

STA 6127: Exercises for Multiple Regression (Chap. 11)

- For students at Walden University, the relationship between Y = college GPA (with range 0–4.0) and X_1 = high school GPA (range 0–4.0) and X_2 = college board score (range 200–800) satisfies $E(Y) = 0.20 + 0.50x_1 + 0.002x_2$.
 - Find the mean college GPA for students having (i) high school GPA = 4.0 and college board score = 800, (ii) $x_1 = 3.0$ and $x_2 = 300$.
 - Show that the relationship between Y and x_1 for those students with $x_2 = 500$ is $E(Y) = 1.2 + 0.5x_1$.
 - Show that when $x_2 = 600$, $E(Y) = 1.4 + 0.5x_1$. Thus, increasing x_2 by 100 shifts the line relating Y to x_1 upward by $100\beta_2 = 0.2$ units.
 - Show that setting x_1 at a variety of values yields a collection of parallel lines, each having slope 0.002, relating the mean of Y to x_2 .
- A regression analysis with recent U.N. data from several nations on Y = percentage of people who use the Internet, X_1 = per capita gross domestic product (in thousands of dollars), and X_2 = percentage of people using cell phones has results shown in the table.
 - Write the prediction equation.
 - Find the predicted Internet use for a country with per capita GDP of 10 thousand dollars and 50% using cell phones.
 - Find the prediction equations when cell-phone use is (i) 0 %, (ii) 100%, and use them to interpret the effect of GDP.
 - Use the equations in (c) to explain the ‘no interaction’ property of the model.

Table 1:

	B	Std. Error	t	Sig
(Constant)	-3.601	2.506	-1.44	0.159
GDP	1.2799	0.2703	4.74	0.000
CELLULAR	0.1021	0.0900	1.13	0.264

R Square .796

ANOVA

	Sum of Squares	DF
Regression	10316.8	2
Residual Error	2642.5	36
Total	12959.3	38

- Refer to the previous exercise.
 - Show how to obtain R -squared from the sums of squares in the ANOVA table. Interpret it.
 - $r^2 = 0.78$ when GDP is the sole predictor. Why do you think R^2 does not increase much when cell-phone use is added to the model, even though it is itself highly associated with Y (with $r = 0.67$)? (Hint: Would you expect X_1 and X_2 to be highly correlated? If so, what’s the effect?)
- Use software with the “2004 statewide crime” data file at the course website, using murder rate as the response variable and poverty rate and percent of high school graduates as the explanatory variables.
 - Construct the partial regression plots, and interpret.
 - Fit the multiple regression model. Report the prediction equation, and explain how to interpret the estimated coefficients.
 - Re-do the analyses after deleting the D.C. observation. Does this observation have much influence on the results?

5. For recent UN data for several nations, a regression of carbon dioxide use (CO₂, a measure of air pollution) on gross domestic product (GDP) has a correlation of 0.786. With life expectancy as a second explanatory variable, the multiple correlation is 0.787.
- Explain how to interpret the multiple correlation.
 - For predicting CO₂, did it help much to add life expectancy to the model? Does this mean that life expectancy is very weakly correlated with CO₂? Explain.
6. For a random sample of 66 state precincts, data are available on

- Y = Percentage of adult residents who are registered to vote
 X_1 = Percentage of adult residents owning homes
 X_2 = Percentage of adult residents who are nonwhite
 X_3 = Median family income (thousands of dollars)
 X_4 = Median age of residents
 X_5 = Percentage of residents who have lived in the precinct at least ten years

The table shows a portion of the printout used to analyze the data.

- Fill in all the missing values in the printout, indicating in each ‘Sig’ space whether $P > 0.05$, $0.01 < P < 0.05$, $0.001 < P < 0.01$, or $P < 0.001$.
- Do you think it is necessary to include all five explanatory variables in the model? Explain.
- To what test does the “F Value” refer? Interpret the result of that test.
- To what test does the t -value opposite x1 refer? Interpret the result of that test.

