PSTAT 126 Final Project: Option 2

The overall project consists of a thorough investigation of regression models that combine concepts and methods of linear regression used throughout the quarter. If you choose this option, you will need to study two data sets.

1 Part I: Real Estate Valuation Data

1.1 Data Description

The market historical data set of real estate valuation are collected from Sindian District, New Taipei City, Taiwan during the period of June 2012 to May 2013. The original data set is available on the *UC Irvine Machine Learning Repository.*

Each line of data set provides information on 8 variables.

Variable Name	Description
TDate	the transaction date (e.g., 2013.250=2013 March, 2013.500=2013 June, etc.)
Age	the house age (unit: year)
Metro	the distance to the nearest MRT station (unit: meter)
Stores	the number of convenience stores in the living circle on foot (integer)
Latitude	the geographic coordinate (unit: degree)
Longitude	the geographic coordinate (unit: degree)
Price	the house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local 7 unit, 1 Ping = 3.3 meter squared)

The file **RealEstateValuation.txt** contains this data and is available on GauchoSpace.

1.2 Project Components

You will investigate the model

$$\texttt{Price} \sim \texttt{TDate + Age + Stores + Latitude} \tag{1}$$

by answering the following questions.

a) What relationships do you expect to see between the response and each of the predictors, and why? What kind of associations, if any, do you expect will be present between the four predictors, and why? Do some exploratory analysis (e.g. plots and/or numerical summaries) to test your intuition.

b) Fit the model in (1) and write down an equation for the fitted regression line. Conduct tests on individual regression coefficients with $\alpha = 0.01$, and interpret the estimated coefficient if it is significant.

c) Conduct tests of whether or not you need to add `Metro` and/or `Longitude` to the model (1) given $\alpha = 0.05$. Give the null and alternative hypotheses (defining any notation that you use), value of the test statistic and its null distribution, the p-value or critical value, and your decision.

d) Consider another model

$$\texttt{Price} \sim \texttt{TDate + Age + Metro + Latitude.} \tag{2}$$

Fit the model in (2), and write the regression line. Between the models in (1) and (2), which one do you prefer? Explain your reasoning.

e) With your preferred model from d), investigate possible transformations predictors and/or response. Did you find any improvement after transforming the variables?

f) Summarize your analysis and comment on any interesting or unexpected findings.

2 Part II: Concrete Compressive Strength Data Set

2.1 Data Description

The data set was collected for the paper, *Modeling of strength of high performance concrete using artificial neural networks*, by Dr. I-Cheng Yeh. You will use this data set to model the relationship between the concrete compressive strength and the 8 concrete components.

Variable Name	Description
X1	Cement (component 1, unit kg/m^3)
X2	Blast Furnace Slag (component 2, unit kg/m^3)
X3	Fly Ash (component 3, unit kg/m^3)
X4	Water (component 4, unit kg/m^3)
X5	Superplasticizer (component 5, unit kg/m^3)
X6	Coarse Aggregate (component 6, unit kg/m^3)
X7	Fine Aggregate (component 7, unit kg/m^3)
X8	Age (Day, $1 \sim 365$)
Y	Concrete compressive strength (MPa)

The file **Concrete.txt** contains this data and is available on GauchoSpace. The original data set is available on the *UC Irvine Machine Learning Repository.*

2.2 Project Components

a) Apply the forward selection algorithm, using BIC as a criterion function. With this final model,

- Do the diagnostic checks to assess whether or not the linear regression assumptions seem to hold. Do you find any influential points that you want to remove? For any data points with large influence, use leverages and/or residuals (standardized or studentized) to explain why they are influential.
- Estimate a mean response for the predictor values of your interest. Compute 95% confidence interval for the mean response, and provide the interpretation.
- Predict a new response for the same predictor values. Compute 95% prediction interval for the individual response, and provide the interpretation.

b) Apply the backward elimination algorithm, using BIC as a criterion function. With this final model,

- Do the diagnostic checks to assess whether or not the linear regression assumptions seem to hold. Do you find any influential points that you want to remove? For any data points with large influence, use leverages and/or residuals (standardized or studentized) to explain why they are influential.
- Do you get the same final model in a)? If they differ, which one do you want to select as your final model. Explain your reasoning.

c) Summarize your analysis and comment on any interesting or unexpected findings.