Key Formulas		
Section		
14.1	**Estimated Multiple Regression Equation** $\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k$ where b_0, b_1, b_2, ..., b_k are estimates of their population counterparts.	**Sum of Squared Errors** $\text{SSE} = \sum\left(y - \hat{y}\right)^2$
14.2	**Coefficient of Determination** $R^2 = \dfrac{\text{SSR}}{\text{Total SS}} = 1 - \dfrac{\text{SSE}}{\text{Total SS}}$	**Adjusted R^2** $R_a^2 = 1 - \left(\dfrac{n-1}{n-k-1}\right)\dfrac{\text{SSE}}{\text{Total SS}}$ where n is the number of observations and k is the number of independent variables in the model.
14.3	***F*-Statistic** $F = \dfrac{\dfrac{\text{Sum of Squares of Regression}}{k}}{\dfrac{\text{Sum of Squared Errors}}{n-(k+1)}} = \dfrac{\dfrac{\text{SSR}}{k}}{\dfrac{\text{SSE}}{n-(k+1)}} = \dfrac{\text{Mean Square Regression}}{\text{Mean Square Error}}$	**100(1 − α)% Confidence Interval for an Individual Coefficient, β_i** $b_i \pm t_{\alpha/2,\, n-(k+1)}s_{b_i} \quad i = 1, \ldots, k$ **Test Statistic for Testing the Hypothesis $\beta_i \neq 0$** $t = \dfrac{b_i - 0}{s_{b_i}} = \dfrac{b_i}{s_{b_i}} \quad i = 1, \ldots, k$ where b_i is the estimated coefficient and s_{b_i} is the standard deviation (standard error) of the estimated coefficient.

Additional Exercises

1. Drew is undecided about whether to go back to school and get his master's degree. He is trying to perform a cost-benefit analysis to determine whether the cost of attending the school of his choice will be outweighed by the increase in salary he will receive after he attains his degree. He does research and compiles data on annual salaries in the industry he currently works in (he has been working for 10 years), along with the years of experience for each employee and whether or not the employee has a master's degree. He has decided that if the multiple regression model shows, with 95% confidence, that earning a master's degree is significant in predicting annual salary, and the estimated increase in salary is at least $20,000, he will enroll in a degree program.

Data

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

Industry Salaries					
Salary ($)	Years of Experience	Master's Degree	Salary ($)	Years of Experience	Master's Degree
75,240	22	No	73,360	22	No
134,160	27	Yes	58,400	11	Yes
62,560	15	No	66,080	18	No
43,000	2	No	60,120	14	No
150,240	28	Yes	106,600	21	Yes
119,640	25	Yes	45,640	7	No
80,360	15	Yes	145,800	31	Yes
162,720	32	Yes	111,840	22	Yes
70,160	19	No	38,560	0	No
72,160	12	Yes	52,000	7	No

a. Create an indicator variable, degree, that is equal to 1 if the employee has a master's degree and equal to 0 if the employee does not have a master's degree.

b. Using statistical software, estimate the following multiple regression model.

$$Salary = \beta_0 + \beta_1\ Experience + \beta_2\ Degree + \varepsilon$$

Write the estimated multiple regression equation.

c. According to the model, how much does salary increase on average with each additional year of experience?

d. Interpret the meaning of the coefficient of the indicator variable *Degree* in the estimated multiple regression equation.

e. According to this model, will Drew decide to enroll in a master's program? Explain your answer.

f. Why should Drew be cautious when using this model to make his decision?

2. The amount of a certain additive injected into a chemical process has a direct effect on the yield. The following table contains data on the amount of additive and yield.

Amount of Additive and Yield											
Additive	12.0	6.7	5.6	13.2	8.9	7.8	12.9	16.4	4.5	9.6	5.8
Yield	96	50	42	82	76	70	89	94	15	75	32

a. Assuming that *Yield* is the dependent variable, plot *Yield* against *Additive*. Does the relationship appear to be linear?

b. Using statistical software, estimate the simple linear regression model. Identify R^2 and s_e^2.

c. In instances such as this where linearity does not hold, polynomial regression can be used to provide a better fit to the data. Polynomial regression is a special case of multiple regression where new predictor variables are formed by raising other predictor variables to integral powers. In this exercise, a new predictor will be formed by squaring the values of additive (*Add_sq*). *Yield* will then be fitted to the predictors *Additive* and *Add_sq*. The prediction equation based upon the polynomial regression is $Yield = -67.53 + 23.04\ Additive - 0.82\ Add\_sq$. R^2 and s_e^2 are 0.95 and 47.53, respectively. Predict the yield when $Additive = 16$. Make this prediction using both the linear and polynomial fits. Compare your results.

d. Compare the linear and polynomial fits to the data by the values for R^2 and s_e^2.

e. Which model do you believe is best to use for estimation and prediction? Explain your answer.

3. The following table contains a list of high-dividend exchange-traded funds (ETFs). Exchange-traded funds are investment funds traded on stock exchanges, much like stocks. ETFs are traditionally index funds, but ETFs can hold assets such as stocks, commodities, or bonds, and trade at approximately the same price as the net asset value of their underlying assets over the course of the trading day. ETFs may be attractive as investments because of their low costs, tax efficiency, and stock-like features. The full data set can be found on the companion site.

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > Exchange-Traded Funds

Exchange-Traded Funds

ETF	Share Price ($)	Dividend Per Share ($)	Dividend Yield (%)	ETF	Share Price ($)	Dividend Per Share ($)	Dividend Yield (%)
1	4.32	0.28	6.49	6	120.55	6.22	5.16
2	15.38	0.92	6.02	7	23.00	1.09	4.74
3	25.25	1.43	5.66	8	13.34	0.62	4.66
4	21.28	1.13	5.31	9	24.82	1.15	4.63
5	698.75	36.88	5.28	10	22.96	1.04	4.53
...							

a. Using the Exchange-Traded Funds data set, can dividend yield be predicted by share price and dividend per share? Is it a useful model? Justify your answers.

b. Which variable explains the greatest amount of variability in dividend yield? Explain your answer.

Data

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

4. The Supplemental Nutrition Assistance Program (SNAP) provides monthly benefits that help eligible low-income households buy the food they need for good health. For most households, SNAP funds account for only a portion of their food budgets, so they must also use their own funds to buy enough food to last throughout the month. Using the SNAP data set, answer the following questions to help predict monthly benefits to eligible households.

SNAP Benefits					
Monthly Benefit ($)	Family Size	Gross Monthly Income ($)	Monthly Benefit ($)	Family Size	Gross Monthly Income ($)
603.41	5	3753	556.42	1	3098
560.69	3	3778	569.05	8	3707
623.24	6	3609	365.80	8	2071
416.12	5	2262	489.08	5	3166
323.90	1	1966	495.86	4	3126
...					

a. Suggest a regression model that will assist SNAP administrators in providing a monthly benefit to eligible households.

b. Fit the model that you suggested in part **a.** Is this model useful in predicting monthly benefits? Justify your answer.

c. Are all independent variables in the model helpful in explaining the variation in monthly benefits? Explain your answer.

d. Give a 95% confidence interval for average monthly benefits for a four-member household with a gross monthly income of $2500. Interpret this interval.

e. Provide a 99% prediction interval for a four-member household with a gross monthly income of $2500. Interpret this interval.

f. What is the difference between the intervals found in parts **d.** and **e.**?

Note

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