Table 2:

	Sum of Squares	DF	Mean Square	F	Sig	R-Square
Regression	----	---	----	----	----	----
Residual	2940.0	---	----			Root MSE
Total	3753.3	---				----

Variable	Parameter Estimate	Standard Error	t	Sig
Intercept	70.0000			
x1	0.1000	0.0450	----	----
x2	-0.1500	0.0750	----	----
x3	0.1000	0.2000	----	----
x4	-0.0400	0.0500	----	----
x5	0.1200	0.0500	----	----

7. Refer to the previous exercise.
- Find a 95% confidence interval for the change in the mean of Y for a 1-unit increase in the percentage of adults owning homes, controlling for the other variables. Interpret.
 - Find a 95% confidence interval for the change in the mean of Y for a 50-unit increase in the percentage of adults owning homes, controlling for the other variables. Interpret.
8. Use software with the “house selling price” data file at the course website to conduct a multiple regression analysis of Y = selling price of home, X_1 = size of home, X_2 = number of bedrooms, X_3 = number of bathrooms.

- a) Use graphics to display the effects of the predictors. Interpret, and explain how the highly discrete nature of x_2 and x_3 affects the plots.
- b) Report the prediction equation and interpret the estimates.
- c) Inspect the correlation matrix, and report the variables having the (i) strongest association, (ii) weakest association.
- d) Report R^2 , and interpret.
- e) Find the F statistic for testing the overall effect of the three predictors, report its df values and its P -value, and interpret.
- f) Find the t test statistic for $H_0: \beta_3 = 0$, report its P -value for $H_a: \beta_3 > 0$, and interpret.
9. Refer to the previous exercise. Now use only number of bathrooms and number of bedrooms as predictors.
- a) Again test the partial effect of number of bathrooms, and interpret.
- b) Construct a 95% confidence interval for the coefficient of number of bathrooms, and interpret.
- c) Find the partial correlation between selling price and number of bathrooms, controlling for number of bedrooms. Compare it to the correlation, and interpret.
- d) Find the estimated standardized regression coefficients for the model, and interpret.
- e) Write the prediction equation using standardized variables. Interpret.
10. A study analyzes relationships among Y = percentage vote for Democratic candidate, X_1 = percentage of registered voters who are Democrats, and X_2 = percentage of registered voters who vote in the election, for several congressional elections in 2006. The researchers expect interaction, since they expect a higher slope between Y and x_1 at larger values of x_2 than at smaller values. They obtain the prediction equation $\hat{Y} = 20 + 0.30x_1 + 0.05x_2 + 0.005x_1x_2$. Does this equation support the direction of their prediction? Explain.
11. Use software with the “house selling price” data file to allow interaction between number of bedrooms and number of bathrooms in their effects on selling price.
- a) Report the prediction equation.
- b) Interpret the fit by showing the equation relating \hat{y} and number of bedrooms for homes with (i) two bathrooms, (ii) three bathrooms.
- c) Use a test to analyze the significance of the interaction term. Interpret.
12. The table shows results of regressing Y = birth rate (BIRTHS, number of births per 1000 population) on x_1 = women’s economic activity (ECON) and x_2 = literacy rate (LITERACY), using UN data for 23 nations.
- a) Report the value of each of the following:
- | | | |
|---|-----------------|------------------------|
| (i) r_{YX_1} | (ii) r_{YX_2} | (iii) R^2 |
| (iv) TSS | (v) SSE | (vi) mean square error |
| (vii) s | (viii) s_Y | (ix) se for b_1 |
| (x) t for $H_0: \beta_1 = 0$ | | |
| (xi) P for $H_0: \beta_1 = 0$ against $H_a: \beta_1 \neq 0$ | | |
| (xii) P for $H_0: \beta_1 = 0$ against $H_a: \beta_1 < 0$ | | |
| (xiii) F for $H_0: \beta_1 = \beta_2 = 0$ | | |
| (xiv) P for $H_0: \beta_1 = \beta_2 = 0$ | | |
- b) Report the prediction equation, and carefully interpret the three estimated regression coefficients.
- c) Interpret the correlations r_{YX_1} and r_{YX_2} .
- d) Report R^2 , and interpret its value.
- e) Report the multiple correlation, and interpret.
- f) Though inference may not be relevant for these data, report the F statistic for $H_0: \beta_1 = \beta_2 = 0$, report its P -value, and interpret.
- g) Show how to construct the t statistic for $H_0: \beta_1 = 0$, report its df and P -value for $H_a: \beta_1 \neq 0$, and interpret.
13. Refer to the previous exercise.

