Lastly, to compare mall and downtown locations, we look at

$$\left(\beta_0 + \beta_1 x_1 + \beta_2\right) - \left(\beta_0 + \beta_1 x_1 + \beta_3\right) = \beta_2 - \beta_3.$$

The estimated difference between β_2 and β_3 is given by $\beta_2 - \beta_3$ (21.1001) which represents the difference between the average annual return for shops in the mall locations having x_1 households in the area and the average annual return for shops in the downtown locations having x_1 households in the area. In terms of the problem, we can say that for any number of households in a given area, the average annual return in a mall area will be \$21,100.10 greater than the average annual return in a downtown location.

There are three potential issues to keep in mind when using these regression results.

1. These results are only meaningful within the relevant range of the data that was used to estimate the regression equation. For example, using the model in Example 14.5.2 to predict annual return for a shop with 1000 households or 1,000,000 households in the surrounding area would likely yield an unreliable point estimate.
2. The regression lines for the three locations in Example 14.5.2 are assumed to have the same slope, but in reality they could have very different slopes. Using a regression model with **interaction terms** allows the slopes for the regression lines to differ.
3. The regression lines estimated in this example are all linear. This implies that annual return increases by the same amount for each additional thousand households within 15 miles of the shops. This assumption is sometimes unrealistic. This issue can be addressed using **polynomial (or nonlinear) regression models**.

Regression models with interaction terms and polynomial regression models are beyond the scope of this text and are not discussed in detail. Multiple regression is a complex topic that involves many methods of estimation. We only present the basics in this text.

14.5 Exercises

Basic Concepts

1. Give three examples of qualitative independent variables that may be of interest to someone performing regression analysis to predict annual salary.

2. Explain how qualitative variables are transformed into quantitative variables in order to estimate a regression model.

3. If a qualitative variable has c classes, how many indicator (dummy) variables will there be in the model? Explain why this is the case.

4. When an indicator (dummy) variable is equal to one, does this represent a difference in the slope or the intercept of the model? Explain.

5. What is a base level variable? Interpret the value of an estimated coefficient for an indicator variable in terms of the base level variable.

6. Identify three potential issues to keep in mind when constructing regression models involving indicator variables. Also suggest how these issues can be addressed.

Exercises

7. **a.** How many indicator variables would it take to construct a qualitative variable with 5 states?

b. Assume the states were: very small, small, medium, large, very large. Develop the indicator variables to represent this qualitative variable in a regression model.

c. Given your design what is the meaning of the constant term in your model?

8. **a.** How many indicator variables would it take to construct a qualitative variable with 4 educational levels?

b. Assume the educational levels are: non high school graduate, high school graduate, some college, and college graduate. Develop the indicator variables to represent this qualitative variable in a regression model.

c. Given the design you have chosen, what is the meaning of the constant term in your model?

9. Consider the following estimated multiple regression model relating GPA to the number of classes attended and the final exam score in a particular class, and if the student is a freshman (= 1 if freshman, = 0 otherwise).

$$\textit{Cumulative GPA} = -0.8777 + 0.0672\ \textit{Attendance} + 0.0678\ \textit{Exam Score} - 0.1436\ \textit{Freshman}$$

a. Are the signs of the estimated coefficients what you would expect for these three independent variables? Explain.

b. Interpret the coefficient for the *Attendance* variable.

c. Interpret the coefficient for the *Exam Score* variable.

d. Interpret the coefficient for the *Freshman* variable.

e. Suppose two students, one a freshman and one a senior, attended the same number of classes and both got a score of 88 on the final exam. What would be the expected difference in GPA for the two students?

