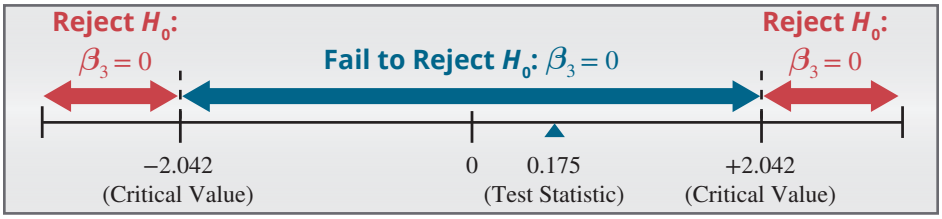
Using the output we find that

$$t = \frac{b_3 - 0}{s_{b_3}} = \frac{b_3}{s_{b_3}} = \frac{1284.92}{7357.66} \approx 0.175.$$

The estimated value of b_3 is 0.175 standard deviations from zero. Is this persuasive evidence that $\beta_3 \neq 0$?



Step 4: Determine the critical value(s) or *P*-value.

This criteria is defined by the critical value of the test statistic. Since the test is two-tailed and $\alpha = 0.05$ then $\alpha/2 = 0.05/2 = 0.025$. The test statistic has a t-distribution with $df = 34 - (3 + 1) = 30$. The critical value corresponds to $t_{0.025,\ 30} = 2.042$. The P-value for the number of bedrooms is given in the output as 0.8625.

Step 5: Choose between the null and alternative hypotheses.

Since the value of the test statistic falls into the *Fail to Reject* region, there is insufficient evidence at the 0.05 level to reject the null hypothesis $\beta_3 = 0$. Alternatively, since the P-value of 0.8625 is greater than $\alpha = 0.05$, we fail to reject the null hypothesis.

Step 6: State the conclusion in terms of the original question.

Since we did not reject the null hypothesis H_0: $\beta_3 = 0$, then the variable *Bedrooms* is not a significant predictor of *List Price*, given the other variables currently in the model.

We apply the exact same t-test to the other variables in the model. Both b_1 and b_2, the coefficients of the variables *Square Footage* and *Age,* are significant. This suggests that a model with only two independent variables, *Square Footage* and *Age*, may produce a model almost as good as the one containing three variables.

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. 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?

5. Explain the significance of the ratio of the mean square regression to the mean square error.
6. 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.
7. How are the degrees of freedom calculated for a multiple regression model?
8. When testing the overall model for significance, do you perform a one or two-tailed test?
9. What is the rejection rule in tests of hypothesis for model significance?
10. What is the expression for a confidence interval for an individual coefficient, β_i?
11. Outline the three pieces of information needed to compute a confidence interval for an individual coefficient.
12. 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?
13. 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.
14. Does a low R^2 imply that a model will not be useful for prediction?

Exercises

Data

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

15. In Lesson 5.3 Exercise 24, we used the Moneyball data set (1962-2001) to look at the individual relationships between runs scored (*RS*) and on-base percentage (*OBP*) and runs scored (*RS*) and slugging percentage (*SLG*). On-base percentage (*OBP*) and slugging percentage (*SLG*) were determined to be two of the most statistically significant variables that contributed to the number of runs scored. Remember that a run differential of 135 runs was identified as necessary to make the MLB playoffs. Use the Moneyball data set, subsetted to only the years 1962-2001, to perform the following.
    a. Build a single model to predict runs scored (*RS*) using on-base percentage (*OBP*) and slugging percentage (*SLG*). Write the estimated regression equation.
    b. Is the overall model significant at the 1% level?
    c. What percent of variation in the runs scored (*RS*) is explained by on-base percentage (*OBP*) and slugging percentage (*SLG*)?
    d. Determine if each independent variable is related to the dependent variable at the 0.01 level of significance.
    e. Should we consider removing any independent variables from this regression model? If yes, identify the variable(s) that should be removed and explain why.
16. Consider the model from Exercise 9 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 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?

17. Consider the printing cost model discussed in Exercise 10 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	133.8201656	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 Cost 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.

18. The following table contains US Census Bureau data from selected cities regarding rental rates of two-bedroom apartments, city populations, and median incomes.[1] 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.

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > City Population Data

Monthly Rent, Population, and Median Income in Selected Cities

City	Monthly Rent ($)	2020 Population (Thousands)	2020 Median Income
Denver, CO	1397	715.522	$72,661
Birmingham, AL	870	200.733	$38,832
San Diego, CA	1770	1386.932	$83,454
Gainesville, FL	965	141.085	$38,028
Winston-Salem, NC	827	249.545	$47,269
Memphis, TN	915	633.104	$41,864
Austin, TX	1346	961.855	$75,752
Seattle, WA	1702	737.015	$97,185
Richmond, VA	1070	226.610	$51,421
Charleston, SC	1318	150.227	$72,071
College Park, MD	1583	34.740	$68,825
Savannah, GA	1049	147.780	$46,149
Minneapolis, MN	1078	429.954	$66,068
Detroit, MI	850	639.111	$32,498
Baton Rouge, LA	886	227.470	$44,177

a. Write the multiple regression model using population and median income to predict rent. 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.

19. Using the information from Exercise 18, estimate the simple linear regression equation using median income to predict rent.

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 18.

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. For a house with 2500 square feet that is ten years old with four bedrooms, the model would predict the price to be

$$\begin{aligned} \textit{List Price} &= 163579.06 + 108.24(2500) - 6318.47(10) + 1284.92(4) \\ &\approx \$376{,}134.04 \end{aligned}$$

This is the point estimate. How good is this estimate? The answer to this question depends on what you are trying to predict. Are you trying to predict the average price for all 2500 square foot homes that are ten years old with four bedrooms, or are you trying to predict the price of a particular home of this type?

Note

The value of \$376,134 for a 2500 square foot home that is 10 years old and has 4 bedrooms that is estimated using the Home Price model differs slightly from the value calculated in the Minitab output (\$376,142) due to rounding.

Confidence Intervals for the Mean Value of *y* Given *x*

In Section 13.3 we discussed a confidence interval for the mean (or average) value of y given $x = x_p$ for the simple linear regression model. In our multiple regression model, the point estimate, \$376,134, is the mean value of y given $x_1 = 2500$, $x_2 = 10$, and $x_3 = 4$. In other words, the price of \$376,134 is the estimated average price for all ten-year-old, 2500 square feet homes with four bedrooms. Since we do not have all homes in the sample, the predicted average price of \$376,134 is only an estimate of the true average value. How good is the estimate? For multiple regression, the expression for the confidence interval of the mean value of y given x is beyond the scope of an introductory text.