Table 3:

	Mean	Std Deviation	N			
BIRTHS	22.117	10.469	23			
ECON	47.826	19.872	23			
LITERACY	77.696	17.665	23			
Correlations						
		BIRTHS	ECON	LITER		
Correlation	BIRTHS	1.00000	-0.61181	-0.81872		
	ECON	-0.61181	1.00000	0.42056		
	LITERACY	-0.81872	0.42056	1.00000		
Sig. (2-tailed)	BIRTHS	.	0.0019	0.0001		
	ECON	0.0019	.	0.0457		
	LITERACY	0.0001	0.0457	.		
ANOVA						
	Sum of Squares	DF	Mean Square	F	Sig	
Regression	1825.969	2	912.985	31.191	0.0001	
Residual	585.424	20	29.271			
Total	2411.393	22				
Root MSE (Std. Error of the Estimate)				5.410	R Square	0.7572
Coefficients						
	Unstandardized Coeff.	Standardized				
	B	Std. Error	Coeff. (Beta)	t	Sig	
(Constant)	61.713	5.2453		11.765	0.0001	
ECON	-0.171	0.0640	-0.325	-2.676	0.0145	
LITERACY	-0.404	0.0720	-0.682	-5.616	0.0001	

- a) Find the partial correlation between Y and X_1 , controlling for X_2 . Interpret both the partial correlation and its square.
- b) Find the estimate of the conditional standard deviation, and interpret its value.
- c) Show how to find the estimated standardized regression coefficient for x_1 using the unstandardized estimate and the standard deviations, and interpret its value.
- d) Write the prediction equation using standardized variables. Interpret.
- e) Find the predicted z -score for a country that is one standard deviation above the mean on both predictors. Interpret.
14. A multiple regression model describes the relationship among a collection of cities between Y = murder rate (number of murders per 100,000 residents) and

$$\begin{aligned} X_1 &= \text{Number of police officers (per 100,000 residents)} \\ X_2 &= \text{Median length of prison sentence given to convicted murderers} \\ &\quad \text{(in years)} \\ X_3 &= \text{Median income of residents of city (in thousands of dollars)} \\ X_4 &= \text{Unemployment rate in city} \end{aligned}$$

These variables are observed for a random sample of thirty cities with population size exceeding 35,000. For these cities, the prediction equation is $\hat{y} = 30 - 0.02x_1 - 0.1x_2 - 1.2x_3 + 0.8x_4$, and $\bar{y} = 15$, $\bar{x}_1 = 100$, $\bar{x}_2 = 15$, $\bar{x}_3 = 13$, $\bar{x}_4 = 7.8$, $s_Y = 8$, $s_{X_1} = 30$, $s_{X_2} = 10$, $s_{X_3} = 2$, $s_{X_4} = 2$.

- a) Can you tell from the coefficients of the prediction equation which explanatory variable has the greatest partial effect on Y ? Explain.
- b) Find the standardized regression coefficients and interpret their values.
- c) Write the prediction equation using standardized variables. Find the predicted z -score on murder rate for a city that is one standard deviation above the mean on x_1 , x_2 , and x_3 , and one standard deviation below the mean on x_4 . Interpret.
15. *The Economist* magazine¹ developed a quality-of-life index for nations as the predicted value obtained by regressing an average of life-satisfaction scores from several surveys on gross domestic product (GDP, per capita, in dollars), life expectancy (in years), an index of political freedom (from 1 = completely free to 7 = unfree), the percentage unemployed, the divorce rate (on a scale of 1 for lowest rates to 5 for highest), latitude (to distinguish between warmer and cold climates), a political stability measure, gender equality defined as the ratio of average male and female earnings, and community life (1 if country has high rate of church attendance or trade-union membership, 0 otherwise). The table shows results of the model fit for 74 countries, for which the multiple correlation is 0.92. The study used the prediction equation to predict the quality of life in 2005 for 111 nations. The top 10 ranks were for Ireland, Switzerland, Norway, Luxembourg, Sweden, Australia, Iceland, Italy, Denmark, and Spain. Other ranks included 13 for the U.S., 14 for Canada, 15 for New Zealand, 16 for Netherlands, and 29 for the U.K.
- a) Which variables would you expect to have negative effects on quality of life? Is this supported by the results?
- b) The study states that “GDP explains more than 50% of the variation in life satisfaction.” How does this relate to a summary measure of association?
- c) The study reported that “Using so-called Beta coefficients from the regression to derive the weights of the various factors, life expectancy and GDP were the most important.” Explain what was meant by this.
- d) Although GDP seems to be an important predictor, in a bivariate sense and a partial sense, the table reports a very small coefficient, 0.00003. Why do you think this is?
- e) The study mentioned other predictors that were not included because they provided no further predictive power. For example, the study stated that education seemed to have an effect mainly through its effects on other variables in the model, such as GDP, life expectancy, and political freedom. Does this mean there is no association between education and quality of life? Explain.

