

HOMEWORK - SLR / MLR 3

ECO 204 (Section 1 and 9)
Instructor: Shaikh Tanvir Hossain

Due: 18th / 19th May in class

§

To help you I have solved some of the problems, please see the RMarkdown or HTML file (✓ means this problem has been solved). Please try to solve all unsolved problems. You should be able to solve R or Excel. This is an **individual assignment**. Submit two files

- 1) For R users, submit RScript
- 2) For Excel users, submit the Excel File

1 §. Basic MLR

1. ✓ (Problem 19 of Anderson) Suppose following estimated regression equation based on 10 observations was presented.

$$\hat{y}_i = 29.1270 + 0.5906 x_{1i} + 0.4980 x_{2i}$$

and the true model is,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

- $SST = 6724.125$,
- $SSR = 6216.375$,
- $\widehat{SE}(\hat{\beta}_1) = 0.0813$,
- $\widehat{SE}(\hat{\beta}_2) = 0.0567$

- (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 β_1 . Use $\alpha = .05$.
- (d) Perform a t test for the significance of $\beta_1 = 0.6$. Use $\alpha = .05$.
- (e) Perform a t test for the significance of β_2 . Use $\alpha = .05$.
2. (Problem 20 of Ch 15.5 from Anderson) Again suppose we have following estimated regression equation based on 10 observations.



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

and the true model is,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

- $SST = 15,182.9$,
- $SSR = 14,052.2$,
- $\widehat{SE}(\hat{\beta}_1) = 0.2471$,
- $\widehat{SE}(\hat{\beta}_2) = 0.9484$

- (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 β_1 . Use $\alpha = .05$.

- (d) Perform a t test for the significance of $\beta_1 = 0.6$. Use $\alpha = .05$.
- (e) Perform a t test for the significance of β_2 . Use $\alpha = .05$.
3. ✓ (slightly adapter from problem 5 of Ch 15 from Anderson) 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)
- (a) 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.
- (b) Now develop an estimated regression equation to predict weekly gross revenue with all other variables as the independent variables.
- (c) 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.
- (d) Print the anova table using , what is SST, SSR and SSE, and MSR and MSE?
- (e) 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 .
- (f) Now assume the true model is the following

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

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

- i. 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.

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

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

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

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

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

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

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

- (g) 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?
- (h) 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.
- (i) Provide a 95% prediction interval for next week's revenue, assuming that the advertising expenditures will be \$3500 on television, and \$2300 on newspaper
- (j) Plot the residuals against the fitted values. Is there any pattern in the residuals? What does this suggest?

2 §. Extension of MLR

4. ✓ (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)
5. ✓ (Extension of MLR - Adding Nonlinear Terms): Here we will use the auto data `Auto_clean.xlsx`
- (a) The model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i$$

where

- Y is mpg
- X_1 is horsepower

Simply estimate the parameter of this model

- (b) 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))`.

(c) This time assume the model is

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

again estimate the parameters of the this model using **R**

- (d) Plot the scatter plot and all fitted line. For plotting after calling the `plot()`, you need to use the `lines` function.
 - (e) Check R^2 and adjusted R^2 , and comment which model do you think fits better with the data?
 - (f) Can you test using `anova()` function which model is overall significant?
6. **(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.
7. **(Extension of MLR - Categorical Predictors):** Again load `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.
8. **✓ (Simulation Example - Bonus Not for Exam):** We will do a simple simulated example. Suppose we have following model

$$Y_i = 3 + 5X_{1i} + 2X_{2i} + \epsilon$$

where

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

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

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

- (a) `set.seed(your id)` so that we can reproduce the results.
- (b) Generate $n = 10, 100, 300, 500$ samples from the above model, and for each sample estimate the model and report the estimated coefficients $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$.
- (c) 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)
- (d) 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?

References: