ST1131 Assignment 2

Chin Zhe Ning, A0255895J

National University of Singapore
April 2023

1 Introduction

This report proposes a linear regression model to predict the resale price of HDB flats in Singapore, based on a dataset of transactions from 2012. The model will consider factors such as location, floor area, flat model, and storey range. We acknowledge that the data's relevance to today's market may be limited due to changes in inflation, housing policy, and socioeconomic factors. The report will provide an overview of the dataset, describe the methodology used to develop the model, present the results, and discuss its strengths and limitations.

2 Data characteristics

The dataset is obtained from https://data.gov.sg/dataset/resale-flat-prices, it is uncertain whether the data is randomized but we will assume the case for the analysis.

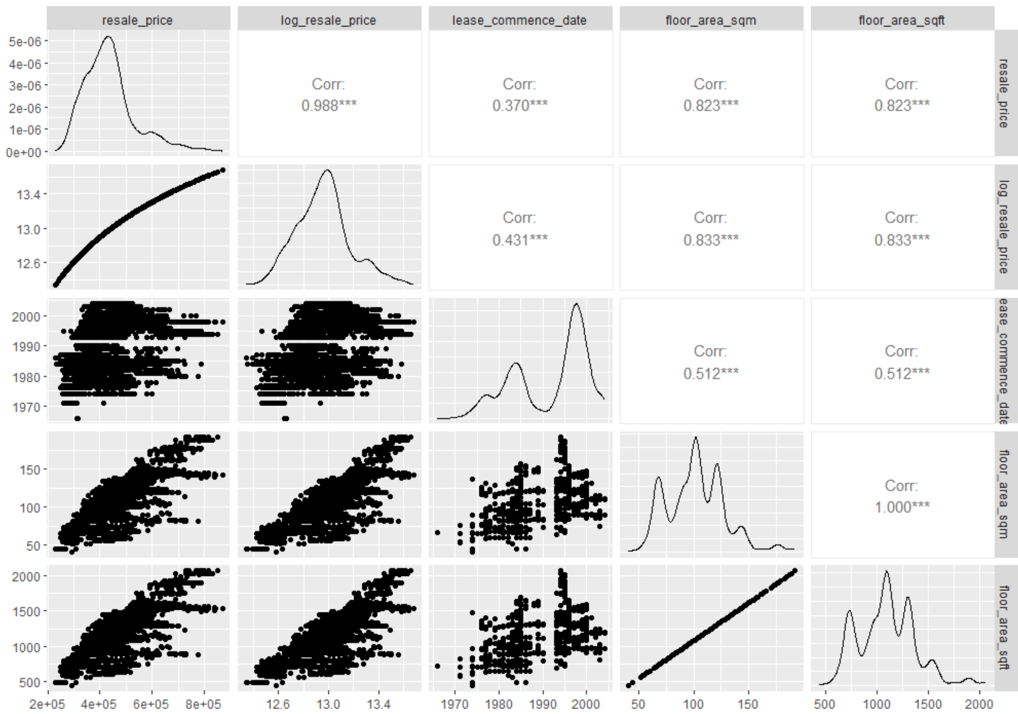


Fig. 1: Scatterplot matrix of quantitative variables.

We observe that `resale_price` is quantitative which is necessary for regression modelling and the distribution is right-skewed. There is weak or no correlation between `lease_commence_date` and `resale_price` as shown by the low correlation of 0.370 so we will not consider `lease_commence_date` as a regressor. There is a strong linear correlation between `floor_area_sqm` and `resale_price`.

```
> head(street_name)
[1] "ROWELL RD"   "ROWELL RD"   "CHANDER RD"   "TG PAGAR PLAZA"   "QUEEN ST"   "KLANG LANE"
> length(unique(street_name))
[1] 66
```

Note that `street_name` has many categorical variables that are not well distributed. In our analysis, we found categories with very few observations, such as "Jlm Berseh" with 3 observations only. The inclusion of `street_name` in the model may lead to overfitting or increased variability of the estimated coefficients. Hence, we will not consider `street_name` as a regressor.

3 Model Selection and Interpretation

3.1 Preliminary Models

Before arriving at the final model, we considered several preliminary models. In this section, we describe each of these models, including the choice of transforms, interactions, and regressors. We also discuss the reduction of non-significant regressors that informed the selection of the final model.

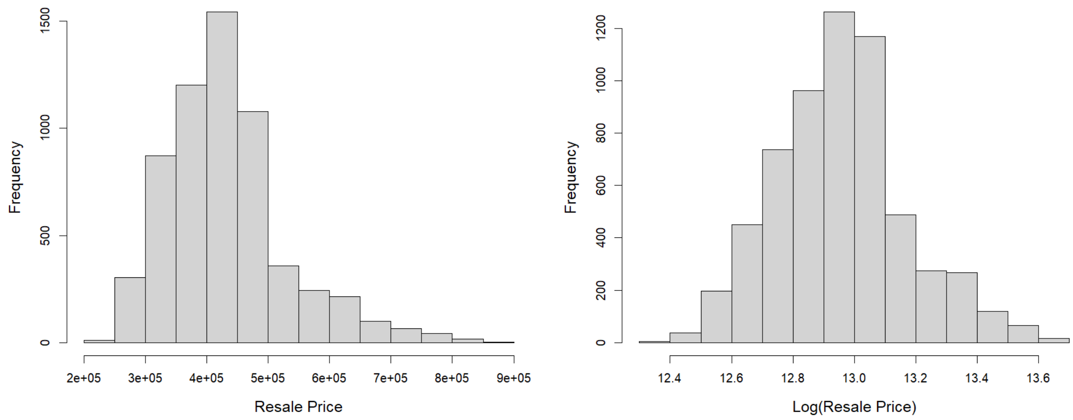


Fig. 2: Histogram of resale_price before and after logarithm transform.

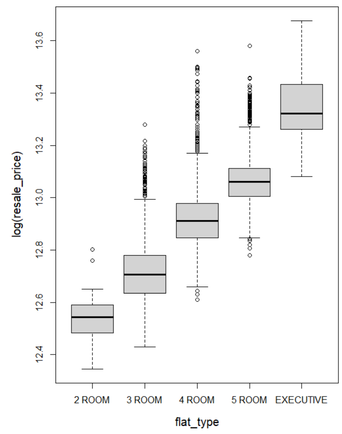


Fig. 3: log(resale_price) against flat_type.

3.1.1 Initial Model

The initial model is a simple linear regression model with a logarithm transform for the response variable `resale_price`. The systematic component consists of the explanatory variable `floor_area_sqm`.

As we identified from our preliminary analysis, `resale_price`, is right-skewed, hence we apply a logarithm transform to `resale_price` to achieve a more symmetrical distribution. This is necessary for `resale_price` to be suitable for linear regression. Since `resale_price` is large, boundary conditions will be avoided, hence the transform is appropriate. The histograms of `resale_price` before and after the transform are shown in Fig. 2.

A high correlation value of 0.833 between `log(resale_price)` and `floor_area_sqm` was also observed. Therefore, we expect `floor_area_sqm` to be a significant predictor of `resale_price`.

To check for linearity, `log(resale_price)` is plotted against `floor_area_sqm` (Fig. 4) A clear and strong linear association is observed.

Verifying these assumptions, we then proceed with the model specification:

$$M_1 : \log(\text{resale\_price}) \sim \text{floor\_area\_sqm}$$

where `log(resale_price)` is the log-transformed response variable, the resale price of HDBs, and `floor_area_sqm` is the explanatory variable representing the floor area of the HDB flat in square meters (m^2).

This model was then fitted in R with the `lm()` function using the Ordinary Least Squares (OLS) Method. The model summary is provided in Fig. 5.

Fig. 4: M_1 Linear Regression Line

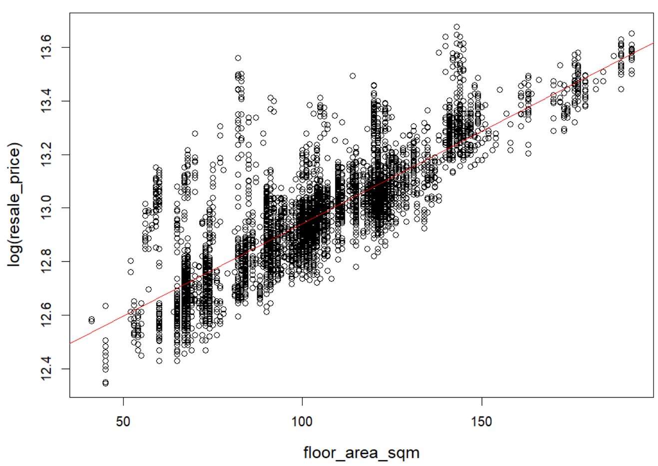


Fig. 5: M_1 Model Summary

```
Call:
lm(formula = log(resale_price) ~ floor_area_sqm)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.28636 -0.07214 -0.01699  0.04164  0.74378 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    1.225e+01  6.203e-03  1975.1   <2e-16 ***
floor_area_sqm 6.893e-03  5.881e-05   117.2   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1185 on 6045 degrees of freedom
Multiple R-squared:  0.6944,	Adjusted R-squared:  0.6944 
F-statistic: 1.374e+04 on 1 and 6045 DF,  p-value: < 2.2e-16
```

The fitted equation for M_1 is $\log(\text{resale\_price}) = 0.006893 \times \text{floor\_area\_sqm} + 12.252$. This shows that for every increase in floor area (m^2), there is an increase of 0.006893 in log resale price. Note that the p-value of the F-test ($< 2.2 \times 10^{-16}$) is very small, and indicates that the overall model is statistically significant. This is of no surprise as `log(resale_price)` and `floor_area_sqm` are highly correlated. However, the R^2 value (0.694) is not very high, so the M_1 does not have strong predictive power. This suggests that the model has poor goodness of fit. The model also has a residual standard error of 0.119 on 6045 degrees of freedom.

We now check the normality assumption. From the normal QQ-plot of the standardised residuals (SR) for M_1 (Fig. 6), we see that the quantile points are not consistent with the theoretical normal line as the right-tail deviates. This means that the SR do not follow a normal distribution.

We also check for constant variance. The residual plot (Fig. 6) of M_1 shows that the spread of the variance varies across the fitted values. For higher fitted values, there is a lower spread present as seen from the funnel shape. This suggests that constant variance is not present. Both the normality and constant variance assumption are not satisfied, therefore, M_1 is not accurate and cannot be used. The next step is to introduce more explanatory variables to improve the accuracy and goodness-of-fit of M_1.

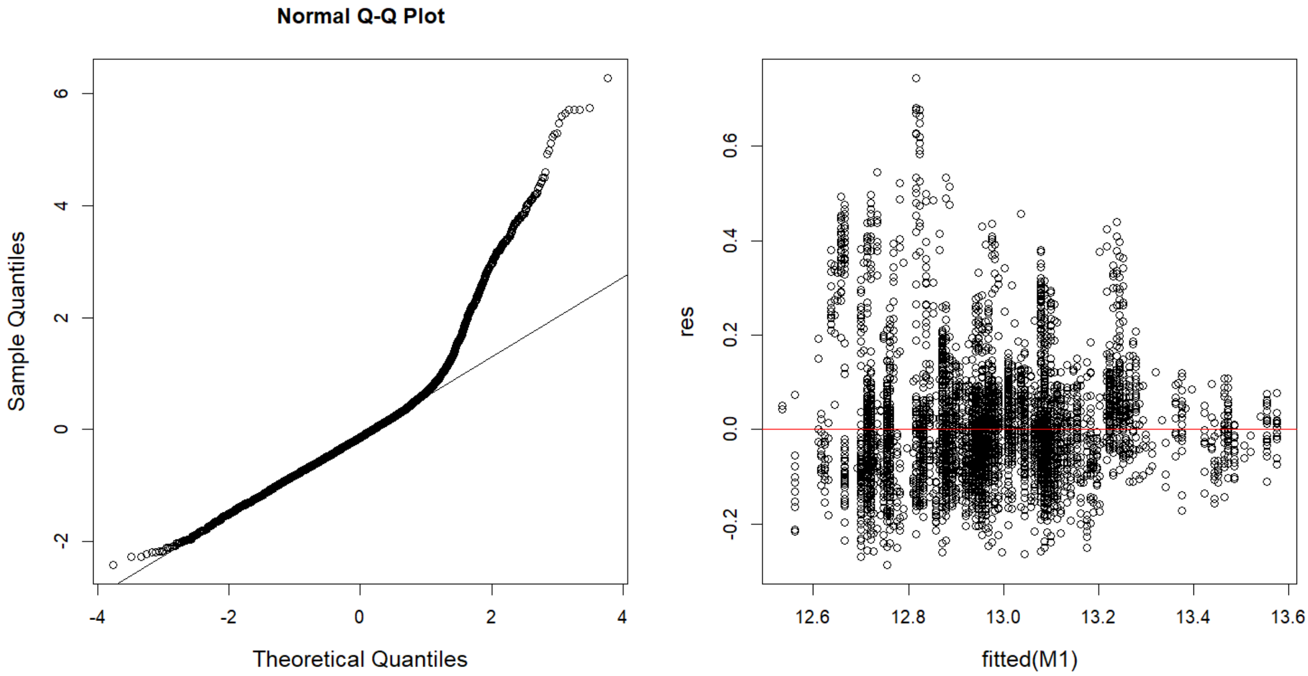


Fig. 6: Normal QQ-Plot and Residual Plot of M_1

3.1.2 Intermediate Model

Similarly, with the initial model, we apply a logarithm transform for the response variable `resale_price`.

Since the explanatory variables `floor_area_sqm` and `floor_area_sqft` are equivalent representations of each other (Fig. 1), we exclude `floor_area_sqft` to avoid multicollinearity in our model. We also exclude `flat_type` as there is some collinearity with `log(floor_area_sqm)` (Fig. 3).

We include `flat_model` and `town` as categorical explanatory variables representing the model and location of the HDB flat respectively. These seem causally related to the resale price of the HDB flat so we included them.

We also include `storey_range`, a categorical explanatory variable to see if it would improve the model's fit and/or accuracy. Since `storey_range` has overlapping categories, we merge overlapping categories to encode. From this, 2 distinct categories are obtained, "01 TO 15" and "16 TO 27".

The interaction terms `floor_area_sqm * town`, `town * flat_model` and `flat_model * floor_area_sqm` were introduced to account for any interaction between these variables. Now the systematic component of the intermediate model includes the explanatory variables: `town`, `floor_area_sqm`, `flat_model`, and `storey_range` and also the interaction terms mentioned above. The intermediate model specification is as follows:

$$M2 : \log(\text{resale\_price}) \sim \text{town} + \text{floor\_area\_sqm} + \text{flat\_model} + \text{storey\_range} + \text{floor\_area\_sqm} * \text{town} \\ + \text{town} * \text{flat\_model} + \text{flat\_model} * \text{floor\_area\_sqm}.$$

The model summary is given in Fig. 7 and the fitted model equation is omitted here for conciseness (see Appendix).

Fig. 7: M_2 Model Summary

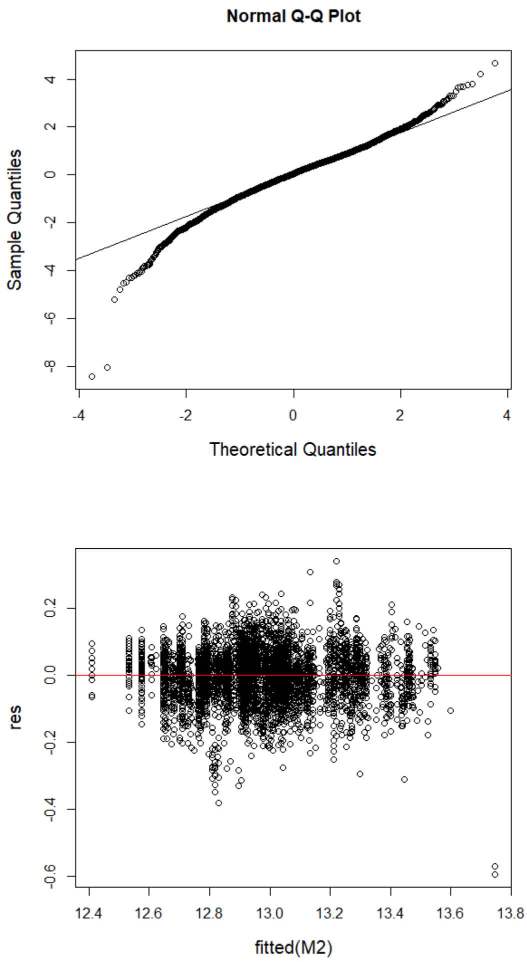
```
Call:
lm(formula = log(resale_price) ~ town + floor_area_sqm + flat_model +
    storey_range + floor_area_sqm * town + town * flat_model +
    flat_model * floor_area_sqm)

Residuals:
     Min       1Q   Median       3Q      Max
-0.59496 -0.04168  0.00339  0.04475  0.34187

Coefficients: (6 not defined because of singularities)
                                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                                  1.257e+01  7.909e-02 158.909  < 2e-16 ***
townJURONG EAST                             -3.700e-01  7.404e-02  -4.997 5.99e-07 ***
townWOODLANDS                               -1.352e-01  7.204e-02  -1.877  0.06054 .
floor_area_sqm                               8.743e-03  3.629e-04  24.090  < 2e-16 ***
flat_modelImproved                          -3.525e-01  7.607e-02  -4.634 3.66e-06 ***
flat_modelMaisonette                         1.881e+00  2.905e-01   6.476 1.02e-10 ***
flat_modelModel A                           -2.244e-01  7.723e-02  -2.906  0.00368 **
flat_modelModel A2                           4.257e-03  1.050e-01   0.041  0.96767
flat_modelNew Generation                    -3.080e-02  8.047e-02  -0.383  0.70193
flat_modelPremium Apartment                 -2.514e-02  4.061e-02  -0.619  0.53589
flat_modelSimplified                        -2.214e-01  1.966e-01  -1.126  0.26019
flat_modelStandard                          -2.047e-01  1.274e-01  -1.607  0.10816
storey_range16 TO 27                         8.686e-02  4.097e-03  21.202  < 2e-16 ***
townJURONG EAST:floor_area_sqm               3.898e-05  3.238e-04   0.120  0.90421
townWOODLANDS:floor_area_sqm                -2.928e-03  3.005e-04  -9.744  < 2e-16 ***
townJURONG EAST:flat_modelImproved           8.765e-03  6.935e-02   0.126  0.89943
townWOODLANDS:flat_modelImproved            -4.166e-02  6.815e-02  -0.611  0.54103
townJURONG EAST:flat_modelMaisonette        -8.465e-02  1.951e-02  -4.338 1.46e-05 ***
townWOODLANDS:flat_modelMaisonette                  NA         NA      NA       NA
townJURONG EAST:flat_modelModel A            1.922e-01  6.979e-02   2.754  0.00590 **
townWOODLANDS:flat_modelModel A              1.735e-01  6.853e-02   2.532  0.01135 *
townJURONG EAST:flat_modelModel A2           1.343e-01  2.449e-02   5.485 4.30e-08 ***
townWOODLANDS:flat_modelModel A2                    NA         NA      NA       NA
townJURONG EAST:flat_modelNew Generation    -4.135e-02  7.227e-02  -0.572  0.56719
townWOODLANDS:flat_modelNew Generation      -1.316e-01  7.115e-02  -1.849  0.06445 .
townJURONG EAST:flat_modelPremium Apartment         NA         NA      NA       NA
townWOODLANDS:flat_modelPremium Apartment           NA         NA      NA       NA
townJURONG EAST:flat_modelSimplified        -1.182e-01  2.927e-02  -4.038 5.45e-05 ***
townWOODLANDS:flat_modelSimplified                  NA         NA      NA       NA
townJURONG EAST:flat_modelStandard          -2.584e-02  2.250e-02  -1.148  0.25095
townWOODLANDS:flat_modelStandard                    NA         NA      NA       NA
floor_area_sqm:flat_modelImproved            2.439e-03  2.236e-04  10.905  < 2e-16 ***
floor_area_sqm:flat_modelMaisonette         -1.286e-02  2.012e-03  -6.391 1.77e-10 ***
floor_area_sqm:flat_modelModel A            -5.812e-04  2.508e-04  -2.317  0.02052 *
floor_area_sqm:flat_modelModel A2           -1.177e-03  1.100e-03  -1.070  0.28446
floor_area_sqm:flat_modelNew Generation     -1.876e-04  2.842e-04  -0.660  0.50921
floor_area_sqm:flat_modelPremium Apartment  -8.402e-05  2.958e-04  -0.284  0.77639
floor_area_sqm:flat_modelSimplified          1.085e-03  2.306e-03   0.471  0.63791
floor_area_sqm:flat_modelStandard            1.847e-04  1.020e-03   0.181  0.85639
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07358 on 6014 degrees of freedom
Multiple R-squared:  0.8829,    Adjusted R-squared:  0.8822
F-statistic:  1416 on 32 and 6014 DF,  p-value: < 2.2e-16
```

Fig. 8: M_2 Normal-QQ Plot and Residual Plot



Like M_1, the p-value of the F-test ($< 2.2 \times 10^{-16}$) for M_2 is very small and suggests that the overall model is statistically significant. Also, the adjusted R^2 value (0.8822) is high and improves significantly from M_1 (adjusted $R^2 = 0.694$). We compare the adjusted R^2 instead of multiple R^2 of M_1 and M_2 because it penalizes excessive model complexity and provides a more accurate measure of the proportion of variance in the dependent variable explained by the independent variables. This suggests that M_2 has stronger predictive power and better goodness-of-fit than M_1. The model also has a lower residual standard error of 0.07358 on 6014 degrees of freedom, improving from M_1.

Note that the normal QQ-plot of the SR of M_2 (Fig. 8) shows the quantile points to be mostly consistent with the theoretical normal line. Hence, we conclude that the SR of M_2 do follow a normal distribution. Therefore, normality is present. From the residual plot (Fig, 8) we observed that the spread of residuals across all levels of fitted values is more or less constant. Therefore, we conclude that the constant variance assumption holds. Since both normality and constant variance assumptions hold, M_2 is an accurate model.

Of the explanatory variables included, `town`, `floor_area_sqm`, `storey_range` are statistically significant, as indicated by the small p-values and *** in the model summary. The interaction terms `flat_model * floor_area_sqm` and `town * flat_model` were not found to be statistically significant in M_2 and thus were discarded. However, the other interaction term `floor_area_sqm * town` was found to be significant, so we include it in our final model M_n.

3.2 Final Model

M_n is specified:

$$M_n : \log(\text{resale\_price}) \sim \text{town} + \text{floor\_area\_sqm} + \text{flat\_model} + \text{storey\_range} \\ + \text{floor\_area\_sqm} * \text{town}$$

The model summary is given in Fig. 9.

