BUSI2045 – Homework 3

Note: Please include both R codes and results in your solutions. (You may use the "Compile Report" function under Menu "File" in RStudio to generate a Word/PDF report of both R codes and results)

Question 1: Linear Regression (50 Points)

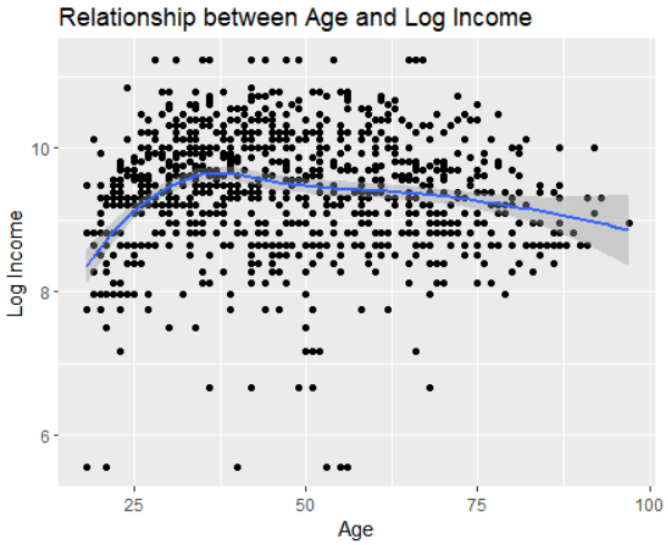
Load the dataset *NILT2012GR _SUBSET.csv* into R. The data contains 9 variables for 1204 citizens, which comes from Queen's University in Belfast (North Ireland) and is based on the Northern Ireland Life and Times Survey (NILT) 2012.

Create a subset named ***Q1*** in which variable ***persinc2*** (personal income) and ***rage*** (age) contains no missing value. Then create a new variable named ***log_income***, which takes log transformation of ***persinc2***, to answer below questions.

(a) We want to explore the effect of age (variable ***rage***) on log income (variable ***log_income***) with a simple linear regression model. In this model, what kind of relationship is assumed between the two variables and how would you present this relationship with an equation formula?

(b) Implement the linear regression in R and interpret the results accordingly.

(c) Visualize the relationship between ***age*** and ***log_income*** with a scatter plot. It should look similar as below.



(d) Create a new variable named ***ragesq***, which is the square of **rage**. Model the effect of ***rage*** and ***ragesq*** together on ***log_income*** with a multiple linear regression (no interaction terms). Check the model summary and interpret the result.

(e) Based on the scatter plot in part (c), explain why the coefficient of ***rage*** in part (b) is not statistically significant, while the coefficients of ***rage*** and ***ragesq*** in part (d) are statistically significant?

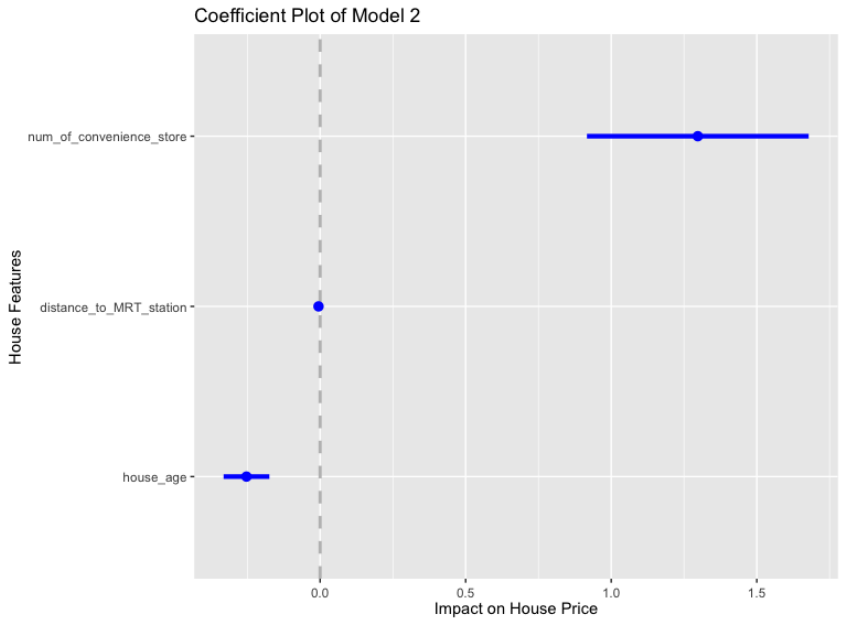
Question 2: Linear Regression and Model Comparison (50 Points)

Read the dataset *real_estate_valuation.csv* in R and answer the following questions. The data set contains information of house price and three different features of 414 houses. We'd like to explore the effect of different house features on house price (variable ***house_price_of_unit_area***).

(a) Model the effect of house age (variable ***house_age***) on house price with a simple linear regression (name the model as ***model_1***). Check the coefficient of ***house_age*** and its 95% confidence interval, interpret them accordingly.

(b) Model the effect of three house features (variable ***house_age***, ***distance_to_MRT_station***, and ***num_of_convenience_store***) on house price with a multiple linear regression (no interaction terms). Name the model as ***model_2***, check the model summary and interpret the result.

(c) Visualize the effect of the three house features on house price by plotting the coefficients of each predictor in ***model_2***. The graph should look similar as below.



(d) Is ***model_2*** significantly different from ***model_1***? Conduct a proper test to compare the two models and interpret the result.