ST3131 Assignment 1 AY25/26, A0308017M Okaggar Bo

Part 0: Introduction

The objective of this report is to develop and propose the best possible linear regression model to predict the resale price of Housing and Development Board (HDB) flats in Singapore. Understanding the factors that influence resale prices is important for assessing housing affordability and property market trends in a nation where over 80% of residents live in HDB flats. The dataset, obtained from Data.gov.sg, contains 11 variables and 11,527 observations representing resale transactions between January and June 2021. Each record includes details such as the month of sale, town, flat type, floor area, flat model, storey range, and lease commencement year, alongside the resale price in Singapore dollars. These variables capture both physical attributes (e.g., flat size, storey height, architectural model) and locational characteristics (e.g., town or region). The analysis proceeds in several stages. Part I (EDA) explores the response and explanatory variables using summary statistics and visualizations such as histograms, scatter plots, and box plots to understand data patterns and relationships. Part II fits and refines multiple linear regression models, starting from an initial model (M_0) with at least six regressors, to test the adequacy of the model, which includes the linearity, normality, and constant-variance assumptions, and to address issues such as multicollinearity and overfitting. Subsequent sections employ model selection criteria (Adjusted R^2, AIC, and plots) to identify the most adequate model. The report concludes with the presentation and interpretation of the final model (M_n) and its practical implications for understanding the key determinants of HDB resale prices in Singapore.

Part I EDA: Exploring the variables and association

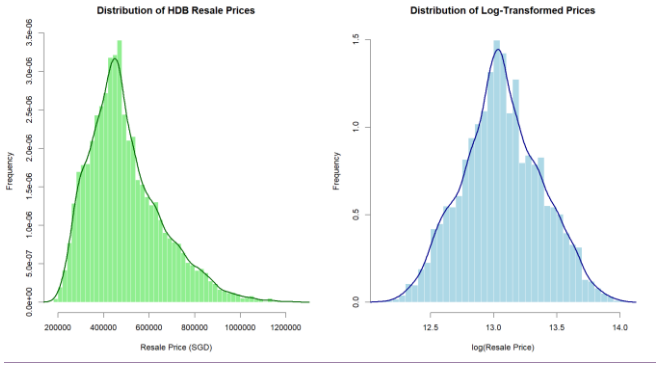


Figure 1.1

The histogram in Figure 1.1 shows that HDB resale prices from January to June 2021 are right-skewed, with most transactions between SGD 300,000 and 1,000,000, and a few high-value flats extending the tail beyond 1 million. The overlaid density curve indicates clear deviation from normality, suggesting that using raw prices may violate linear regression assumptions. To correct this, the response variable was log-transformed, yielding a roughly symmetric, bell-shaped distribution (Figure 1.1). This transformation effectively reduces skewness and stabilizes variance, making log(resale price) a more appropriate response variable for subsequent regression modeling.

To prepare the data for analysis, storey range was converted from a categorical predictor (e.g., "07 TO 09") to numeric midpoints (e.g., 8), allowing storey height to be treated as a quantitative predictor. The 26 towns were also grouped into three regional categories, Core Central Region (CCR), Rest of Central Region (RCR), and Outside Central Region (OCR), to reduce dimensionality and enhance interpretability of location effects, as well as to prevent excessive coefficients in our models. Variables such as block and street name were removed because they capture overly specific location details that do not generalize well. Since the regional grouping already represents geographic variation, including these variables would add redundancy and risk multicollinearity without improving explanatory power.

In summary, the dataset comprises both quantitative variables (e.g., floor area, storey range, lease commencement date, remaining lease) and categorical variables (e.g., month, flat type, flat model, region), allowing for a comprehensive analysis of how both continuous and qualitative attributes influence HDB resale prices

The scatterplots in Figure 1.2 illustrate the relationships between log(resale price) and four quantitative predictors: floor area, lease commencement date, storey range, and remaining lease. Floor area shows the strongest positive linear relationship, indicating that larger flats generally command higher resale prices. A moderate positive trend is also observed for storey height, as higher floors tend to be valued more for better views and ventilation. Lease commencement date and remaining lease both exhibit mild positive relationships with price, reflecting the depreciation effect of Singapore's 99-year leasehold system, as newer flats or those with longer leases sell for more. Although some dispersion and mild heteroscedasticity remain, the overall trends are approximately linear, making these predictors suitable for multiple linear regression analysis.

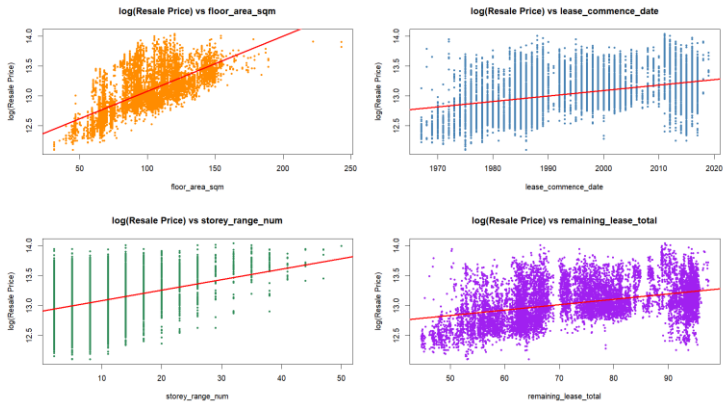


Figure 1.2

The boxplots in Figure 1.3 below show variations in log(resale price) across the categorical variables: month, flat type, flat model, and region. Prices remain relatively stable across months from January to June 2021, indicating minimal temporal effects within this period. In contrast, flat type displays a clear upward trend, where larger units (4-room, 5-room, Executive, and Multi-Generation) command higher resale prices, consistent with the earlier positive association between floor area and price. The flat model variable also shows notable price variation across architectural designs, older models such as Improved or Standard have lower median prices, while premium designs like Maisonette and Terrace achieve higher valuations, capturing structural differences not explained by size alone. Finally, region exhibits the expected location gradient: flats in the Core Central Region (CCR) are most expensive, followed by the Rest of Central Region (RCR) and Outside Central Region (OCR). Overall, flat type, flat model, and region emerge as key categorical predictors of resale price, while month plays a minor role during this six-month window.

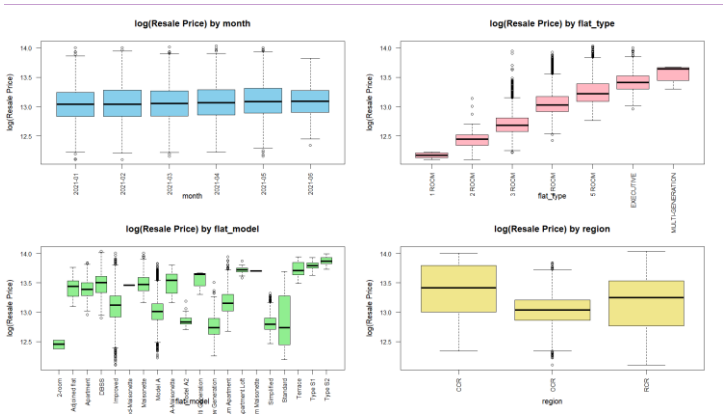
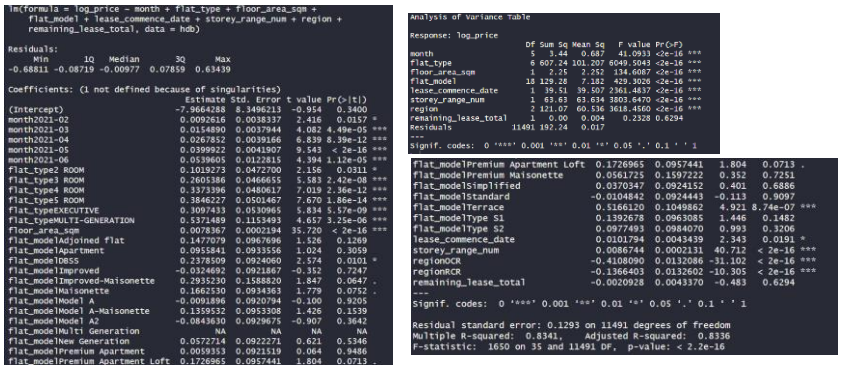


Figure 1.3

Some predictors in the dataset are likely interrelated. For example, floor area correlates with flat type, as larger flats typically have more rooms, while remaining lease is closely tied to lease commencement date, since newer flats naturally retain longer leases. These relationships suggest potential multicollinearity, where overlapping information among predictors may distort coefficient estimates. This issue will be formally assessed using Variance Inflation Factors (VIF) to identify and mitigate redundancy, ensuring a more stable and parsimonious model. Additionally, model selection techniques such as AIC, BIC, and stepwise regression will be employed to determine an optimal subset of variables that balances goodness of fit with model simplicity, ensuring that the final model is both parsimonious and statistically robust.

