

White Be Boundless tagline graphic­­­­­DOCUMENT APPROVAL SHEET

DATA SCIENCE PROJECT

MULTI-REGRESSION

gold boundless bar

**DOCUMENT IDENTIFICATION:**

MSBA AY 21-22/MULTI-REGRESSION LAB

**DOCUMENT AUTHOR:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **APPROVED BY/ ID** | **DEPT.** | **SCHOOL** | **SIGNATURE** | **DATE** |
| Teja Alluru (1150077) | MSBA/ Student | University of Washington, Tacoma, Milgard School of Business |  | 24-Oct-2021 |

*AUTHORED BY: Authored by mean of a person/entity responsible for the creation of the document.*

INDEX OF CONTENTS

gold boundless bar

[DOCUMENT APPROVAL SHEET 1](#_Toc86018516)

[INDEX OF CONTENTS 3](#_Toc86018517)

[INDEX OF FIGURES 4](#_Toc86018518)

[1 BUSSINESS UNDERSTANDING 5](#_Toc86018519)

[2 DATA GATHERING/PREPARATION 5](#_Toc86018520)

[3 PRELIMINARY DATA ANALYSIS – PLOTS & INTERPRETATION 6](#_Toc86018521)

[3.1 Statistics 6](#_Toc86018522)

[3.2 Visualizations & Insights 8](#_Toc86018523)

[3.2.1 Bivariate Scatter Plots 8](#_Toc86018524)

[3.2.2 Univariate Histograms 10](#_Toc86018525)

[4 REGRESSION ANALYSIS/ OUTPUT 12](#_Toc86018526)

[5 POST REGRESSION ANALYSIS 15](#_Toc86018527)

[6 PREDICTION 23](#_Toc86018528)

[7 CONCLUSION 24](#_Toc86018529)

INDEX OF FIGURES

gold boundless bar

[Figure 1‑1: Richest Neighbourhood – In/Around Seattle 5](#_Toc86018550)

[Figure 3‑1: Variable Statistics 6](#_Toc86018551)

[Figure 3‑2: Variable Statistics – Null Values Removed 7](#_Toc86018552)

[Figure 3‑3: Bivariate Analysis Visualizations Price Vs Other Variables 8](#_Toc86018553)

[Figure 3‑4: Bivariate Analysis Visualizations Price Vs Other Variables – With Trendline 9](#_Toc86018554)

[Figure 3‑5: Univariate Properties 10](#_Toc86018555)

[Figure 3‑6: Univariate Properties – With Log Transformation 11](#_Toc86018556)

[Figure 4‑1: Regression Output – All Variables Included 12](#_Toc86018557)

[Figure 4‑2: Regression Output – Only Significant Variables Included 13](#_Toc86018558)

[Figure 4‑3: Regression Output – Only Most Significant Variable Included 13](#_Toc86018559)

[Figure 4‑4: Regression Output – Only All Variable Log Transformation Included 14](#_Toc86018560)

[Figure 5‑1: Residual Plots – All Variable Included 15](#_Toc86018561)

[Figure 5‑2: Residual Plots – All Variable Included-2 16](#_Toc86018562)

[Figure 5‑3: Residual Plots – Only Significant Variables Included 17](#_Toc86018563)

[Figure 5‑4: Residual Plots – Only Significant Variables Included-2 18](#_Toc86018564)

[Figure 5‑5: Residual Plots – Only Most Significant Variable Included 19](#_Toc86018565)

[Figure 5‑6: Residual Plots – Only Most Significant Variable Included-2 20](#_Toc86018566)

[Figure 5‑7: Residual Plots – All Variable Log Transformation Included 21](#_Toc86018567)

[Figure 5‑8: Residual Plots – All Variable Log Transformation Included-2 22](#_Toc86018568)

[Figure 6‑1: Predicted Vs Actual Values 23](#_Toc86018569)

# BUSINESS UNDERSTANDING

The business problem, I have chosen for the project is to do a multi regression analysis to identify the most important features that effect the pricing of houses in the richest district in and around Seattle. Here for our business problem, we will be focusing on determining housing prices by its characteristics like size, bedroom, baths, and other features of housing. To summarize we would like to understand whether the rich are interested more in bedrooms, baths, or size or anything else. The neighborhood chosen is shown in Figure 1‑1.

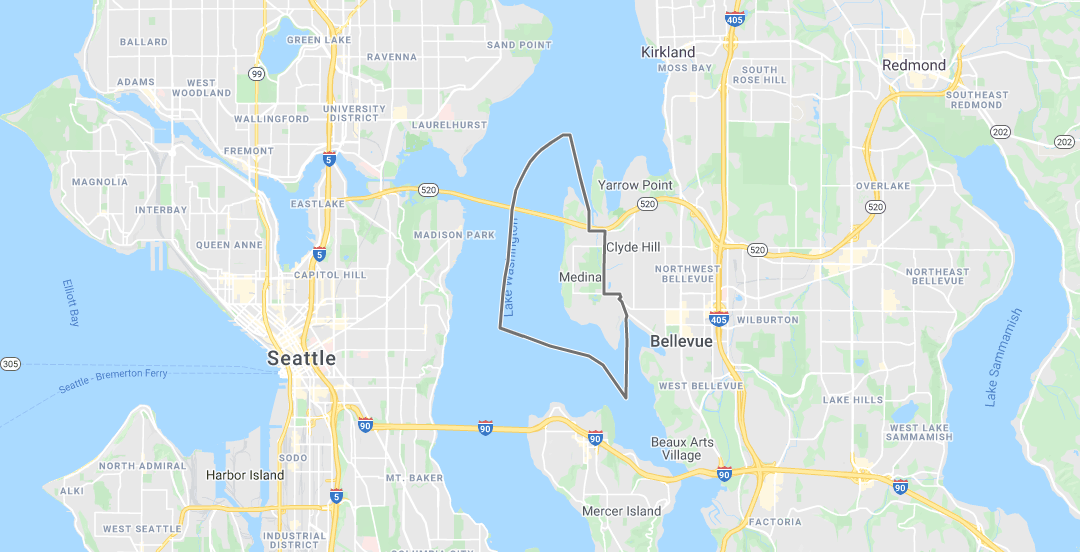


Figure 1‑1: Richest Neighbourhood – In/Around Seattle

# DATA GATHERING/PREPARATION

The data for our business problem is downloaded from refin.com for the neighborhood 98039, which includes Medina and lakeside where the richest people like Bill Gates & Jeff Bezos live. We have collected the data for the last five years. A lot of attributes were collected and downloaded as “Billionaire-Suburb-980039.csv” from redfin.com. To streamline our business problem, we have retained PRICE, BEDS, BATHS, SQUARE FEET, LOT SIZE, and YEAR BUILT features only and the rest of the features were deleted. The .csv file was then converted to .xlsx format. The statistics, visualizations, univariate & Bivariate analysis of the variables are shown in section 3.

# PRELIMINARY DATA ANALYSIS – PLOTS & INTERPRETATION

## Statistics

The variable statistics from our dataset are presented in Figure 3‑1 & Figure 3‑2 below. In Figure 3‑1 the null values for variables are not removed while in Figure 3‑2 the null values for the variables are removed. From the figure, we can see that the median is around 3.3 mil & the mean is 4.7 mil. We need to also watch for multi-billionaire homes which are quite evident from the fact that the difference between the 3rd quartile of the price and the max price is quite huge. This trend of the big difference between the 3rd quartile and max does provide insights that there is some house built by ultra-rich people that might tweak the entire neighborhood analysis data.

