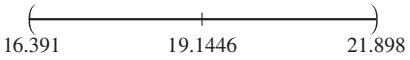
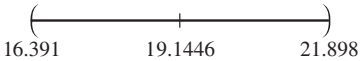
According to the Minitab output given in Figure 14.4.1, the 95% confidence interval for the mean delivery time for 5 pizzas being delivered 6 miles away is 16.391 minutes to 21.898 minutes.



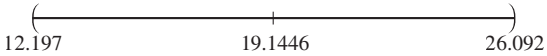
Confidence Interval for the Predicted Value of *y* Given *x*

A caller asks to speak to the manager of the pizza restaurant. The caller wants to know how long it would take to deliver five pizzas to his specific location, which is six miles away. As the manager, you would like to guarantee how long it will take to make this delivery of 5 pizzas to this customer. You are not especially interested in the *average* delivery time for such a delivery. Instead, it would be preferable to create a confidence interval for the time it is going to take for this particular order to be delivered. Once again, for multiple regression, the expression for the prediction interval is beyond the scope of this text. Fortunately, statistical analysis programs such as Minitab will also produce a prediction interval. Using the output shown in Figure 14.4.1, the 95% prediction interval for the delivery of 5 pizzas to a location 6 miles away is 12.197 minutes to 26.092 minutes. As we observed in Section 13.6, to account for individual variation, the prediction interval for y given x is substantially wider than the confidence interval for the mean value of y given x.

Confidence Interval for Average Delivery Time



Prediction Interval for Individual Delivery Time



Can the model make a useful prediction of the delivery time? Although the model has an R^2 of 0.9640, the prediction interval is fairly wide. This indicates that not a great deal of confidence can be placed in the estimated value of 19.1446 minutes as a delivery time for 5 pizzas traveling 6 miles.

14.4 Exercises

Basic Concepts

1. What is a point estimate for a multiple regression model?
2. Explain how a point estimate is interpreted as an "average" value.
3. Distinguish between a confidence interval and a prediction interval for a multiple regression model.
4. What is the price that is paid when making predictions regarding individual values?
5. Suppose an estimated multiple regression model, $\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i}$, produces a 95% confidence interval of (3.292, 7.072) and a 95% prediction interval of (0.364, 10.000) when $x_1 = 6$ and $x_2 = 6$. Interpret both of these intervals.

Exercises

6. Consider the multiple regression model predicting graduating GPA using both the SAT critical reading score and the SAT math score. Computer output of the model

$$\text{GPA} = \beta_0 + \beta_1 (\text{SAT Reading}) + \beta_2 (\text{SAT Math}) + \varepsilon_i$$

is given.

Session

Regression Analysis: Graduating GPA versus SAT Critical Reading, SAT Math

The regression equation is
Graduating GPA = 1.35 + 0.00151 SAT Critical Reading + 0.00134 SAT MATH

Predictor	Coef	SE Coef	T	P
Constant	1.3480	0.5760	2.34	0.028
SAT Critical Reading	0.001510	0.001757	0.86	0.399
SAT Math	0.001341	0.001337	1.00	0.326

S = 0.452511 R-Sq = 19.5% R-Sq (adj) = 12.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	1.1869	0.5934	2.90	0.075
Residual Error	24	4.9144	0.2048		
Total	26	6.1013			

Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI
1	2.7736	0.1416	(2.4814, 3.0659)	(1.7950, 3.7522)

Values of Predictors for New Observations

New Obs	SAT Critical Reading	SAT Math
1	500	500

a. Find the standard deviation of the error terms in the output.

b. Interpret the coefficient of SAT Critical Reading. What would it mean if the coefficient was negative?

c. Determine if the overall model is useful in explaining GPA. Test at the 0.05 level.

d. What fraction of the variation in GPA is explained by the model?

e. Determine if the SAT Critical Reading variable is a useful predictor of GPA. Test at the 0.05 level.

f. The output includes a predicted GPA for someone scoring 500 on both the SAT Critical Reading and SAT Math portions. Find the predicted value in the output.

g. What is the average GPA for an individual who scored 500 on both the SAT Critical Reading and SAT Math sections? Find the 95% confidence interval for this average. Interpret this interval.

h. Suppose your nephew scored 500 on both the critical reading and math sections. What would be the model's prediction for his graduating GPA? Find the 95% prediction interval for your nephew in the output. Interpret this interval.

i. Why is the prediction interval so much wider than the confidence interval in part **g.**?

j. Summarize the strengths and weaknesses of the estimated model.

7. How tall will your child be? A researcher has collected a random sample of heights of parents and their female children (all heights are in inches). The heights of the mother, father, and daughter are recorded in the following table.

Heights of Parents and Daughters (Inches)													
Mother	64	66	62	70	70	58	66	66	64	67	65	66	68
Father	73	70	72	72	72	63	75	75	72	69	77	70	74
Daughter	65	65	61	69	67	59	69	70	68	70	70	65	70

a. Create two scatterplots using the mother with the daughter and the father with the daughter. Does there appear to be a linear relationship in either of the plots?

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

$$\text{Daughter Height} = \beta_0 + \beta_1(\text{Mother Height}) + \beta_2(\text{Father Height}) + \varepsilon_i$$

c. Is the overall model useful in explaining the variation in daughter height? Test at the 0.05 level.

d. Is the father's height useful in explaining the daughter's height? Test at the 0.05 level.

e. Is the mother's height useful in explaining the daughter's height? Test at the 0.01 level.

f. Interpret each of the regression coefficients.

g. Construct and interpret 95% confidence intervals for β_1 and β_2. Interpret these intervals.

h. Predict the height of a daughter whose father is six feet two inches tall and whose mother is five feet four inches tall.

i. Find a 95% prediction interval for the height of a daughter whose father is six feet two inches tall and whose mother is five feet four inches tall. Interpret this interval.

j. Find a 95% confidence interval for the average height of a daughter whose father is six feet two inches tall and whose mother is five feet four inches tall.

8. On Sunday, January 2, 2011, 16 games were played in the National Football League. The number of rushing yards, passing yards, first downs, and points for the 32 teams participating in these games is given in the table.

NFL Teams 2011				
Team	**Rushing Yards**	**Passing Yards**	**First Downs**	**Points**
Miami	44	240	16	7
New England	181	321	24	38
Buffalo	37	130	6	7
New York (Jets)	276	119	17	38
Cincinnati	90	305	20	7
Baltimore	98	125	10	13
Pittsburgh	100	325	24	41
Cleveland	43	210	17	9
Oakland	209	160	21	31
Kansas City	115	142	17	10
Minnesota	74	145	16	13
Detroit	107	258	22	20
Carolina	137	182	12	10
Atlanta	99	256	24	31

Data

This data set can be found at stat.hawkeslearning.com under **Discovering Business Statistics, Second Edition > Data Sets > NFL Teams 2011**.

NFL Teams 2011 (cont.)

Team	Rushing Yards	Passing Yards	First Downs	Points
Tampa Bay	84	255	18	23
New Orleans	106	212	20	13
Jacksonville	198	140	23	17
Houston	244	253	22	34
Dallas	159	127	14	14
Philadelphia	121	162	14	13
New York (Giants)	82	243	14	17
Washington	67	336	20	14
San Diego	164	313	20	33
Denver	146	205	18	28
Arizona	78	242	19	7
San Francisco	100	276	16	38
Chicago	110	168	13	3
Green Bay	60	229	14	10
Tennessee	51	300	17	20
Indianapolis	101	264	24	23
Saint Louis	47	155	10	6
Seattle	141	192	19	16

