

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

- The Multiple Regression model is given by:  $E(Y) = 0.20 + 0.50x_1 + 0.002x_2$ .

  - (i) The mean college GPA is  $0.20 + 0.50(4) + 0.002(800) = 3.8$ . (ii) 2.3.
  - Fixing  $x_2$  at 500, we have the partial relationship between  $Y$  and  $x_1$  to be:  $E(Y) = 0.20 + 0.50x_1 + 0.002(500) = 1.2 + 0.50x_1$ .
  - Similarly as above, we have, on fixing  $x_2$  at 600,  $E(Y) = 0.20 + 0.50x_1 + 0.002(600) = 1.4 + 0.50x_1$ .
  - If we set  $x_1$  at different values, then the partial relationship between  $Y$  and  $x_2$  would be expressible by regression equations having the same slopes i.e 0.002, the coefficient of  $x_2$ , but each having a different  $Y$  intercept. So, we will get an array of parallel lines each having slope 0.002.
  
- The prediction equation is given by:  $\hat{y} = -3.601 + 1.2799x_1 + 0.1021x_2$ .
  - The predicted internet use for the country with the given specifications is 14.3%.
  - (i) The prediction equation when cell-phone use is 0% is  $\hat{y} = -3.601 + 1.280x_1$ . (ii) The prediction equation when cell-phone use is 100% is  $\hat{y} = 6.609 + 1.280x_1$ .  
From the above two equations we conclude that for 100 unit change in the percentage of cell-phone users (from a minimum of 0% to a maximum of 100%), the estimated percentage of people using the internet increases by 100 (0.102) = 6.609 - (-3.601) = 10.2 percent.
  - In the above two equations we see that slope of the partial relationship between  $Y$  and  $x_1$  is unaffected by the different values ascribed to  $x_2$ . So, the given model is a "no-interaction" model.
  
- $R^2 = (12959.3 - 2642.5)/12959.3 = 0.796$ . So, we conclude that using GDP and percentage of cell phone users to predict  $Y$  results in a 79.6% reduction in prediction error relative to using only  $\bar{y}$ .
  - One reason why this happens may be because  $X_1$  and  $X_2$  are highly correlated (this is plausible because a country which has a high per capita GDP would have a large number of individuals who can afford cell-phones and so the percentage of cell phone users would be high). So, when GDP is already in the model "percentage of cell phone users" would explain little extra variation in  $Y$ .
  
- The Partial Regression plots are given below:

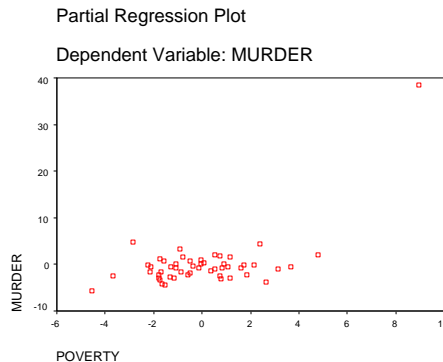


Figure 1: Partial Regression plot of Murder vs Poverty

- The first Partial Regression plot for  $Y =$  Murder Rate and  $X_1 =$  Poverty Rate shows that the partial effect of Poverty Rate is almost constant; i.e., Murder rate remains almost constant as Poverty Rate increases and there is one outlier(D.C).
- The second Partial Regression plot for  $Y =$  Murder Rate and  $X_2 =$  percent of High School graduates also shows that the partial effect of the percent of High School graduates remains more

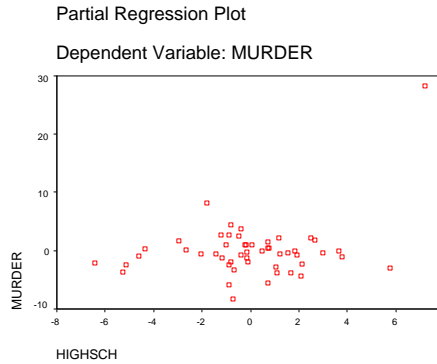


Figure 2: Partial Regression plot of Murder vs H.S graduation rate

or less at a horizontal level all through.

**b)** The prediction equation is:  $\hat{y} = -60.498 + 1.605x_1 + 0.588x_2$ . We observe the following: (i) Controlling for  $X_1$ , as percent of High School graduates increases by 1%, the predicted murder rate increases by 0.588%.

(ii) Controlling for  $X_2$ , as poverty rate increases by 1%, the predicted murder rate increases by 1.605%.

**c)** Deleting D.C we notice that there has been a huge change in the Partial Regression plots of  $Y$  on  $X_1$  and  $X_2$ . The new Partial Regression plot for  $Y =$  Murder Rate and  $X_1 =$  Poverty Rate shows that the partial effect of Poverty Rate is linear and positive indicating that as Poverty rate increases, murder rate increases too. The second Partial Regression plot for  $Y =$  Murder Rate and  $X_2 =$  percent of High School graduates also shows that the partial effect of the percent of High School graduates is also linear but now it has a negative direction indicating that as the percent with high school degree increases, the murder rate drops. Both these observations are quite realistic. The new prediction equation is  $\hat{y} = 18.9 + 0.304x_1 - 0.196x_2$ .

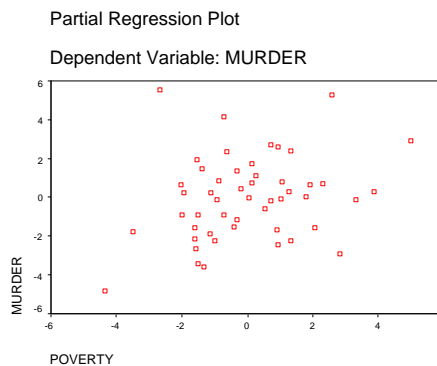


Figure 3: Partial Regression plot of Murder vs Poverty w/o D.C

5. **a)** 0.787 is the correlation between the observed  $Y$  and the predicted  $\hat{Y}$  values (which are based on the regression of the expected value of  $Y$  on GDP and Life Expectancy). This shows relatively strong association.
- b)** On adding life-expectancy, the correlation remained nearly unchanged thus indicating that in presence of GDP, life-expectancy is not necessary as a predictor of CO2. This happens because

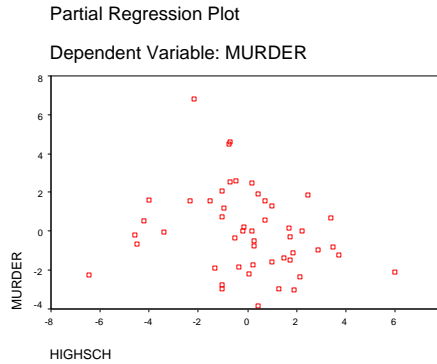


Figure 4: Partial Regression plot of Murder vs H.S graduation rate w/o D.C

life expectancy is itself highly correlated with GDP.

6. a)

Table 1:

	Sum of Squares	DF	Mean Square	F	Sig	R-Square
Regression	813.3	5	162.7	3.3	0.01	0.217
Residual	2940.0	60	49.0			Root MSE
Total	3753.3	65				7.0

Variable	Parameter Estimate	Standard Error	t	Sig
Intercept	70.0000			
x1	0.1000	0.0450	2.22	0.03
x2	-0.1500	0.0750	-2.00	0.05
x3	0.1000	0.2000	0.50	0.62
x4	-0.0400	0.0500	-0.80	0.43
x5	0.1200	0.0500	2.40	0.02

b) No,  $X_3$  or  $X_4$  or both can be dropped since the P-values corresponding to their partial tests are quite large.

c) The "F-value" refers to the test of the hypothesis  $H_0 : \beta_1 = \beta_2 = \dots = \beta_5 = 0$ . Since the P-value is very small we conclude that there is very strong evidence that at least one predictor has an effect on  $Y$ .

d) The "t-value" opposite  $x_1$  refers to the test of the hypothesis  $H_0 : \beta_1 = 0$ . Since the P-value is very small we conclude that there is very strong evidence that  $X_1$  has an effect on  $Y$ , controlling for the other predictors.