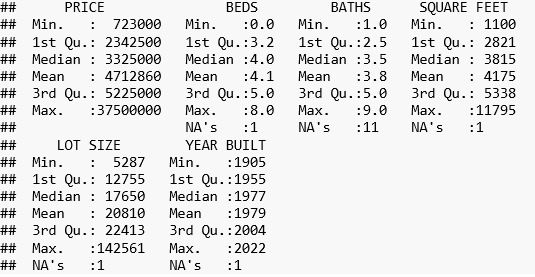


Figure 3‑1: Variable Statistics

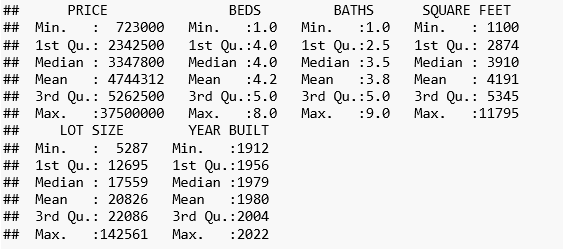


Figure 3‑2: Variable Statistics – Null Values Removed

## Visualizations & Insights

### Bivariate Scatter Plots

Figure 3‑3 gives the bivariate properties of the variables vs the price variable & Figure 3‑4 gives the same properties with a trendline. As you can observe from both the figures there are a few values that are quite further away which are influencing the price/data a lot. This is because the multi-billionaires’ homes are skewing the data a lot.