Fig. 9: M_n Model Summary

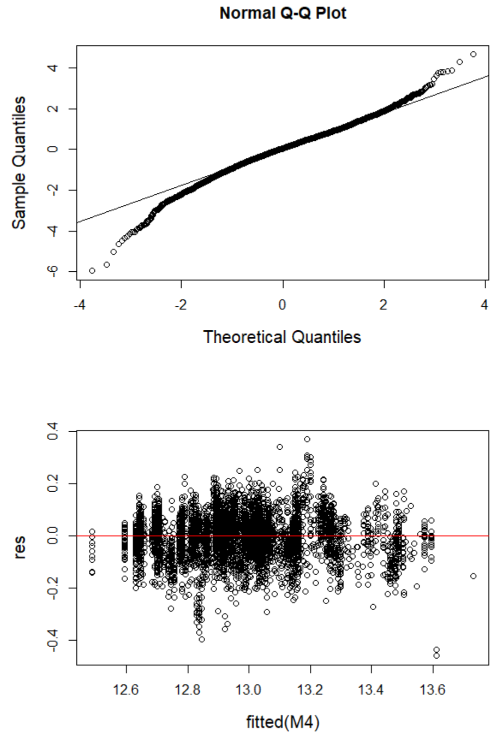
```
Call:
lm(formula = log(resale_price) ~ town + floor_area_sqm + flat_model + 
    storey_range + floor_area_sqm * town)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.46034 -0.04534  0.00466  0.04994  0.36986 

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 12.3977836  0.0216059 573.816  < 2e-16 ***
townJURONG EAST             -0.2148253  0.0239793  -8.959  < 2e-16 ***
townWOODLANDS               -0.1642640  0.0225781  -7.275 3.89e-13 ***
floor_area_sqm               0.0093359  0.0002859  32.658  < 2e-16 ***
flat_modelImproved          -0.0645486  0.0057373 -11.251  < 2e-16 ***
flat_modelMaisonette         0.0136404  0.0074771   1.824 0.068157 .  
flat_modelModel A           -0.0452463  0.0060737  -7.450 1.07e-13 ***
flat_modelModel A2          -0.0172205  0.0080031  -2.152 0.031458 *  
flat_modelNew Generation    -0.0677977  0.0071862  -9.434  < 2e-16 ***
flat_modelPremium Apartment  0.0214210  0.0061369   3.491 0.000485 ***
flat_modelSimplified        -0.0485969  0.0087918  -5.528 3.38e-08 ***
flat_modelStandard          -0.1306309  0.0093642 -13.950  < 2e-16 ***
storey_range16 TO 27         0.0919716  0.0042607  21.586  < 2e-16 ***
townJURONG EAST:floor_area_sqm -0.0007321  0.0003080  -2.377 0.017481 *  
townWOODLANDS:floor_area_sqm   -0.0022456  0.0002970  -7.561 4.59e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07947 on 6032 degrees of freedom
Multiple R-squared:  0.8629,	Adjusted R-squared:  0.8626 
F-statistic:  2713 on 14 and 6032 DF,  p-value: < 2.2e-16
```

Fig. 10: M_n Normal-QQ Plot and Residual Plot



The model has an adjusted R^2 of 0.8626, indicating that the model explains 86.26% of the variance in the log of resale price, after adjusting for the number of predictors. The residual standard error is 0.07947, suggesting that the model's predictions are, on average, within 7.95% of the true log resale price. The adjusted R^2 value for M_n is lower than the adjusted R^2 value for M2 (0.8822) but is still very high. Like M1 and M2, the p-value of the F-test ($< 2.2 \times 10^{-16}$) for M_n is very small, indicating that the overall model is statistically significant and has high goodness of fit. We prefer this model despite the slightly lower adjusted R^2 but with fewer regressors and interaction terms because it would have better generalization performance, meaning it would be less likely to overfit the training data and would better capture the true underlying relationship between the variables. Almost all of the regressors in M3 are statistically significant at the 99.9% level as indicated by the p-value (< 0.001).

To check for normality and constant variance, we consider the normal QQ-plot of the SR of M_n and the residual plot in Fig. 10. Note that the quantile points are very consistent with the theoretical normal line hence the SR of M_n follow a normal distribution. So normality is present. From the residual plot, it is observed that the spread of residuals across all levels of fitted values is consistent. Therefore, we conclude that constant variance is present. Since the assumptions are satisfied, M_n is an accurate model and therefore can be used. Finally, we check for outliers and influential points in R.

```
> length(which(SR > 3 | SR < -3))
[1] 52
> C <- cooks.distance(Mn)
> which(C > 1)
named integer(0)
```

The output tells us that there are 52 outliers (of 6047 observations) and no influential points as there are 0 points beyond the Cook's distance. Then the fitted equation for M_n is given by:

$$
\begin{aligned}
\log(\text{resale\_price}) = {} & 12.398 + 0.00934 * \text{floor\_area\_sqm} + 0.0920 * I(\text{storey\_range} = 16 \text{ TO } 27) \\
& - 0.0645 * I(\text{flat\_model} = \text{Improved}) + 0.0136 * I(\text{flat\_model} = \text{Maisonette}) \\
& - 0.0452 * I(\text{flat\_model} = \text{Model A}) - 0.0172 * I(\text{flat\_model} = \text{Model A2}) \\
& - 0.0678 * I(\text{flat\_model} = \text{New Generation}) + 0.0214 * I(\text{flat\_model} = \text{Premium Apartment}) \\
& - 0.0486 * I(\text{flat\_model} = \text{Simplified}) - 0.131 * I(\text{flat\_model} = \text{Standard}) \\
& + I(\text{town} = \text{JURONG EAST})(-0.215 - 0.000732 \times \text{floor\_area\_sqm}) \\
& + I(\text{town} = \text{WOODLANDS})(-0.164 - 0.00225 \times \text{floor\_area\_sqm})
\end{aligned}
$$

