In order to compute the t-value, the degrees of freedom must be determined.

$$df = n - (k + 1) = 25 - (2 + 1) = 22$$

For a 95% confidence interval, $t_{\alpha/2,df}$ will be $t_{0.025,22} = 2.074$. The resulting confidence interval will be

$$1.5891 \pm 2.074(0.1563)$$
$$1.5891 \pm 0.3242$$
$$1.2649 \text{ to } 1.9133$$

We are 95% confident that the true value of β_1, the increase in the delivery time for each additional pizza (given that distance is held constant), will be between 1.2649 and 1.9133 minutes.

Notice that JMP automatically generates confidence intervals for each individual coefficient. The endpoints for the 95% confidence intervals are given in the Lower 95% and Upper 95% columns of the output. Compare the upper and lower limits given by JMP to the ones just calculated by hand. JMP has the ability to calculate confidence intervals for individual coefficients for any level of significance.

14.3 Exercises

Basic Concepts

1. If the overall multiple regression model is not useful, what does this tell us about the coefficients of the independent variables?

2. What is the hypothesis being tested when we test to determine if the overall multiple regression model is useful?

3. When testing the overall model, describe the null and alternative hypotheses in plain English.

4. Why is the R^2 value not used in the test statistic for a hypothesis test to determine if a multiple regression model is significant?

5. What is the test statistic used in a hypothesis test to determine if an overall model is significant? What is the distribution of this test statistic?

6. Give two equivalent formulas for the test statistic in a hypothesis test about an overall regression model.

7. Explain the significance of the ratio of the mean square regression to the mean square error.

8. True or false: Even if there is no relationship between any of the independent variables and the dependent variable, sampling variation will explain some portion of the variation in the dependent variable.

9. How are the degrees of freedom calculated for a multiple regression model?

10. When testing the overall model for significance, do you perform a one or two-tailed test?

11. What is the rejection rule in tests of hypothesis for model significance?

12. What is the expression for a confidence interval for an individual coefficient, β_i?

13. Outline the three pieces of information needed to compute a confidence interval for an individual coefficient.

14. What is the test statistic used to test a hypothesis about an individual coefficient in a multiple regression model? How many degrees of freedom are associated with this test statistic?

15. If we fail to reject the null hypothesis in a hypothesis test about an individual coefficient, should this variable remain in the regression model? Explain.

Exercises

16. An article appearing in the *Journal of Wildlife Management* summarized the percent fat of 75 arctic foxes. According to the authors, "Storage of fat to provide energy during regular periods of food shortage, or to insulate against low ambient temperatures, is essential for survival in severe arctic homeotherms." Computing percent fat was a laborious process. In one analysis, the author regressed y = percent fat on rump fat thickness (RFT), which was measured in millimeters and was much easier to determine than percent fat. It was noted that a plot of percent fat versus RFT was indeed linear, and the resulting regression equation was $y = 7.40 + 1.36(\text{RFT})$. R^2 and s_e^2 were determined to be 0.88 and 3.70, respectively.

ANOVA

	df	*SS*	*MS*	*F*
Regression	k	SSR	SSR/k	MSR/MSE
Residual	$n-(k+1)$	SSE	SSE/$(n-(k+1))$	
Total	$n-1$	SST		

Use the table shown to help answer the following questions.

a. Compute the sum of squares of regression and the sum of squared errors. Note that R^2 is SSR/TSS and s_e^2 is the same as MSE.

b. Give the degrees of freedom for the regression and for the error (residual).

c. Compute MSR and MSE.

d. Compute the F ratio for testing the significance of the regression line. With $\alpha = 0.05$, can we conclude that the relationship between percent fat and RFT is significant?

e. Compute a point estimate for percent fat if RFT = 10 millimeters.

f. For a 2-millimeter increase in RFT, what would be the expected change in percent fat?

17. Consider the model from Exercise 10 in Section 14.1 relating annual salary to years of work experience and years of education.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.566946595
R Square	0.321428441
Adjusted R Square	0.29192533
Standard Error	10909.996
Observations	49

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	2593556200	1296778100	10.89473033	0.000133875
Residual	46	5475288584	119028012.7		
Total	48	8068844784			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	11214.19915	5625.172956	1.993574106	0.052147881	-108.6867382	22537.08504
Education (Years)	2854.891271	689.6666061	4.139523715	0.000146836	1466.664395	4243.118147
Experience (Years)	839.6360369	261.7094444	3.208275646	0.002433357	312.842248	1366.429826

a. Formulate the hypotheses for testing the multiple regression model for overall significance.

b. Find the value of the test statistic for a hypothesis test about the overall model.

c. Is there evidence at the 5% level of significance that the overall model is useful in predicting annual salary?

d. Consider the coefficient for years of education. Find a 95% confidence interval for the value of β_1. Interpret this interval.

e. Formulate the hypotheses for testing the significance of the coefficient β_1.

f. Is there sufficient evidence at the 0.05 level that years of education is useful in predicting annual salary?