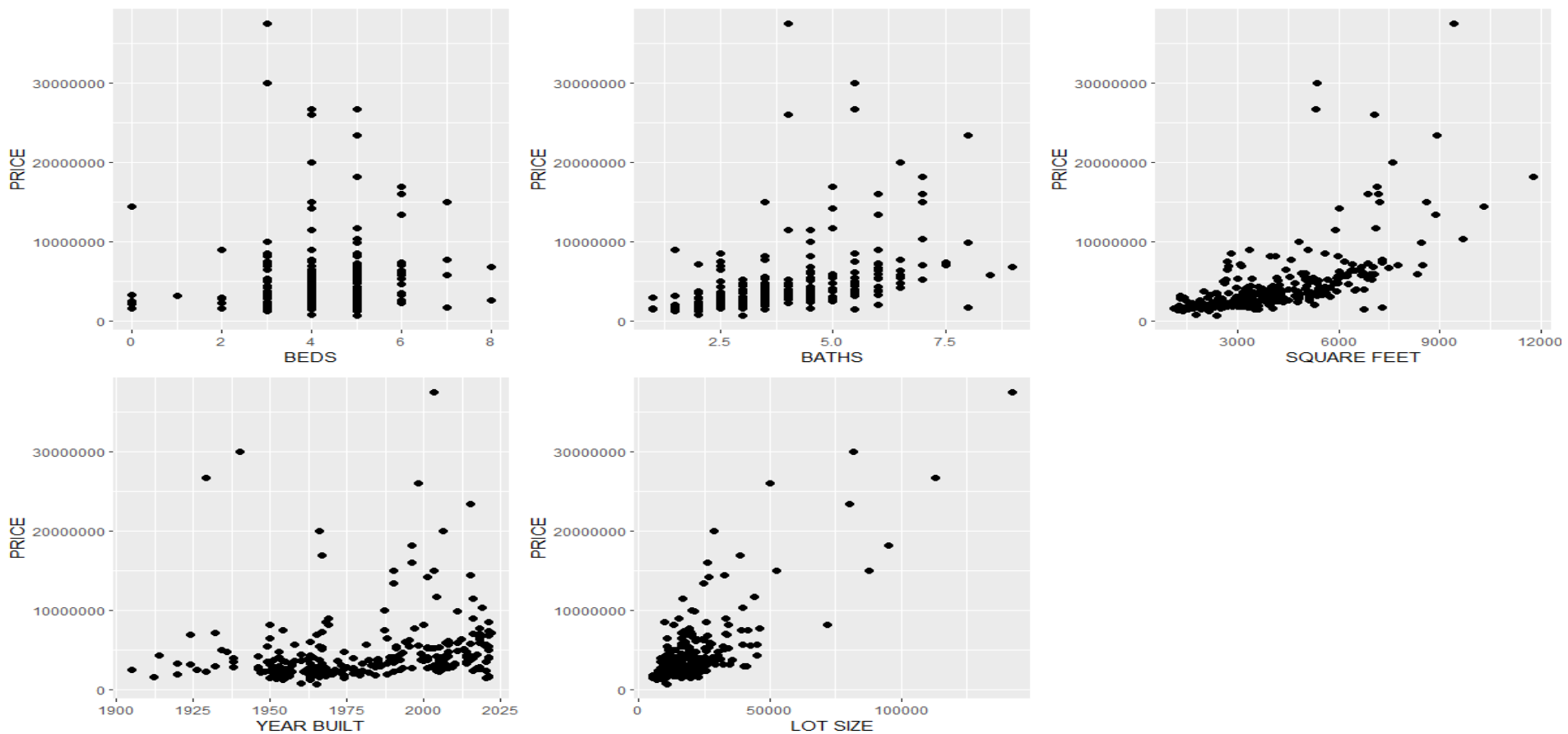


Figure 3‑3: Bivariate Analysis Visualizations Price Vs Other Variables

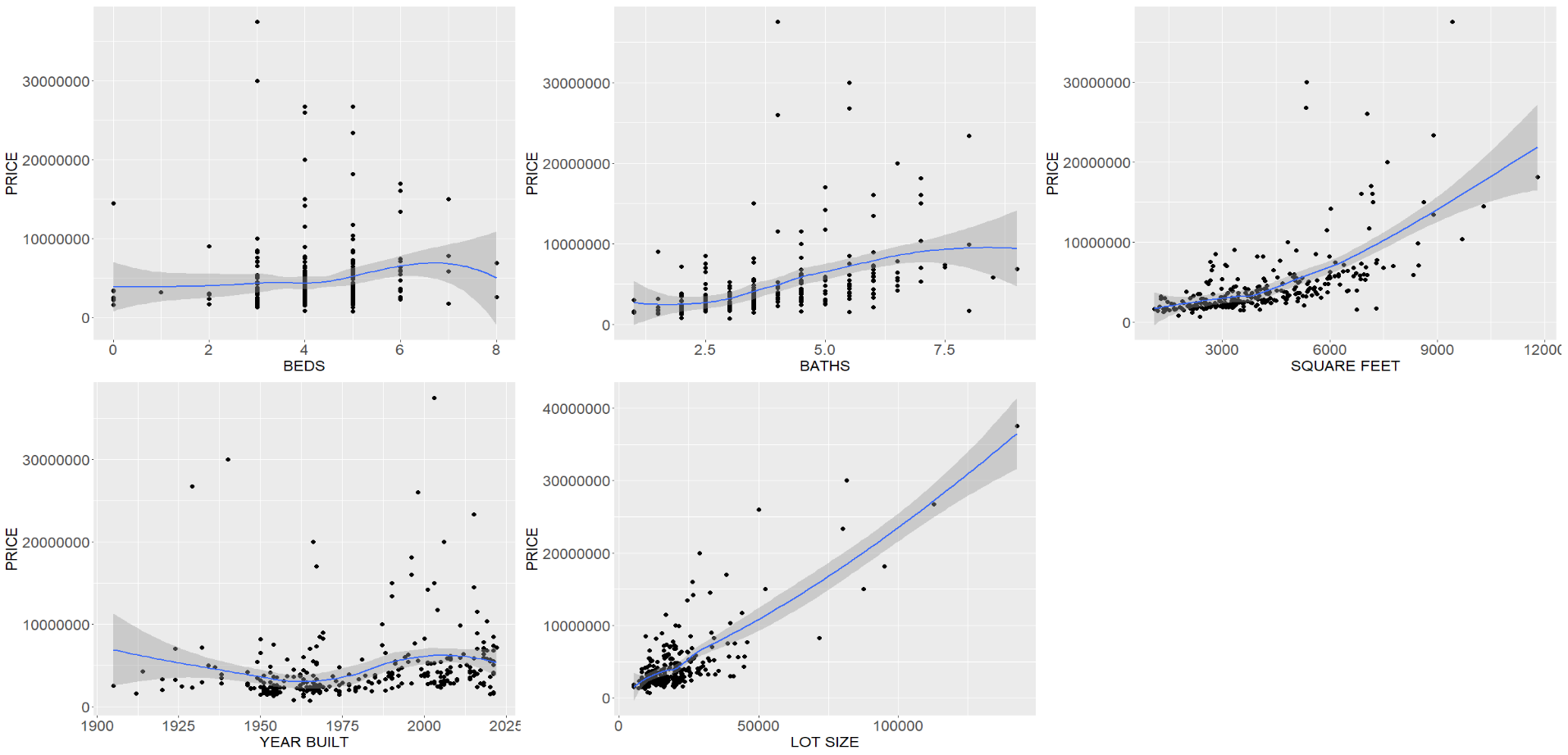


Figure 3‑4: Bivariate Analysis Visualizations Price Vs Other Variables – With Trendline

### Univariate Histograms

Figure 3‑5 gives the univariate properties of the variable’s & Figure 3‑6 gives the same properties with log transformation