10. Consider the following computer output for the multiple regression model discussed in the previous exercise.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.714589997
R Square	0.510638864
Adjusted R Square	0.508467143
Standard Error	0.516416069

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	188.1180981	62.70603271	235.1309671	1.8485E-104
Residual	676	180.2794359	0.266685556		
Total	679	368.397534			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	-0.877712645	0.138557037	-6.334666683	4.34037E-10	-1.149766538
Attendance	0.067163994	0.003669275	18.3044333	4.30384E-61	0.059959449
Exam Score	0.067820136	0.004265782	15.89864106	1.37161E-48	0.059444361
Freshman	-0.143623671	0.047077779	-3.050774156	0.002371853	-0.236059922

a. Test the usefulness of the overall model in predicting *Cumulative GPA* using a 5% significance level.

b. What percentage of the variation in *Cumulative GPA* is explained by the three independent variables?

c. Is the qualitative independent variable, *Freshman*, useful in predicting Cumulative GPA? Use $\alpha = 0.05$.

d. Can you think of other variables that could be added to the model? Name one quantitative variable and one qualitative variable that might be useful.

11. A personnel director is interested in studying the effects which age, experience, and education level have on salary. Eight employees are randomly selected and each employee's salary, age, experience, and education level (0 if high school degree or below, 1 if college degree or above) are recorded.

Employee Data

Salary ($)	Age	Experience (Years)	Education Level
48,600	25	2	1
90,000	55	20	0
86,400	27	5	0
63,000	30	7	1
52,200	22	3	1
104,400	33	8	1
41,400	19	1	0
77,400	45	15	1

a. Create three scatterplots using salary with age, salary with experience, and salary with education level. Does each of the plots have a linear relationship?

b. Using statistical software, estimate the parameters of the following regression model:

$$Salary = \beta_0 + \beta_1\, Age + \beta_2\, Experience + \beta_3\, Education\ Level + \varepsilon.$$

c. Is the overall model useful in explaining salary? Test at the 0.05 level.

d. Is age useful in explaining salary? Test at the 0.05 level.

e. Is experience useful in explaining salary? Test at the 0.01 level.

f. Is education level useful in explaining salary? Test at the 0.10 level.

g. Interpret each of the regression coefficients.

h. Predict the salary of an employee with a college degree who is 35 years old with 10 years of experience.

i. Construct and interpret a 95% prediction interval for an employee with a Master's degree who is 35 years old with 10 years of experience. How useful is this interval?

j. Construct and interpret a 95% confidence interval for the average salary of an employee with a PhD who is 35 years old with 10 years of experience. How useful is this interval?

Note

To complete parts **i.** and **j.**, you will need to use Minitab, R, or Rguroo.

12. Consider the following crime data from select college campuses. The table contains the number of crimes committed, the number of campus police employed on campus, the total enrollment of the college, and whether or not the college is private. The full data set is available on the companion site.

Data

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

Campus Crime Data				
School	Number of Crimes	Number of Police	Total Enrollment	Private School
1	64	12	1131	Yes
2	138	21	12,954	No
3	141	32	16,009	No
4	84	22	1682	Yes
5	86	35	2888	Yes
...				

a. Create an indicator (dummy) variable for whether or not the college is private. Let *Private* = 1 if the school is private and *Private* = 0 if the school is public.

b. Suppose education officials wish to predict the number of crimes on college campuses based on the number of police employed and total enrollment. They would also like to know whether there are fewer crimes committed on private campuses than public ones. Use statistical software to estimate the following regression model.

$$Crimes = \beta_0 + \beta_1\, Police + \beta_2\, Enrollment + \beta_3\, Private + \varepsilon$$

Write the estimated multiple regression equation.

c. Is the overall model useful in predicting the number of crimes? Use $\alpha = 0.05$.

d. Are the signs of the coefficients of the independent variables what you would expect for these data? Explain.

e. Is there evidence to support the officials' belief that there are fewer crimes committed at private schools than at public schools? Test using $\alpha = 0.05$. Would this decision change if $\alpha = 0.01$?

13. You wish to develop a model to analyze if the manufacturer influences the price of used cars, using data on six-year-old cars produced by three Japanese manufacturers: Honda, Nissan, and Toyota. Your data consists of number of doors (2 or 4), curb weight, engine size, city mpg, highway mpg, and price.