Part II Building Model:



```
lm(formula = log_price ~ month + flat_type + floor_area_sqm +
    flat_model + lease_commence_date + storey_range_num + region +
    remaining_lease_total, data = hdb)

Residuals:
      Min        1Q    Median        3Q       Max
-0.68811 -0.08719 -0.00977  0.07859  0.63439

Coefficients: (1 not defined because of singularities)
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -7.9664288  8.3496213  -0.954   0.3400
month2021-02                    0.0092616  0.0038337   2.416   0.0157 *
month2021-03                    0.0154890  0.0037944   4.082 4.49e-05 ***
month2021-04                    0.0267852  0.0039166   6.839 8.39e-12 ***
month2021-05                    0.0399922  0.0041907   9.543  < 2e-16 ***
month2021-06                    0.0539605  0.0122815   4.394 1.12e-05 ***
flat_type2 ROOM                 0.1019273  0.0472700   2.156   0.0311 *
flat_type3 ROOM                 0.2605386  0.0466655   5.583 2.42e-08 ***
flat_type4 ROOM                 0.3373396  0.0480617   7.019 2.36e-12 ***
flat_type5 ROOM                 0.3846227  0.0501467   7.670 1.86e-14 ***
flat_typeEXECUTIVE              0.3097433  0.0530965   5.834 5.57e-09 ***
flat_typeMULTI-GENERATION       0.5371489  0.1153493   4.657 3.25e-06 ***
floor_area_sqm                  0.0078367  0.0002194  35.720  < 2e-16 ***
flat_modelAdjoined flat         0.1477079  0.0967696   1.526   0.1269
flat_modelApartment             0.0955841  0.0933556   1.024   0.3059
flat_modelDBSS                  0.2378509  0.0924060   2.574   0.0101 *
flat_modelImproved              -0.0324692  0.0921887  -0.352   0.7247
flat_modelImproved-Maisonette   0.2935230  0.1588820   1.847   0.0647 .
flat_modelMaisonette            0.1662530  0.0934363   1.779   0.0752 .
flat_modelModel A               -0.0091896  0.0920794  -0.100   0.9205
flat_modelModel A-Maisonette    0.1359532  0.0953308   1.426   0.1539
flat_modelModel A2              -0.0843630  0.0929675  -0.907   0.3642
flat_modelMulti Generation              NA         NA      NA       NA
flat_modelNew Generation        0.0572714  0.0922271   0.621   0.5346
flat_modelPremium Apartment     0.0059353  0.0921519   0.064   0.9486
flat_modelPremium Apartment Loft  0.1726965  0.0957441   1.804   0.0713 .
```

```
Analysis of Variance Table

Response: log_price
                     Df Sum Sq Mean Sq   F value Pr(>F)
month                 5   3.44   0.687   41.0933 <2e-16 ***
flat_type             6 607.24 101.207 6049.5043 <2e-16 ***
floor_area_sqm        1   2.25   2.252  134.6087 <2e-16 ***
flat_model           18 129.28   7.182  429.3026 <2e-16 ***
lease_commence_date   1  39.51  39.507 2361.4837 <2e-16 ***
storey_range_num      1  63.63  63.634 3803.6470 <2e-16 ***
region                2 121.07  60.536 3618.4560 <2e-16 ***
remaining_lease_total 1   0.00   0.004    0.2328 0.6294
Residuals         11491 192.24   0.017
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 2.1

Initial Model (M_0): Figure 2.1 shows that the initial multiple linear regression model (M_0) was fitted using all major categorical and quantitative predictors. M_0 explains a high proportion of the variance in log(resale price) ($R^2 = 0.8341$, Adjusted $R^2 = 0.8336$), with the overall F-statistic ($p < 2.2e\text{-}16$) confirming strong model significance. Among predictors, floor area and storey range are highly significant ($p < 0.001$), indicating that larger flats and higher floors are associated with higher resale prices, while lease commencement date also has a positive effect ($p = 0.0191$). In contrast, several levels of flat model show large p-values , suggesting redundancy or possible overlap with flat type and floor area. The ANOVA results reinforce that flat type, region, and storey range explain substantial variance, whereas remaining lease total contributes little, as its effect is already captured by lease commence date. Overall, M_0 demonstrates strong explanatory power but exhibits signs of over-parameterization and multicollinearity, motivating refinement through model simplification and diagnostic analysis.

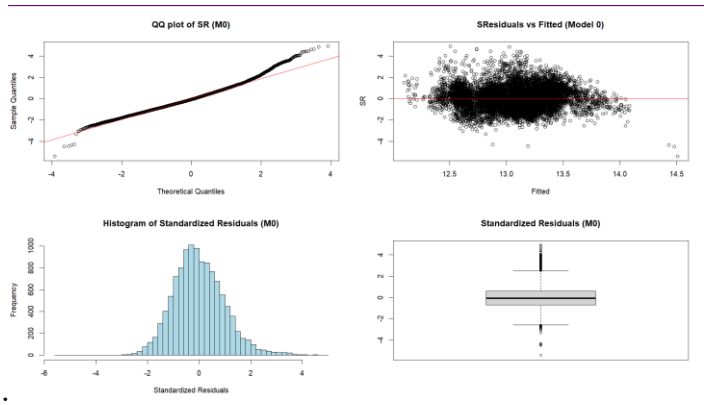


Figure 2.2

Model Adequacy Check: Model adequacy for M_0 was evaluated using diagnostic plots of standardized residuals as shown in Figure 2.2. The QQ plot and histogram indicate approximate

normality, with only slight tail deviations likely caused by a few outliers (≈203 cases out of 11,527) as shown in the boxplot, which have minimal impact on model fit. The residuals-versus-fitted y-values plot also show generally linear trend, suggesting the linearity assumption is reasonably satisfied. However, the residuals-versus-fitted y-values plot displays a clear funnel shape, indicating violation of the constant variance assumption. Overall, while linearity and normality hold adequately, the non-constant variance suggests the need for transformation or a Weighted Least Squares (WLS) approach to improve model adequacy.

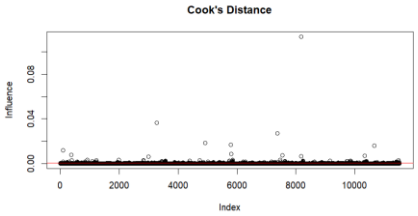


Figure 2.3

Check for Influential Points: In Figure 2.3 shown, the Cook's Distance plot was used to detect influential observations that might disproportionately affect the model. Almost all points have values near zero, and none exceed the conventional threshold of 1, indicating no influential observations. Although a few data points show slightly higher values, they do not significantly impact the regression results. Thus, while M_0 contains minor outliers, there is no evidence of undue influence compromising model stability or reliability.

