**ASSIGNMENT 3**

**Dataset :** **http://www.nyc.gov/html/tlc/html/about/trip\_record\_data.shtml**

**Is the relationship significant?**

# As the p-value is much less than 0.05, we reject the null hypothesis that β = 0.

# Hence there is a significant relationship between the variables in the linear

# regression model of the dataset.

# Hence,the model has a significant relationship.

**Are any model assumptions violated?**

1. Linear relationship :

# Plotting a Scatter plot

# Assumption 1: To check if their is a linear relationship in the model.

# Linear regression needs the relationship between the independent and dependent variables to be linear.

# The linearity assumption can best be tested with scatter plots,

# the following examples depict all the cases, where no and little linearity is present.

1. Multivariate Normality:

Plotting a Q-Q plot for the model

# Assumption 2: To check for multivariate normality in the model

# This assumption can best be checked with a Q-Q-Plot.

# Normality can be checked with a goodness of fit test

# Below graphs shows that the variables are multivariate normal

1. No Multicollinearity:

Multicollinearity may be tested with three central criteria:

1. Correlation Matrix
2. Tolerence
3. VIF (Variance Inflation Factor)

# 1) Correlation matrix : All the values are <=1

# Hence, the model is near to accuracy

# 2)Test for tolerance:

# T = 1 – R²

# = 1 – 1

# = 0

# With T < 0.1 there might be multicollinearity in the data. Hence, there is multicolinearity in the model.

# Checking for Multicolinearity

# 3) Variance Inflation Factor (VIF) – the variance inflation factor of the linear regression

# is defined as VIF = 1/T. With VIF > 10 there is an indication that multicollinearity may be present;

# with VIF > 100 there is certainly multicollinearity among the variables.

# In the above model, Multicollinearity is present. Hence to remove the multicollinearity we perform the setps to remove multicolinearity.

# Calculating VIF

# The value should be near to 4

# This suggests that the multicollinearity has been removed

1 / (1-r\_square)

1. Homoscedasticity

Scatter plots and Goldfeld-Quandt test is done for homoscedasticity

Durbin-Watson’s d tests the null hypothesis while performing auto-correlation

# As a rule of thumb values of 1.5 < d < 2.5 show that there is no auto-correlation in the data