-2001 to estimate the equation. Round values to 5 decimal places.

e. Using the regression equation from part **a.**, complete the following table. Round values to the nearest whole number.

Runs Scored using On-Base Percentage

Observed *RS*	Observed *OBP*	Predicted *RS*	Error	Squared Error
687	0.319			
897	0.350			
724	0.320			
923	0.354			
642	0.323			

f. Calculate the sum of squared errors and the standard error for the table in part **e.** Round answers to 3 decimal places.

g. Using the regression equation from part **d.**, complete the following table. Round your answer to the nearest whole number.

Predicted Runs Allowed using Opponent Slugging Percentage				
Observed *RA*	Observed *OSLG*	Predicted *RA*	Error	Squared Error
713	0.398			
806	0.440			
627	0.378			
968	0.494			
766	0.437			

h. Calculate the sum of squared errors and the standard error for the table in part **g.** Round answers to 3 decimal places.

24. A digital marketing company has been experimenting with the effect of price on sales. Five different product prices have been sent to different sets of customers. They have carefully tracked the customers from each group and have recorded the proportion from each price category that purchased the product. The results are given in the following table.

Product Price Marketing Experiment Results					
Proportion That Purchased Product	0.032	0.028	0.026	0.015	0.009
Price of Product ($)	29.95	34.95	39.95	44.95	49.95

a. What level of measurement do the two variables in the table possess?

b. Specify the model that the marketing manager would be interested in estimating.

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

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

e. Draw a scatterplot of the data.

f. Use the data in the table to estimate the model.

g. Predict the proportion that will buy the product if the price is $35.00.

h. Compute the mean error for the model you estimated in part **f.**

i. Determine the variance of the error term.

j. What is the coefficient of determination? Interpret this value in terms of the problem.

25. An economist is studying the relationship between income and savings. He has randomly selected seven subjects and obtained income and savings data from them. He wishes to use a simple linear regression model to predict savings based on annual income.

Income and Savings							
Income (Thousands of Dollars)	28	25	34	43	48	39	74
Savings (Thousands of Dollars)	0.2	0	0.8	1.2	3.1	2.1	8.3

a. What level of measurement do the two variables in the table possess?

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

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

d. Draw a scatterplot of the data. Does the scatterplot suggest that a linear model is appropriate? Explain.

e. Use the data to estimate the appropriate model.

f. Predict the savings for someone who earns fifty thousand dollars annually.

g. Interpret the meaning of the slope coefficient in the problem.

h. What fraction of the variation in savings is explained by income?

26. Since 2009, the average term for a new-car loan was nearly 64 months. This leaves the buyer vulnerable to owing more on the car than it is worth. When applying for an automobile loan, it is oftentimes recommended to sign up for the shortest term you can afford. It is believed that along with one's credit rating, the length of the loan will help the buyer get a favorable interest rate. The following table contains interest rates and lengths of loans for 20 randomly selected auto purchases. Using the data in the table, answer the following questions.

Lengths of Loans and Interest Rates

Months Financed	Interest Rate (%)	Months Financed	Interest Rate (%)
12	4.00	48	6.51
24	4.40	48	6.68
36	5.24	60	7.13
12	3.43	60	7.48
24	4.40	72	8.31
36	5.79	60	7.85
36	5.98	72	8.07
48	6.58	72	8.48
36	5.31	48	6.12
36	5.91	72	8.07

a. Using statistical software, estimate the coefficients of the least squares regression equation.

b. Interpret the meaning of the slope and the intercept in part **a.**

c. Predict the interest rate for a person interested in a four-year auto loan.

d. Should you use the model to predict interest rates for an eight-year loan? Justify your answer.

e. Determine the coefficient of determination and explain its meaning in terms of the problem.

f. Calculate the correlation coefficient for this model. What does it mean?

g. What interest rate would one expect to get if they were planning to apply for a five-year auto loan?