Source: National Football League

a. In order to predict a team's points from rushing yards, passing yards, and first downs, a multiple regression analysis is performed on the data with points as the dependent variable. The associated regression output is given. Write the estimated regression equation for predicting points based on the three predictor variables.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.774540403
R Square	0.599912836
Adjusted R Square	0.557046354
Standard Error	7.508433133
Observations	32

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	2366.956093	788.9853643	13.99492	9.23535E-06
Residual	28	1578.543907	56.37656811		
Total	31	3945.5			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-15.6150327	5.982932568	-2.609929582	0.014378
Rushing Yards	0.127133936	0.029615758	4.292780033	0.000191
Passing Yards	0.081411959	0.029307161	2.777886231	0.009655
First Downs	0.121492053	0.460140758	0.264032366	0.793689

b. Find the standard deviation of the error terms in the output.

c. Determine if the overall model is useful in predicting points scored. Use $\alpha = 0.05$.

d. What fraction of the total variation in points is explained by the model?

e. Is the rushing yards variable useful in predicting points scored at the 0.01 level?

f. Is the passing yards variable useful in predicting points scored at the 0.01 level?

g. Is the first downs variable useful in predicting points scored at the 0.01 level?

h. The coefficient of rushing yards in the regression equation is 0.1271. Interpret this value.

i. Should any variables be removed from this model? Explain.

9. In the previous exercise, total points was predicted based on rushing yards, passing yards, and first downs. It is noted from the summary output that both rushing yards and passing yards have *P*-values of less than 0.01. However, first downs does not appear to be significant as an independent variable. Perhaps a simpler model would be better.

a. Using the data from the previous exercise, estimate the regression equation

$$\text{Points} = \beta_0 + \beta_1(\text{Rushing Yards}) + \beta_2(\text{Passing Yards}) + \varepsilon_i.$$

b. Is the overall model significant in predicting total points? Test at $\alpha = 0.01$.

c. What percentage of the variation in total points is explained by rushing yards and passing yards? Compare this to the percentage of the variation in total points that was explained by the three independent variables rushing yards, passing yards, and first downs.

d. Which model do you think would be better to use for estimation and prediction of total points; the model from Exercise 8 or the model in this exercise? Explain your answer.

e. Suppose that in preparation for the upcoming game against Miami, the coach of Buffalo wishes to predict the points that will be scored. He has studied Miami's defense in previous games, and predicts that the Buffalo offense will have approximately 102 rushing yards and 63 passing yards. How many points, according to the model, should Buffalo score in the next game?

f. Construct at 95% confidence interval for the average number of points that will be scored in the game against Miami. Interpret this interval.

g. Construct a 95% prediction interval for the number of points that will be scored in the game against Miami. Interpret this interval.

14.5 Models with Qualitative Independent Variables

Throughout Chapter 13 and this chapter, we have discussed quantitative variables in the regression models. Quantitative variables take on values on a well-defined scale, such as number of pizzas, miles to destination, income, age, and temperature. Many variables of interest in business and economics, however, are not quantitative, but qualitative. Examples of qualitative variables are gender (male or female), firm size (small, medium, or large), and type of investment (stock or mutual fund).

Definition

Indicator (or Dummy) Variable

An **indicator (or dummy) variable** is created to assign numerical values to classes of a categorical variable. The dummy variable allows one to use a single variable to represent multiple categories.

In order to use qualitative variables in regression analysis, we need to identify the classes of the qualitative variable quantitatively. To do this, we will use **indicator** (or **dummy**) **variables** that take on values of 0 and 1. If we have a qualitative variable with c classes, that variable will be represented by $c - 1$ indicator variables in the regression model, with each indicator variable taking on a value of 0 or 1.

Suppose we added a variable to our pizza model that asked the customer if they lived in town or out of town. The variable, let's call it town (in or out), has $c = 2$ classes. Thus, the variable town will be represented by $c - 1 = 1$ indicator variable in the model. Town could be modeled as follows.

$$x_3 = \begin{cases} 1 & \text{if In Town} \\ 0 & \text{otherwise} \end{cases}$$