Data

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

Car Price Data							
Car	Manufacturer	Number of Doors	Curb Weight	Engine Size	City MPG	Highway MPG	Price
1	Toyota	2	1985	92	35	39	5348
2	Honda	2	1837	79	38	42	5399
3	Nissan	2	1889	97	31	37	5499
4	Toyota	2	2040	92	31	38	6338
5	Honda	2	1713	92	49	54	6479
6	Toyota	4	2015	92	31	38	6488
©HAWKES LEARNING ...							

a. Define indicator variable(s) that could be used to code the 3 Japanese manufacturers.

b. Write the proposed model using all the variables in the data including the indicator variable(s) developed in part **a.**

c. What variables in the model are significant at the 0.05 level? Which are not significant?

d. What proportion of the variation in car prices is explained by the model?

e. If you eliminate the variables that are not significant and rebuild the model, how much of the variation in price will you explain?

14. Consider the following sales data regarding weekly sales, the number of sales reps, and whether or not the sales were made in the first, second, third, or fourth quarter of the year. For each column containing an indicator variable, the variable is equal to 1 if that particular week was in that particular quarter, and equal to zero otherwise. For example, if the weekly data were recorded in January, the 1st quarter indicator variable would be equal to 1 and the indicator variables for the 2nd, 3rd, and 4th quarters would be equal to zero. The first quarter comprises January through March, the second quarter April through June, the third quarter July through September, and the fourth quarter October through December.

Weekly Sales by Quarter

Weekly Sales ($)	Number of Sales Reps	Quarter
4272.90	3	1
5069.70	9	1
6067.70	11	1
6680.55	17	1
9725.05	20	1
4107.10	3	2
7520.25	9	2
12,135.00	11	2
13,016.55	17	2
13,673.90	20	2
3272.05	3	3
5074.40	9	3
7505.45	11	3
8272.75	17	3
10,020.40	20	3
4925.75	3	4
10,018.10	9	4
12,505.85	11	4
15,329.05	17	4
19,477.20	20	4

a. How many indicator variables should be included in the multiple regression model relating weekly sales to the number of sales reps and the quarter of the year? Explain why.

b. What sign would you expect the coefficient for the sales reps variable to have? Explain your reasoning.

c. Using statistical software, estimate the following multiple regression model.
$\text{Sales} = \beta_0 + \beta_1(\text{Reps}) + \beta_2(\text{Quarter 1}) + \beta_3(\text{Quarter 2}) + \beta_4(\text{Quarter 3}) + \varepsilon_i$
Write the estimated multiple regression equation.

d. Interpret the coefficient of the indicator variable representing the first quarter.

e. Is there sufficient evidence that sales in the second quarter tend to be different from the sales in the fourth quarter? Use $\alpha = 0.05$.

f. What concerns should we have when predicting weekly sales using this model?

CR Chapter Review

Key Terms and Ideas

- Multiple Regression
- Multiple Regression Model
- Method of Least Squares
- Estimated Multiple Regression Equation
- Coefficient of Determination (Multiple Coefficient of Determination)
- Adjusted R^2
- F-Distribution
- Numerator Degrees of Freedom
- Denominator Degrees of Freedom
- F-Statistic
- Sum of Squares of Regression
- Sum of Squared Errors
- Total Sum of Squares
- Mean Square Regression
- Mean Square Error
- Calculating Degrees of Freedom in Multiple Regression Models
- Hypothesis Tests Concerning Individual Coefficients
- Test Statistic for Testing the Hypothesis $\beta_i \neq 0$
- Confidence Intervals for Individual Coefficients
- Confidence Interval for the Mean Value of y Given x
- Confidence Interval for the Predicted Value of y Given x
- Indicator (Dummy) Variable
- Base Level Variable
- Interaction Terms
- Polynomial (Nonlinear) Regression Models

Key Formulas

Section	
14.1	**Multiple Regression Model** $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$ where β_0, β_1, β_2, ..., β_k are the model's parameters, x_1, x_2, ..., x_k are the independent variables, and ε is a random error.