7. a) This is a confidence interval for  $\beta_1$ , which is given by  $0.10 \pm 1.96(0.0450)$  or  $(0.01, 0.19)$ . Thus, controlling for all the other explanatory variables in the model, we infer that the change in the

mean of  $Y$  for a 1-unit increase in  $X_1$  falls between 0.01 and 0.19, with 95% confidence.

b) The change in the mean of  $Y$  for a 50-unit increase in the percentage of adults owning homes is  $50(\beta_1)$ . So, the required 95% C.I. would be  $50(0.01, 0.19) = (0.5, 9.5)$ . So, controlling for all the other explanatory variables in the model, the change in the mean of  $Y$  for a 50-unit increase in  $X_1$  falls between 0.5 and 9.5, with 95% confidence.

8. a) The Partial regression plots between (price, size), (price, bedroom) and (price, bathroom) are given below:

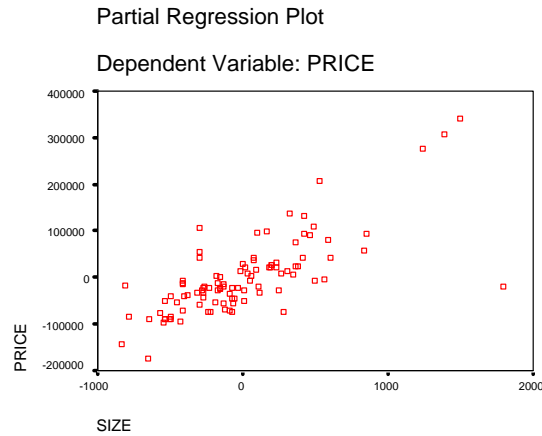


Figure 5: Partial Regression plot of Price vs size of houses

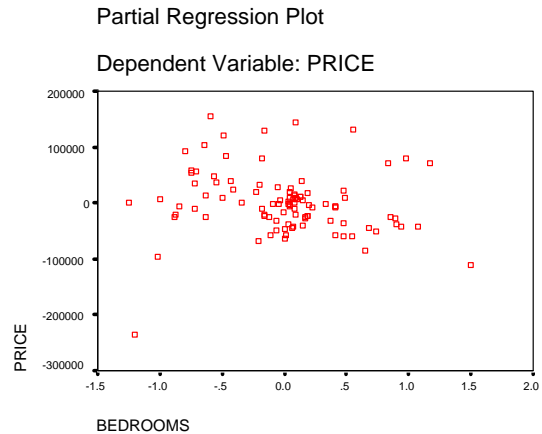


Figure 6: Partial Regression plot of Price vs number of bedrooms

We conclude that controlling for other predictors, the partial effect of size is clearly linear and positive. There is no clear partial effect for number of bedrooms or for number of bathrooms. Looking at the bivariate scatterplots predicting  $Y$  shows the effects of number of bedrooms and number of bathrooms being highly discrete, as the points occur at only a few levels on the predictor

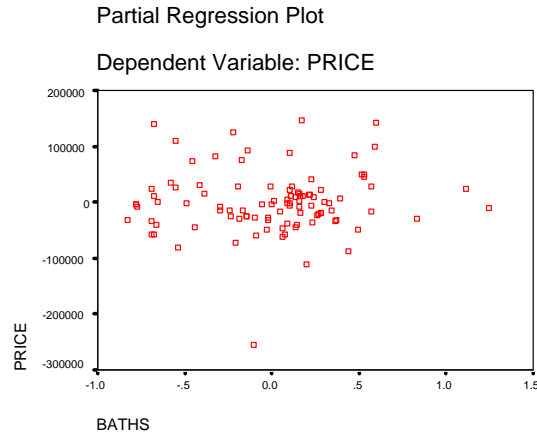


Figure 7: Partial Regression plot of Price vs number of bathrooms

scale.