18. Consider the printing cost model discussed in Exercise 11 of Section 14.1.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.987606014
R Square	0.975365639
Adjusted R Square	0.972467479
Standard Error	0.445885396
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	133.8204656	66.91008281	336.5464936	2.12863E-14
Residual	17	3.379834375	0.198813787		
Total	19	137.2			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	6.134155476	3.993435752	1.536059638	0.142925974	-2.291257484	14.55956844
Number of Pages	0.010801	0.004147682	2.604105041	0.018522101	0.002050156	0.019551845
Number of Copies	-0.009954478	0.005271436	-1.888380579	0.07616193	-0.021076236	0.00116728

a. What percentage of the variation in printing price is explained by the two independent variables number of pages and number of copies?

b. Is the overall model significant at the 1% level?

c. Consider the estimated regression coefficient for the number of pages. Construct a 99% confidence interval for β_1. Interpret this interval.

d. Is the number of pages variable useful in predicting printing cost at the 5% level? Would the decision change at the 1% level?

e. Construct a 95% confidence interval for β_2. Interpret this interval.

f. Is the number of copies useful in explaining the variation in printing cost at the 5% level of significance? Do you think the publisher should consider removing this variable from the model? Explain your answer.

19. The following table contains data from selected cities regarding rental rates of two-bedroom apartments, city populations, and median incomes. Monthly rent is given in dollars, population is given in thousands of people, and median income is given in thousands of dollars. Suppose we wish to build a multiple regression model to predict the cost of rent based on population and median income.

Monthly Rent, Population, and Median Income in Selected Cities

City	Monthly Rent ($)	2010 Population (Thousands)	2010 Median Income (Thousands of Dollars)
Denver, CO	868	600.158	45.438
Birmingham, AL	711	212.237	31.704
San Diego, CA	1414	1307.402	61.962
Gainesville, FL	741	124.354	28.653
Winston-Salem, NC	707	229.617	41.979
Memphis, TN	819	646.889	36.535

Monthly Rent, Population, and Median Income in Selected Cities (cont.)

City	Monthly Rent ($)	2010 Population (Thousands)	2010 Median Income (Thousands of Dollars)
Austin, TX	966	790.390	50.236
Seattle, WA	1219	608.660	58.990
Richmond, VA	735	204.214	37.735
Charleston, SC	812	120.083	47.799
College Park, MD	1407	30.413	66.900
Savannah, GA	789	136.286	33.778
Minneapolis, MN	988	382.578	45.625
Detroit, MI	805	713.777	29.447
Baton Rouge, LA	827	229.493	35.436

Source: U.S. Census Bureau

a. Write the multiple regression model in terms of rent, population, and income. Assume the regression coefficients have not yet been estimated.

b. Predict the signs of the coefficients β_1 and β_2. Explain your answers.

c. Using statistical software, estimate the multiple regression equation. Identify the values of b_0, b_1, and b_2 and write the estimated multiple regression equation. Interpret the estimated coefficients.

d. At the 1% level of significance, is the overall model useful in predicting monthly rent? Identify the test statistic for this test.

e. Find a 95% confidence interval for β_2. Interpret this interval.

f. Determine if each independent variable is related to the dependent variable at the 0.05 level of significance.

g. Should we consider removing any independent variables from this regression model? If yes, identify the variable(s) that should be removed and explain why.

20. Using the information from Exercise 19, estimate the simple linear regression equation relating monthly rent to median income only.

a. Write the estimated simple regression equation.

b. Is the simple linear regression model significant at $\alpha = 0.01$?

c. Is median income related to the monthly rental rate at $\alpha = 0.01$? Identify the test statistic used in this hypothesis test.

d. What percent of the variation in monthly rent is explained by median income? Compare this to the percent of variation in monthly rent explained by both population and median income in Exercise 19.

e. Which model do you think is a better model to use to predict monthly rental rates? Explain your answer.

14.4 Inference Concerning the Model's Prediction

Many regression models are developed solely to predict the dependent variable. To use the multiple regression model for prediction, insert the values of the independent variables in the model and calculate the predicted value. Recall the estimated multiple regression equation for our pizza delivery model:

$$\text{Delivery Time} = 1.7929 + 1.5891(\text{Number of Pizzas}) + 1.5677(\text{Distance}).$$