```
> vif_vals
                          GVIF Df GVIF^(1/(2*Df))
month                 1.217155  5        1.019846
flat_type           268.960155  5        1.749721
floor_area_sqm       18.654808  1        4.319121
flat_model          115.838962 18        1.141115
lease_commence_date 2691.134285  1       51.876144
storey_range_num       1.238637  1        1.112941
region                 1.878900  2        1.170782
remaining_lease_total 2695.136244  1       51.914702
```

Figure 2.4

Check for Multicollinearity using VIF: As shown in Figure 2.4, to evaluate multicollinearity, the Variance Inflation Factor (VIF) was computed for all predictors. Most variables showed low VIF values (<5), indicating minimal collinearity. However, lease commence date and remaining lease total displayed extremely high adjusted GVIFs (≈52), reflecting severe redundancy since both capture similar lease-age information. Consequently, remaining lease total was removed through backward stepwise regression based on the Akaike Information Criterion (AIC), resulting in a refined model (M_1) with a slightly lower AIC and marginally improved adjusted R^2. This confirms that eliminating redundant variables enhances parsimony without sacrificing explanatory power.

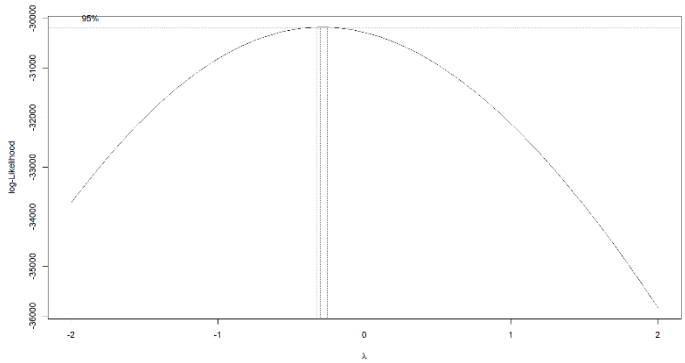


Figure 2.5