The intercept coefficient is 12.398. This means that the mean value of `resale_price` of HDB flats is $e^{12.398} = 242,000$ when the floor area is 0 square metres, storey is within 1 to 15, flat model is Apartment, and location is CENTRAL AREA. Furthermore, the coefficient of indicator variable `storey_range` is 0.0920 so we expect, holding all other variables constant, the `resale_price` to be higher by $(e^{0.0920} - 1) \times 100 = 9.64\%$ The coefficient of `floor_area_sqm` is 0.00934. This means that for every 1 square meter increase in the floor area, the average percentage increase of the HDB flat `resale_price` is $(e^{0.00934}–1) \times 100 = 0.938\%$ when the location is CENTRAL AREA and all other variables are held constant.

The reference group is Apartment flats for analysis of `flat_model`. The coefficient is

1. -0.0645 when the flat model is Improved so we expect, holding all other variables constant, the mean `resale_price` to be lower by $(e^{-0.0645} - 1) \times 100 = -6.25$ %
2. +0.0136 when the flat model is Maisonette so we expect, holding all other variables constant, the mean `resale_price` to be higher by $(e^{0.0136} - 1) \times 100 = 1.37$ %
3. -0.0452 when the flat model is Model A so we expect, holding all other variables constant, the mean `resale_price` to be lower by $(e^{-0.0452} - 1) \times 100 = -4.42$ %
4. -0.0172 when the flat model is Model A2 so we expect, holding all other variables constant, the mean `resale_price` to be lower by $(e^{-0.0172} - 1) \times 100 = -1.71$ %
5. -0.0678 when the flat model is New Generation so we expect, holding all other variables constant, the mean `resale_price` to be lower by $(e^{-0.0678} - 1) \times 100 = -6.56$ %
6. +0.0214 when the flat model is Premium Apartment so we expect, holding all other variables constant, the mean `resale_price` to be higher by $(e^{0.0214} - 1) \times 100 = 2.16$ %
7. -0.0486 when the flat model is Simplified so we expect, holding all other variables constant, the mean `resale_price` to be lower by $(e^{-0.0486} - 1) \times 100 = -4.74$ %
8. -0.131 when the flat model is Standard so we expect, holding all other variables constant, the mean `resale_price` to be lower by $(e^{-0.131} - 1) \times 100 = -12.3$ %

when compared to HDBs of flat model Apartment.

For analysis of `town`, the reference group is flats in CENTRAL AREA. The coefficient of `town` is -0.215 and -0.164 when the flat is in JURONG EAST and WOODLANDS respectively. This means that we expect the mean `resale_price` to be lower by

1. $(e^{-0.215} - 1) \times 100 = -19.3\%$ on average for HDBs in JURONG EAST compared to CENTRAL AREA.
2. $(e^{-0.164} - 1) \times 100 = -15.1\%$ on average for HDBs in WOODLANDS compared to CENTRAL AREA.

The coefficients for the interaction term `flat_model * floor_area_sqm` are -0.000732 and -0.00225. Hence, given a 1 square meter increase in `floor_area_sqm`, the average percentage increase in mean `resale_price` is $(e^{0.00934-0.000732} - 1) \times 100 = 0.865\%$ and $(e^{0.00934-0.00225} - 1) \times 100 = 0.712\%$ which is lower by 0.073% and 0.226% for flats in JURONG EAST and WOODLANDS, respectively, when compared to HDB flats in CENTRAL AREA.

4 Conclusion

In this analysis, we developed a linear regression model to predict HDB resale prices in Singapore. We started with a preliminary model that included several explanatory variables and iteratively refined the model through diagnostic tests and statistical analysis. Our final model includes significant regressors such as floor area, town, flat model, storey range, and the interaction terms between floor area and town, providing a good fit to the data with reasonable diagnostic results.