Assignment 2 Statistical Regression

Instructions:

- Submit your answers to the task below as a clean and well-structured report, not exceeding four (4) pages. The four pages include both the cover page and conclusion.

  Note: If your submission exceeds four pages, only the first four pages will be graded.

- Submit a separate R script file containing your implementation code. The tutors will verify your results by running your R script.

- The written report must be in your own words and should not include any code within the report.

Important: Failure to follow the instructions will result in heavy penalties.

Tasks

Consider the dataset available for download at

https://archive.ics.uci.edu/dataset/477/real+estate+valuation+data+set

The data set is about the price of the houses in a district in Taiwan.

Before exporting your file in R, we suggest to remove the first column in the Excel file and change the headers of the other columns to get shorter ones (use the ones below on the left) instead of the original ones (below on the right). The variables are:

`Date`	X1 transaction date
`Age`	X2 house age
`MRT-distance`	X3 distance to the nearest MRT station
`Convenience`	X4 number of convenience stores
`Latitude`	X5 latitude
`Longitude`	X6 longitude
`Price`	Y house price of unit area

1. Compute the correlations between each pair of variables and plot each variable against the others. Using correlations and plots, comment on the relations that you find more interesting. Consider in particular the correlation between `Price` and the other variables and comment.

2. Assuming `Price` as response variable, consider six multiple linear regressions where you start with `Date`, then you add `Age`, then add `MRT-distance`, then `Convenience`, then `Latitude` and, finally, you get all the variables adding `Longitude`. For each of the six models tell which coefficients are significant and if there are changes from one model to another. Compare the various R^2 and R^2_{Adj} and comment.

3. Find the best model according to AIC (Akaike Information Criterion) using the appropriate R command and comment.

4. Plot the residuals for the regression with just `Date` and `Age` and for the regression with all the variables. Compare the two plots and comment (we strongly suggest to use the same limits for the Y axis, to favour comparison).

5. Try to add powers and/or interactions and/or transformations of the X data, e.g. consider `Age*Age` or `Age * Convenience` or `log(MRT-distance)`. Out of your attempts, report the three which are better, in terms of R^2_{Adj}.

End