```
> AIC(M2, M3, M4)
   df       AIC
M2 35 -14528.99
M3 35 -14514.61
M4 36 -14280.84
```

Figure 2.6

Transformations: To confirm the suitability of transforming the response variable, a Box–Cox transformation was applied with resale price as the dependent variable. The resulting plot showed an optimal λ near zero, with the 95% confidence interval tightly surrounding λ = 0, validating the use of log(resale price) as the response (Figure 2.5). This transformation effectively stabilized variance and improved residual normality. To further address mild non-linearity and heteroscedasticity, floor area per square metre and lease commence date were also log-transformed, producing Model M_2, which achieved an improved fit (AIC = –14528.99) (Figure 2.6). A Weighted Least Squares (WLS) model (M_3) was then fitted using weights inversely proportional to the squared fitted values from M_2, reducing variance non-constancy though yielding a slightly higher AIC (–14514.61) (Figure 2.6). A subsequent Generalized WLS model (M_4) with a varPower variance structure produced a much higher AIC (–14280.84) (Figure 2.6), indicating that its added complexity did not enhance model performance. Hence, M_4 was rejected due to over-parameterization and inferior fit.

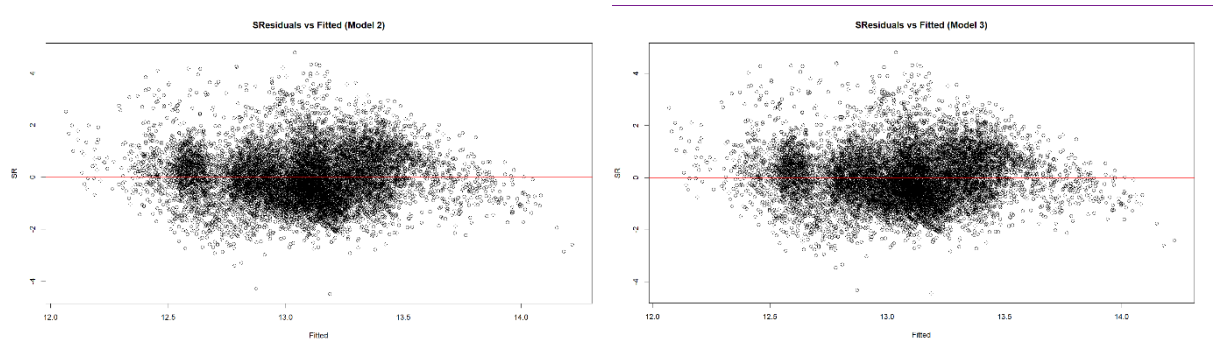


Figure 2.7

Check for Correction of Inadequacy: As shown in Figure 2.7, residual plots for Model 2 (OLS) and Model 3 (WLS) were compared to assess the constant-variance assumption. In M_2, standardized residuals show a mild funnel pattern, indicating slight heteroscedasticity despite the earlier log transformations. Applying inverse-squared fitted value weights in M_3 produced a more uniform residual spread, substantially reducing this pattern and satisfying the homoscedasticity assumption. Although M_3 has a slightly higher AIC (–14514.61 vs –14528.99), its improved variance stability makes it statistically more adequate overall.

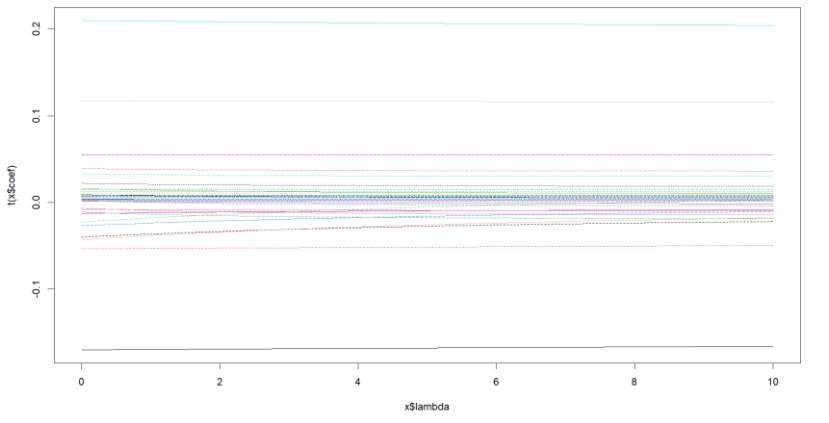
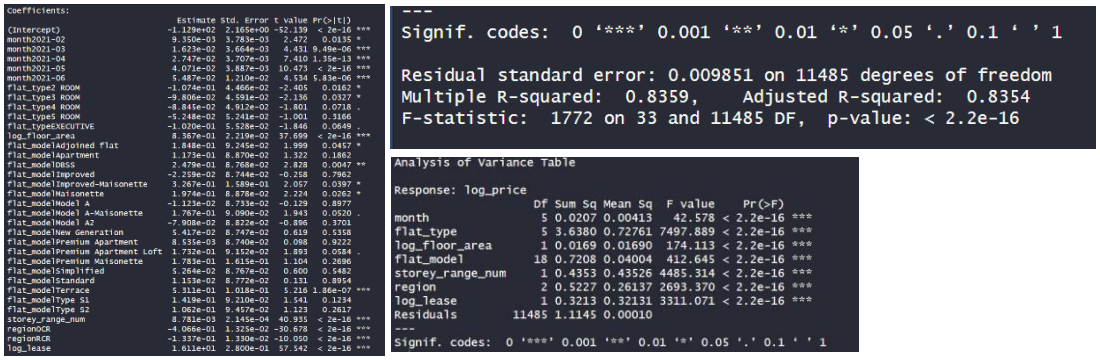


Figure 2.8

Check for Multicollinearity using Ridge Regression: To evaluate multicollinearity in the final weighted model (M_3), ridge regression diagnostics were performed. The ridge trace plot shows

stable, nearly flat coefficient paths across penalty values (λ = 0–10), with no major fluctuations or crossovers (Figure 2.8). This indicates that the coefficient estimates are stable and multicollinearity is minimal within M_3. Hence, since M_3 already achieves low collinearity, regularization was unnecessary.

Chosen Model (M_3):