¹<http://www.economist.com/media/pdf/QUALITYOFLIFE.pdf>

Table 4:

	Coefficients	Standard error	t statistic
Constant	2.796	0.789	3.54
GDP per person	0.00003	0.00001	3.52
Life expectancy	0.045	0.011	4.23
Political freedom	-0.105	0.056	-1.87
Unemployment	-0.022	0.010	-2.21
Divorce rate	-0.188	0.064	-2.93
Latitude	-1.353	0.469	-2.89
Political stability	0.152	0.052	2.92
Gender equality	0.742	0.543	1.37
Community life	0.386	0.124	3.13

16. In Exercise 1 on $Y = \text{college GPA}$, $X_1 = \text{high school GPA}$, and $X_2 = \text{college board score}$, $E(Y) = 0.20 + 0.50x_1 + 0.002x_2$. True or false: Since $\beta_1 = 0.50$ is larger than $\beta_2 = 0.002$, this implies that X_1 has the greater partial effect on Y . Explain.
17. The table shows results of fitting various regression models to data on $Y = \text{college GPA}$, $X_1 = \text{high school GPA}$, $X_2 = \text{mathematics entrance exam score}$, and $X_3 = \text{verbal entrance exam score}$. Indicate which of the following statements are false. Give a reason for your answer.

Table 5:

Estimates	Model		
	$E(Y) = \alpha + \beta x_1$	$E(Y) = \alpha + \beta_1 x_1 + \beta_2 x_2$	$E(Y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
Coefficient of x_1	0.450	0.400	0.340
Coefficient of x_2		0.003	0.002
Coefficient of x_3			0.002
R^2	0.25	0.34	0.38

- a) The correlation between Y and X_1 is positive.
- b) A one-unit increase in x_1 corresponds to a change of 0.45 in the estimated mean of Y , controlling for x_2 and x_3 .
- c) The value of SSE increases as we add additional variables to the model.
- d) It follows from the sizes of the estimates for the third model that X_1 has the strongest partial effect on Y .
- e) The value of $r_{YX_3}^2$ is 0.40.
- f) The partial correlation $r_{YX_1 \cdot X_2}$ is positive.
- g) The partial correlation $r_{YX_1 \cdot X_3}$ could be negative.
- h) Controlling for X_1 , a 100-unit increase in X_2 corresponds to a predicted increase of 0.3 in college GPA.
- i) For the first model, the estimated standardized regression coefficient equals 0.50.
18. In regression analysis, which of the following statements must be false? Why?
- a) For the model $E(Y) = \alpha + \beta_1 x_1$, Y is significantly related to x_1 at the 0.05 level, but when x_2 is added to the model, Y is not significantly related to x_1 at the 0.05 level.
- b) The estimated coefficient of x_1 is positive in the bivariate model, but negative in the multiple regression model.
- c) When the model is refitted after Y is multiplied by 10, R^2 , r_{YX_1} , $r_{YX_1 \cdot X_2}$, b_1^* , the F statistics and t statistics do not change.
- d) The F statistic for testing that all the regression coefficients equal 0 has $P < 0.05$, but none of the individual t tests have $P < 0.05$.
- e) The correlation between Y and \hat{Y} equals -0.10 .

For Problems 19–20, select the correct answer(s) and indicate why the other responses are inappropriate. (More than one response may be correct.)

19. If $\hat{Y} = 2 + 3x_1 + 5x_2 - 8x_3$, then controlling for x_2 and x_3 , the predicted mean change in Y when x_1 is increased from 10 to 20 equals
a) 3 b) 30 c) 0.3 d) Cannot be given—depends on specific values of x_2 and x_3 .
20. If $\hat{Y} = 2 + 3x_1 + 5x_2 - 8x_3$,
a) $r_{YX_3} < 0$
b) $r_{YX_3 \cdot X_1} < 0$
c) $r_{YX_3 \cdot X_1, X_2} < 0$
d) Insufficient information to answer.
e) Answers (a), (b), and (c) are all correct.