

## PROBLEM SET - 4 (MULTIPLE LINEAR REGRESSION - 1)

ECO 204 (Section 6)  
Instructor: Shaikh Tanvir Hossain

Due: 15th Dec (before 10.00 PM), submit in Google Classroom

§

To help you some of the problems have been solved (✓ means this problem has been solved in a separate Rmarkdown file). Please try to solve all unsolved problems (this means solve the problems where you don't have ✓). You must solve all problems using R so that you have a good practice. This is an **individual assignment**. Submit two files

- 1) A RMarkdown file with code and explanations and
- 2) A HTML file or a PDF file that you could generate using Rmarkdown file.

### 1 §. Basic MLR

1. ✓ (Writing your own function in R) Suppose we have following estimated regression equation based on 10 observations

$$\hat{y}_i = 25 + 10x_{1i} + 8x_{2i} + 9x_{3i}$$

- (a) Interpret  $\hat{\beta}_1$ ,  $\hat{\beta}_2$  and  $\hat{\beta}_3$  in this estimated regression equation.
  - (b) Write a user defined R function that will return the predicted value of  $\hat{y}$  given  $x_1, x_2$  and  $x_3$ .
  - (c) Now use your function to predict  $Y$  resulting from a  $X_1 = 15, X_2 = 10, X_3 = 5$ .
2. (Problem 4 of Ch 15.2 from Anderson et al. (2020)) A shoe store developed the following estimated regression equation relating sales to inventory investment and advertising expenditures.

$$\hat{y}_i = 25 + 10x_{1i} + 8x_{2i}$$

where

$$\begin{aligned}x_1 &= \text{inventory investment (\$1000 s)} \\x_2 &= \text{advertising expenditures (\$1000 s)} \\y &= \text{sales (\$1000 s)}\end{aligned}$$

- (a) Predict the sales resulting from a \$15,000 investment in inventory and an advertising budget of \$10,000.
  - (b) Interpret  $\hat{\beta}_1$  and  $\hat{\beta}_2$  in this estimated regression equation.
3. ✓ (Problem 19 of Ch 15.5 from Anderson et al. (2020)) Suppose following estimated regression equation based on 10 observations was presented.

$$\hat{y}_i = 29.1270 + .5906x_{1i} + .4980x_{2i}$$

Here SST = 6724.125, SSR = 6216.375, SE( $\hat{\beta}_1$ ) = .0813, and SE( $\hat{\beta}_2$ ) = .0567.

- (a) Compute MSR and MSE.
- (b) Compute  $F$  and perform the appropriate  $F$  test. Use  $\alpha = .05$ .
- (c) Perform a  $t$  test for the significance of  $\beta_1$ . Use  $\alpha = .05$ .
- (d) Perform a  $t$  test for the significance of  $\beta_2$ . Use  $\alpha = .05$ .

4. **(Problem 20 of Ch 15.5 from Anderson et al. (2020))** Again suppose we have following estimated regression equation based on 10 observations.

$$\hat{y}_i = -18.37 + 2.01x_{1i} + 4.74x_{2i}$$



Here  $SST = 15,182.9$ ,  $SSR = 14,052.2$ ,  $SE(\hat{\beta}_1) = .2471$ , and  $SE(\hat{\beta}_2) = .9484$ .

- Test for a significant relationship among  $x_1$ ,  $x_2$ , and  $y$ . Use  $\alpha = .05$ .
  - Is  $\beta_1$  significant? Use  $\alpha = .05$ .
  - Is  $\beta_2$  significant? Use  $\alpha = .05$ .
5. **(Problem 21 of Ch 15.5 from Anderson et al. (2020))** The following estimated regression equation was developed for a model involving two independent variables.

$$\hat{y}_i = 40.7 + 8.63x_{1i} + 2.71x_{2i}$$

After  $x_2$  was dropped from the model, the least squares method was used to obtain an estimated regression equation involving only  $x_1$  as an independent variable.

$$\hat{y}_i = 42.0 + 9.01x_{1i}$$

- Give an interpretation of the coefficient of  $x_1$  in both models.
  - why do you think the coefficient of  $x_1$  differs in the two models?
6. **✓ (slightly adapter from problem 5 of Ch 15 from Anderson et al. (2020))** The owner of Showtime Movie Theaters, Inc., would like to predict weekly gross revenue as a function of advertising expenditures. Historical data for a sample of eight weeks is given in  Showtime.xlsx. The variables are
- revenue: gross revenue (\$1000)
  - tv: television advertising (\$1000)
  - newspaper: newspaper advertising (\$1000)
  - magazines: magazine advertising (\$1000)
  - leaflets: leaflets advertising (\$1000)
- Suppose we want to predict weekly gross revenue as a function of television advertising expenditures. Develop an estimated regression equation with television advertising as the independent variables.
  - Now develop an estimated regression equation to predict weekly gross revenue with all other variables as the independent variables.
  - Is the estimated regression equation coefficient for television advertising expenditures the same in part (a) and in part (b)? Interpret the coefficient in each case.
  - Print the anova table using , what is SST, SSR and SSE, and MSR and MSE?
  - What is  $R^2$  and Adjusted  $R^2$  in the multiple linear regression model. Did  $R^2$  increase when you add one more variable, what about Adjusted  $R^2$ .
  - Now assume the true model is the following

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \epsilon$$

where

- $Y$  is gross revenue (\$1000),
- $X_1$  is television advertising (\$1000) and
- $X_2$  newspaper advertising (\$1000).
- $X_3$  magazine advertising (\$1000).