```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -1.128e+02  2.165e+00 -52.139  < 2e-16 ***
month2021-02                     9.350e-03  3.783e-03   2.472   0.0135 *
month2021-03                     1.623e-02  3.664e-03   4.431 9.49e-06 ***
month2021-04                     2.747e-02  3.707e-03   7.410 1.35e-13 ***
month2021-05                     4.071e-02  3.887e-03  10.473  < 2e-16 ***
month2021-06                     5.487e-02  1.210e-02   4.534 5.83e-06 ***
flat_type2 ROOM                 -1.074e-01  4.466e-02  -2.405   0.0162 *
flat_type3 ROOM                 -9.806e-02  4.591e-02  -2.136   0.0327 *
flat_type4 ROOM                 -8.845e-02  4.912e-02  -1.801   0.0718 .
flat_type5 ROOM                 -5.248e-02  5.241e-02  -1.001   0.3166
flat_typeEXECUTIVE              -1.020e-01  5.528e-02  -1.846   0.0649 .
log_floor_area                   8.367e-01  2.219e-02  37.699  < 2e-16 ***
flat_modelAdjoined Flat          1.848e-01  9.245e-02   1.999   0.0457 *
flat_modelApartment              1.175e-01  8.870e-02   1.322   0.1862
flat_modelDBSS                   2.479e-01  8.768e-02   2.828   0.0047 **
flat_modelImproved              -2.259e-02  8.744e-02  -0.258   0.7962
flat_modelImproved-Maisonette    1.267e-01  1.589e-01   2.057   0.0397 *
flat_modelMaisonette             1.974e-01  8.878e-02   2.224   0.0262 *
flat_modelModel A               -1.123e-02  8.733e-02  -0.129   0.8977
flat_modelModel A-Maisonette     1.767e-01  9.090e-02   1.945   0.0520 .
flat_modelModel A2              -7.908e-02  8.822e-02  -0.896   0.3701
flat_modelNew Generation         5.417e-02  8.747e-02   0.619   0.5358
flat_modelPremium Apartment      8.535e-03  8.740e-02   0.098   0.9222
flat_modelPremium Apartment Loft 1.732e-01  9.152e-02   1.893   0.0584 .
flat_modelPremium Maisonette     1.783e-01  1.615e-01   1.104   0.2696
flat_modelSimplified             5.264e-02  8.767e-02   0.600   0.5482
flat_modelStandard               3.153e-02  8.772e-02   0.131   0.8954
flat_modelTerrace                5.311e-01  1.018e-01   5.216 1.86e-07 ***
flat_modelType S1                1.419e-01  9.210e-02   1.541   0.1234
flat_modelType S2                1.062e-01  9.457e-02   1.123   0.2617
storey_range_num                 8.781e-03  2.145e-04  40.935  < 2e-16 ***
regionOCR                       -4.066e-01  1.325e-02 -30.678  < 2e-16 ***
regionRCR                       -1.337e-01  1.330e-02 -10.050  < 2e-16 ***
log_lease                        1.611e+01  2.800e-01  57.542  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.009851 on 11485 degrees of freedom
Multiple R-squared:  0.8359,    Adjusted R-squared:  0.8354
F-statistic:  1772 on 33 and 11485 DF,  p-value: < 2.2e-16

