Modelling linear relationship between Ozone Concentration and Meteorology, LA Basin, 1976

March 26, 2021

Apoorva Singh	apoorvac@iitk.ac.in	17816140
Arkajyoti Bhattacharjee	arkajyoti20@iitk.ac.in	201277
Jetti Ramya Chowdary	ramyaj@iitk.ac.in	17817327
Saurab Jain	jnsaurab@iitk.ac.in	170642
Vishweshwar Tyagi	tyagi@iitk.ac.in	191173

Introduction

It is often the case that nowadays data analysts while implementing linear regression via inbuilt libraries often overlook model inadequacies and violation of basic assumptions upon which the whole theory of linear regression rests. In this project, we'll try to study these inadequacies as well as demonstrate how rectifying them can massively improve the overall performance of our linear model.

Dataset

We'll make use of the **ozone: Ozone in LA in 1976**[1] dataset in order to study the **linear** relationship between atmospheric ozone concentration and meteorology in the Los Angeles Basin in 1976.

This dataset consists of 330 observations on the following 10 variables:

O3
: Ozone conc., ppm, at Sandbug AFB.

vh
: Vandenburg 500 millibar height (m)

wind
: wind speed (mph)

humidity
: a numeric vector

temp
: temperature (C°)

ibh
: inversion base height (ft.)

dpg
: Daggett pressure gradient (mmhg)

ibt
: inversion base temperature (F°)

vis
: visibility (miles)

doy
: day of the year

Here, **O3** is the response variable and the remaining 9 are the regressors, hence, this is the case of **Multiple Linear Regression**.

Aim

The aforementioned dataset is a historical time-series data set. Our principal objective in this project would be to understand how the various meteorological variables affect the ozone concentration. We would first understand the regressors and see how they behave with respect to the response variable as well as with each other through basic exploratory data analysis. Further, we would deal with the violation of basic assumptions such as multicollinearity, auto-correlation, heteroscedasticity and normality of the errors as well as take necessary measures to rectify them. This will constitute a large chunk of feature engineering. Finally, we plan to carry out model selection across various models obtained in process while making use of adequate metrics.

References

[1] ozone: Ozone in LA in 1976

[2] Access to downloaded dataset

[3] Introduction to Linear Regression Analysis Book by Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining