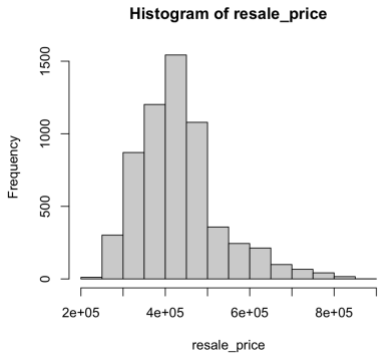
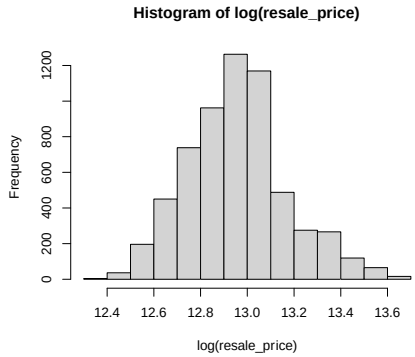
ST1131 Assignment 2

The purpose of this report is to investigate how different variables affect the response variable: the resale price of flats. I will first compare the suitability of the different categorical and quantitative variables to determine their suitability in a linear regression model. Then, I will propose a linear regression model taking into account its adequacy and goodness-of-fit.

Summary of Resale Price(SGD)

Min	230,000
1st Quantile	367000
Median	422000
Mean	433652
3rd Quantile	475000
Max	870000

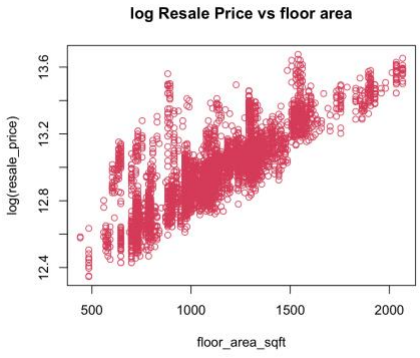




Given that resale price is a quantitative variable and its log resembles a normal distribution, log(Resale Price) is suitable to fit a linear regression model.

Explanatory Variables to be Tested

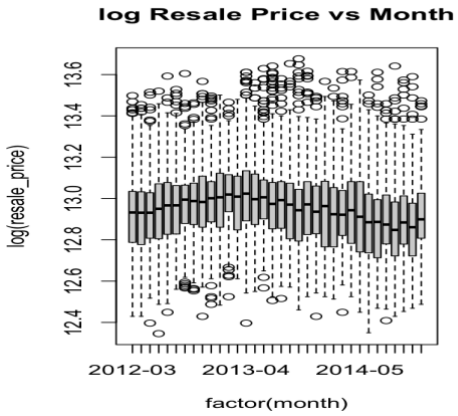
1. Floor area in square feet



The graph shows a linear relationship between log(resale price) and floor area. The t-value is extremely small, indicating the regressor is significant. Moreover, it has a relatively high R-squared value of 0.6944 and correlation of 0.8333. Thus, it is suitable to be the main regressor.

2. Month

According to the t-values, the range of months that have significance to the log of the resale price are those between August 2012 to November 2013. It has small R-squared value of 0.03557.

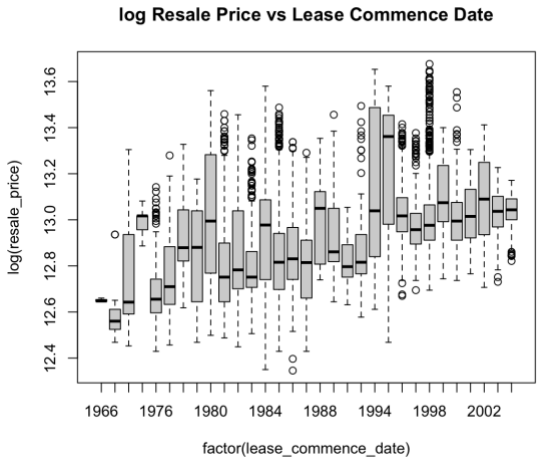


The boxplot shows a lot of overlaps between months.

Thus, the variable “months” might not be a suitable regressor.

3. Lease Commence Date

According to the t-values, the range of dates that have significance to the log of the resale price are those between 199 4 to 2004. It has small R-squared value of 0.2921.



