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.

Definition

Polynomial Regression

Polynomial regression models are used when the relationship between the independent variable and the dependent variable are modeled using an n^{th} degree polynomial in the independent variable. For example, the regression model could resemble $y = \beta_0 + \beta_1 x^3 + \varepsilon$.

14.5 Exercises

Basic Concepts

1. Explain why qualitative variables have not been used in the regression models we have discussed in previous sections.

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

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

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

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

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

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

Exercises

8. 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).

$$\text{GPA} = -0.8777 + 0.0672(\text{Attendance}) + 0.0678(\text{Exam Score}) - 0.1436(\text{Freshman})$$

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?

9. 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 GPA using a 5% significance level.

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

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

d. Do you believe this is a good model to use to predict GPA? Why or why not?

e. 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.

10. A personnel director is interested in studying the effects which age, experience, and gender have on salary. Eight employees are randomly selected and each employee's salary, age, experience, and gender (= 0 if male, = 1 if female) are recorded.

Employee Data

Salary ($)	Age	Experience (Years)	Gender
27,000	25	2	1
50,000	55	20	0
48,000	27	5	0
35,000	30	7	1
29,000	22	3	1
58,000	33	8	1
23,000	19	1	0
43,000	45	15	1

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

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

$$\text{Salary} = \beta_0 + \beta_1(\text{Age}) + \beta_2(\text{Experience}) + \beta_3(\text{Gender}) + \varepsilon_i.$$

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 gender useful in explaining salary? Test at the 0.10 level.

g. Interpret each of the regression coefficients.

h. Construct and interpret 95% confidence intervals for each of the regression coefficients. Do you think that the company discriminates in the salary paid based on gender?

i. Predict the salary of a female employee who is 35 years old with 10 years of experience.

j. Construct and interpret a 95% prediction interval for a female employee who is 35 years old with 10 years of experience. How useful is this interval?

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

Data

This data set can be found at stat.hawkeslearning.com by navigating to **Discovering Business Statistics, Second Edition > Data Sets > Campus Crime**.

11. 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.

Campus Crime Data

Number of Crimes	Number of Police	Total Enrollment	Private School
64	12	1131	Yes
138	21	12,954	No
141	32	16,009	No
84	22	1682	Yes
86	35	2888	Yes
141	45	17,407	No
135	42	3028	Yes
174	50	4306	Yes
201	75	34,511	No
203	84	37,240	No
125	36	2918	Yes
234	109	39,414	No
143	45	4000	Yes
148	50	20,950	No
152	48	4277	Yes
158	52	26,519	No
174	69	27,687	No
84	26	2810	Yes
173	58	27,619	No
193	56	4563	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.

$$\text{Crimes} = \beta_0 + \beta_1(\text{Police}) + \beta_2(\text{Enrollment}) + \beta_3(\text{Private}) + \varepsilon$$

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 $\alpha = 0.01$?

12. 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.

Data

This data set can be found at stat.hawkeslearning.com by navigating to **Discovering Business Statistics, Second Edition > Data Sets > Weekly Sales by Quarter**.

Weekly Sales by Quarter

Weekly Sales ($)	Number of Sales Reps	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
4272.90	3	1	0	0	0
5069.70	9	1	0	0	0
6067.70	11	1	0	0	0
6680.55	17	1	0	0	0
9725.05	20	1	0	0	0
4107.10	3	0	1	0	0
7520.25	9	0	1	0	0
12,135.00	11	0	1	0	0
13,016.55	17	0	1	0	0
13,673.90	20	0	1	0	0
3272.05	3	0	0	1	0
5074.40	9	0	0	1	0
7505.45	11	0	0	1	0
8272.75	17	0	0	1	0
10,020.40	20	0	0	1	0
4925.75	3	0	0	0	1
10,018.10	9	0	0	0	1
12,505.85	11	0	0	0	1
15,329.05	17	0	0	0	1
19,477.20	20	0	0	0	1

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.

$$\text{Sales} = \beta_0 + \beta_1(\text{Reps}) + \beta_2(\text{Quarter 1}) + \beta_3(\text{Quarter 2}) + \beta_4(\text{Quarter 3}) + \varepsilon$$

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?