Analysis of Variance Table

Response: log_price
                  Df Sum Sq Mean Sq  F value    Pr(>F)
month              5 0.0207 0.00413   42.578 < 2.2e-16 ***
flat_type          5 3.6380 0.72761 7497.889 < 2.2e-16 ***
log_floor_area     1 0.0169 0.01690  174.113 < 2.2e-16 ***
flat_model        18 0.7208 0.04004  412.645 < 2.2e-16 ***
storey_range_num   1 0.4353 0.43526 4485.314 < 2.2e-16 ***
region             2 0.5227 0.26137 2693.370 < 2.2e-16 ***
log_lease          1 0.3213 0.32131 3311.071 < 2.2e-16 ***
Residuals      11485 1.1145 0.00010
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3

Conclusion: Model 3 (WLS) demonstrates the most adequate performance among all competing models. The application of weighted least squares effectively stabilized residual variance, as shown by the residual–fitted plot, which no longer exhibits the funnel pattern seen in M_2. Although its AIC (–14514.61) is slightly higher than M_2's (–14528.99) (Figure 2.6), this small trade-off is justified by superior compliance with the constant-variance assumption. The model also avoids over-parameterization, maintaining a balanced set of predictors that each contribute meaningfully to explaining variation in log(resale price). Statistically, M_3 exhibits strong explanatory power (R^2 = 0.8359, Adjusted R^2 = 0.8354), with the overall F-statistic (F = 1772, p < 2.2e–16) confirming high model significance (Figure 3). The residual standard error of 0.00985 further indicates excellent model precision. Overall, M_3 achieves both strong explanatory performance and full compliance with regression assumptions, linearity, normality, homoscedasticity, and low multicollinearity, making it the most statistically robust and practically interpretable model for predicting HDB resale prices.

In practical terms, the final model highlights how physical characteristics (flat size, storey height, design type) and locational attributes (region and lease recency) jointly shape HDB resale values in Singapore. These findings reinforce that newer, larger, and more centrally located flats command significant price premiums. However, the model is limited to transactions from the first half of 2021 and does not account for external factors such as proximity to amenities, renovations, or macroeconomic changes. Future analyses could incorporate these variables or extend the dataset across multiple years to enhance predictive generalizability.

Final Model Equation:

$$\widehat{\log(\text{Resale Price})} = \hat{\beta}_0 + \sum_{i=1}^{5} \hat{\beta}_{1i}(\text{Month}_i) + \sum_{j=1}^{6} \hat{\beta}_{2j}(\text{Flat Type}_j) + \hat{\beta}_3 \log(\text{Floor Area})$$

$$+ \sum_{k=1}^{18} \hat{\beta}_{4k}(\text{Flat Model}_k) + \hat{\beta}_5(\text{Storey Range}_{\text{num}}) + \sum_{m=1}^{2} \hat{\beta}_{6m}(\text{Region}_m) + \hat{\beta}_7 \log(\text{Lease})$$