- $X_4$  leaflets advertising (\$1000).

Now based on the multiple linear regression that you did in (b) do following tests

- With  $\alpha = .05$  test individual significance, and comment on whether should we drop any of the variable  $X_1$ , or  $X_2$ , or  $X_3$ , or  $X_4$ . This means you need to do four different tests.

$$H_0 : \beta_j = 0 \text{ for } j = 1, 2, 3, 4$$

- With  $\alpha = .05$  test the hypotheses, and comment on whether should we drop all the variables.

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

$$H_a : \text{at least one of } \beta_j \text{ for } j = 1, 2, 3, 4 \text{ is not zero}$$

- With  $\alpha = .05$  test the hypotheses, and comment on whether should we drop both magazine and leaflets advertising.

$$H_0 : \beta_3 = \beta_4 = 0$$

$$H_a : \text{at least one of } \beta_j \text{ for } j = 1, 2, 3, 4 \text{ is not zero}$$

- What is the gross revenue expected / predicted for a week when \$3500 is spent on television advertising and \$2300 is spent on newspaper advertising, and \$1000 is spent on magazine advertising and \$500 is spent on leaflets advertising?
- Provide a 95% confidence interval for the mean revenue of all weeks where \$3500 is spent on television advertising and \$2300 is spent on newspaper advertising.
- Provide a 95% prediction interval for next week's revenue, assuming that the advertising expenditures will be \$3500 on television, and \$2300 on newspaper
- Plot the residuals against the fitted values. Is there any pattern in the residuals? What does this suggest?

7. ✓ **(Simulation Example):** Again we will do a simple simulated example. Suppose we have following model

$$Y = 3 + 5X_1 + 2X_2 + \epsilon$$

where

$$X_1 \sim \text{Unif}(0, 1)$$

$$X_2 \sim \mathcal{N}(0, .35)$$

$$\epsilon \sim \mathcal{N}(0, .25)$$

- set.seed(your id) so that we can reproduce the results.
- Generate  $n = 10, 100, 300, 500$  samples from the above model, and for each sample estimate the model and report the estimated coefficients  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ .
- Now for sample size  $n = 50$  generate 1000 samples and for each sample estimate the model store the estimated coefficients  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$  in three different vectors. Plot the histogram to see the distribution of the estimated coefficients. You should have three histograms. These histograms will give you an idea about the sampling distribution of the estimated coefficients (Note: You need to use *for loop* or *while loop* in this case)
- Now do the same task for sample size  $n = 100$ . You should have three histograms. Compare the histograms with the previous one. What do you observe?

## 2 §. Extension of MLR

8. ✓ **(Extension of MLR - Interaction Effects):** Suppose in the advertisement data in problem 6, we run following regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

where

- $Y$  is gross revenue (\$1000),
- $X_1$  is television advertising (\$1000) and
- $X_2$  is newspaper advertising (\$1000).

This is called an *interaction effect* model, where we are allowing the effect of  $X_1$  to depend on the value of  $X_2$  and vice versa. First note, why we are doing this? It could be that there is a synergy between television and newspaper advertising, and the effect of television advertising depends on the level of newspaper advertising. For example, if there is no newspaper advertising, then television advertising may not be very effective. On the other hand, if there is a lot of newspaper advertising, then television advertising may be more effective. You can think the people who read newspaper are more likely to watch television.

- (a) Is this a linear model in variables?
  - (b) Fit this model using **R**, you should use the syntax `lm(revenue ~ tv*newspaper)`, and check the significance (the interpretation is difficult here, we need to think about for a fixed value of one variable)
9. ✓ **(Extension of MLR - Adding Nonlinear Terms):** Here we will use the auto data in PS - 3.
- (a) As you pointed out the auto data has some problem, clean the data (you can skip this question for exam!), you will have the clean data in your exam, so no worries.
  - (b) We have already run the linear model in PS - 2. Recall the linear regression model was,

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

where

- $Y$  is mpg
- $X_1$  is horsepower

Again simply estimate the parameter of this model (this is the problem as PS - 2)

- (c) Now assume the model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \epsilon$$

where

- $Y$  is mpg
- $X_1$  is horsepower
- $X_1^2$  is horsepower squared

This is a *nonlinear model*, in particular *quadratic model*, estimate the parameters of the this model using **R**. You can do this by using the syntax `lm(mpg ~ horsepower + I(horsepower^2))`.

(d) This time assume the model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3 + \epsilon$$

This is again estimate the parameters of the this model using 

- (e) Plot the scatter plot and all fitted line. For plotting after calling the `plot()`, you need to use the `lines` function.
  - (f) Check  $R^2$  and adjusted  $R^2$ , and comment which model do you think fits better with the data?
  - (g) Can you test using `anova()` function which model is overall significant?
10. **(Extension of MLR - Adding Nonlinear Terms):** Do the same task as last one, but this time predict the `mpg` using `weight`. Use the same three models as last one and compare in a similar manner.
11. **(Extension of MLR - Categorical Predictors):** Again load the `Auto.xlsx`, but this time load the clean data `Auto_clean.xlsx`. We will use the `origin` variable as a categorical variable. The `origin` variable has three categories, 1 = American, 2 = European, 3 = Japanese. We will use this variable to predict `mpg`.

## References:

Anderson, D. R., Sweeney, D. J., Williams, T. A., Camm, J. D., Cochran, J. J., Fry, M. J. and Ohlmann, J. W. (2020), *Statistics for Business & Economics*, 14th edn, Cengage, Boston, MA.