profits if trades were made using the model's prediction of future prices of a given stock. In case you are wondering, the model eventually worked. Before the company was sold it was trading six percent of all trades on the New York Stock Exchange and the NASDAQ. Regression modeling is an incredibly effective tool for modeling real-world phenomena, but if you're interested in a career in model building, it's also worth exploring machine learning. Both machine learning and statistical modeling are data-driven, but what sets machine learning apart, particularly in terms of modeling, is its emphasis on a model's ability to perform predictive tasks and its capacity to utilize multiple modeling techniques to accomplish this objective. The ultimate goal of machine learning is to determine the optimal model for a given predictive task.

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, $\hat{y} = b_0 + b_1 x_1 + b_2 x_2$, 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

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > SAT Scores and Graduating GPA

6. Use the SAT Scores and Graduating GPA data set of 30 students that was discussed in Example 13.2.3. In that example, we estimated a model using only total SAT score to predict graduating GPA. However, with multiple regression, the *SAT Verbal* and *SAT Math* scores can be treated as separate variables in the model. Computer output of the model

$$College\ GPA = \beta_0 + \beta_1\, SAT\ Verbal + \beta_2\, SAT\ Math + \varepsilon$$

is given.

Regression Analysis: College GPA versus SAT Verbal, SAT Math

Analysis of Variance

Source	DF	Adj SS	Adj Ms	F-Value	P-Value
Regression	2	0.8763	0.4382	2.643	0.0895
Error	27	4.4760	0.1658		
Total	29	5.3524			

Model Summary

S	R-sq	R-sq(adj)
0.407160	16.37%	10.18%

Coefficient

Term	Coef	SE Coef	T-Value	P-Value
Constant	0.13128	1.15417	0.114	0.9103
SAT Verbal	0.00179	0.00137	1.311	0.2009
SAT Math	0.00273	0.00187	1.457	0.1566

Regression Equation

College GPA = 0.13128 + 0.00179 SAT Verbal + 0.00273 SAT Math

Prediction for College GPA

Settings

Variable	Setting
SAT Verbal	500
SAT Math	500

Prediction

Fit	SE Fit	95% CI	95% PI
2.39	0.18	(2.03, 2.75)	(1.48, 3.30)

a. What is the estimated regression model?

b. Use the output provided to determine the standard deviation of the error terms.

c. Interpret the coefficient of *SAT Verbal*. What would it mean if the coefficient was negative?

d. Determine if the overall model is useful in explaining *College GPA*. Test at the 0.05 level.

e. What proportion of the variation in GPA is explained by the model?

f. Determine if the *SAT Verbal* variable is a useful predictor of *College GPA*. Test at the 0.05 level.

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

h. What is the model's estimate of the average GPA for individuals who scored 500 on both the SAT Verbal and SAT Math sections? Find the 95% confidence interval for this average. Interpret this interval.

i. Suppose your nephew scored 500 on both the SAT Verbal and SAT 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.

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

k. 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.

$$Daughter\ Height = \beta_0 + \beta_1\ Mother\ Height + \beta_2\ Father\ Height + \varepsilon$$

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, September 12, 2021, 14 games were played in the National Football League. The number of rushing yards, passing yards, first downs, and points for the 28 teams participating in these games is given in the table.[2]

Note

As of the date of this publication, Excel does not perform 95% confidence intervals for the mean of y given x nor will it perform 95% prediction intervals. Minitab, R, and Rguroo can perform these analyses.

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > NFL Statistics 09/12/2021

Team Data: September 12, 2021									
Team	**Rushing Yards**	**Passing Yards**	**First Downs**	**Points**	**Team**	**Rushing Yards**	**Passing Yards**	**First Downs**	**Points**
Jacksonville Jaguars	76	319	20	21	**Houston Texans**	160	289	22	37
Seattke SeaHawks	140	241	18	28	**Indianapolis Colts**	113	223	23	16
Los Angeles Chargers	90	334	27	20	**Washington**	126	133	15	16
NY Jets	45	207	16	14	**Carolina Panthers**	111	270	18	19
Minnesota Vikings	67	336	24	24	**Cincinnati Bengals**	149	217	20	27
Arizona Cardinals	136	280	22	38	**Tennessee Titans**	86	162	17	13
San Francisco 49ers	131	311	21	41	**Detroit Lions**	116	314	31	33
Pittsburgh Steelers	75	177	16	23	**Buffalo Bills**	117	254	22	16
Philadelphia Eagles	173	261	24	32	**Atlanta Falcons**	124	136	19	6
Cleveland Browns	153	304	24	29	**Kansas City Chiefs**	73	324	21	33
Green Bay Packers	43	186	14	3	**New Orleans Saints**	171	151	22	38
Denver Broncos	165	255	24	27	**NY Giants**	60	254	19	13
Miami Dolphins	74	185	16	17	**New England Patriots**	125	268	24	16
Chicago Bears	134	188	24	14	**Los Angeles Rams**	74	312	18	34

a. In order to predict a team's points from rushing yards, passing yards, and first downs, a multiple regression model is constructed. 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	0.7545798
R Square	0.5693906
Adjusted R Square	0.5155645
Standard Error	7.0592337
Observations	28

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	1581.442	527.1473	10.57832	0.000127
Residual	24	1195.987	49.83278		
Total	27	2777.429			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-8.465112	7.655233	-1.10579	0.279778	-24.2647	7.33451	-24.2647	7.33451
Rushing Yards	0.1723691	0.042912	4.016838	0.000505	0.083804	0.26093	0.0838	0.26093
Passing Yards	0.1128868	0.028515	3.958808	0.000585	0.054034	0.17174	0.05403	0.17174
First Downs	-0.737402	0.515136	-1.43147	0.16519	-1.80059	0.32579	-1.80059	0.32579

b. Use the regression equation to predict the points scored by a team that rushed for 152 yards, had 190 passing yards, and 21 first downs.

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

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

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

f. Is the *Rushing Yards* variable useful in predicting points scored at the 0.01 level?

g. Is the *Passing Yards* variable useful in predicting points scored at the 0.01 level?

h. Is the *First Downs* variable useful in predicting points scored at the 0.01 level?

i. The coefficient of *Rushing Yards* in the regression equation is 0.1724. Interpret this value.

j. 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

$$Points = \beta_0 + \beta_1\, Rushing\ Yards + \beta_2\, Passing\ Yards + \varepsilon.$$

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 263 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.

10. A personnel director is interested in studying the effects that age and work experience have on annual salary. Eight employees are randomly selected, and each employee's salary, age, and years of work experience are recorded.

Employee Data		
Salary	**Age**	**Experience**
$43,500	25	3
$75,000	55	20
$72,000	47	15
$52,500	30	7
$40,500	22	2
$87,000	62	26
$34,500	19	1
$64,500	44	10

a. Using statistical software, estimate the multiple regression equation for $Salary = \beta_0 + \beta_1 Age + \beta_2 Experience + \varepsilon$.

b. Determine if the overall model is useful in explaining salary at the 0.05 level of significance. What is the test statistic for the hypothesis test?

c. Is the *Experience* variable useful in predicting annual *Salary* at the 0.05 significance level?

14.5 Multiple Regression 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, however, are not quantitative, but qualitative. Examples of qualitative variables are type of college (public or private), season of the year (spring, summer, fall, and winter), and type of investment (stocks, mutual funds, or bonds).