MSCA31010: Linear & Non-Linear Models

Winter 2022 Assignment 1

A medical doctor has a habit to measure his weight in the morning. The WeightDiary.xlsx file contains his weights between March 4, 2017 and December 31, 2019. The Excel sheet contains two fields.

1. Date. The dates when the body weights are measured.
2. Weight. The body weight in number of pounds.

Please beware that, due to various reasons, the medical doctor forgot to measure his weight on some random dates. The medical doctor is curious to know whether his weights are influenced by the months and the days of a week of the measurements.

You will need to calculate these two categorical variables.

1. Month. This is the calendar month of the measurement date. The categories are, in this order, January, February, March, April, May, June, July, August, September, October, November, and December.
2. DayOfWeek. This is the day of a week of the measurement date. The categories are, in this order, Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, and Saturday.

# Question 1 (50 points)

You will use this linear model Weight ~ Intercept + Month + DayOfWeek to study the effect of Month and DayOfWeek on Weight. You are required to implement the SWEEP Operator in Python.

1. (5 points). Provide a frequency table for the Month, and another frequency table for the DayOfWeek.
2. (5 points). What is the Residual Sum of Squares for this model Weight ~ Intercept? Give your answer using the “.7E” scientific notation.
3. (5 points). What is the Residual Sum of Squares for this model Weight ~ Intercept + Month? Give your answer using the “.7E” scientific notation.
4. (5 points). What is the Residual Sum of Squares for this model Weight ~ Intercept + DayOfWeek? Give your answer using the “.7E” scientific notation.
5. (5 points). What is the generalized inverse that the SWEEP Operator gives for this model Weight ~ Intercept + DayOfWeek? Give your answer using the “.7E” scientific notation.
6. (5 points). What is the Residual Sum of Squares for this model Weight ~ Intercept + Month + DayOfWeek? Give your answer using the “.7E” scientific notation.
7. (5 points). Which model yields the smallest Residual Sum of Squares?
8. (5 points). How many regression parameters (including the aliased parameters) are in this model Weight ~ Intercept + Month + DayOfWeek?
9. (10 points). What are the regression coefficients (including the aliased parameters) of this model Weight ~ Intercept + Month + DayOfWeek? Give your answer using the “.7E” scientific notation.

# Question 2 (50 points)

Let us focus on this model Weight ~ Intercept + Month + DayOfWeek.

1. (10 points). Generate a Boxplot of the residuals versus Month. The residuals are on the vertical axis and the Month categories are on the horizontal axis. Also, generate another Boxplot of the residuals versus DayOfWeek. Comment on the evidence of heteroskedasticity of the residuals.
2. (10 points). Calculate the Anderson-Darling Test statistic and generate a Normality Q-Q Plot for the residuals. Comment on the evidence of normality (or non-normality) of the residuals.
3. (10 points). Perform the Breusch-Pagan Test and the White Test of Heteroskedasticity. Provide the Chi-square statistics, the degrees of freedom, and the significance values. Comment on the evidence of non-homogenous variance.
4. (10 points). Calculate the Durbin-Watson Test statistic. Comment on the evidence of autocorrelation among observations.
5. (10 points). Calculate the Shapley values of the two predictors Month and DayOfWeek. Also, provide the Percent Shapley values of the two predictors. Among these two predictors, which one influence the weight more?