**b)** The required prediction equation of  $Y$  on  $X_1$ ,  $X_2$  and  $X_3$  is given by:  $\hat{y} = -27290.1 + 130.43x_1 - 14465.8x_2 + 6890.3x_3$ .

Interpretation: (i) Controlling for all the other variables, the predicted selling price of a house increases by 130.43 dollars for a one square-foot increase in the size of the house.

(ii) Controlling for all the other variables, the predicted selling price of a house decreases by 14465.8 dollars for an increase of one in the number of bedrooms.

(iii) Controlling for all the other variables, the predicted selling price of a house increases by 6890.3 dollars for an increase of one in the number of bathrooms.

**c)** Price and size have the highest correlation (0.834) and price and bedrooms have the weakest association (0.394).

**d)** The  $R^2$  value is 0.701, implying that using  $X_1$ ,  $X_2$  and  $X_3$  to predict the price of houses produces a 70% reduction in prediction error relative to using only  $\bar{y}$ .

**e)** The F-test statistic for testing the hypothesis:  $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$  is 75.1. Its  $df$  are (3, 96) and the P-value is 0.000. So, we conclude that at least one of the three explanatory variables have an effect on  $Y$ .

**f)** The t-test statistic has value 0.51. The reported P-value (0.61) is the two sided p-value. So, the p-value corresponding to the alternative hypothesis  $H_a = \beta_3 > 0$  will be  $0.61/2 = 0.31$ . Since this value is quite large, controlling for the other two variables, there is not strong evidence that number of bathrooms has an influence in predicting the price of a house.

9. **a)** In this case the  $t$  test statistic for testing the partial effect of the number of bathrooms has value 5.04 and P-value 0. So we conclude that the number of bathrooms has a significant effect on the price of houses, controlling for number of bedrooms.

**b)** The 95% C.I is given by  $85857.22 \pm 1.985(17043.16) = (52031.3, 119683.2)$ . Thus, controlling for the number of bedrooms, the change in the mean of  $Y$  for an increase of 1 in the number of bathrooms falls between 52,031.3 dollars and 119,683.2 dollars, with 95% confidence.

**c)** The partial correlation is given by 0.455. The ordinary correlation is 0.57. There is a moderate positive association and partial association between these variables.

**d)** The estimated standardized regression coefficient corresponding to bedrooms is 0.16. This implies that controlling for bathrooms, for a 1 standard deviation increase in the number of bedrooms, the estimated change in the mean of  $Y$  in  $Y$  standard deviations is 0.16.

Similarly, the estimated standardized regression coefficient corresponding to bathrooms is 0.48. This implies that controlling for bedrooms, for a 1 standard deviation increase in the number of bathrooms, the estimated change in the mean of  $Y$  in  $Y$  standard deviations is 0.48.

e) The standardized prediction equation relating the  $z$ -score of  $Y$  to the  $z$  score of  $X_1$  and  $X_2$  is given by  $z\hat{Y} = 0.16z_{X_1} + 0.48z_{X_2}$ .