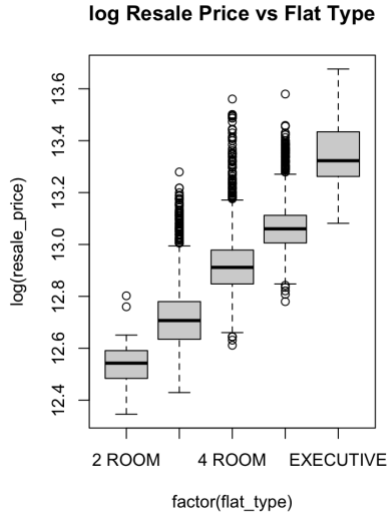
The boxplot shows minimal overlaps between years.

.

Thus, the variable “lease commence date” might be a suitable regressor.

4. Flat Type

According to the t-values, the types of flat are significant. It has relatively high R-squared value of 0.6948.

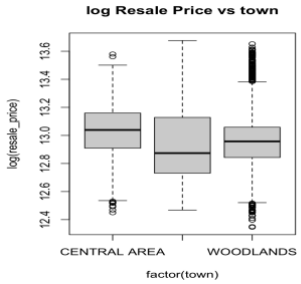


The boxplot shows almost no overlaps between flat types.

Thus, the variable “flat type” might be a suitable regressor.

5. Town

According to the t-values, the towns where the flats are located are significant. However, It has relatively low R-squared value of 0.007081.

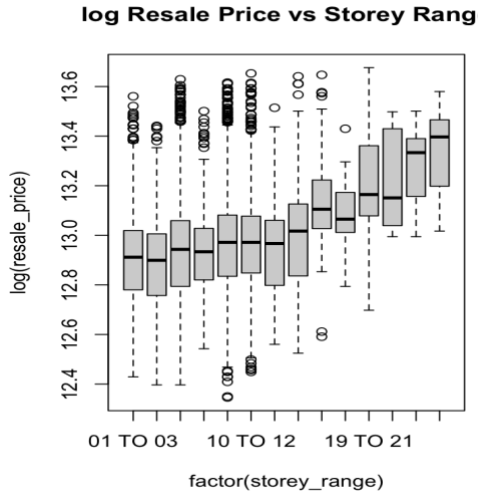


The boxplot shows overlaps between flat types but there are differences in the mean.

Thus, the variable “town” might not be a suitable regressor.

6. Storey Range

According to the t-values, the stories where the flats are located are significant. However, It has relatively low R-squared value of 0.05517.

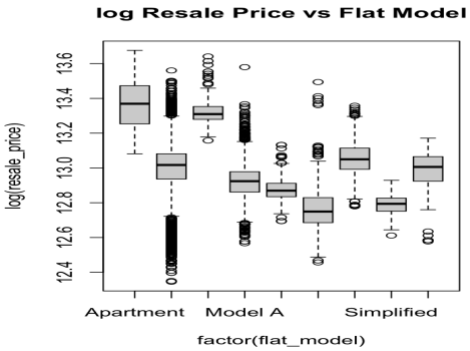


The boxplot shows little overlap between storey levels.

However, due to the disorganised data, the variable “storey range” might not be a suitable regressor.

7. Flat Model

According to the t-values, the flat models are significant. However, It has relatively low R-squared value of 0.5067.



