BU510.650 – Data Analytics

Assignment # 5

Please submit two documents: Your answers to each part of every question in .pdf or .doc format, and your R script, in .R format. In your document with answers, please do **not** respond with R output only. While it is okay to include R output in that document, please make sure you spell out the response to the question asked and also include all the plots. Please submit your assignment through Blackboard and name your files using the convention LastName_FirstName_AssignmentNumber. For example, Yazdi_Mohammad_5.pdf and Yazdi_Mohammad_5.R.

Conceptual Part

For this conceptual section, there is no requirement to use any R coding.

Question 1: How does a one-unit increase in X_1 affect the average value of $\hat{Y}$?

$$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X_1 + \widehat{\beta_2} {X_1}^2$$

Programming Part

1. This question involves the use of simple linear regression on the Bikeshare data set (adapted from a data set of bike rentals from DC's Capital Bikeshare system – see the following url for details: https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset). The following is a brief description of the data, which is in the file `Bikeshare.csv` on Blackboard.

   - Temperature – normalized temperature in Celsius, derived according to: (temperature on that day - t_min)/(t_max - t_min), where t_min = -8, t_max = +39 (minimum and maximum temperatures encountered during the time period the data was collected).
   - Humidity – normalized humidity, derived according to: Humidity (measured on a scale of 0 to 100) on that day / 100.
   - Windspeed – normalized windspeed in km/h, derived according to: Windspeed on that day / wind_max, where wind_max = 67, the fastest wind encountered during the time period the data was collected.
   - Rentals – number of bikes rented on that day.

Hint: Keep the dataset in the normalized values and do NOT change the normalized to original values.

a) First, read the data in `Bikeshare.csv` to a data frame called `Bikeshare`. Use the `lm()` function to run a simple linear regression with "Rentals" as the output variable and "Temperature" as the input variable. Use the `summary()` function to print the results.

   - Comment on the output. Specifically: Does temperature have a statistically significant effect on the number of rentals?

- What is the effect of a one degree (Celsius) change in temperature on the rentals? Hint: The answer to this question is the same as the answer to the following question: what is the effect of a 1/47 degree Celsius change in normalized temperature on the rental?

b) Repeat part (a), but this time with "Humidity" as the input variable.

c) Repeat part (a), but this time with "Windspeed" as the input variable.

d) Check the R^2 value you obtained in part (c). You will notice that it is very small. How do you reconcile the small R^2 value with your answer for part (c)?

e) Plot "Rentals" versus "Temperature", and display the "regression line" on the plot, that is, the line that shows how "Rentals" changes with respect to "Temperature" according to your regression. The following command will produce such a line: `abline(..., lwd = 5, col = "red")`. Here, "…" should be replaced with the name of the variable where you stored your regression results, "lwd = 5" specifies the width of the line, and "col = "red"" makes it a red line.

f) The goal of this part is to introduce you to a useful plot type, called "scatter plot matrix". Obtain a scatter plot matrix of all variables (except the variable "Day") using the following command:
`pairs(~ Rentals + Temperature + Humidity + Windspeed, data=Bikeshare)`

Study the graph you obtained. Which input variables appear to have an effect on "Rentals"?

g) Run multiple linear regression using all variables, except "Day", as input variables. Provide the summary information. Which input variables have a statistically significant effect on "Rentals"? Justify your answer.

h) What is the predicted number of rentals on a day when the temperature is 15 degrees Celsius, humidity is 50 (out of 100), and the windspeed is 5 km/h?

2. In this question, you will work on the updated Bikeshare dataset. In particular, you will check whether weekends, in addition to weather conditions, affect rental patterns. In addition to all the previous data, the updated Bikeshare dataset has the following data:

- Weekday – goes from 0 to 6, with 0 indicating that the day was Sunday, 1 indicating that the day was Monday, etc.
- Registered – number of bikes rented by registered users on that day.
- Casual – number of bikes rented by casual users on that day.

To start your work on this question, read the data in `Bikeshare_updated.csv` to a data frame called `BikeshareUpdated`. Then, create a new column in your data frame called "Weekend," which shows 1 if the day is a Saturday or Sunday, and 0 otherwise. (R Hint: In R, the "or" operator is the symbol `|`. For example, `(x == 5) | (x == 6)` will return `TRUE` if x is 5 or 6.)

(a) Run a multiple linear regression with "Rentals" as the output variable and "Temperature," "Humidity," "Windspeed," and "Weekend" as input variables. Comment on the output: Which input variables have a statistically significant effect on the number of rentals?

(b) Run a multiple linear regression with “Registered” as the output variable and “Temperature,” “Humidity,” “Windspeed,” and “Weekend” as input variables. Comment on the output: Which input variables have a statistically significant effect on the number of rentals by registered users?

(c) Run a multiple linear regression with “Casual” as the output variable and “Temperature,” “Humidity,” “Windspeed,” and “Weekend” as input variables. Comment on the output: Which input variables have a statistically significant effect on the number of rentals by casual users?

(d) Compare and contrast your results from the previous three parts to answer the following question: How does the weekend affect rental patterns?