10. Let us check this using three different values of  $x_2$ . When  $x_2 = 0$ , the prediction equation becomes  $\hat{y} = 20 + 0.30x_1 + 0.05(0) + 0.005x_1(0)$  or  $\hat{y} = 20 + 0.30x_1$ . So, the slope between  $Y$  and  $x_1$  is 0.30. Similarly, when  $x_2 = 50$ , the slope between  $Y$  and  $x_1$  is 0.55 which is greater than 0.30. Lastly, when  $x_2$  is 100, the slope between  $Y$  and  $x_1$  becomes 0.80 which is even higher than 0.55. So, we conclude that yes, the slope between  $Y$  and  $x_1$  is an increasing function of the values of  $x_2$ .
11. a) The required prediction equation is given by:  
 $\hat{y} = 161497.2 - 59113.7x_1 - 31866x_2 + 38553.17x_1x_2$ . Here  $X_1$  denotes no. of Bedrooms and  $X_2$  denotes the no. of bathrooms.
- b) (i) The equation relating  $\hat{y}$  to  $X_1$  when  $X_2 = 2$  is given by:  $\hat{y} = 161497.2 - 59113.7x_1 - 31866(2) + 38553.166x_1(2)$  or  $\hat{y} = 97765.2 + 17992.6x_1$ . This means that for houses with 2 bathrooms, the predicted price increases by 17992.6 Dollars for an increase in one bedroom.  
(ii) Similarly, the equation relating  $\hat{y}$  to  $X_1$  when  $X_2 = 3$  is given by:  $\hat{y} = 161497.2 - 59113.7x_1 - 31866(3) + 38553.166x_1(3)$  or  $\hat{y} = 65899.2 + 56545.8x_1$ . This means that for houses with 3 bathrooms, the predicted price increases by 56545.8 Dollars for an increase in one bedroom. Since the slope of  $Y$  on  $X_1$  markedly changes with values of  $X_2$ , we conclude that there may be considerable interaction between the number of bedrooms and the number of bathrooms in their effects on the price of a house.
- c) The test for the significance of interaction between  $X_1$  and  $X_2$  can be formulated as  $H_0 : \beta_3 = 0$  against  $H_a : \beta_3 \neq 0$  where  $\beta_3$  is the coefficient of  $x_1x_2$  in the above prediction equation. The t-statistic and the corresponding P-value are given in the output. Since the P-value is quite low (0.012), we conclude that there is strong evidence of interaction between  $X_1$  and  $X_2$  in their effects on  $Y$ .
12. a) (i)  $r_{YX_1} = -0.612$ , (ii)  $r_{YX_2} = -0.819$  (iii)  $R^2 = 0.7572$ , (iv) TSS = 2411.393, (v) SSE = 585.424, (vi) MSE = 29.271, (vii)  $s = 5.41$ , (viii)  $s_Y = 10.469$ , (ix)  $se$  for  $b_1 = 0.0640$ , (x)  $t = -2.676$ , (xi)  $P = 0.0145$ , (xii)  $P = 0.0145/2 = 0.007$ , (xiii)  $F = 31.191$ , (xiv)  $P = 0.001$ .
- b) The required prediction equation is given by  $\hat{y} = 61.71 - 0.171x_1 - 0.404x_2$ .  
-0.171 is the change in the predicted birth rate for 1 unit increase in ECON, controlling for LITER and -0.404 is the change in the predicted birth rate for 1 unit increase in LITER controlling for ECON.
- c)  $r_{YX_1} = -0.612$ . Thus, there is a moderate negative association between birth rate and ECON.  $r_{YX_2} = -0.819$ . This implies that there is a strong negative association between birth rate and literacy.
- d)  $R^2 = (2411.4 - 585.4)/2411.4 = 0.76$ . Thus, there is a 76% reduction in error in using ECON and LITER instead of  $\bar{y}$  in predicting birth rate.
- e)  $R = \sqrt{0.76} = 0.87$ . Thus, the correlation between the observed  $Y$  values and the predicted  $\hat{y}$  values is 0.87.
- f) The required F value is given by 31.2 (from the output) with  $df_1 = 2$  and  $df_2 = 20$  and P-value = 0.0001. Since the P-value is very small, we conclude that at least one of ECON and LITER has a significant influence on  $Y$ .
- g) The required t-statistic is given by  $(-0.171/0.064) = -2.676$  with  $df = 20$  and P-value = 0.0145. This means that there is a strong evidence of a negative relationship between birth rate and ECON, controlling for LITER.
13. a)  $r_{YX_1.X_2} = [(-0.612) - (-0.819)(0.421)]/\sqrt{[1 - (-0.819)^2][1 - (0.421)^2]} = -0.51$ . Thus, controlling for  $X_2$ , there is a moderate tendency for  $Y$  to decrease as  $X_1$  increases.  
 $r_{YX_1.X_2}^2 = 0.26$ . This is the proportion of variation in  $Y$  explained by  $X_1$ , out of that part which is unexplained by  $X_2$ .

- b) The estimate of the conditional standard deviation is given by  $\hat{\sigma} = \sqrt{585.4/20} = \sqrt{29.27} = 5.4$ . This is the estimated standard deviation of  $Y$  values at fixed values for  $X_1$  and  $X_2$ .
- c) The estimated standardized regression coefficient for  $x_1$  is given by  $b_1^* = -0.171(19.87/10.47) = -0.325$ . This implies that the birth rate is predicted to decrease 0.325 standard deviations for each standard deviation increase in economic activity, controlling for literacy.
- d) The required prediction equation is given by  $\hat{z}_Y = -0.325z_{X_1} - 0.682z_{X_2}$ . This means that birth rate is predicted to decrease by 0.325 standard deviations for each standard deviation increase in economic activity, controlling for literacy and is predicted to decrease 0.682 standard deviations for each standard deviation increase in literacy, controlling for economic activity.
- e) The predicted z score is given by  $-0.325(1) - 0.682(1) = -1.0$ . Thus the country is predicted to be one standard deviation below the mean in birth rate.
14. a) No, cannot determine the relative partial influence of any variable because the units of measurement of each variable is different.
- b) (i)  $b_1^* = -0.02(30/8) = -0.075$ . This implies that murder rate is predicted to decrease 0.075 standard deviations for each standard deviation increase in the numbers of police officers, controlling for other predictors.
- (ii)  $b_2^* = -0.1(10/8) = -0.125$ . This implies that murder rate is predicted to decrease 0.125 standard deviations for each standard deviation increase in  $X_2$ , controlling for other predictors.
- (iii)  $b_3^* = -1.2(2/8) = -0.3$ . This implies that murder rate is predicted to decrease 0.3 standard deviations for each standard deviation increase in  $X_3$ , controlling for other predictors.
- (iv)  $b_4^* = 0.8(2/8) = 0.2$ . This implies that murder rate is predicted to increase 0.2 standard deviations for each standard deviation increase in  $X_4$ , controlling for other predictors.
- c) The required prediction equation is given by  $\hat{z}_Y = -0.075z_{X_1} - 0.125z_{X_2} - 0.3z_{X_3} + 0.2z_{X_4}$ . Accordingly, the predicted z score is given by  $\hat{z}_Y = -0.075(1) - 0.125(1) - 0.3(1) + 0.2(-1) = -0.7$ . Thus, the predicted murder rate for that city would be 0.7 standard deviations below the mean.
15. a) Results suggest that Political freedom, percentage unemployed, divorce rate, and climate have negative effect.
- b) 50% signifies the square of the correlation coefficient between  $Y$  and  $X_1$  (GDP) is 0.50.
- c) The estimated standardized regression coefficients are highest for these predictors.
- d) GDP is measured in dollars per capita, so a one-unit change is trivial. The effect would be 1000 times as large if it were measured in thousands of dollars per capita.
- e) No, this does not mean that education has no effect on  $Y$ . It may have a considerable effect but as it is highly correlated with GDP, life-expectancy and Political freedom, including it in the prediction equation does not lead to a significant increase in the value of  $R^2$ .
16. This statement is false since the two variables have different units. So, we cannot compare the values of the partial slope.
17. b) The partial change is 0.34 (see the third model) not 0.45 which is the overall change ignoring, rather than controlling for  $x_2$  and  $x_3$ .
- c) The worst that can happen to the SSE on adding additional variables is that it can remain constant. But SSE can never increase as we go from simpler to more complex models.
- d) We cannot say this because in the third model, the variables have different units.
- e) False, since this  $r_{YX_3}^2$  cannot exceed 0.38, the  $R^2$  value for the model with  $X_3$  and other variables.
18. e) False, since this, being the multiple correlation, cannot be negative.
19. (b) since  $(20-10)3 = 30$ .
20. (c) since the coefficient of  $X_3$  is -8 and since the partial slope has the same sign as the partial correlation. We need the sign of the slope for the bivariate model to tell the sign of the correlation

in (a) and the sign of the partial slope of  $X_3$  for the model with it and  $X_1$  alone as predictors to tell the sign of the partial correlation in (b).