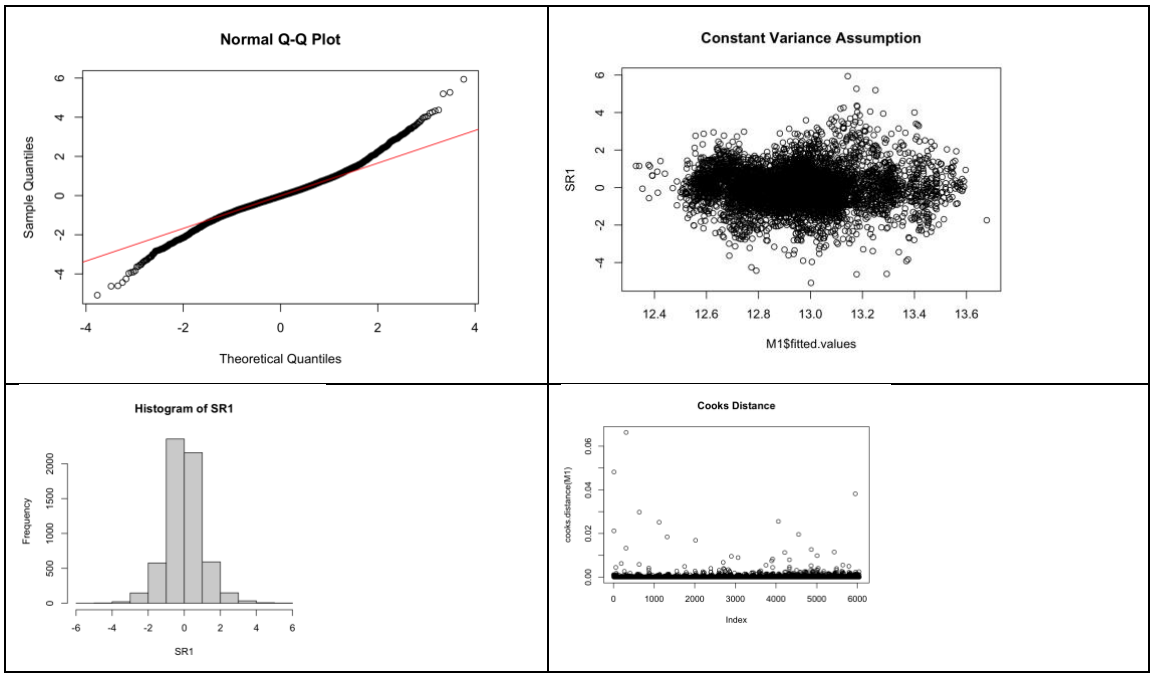
The boxplot shows little overlap between flat models

Thus, the variable “Flat Model” might be a suitable regressor.

<u>Linear Model 1 (M1):</u>

The first linear model takes into account all the variables stated above, with the assumption that floor area and flat type are associated to one another. The response variable is log(Resale Price).

Adj. R-squared value: 0.9198
Degrees of freedom: 5981
F – Statistic: < 2.26 e-16
SR1 is the standardisation of the residuals of this model



From these four diagrams, we see that this model fails the checks of normality, linearity and there are many influential points. Thus, despite model the low F – Statistic and that 91.98% of the variation in log(resale price) is explained by this regression model, the model is not accurate and hence needs improvement.

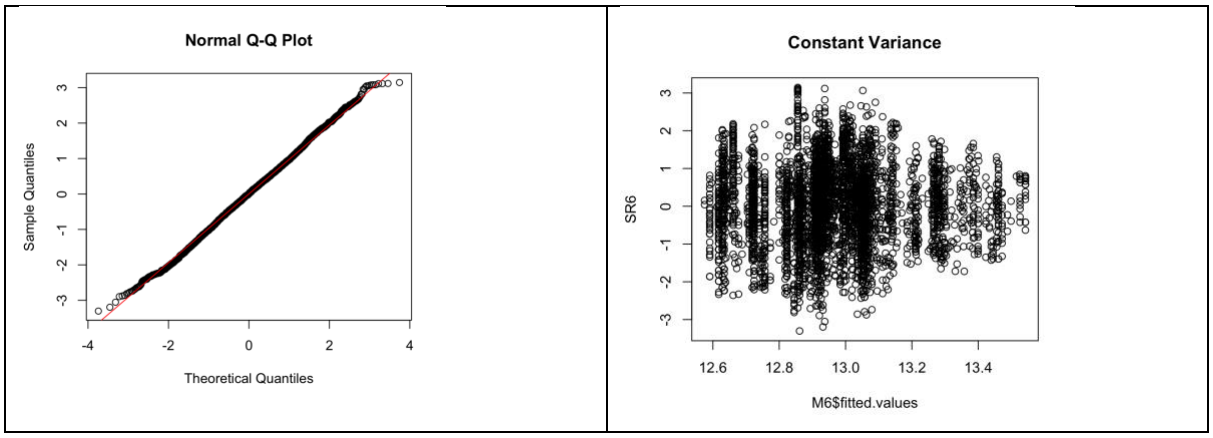
Adjustments:

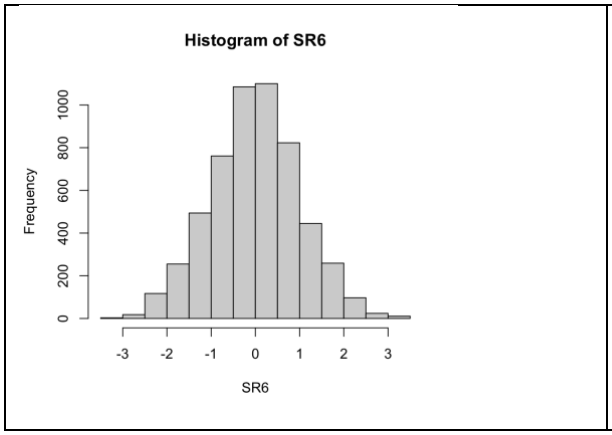
Using the anova function, variables were removed sequentially depending on their F – values and significance on the model. This was done until the best QQ plot with the highest p-value in the Shapiro test could be found. The variable storey range was removed as the data was disorganised which may affect the accuracy of the model by over complicating it. In addition, influential points with a Cooks Distance of more than twice the mean Cook's Distance were removed. Although flat area and flat type are associated, it is found that the model is more accurate without the associated variable.

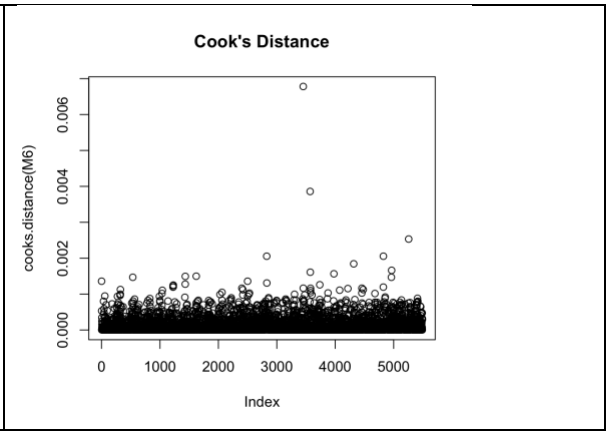
Linear Model 6 (LM6)

LM6 is tested against the same response variable, log(Resale Price) against the regressors floor area, flat type, town and storey range. The data that it draw from excludes the influential points described above.

Adj. R-squared value: 0.8899
Degrees of freedom: 5484
F – Statistic: < 2.26 e-16
SR6 is the standardisation of the residuals of this model







From these four diagrams, we see that this model passes the checks of normality, linearity and constant variance while having no influential points with a Cook's Distance of more than twice its mean. Thus, despite the lower Adj. R-squared value compared to previous models, the model is accurate and still has "goodness-of-fit". 88.99% of the variation in log(resale price) is explained by this regression model. Thus, it is a suitable model.

Fitted regression model:
Reference categories: 2 Room (flat type) and Central Area (Town)

Log(resale price) = 12.353 + 0.000504 x (floor area in sqft) + 0.293 x (I = 3 room) + 0.396 x (I = 4 room) + 0.424 x (I = 5 room) + 0.529 x (I = Executive) - 0.290 x (I = Jurong East) - 381 x (I = Woodlands)

Impact:
Effect of area on resale price:
For every increase in area by 1 unit, the resale price increases by e^0.000504 times.

Effect of flat type on resale price:
Given the same area and town, the difference in price in the flat type of any flat (excluding 2 room) and the 2 room flat is the multiple of e^(coefficient of that flat type). For example, the difference in price between a 2 room flat and a 3 room flat is e^0.293 times.

Effect of town area on resale price:
Given the same area and flat type, the difference in price of a flat in one town (excluding Central Area) and a Central Area flat is the multiple of e^(coefficient of that town). For example, the difference in price between a Central Area flat and a Jurong East flat is e^-0.290 times.

Conclusion:
The log of the resale price of HDB flats has a strong positive linear relationship with the flat area, flat type, and town the flats are located in. The larger the areas and number of rooms of the flats, the higher the resale price while flats located away from the central area tend to be cheaper. The. Model meets all its assumptions. In this model, 88.99% of the variation in log(resale price) is explained by this regression model. In future, the model can be further experimented on using a more structured storey range dataset.