Linear Regression Models
Statistics GR5205/GU4205 — Fall 2020

Homework 6: Data Analysis Project

1 Background and Dataset

Your instructor used to travel frequently attending different conferences. In order to arrive at the conference meeting on time, she always needs to take the flight delays into consideration.
You are going to help your instructor with the prediction using dataset `pnwflights14` from:

`https://raw.githubusercontent.com/ismayc/pnwflights14/master/data/flights.csv`.

In class we have explored that the dataset contains 16 columns, including

`year, month, day`
`dep_time, dep_delay, arr_time, arr_delay, air_time`
`carrier, tailnum, flight, origin, dest, distance, hour, minute`.

2 Data Analysis Report

For this assignment, perform the following four steps. Write a section with one section per step.

1. Familiarize yourself with the data, including

   (a) What does each column stand for? Which cleaning/pre-processing steps you are going to apply.

   (b) Which column should be chosen as potential predictors and which should be chosen as response variables so that you will have a meaningful model? (Hint: what are the relationship between the variables in the set {`dep_time, dep_delay, arr_time, arr_delay, air_time`}? Are there any redundant variables that would not contribute to your analysis hence should be deleted?)

2. Formulate at least two potential regression models to help your instructor to predict the travel time of future trips.

3. Apply the regression analysis to get parameter estimation for your proposed models and explain the physical meaning of your models.

4. Which model is the best among your analysis, and give the reason of your choice.

Useful References for Step 1.

1. https://www.datacamp.com/community/tutorials/categorical-data
2. https://github.com/ismayc/pnwflights14
3. https://cran.r-project.org/web/packages/nycflights13/nycflights13.pdf