applied to the variables. As you can observe from both the figures there are a few values that are quite further away which are skewing the data. The log transformation has improved the distribution a bit, but the lot size still seems to be skewed tough better than before log transformation was applied.

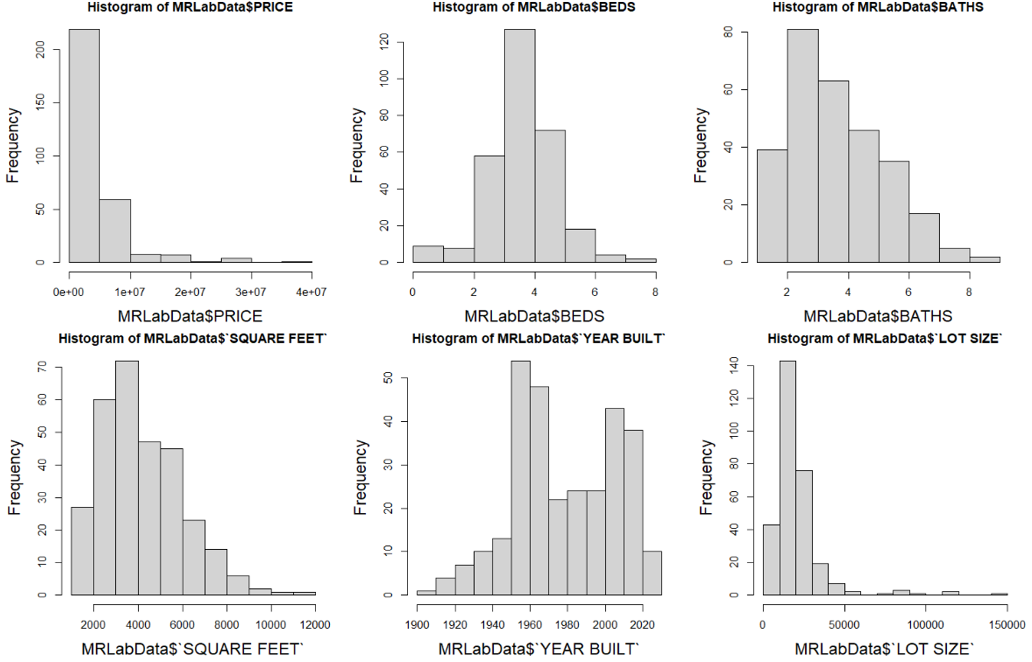


Figure 3‑5: Univariate Properties

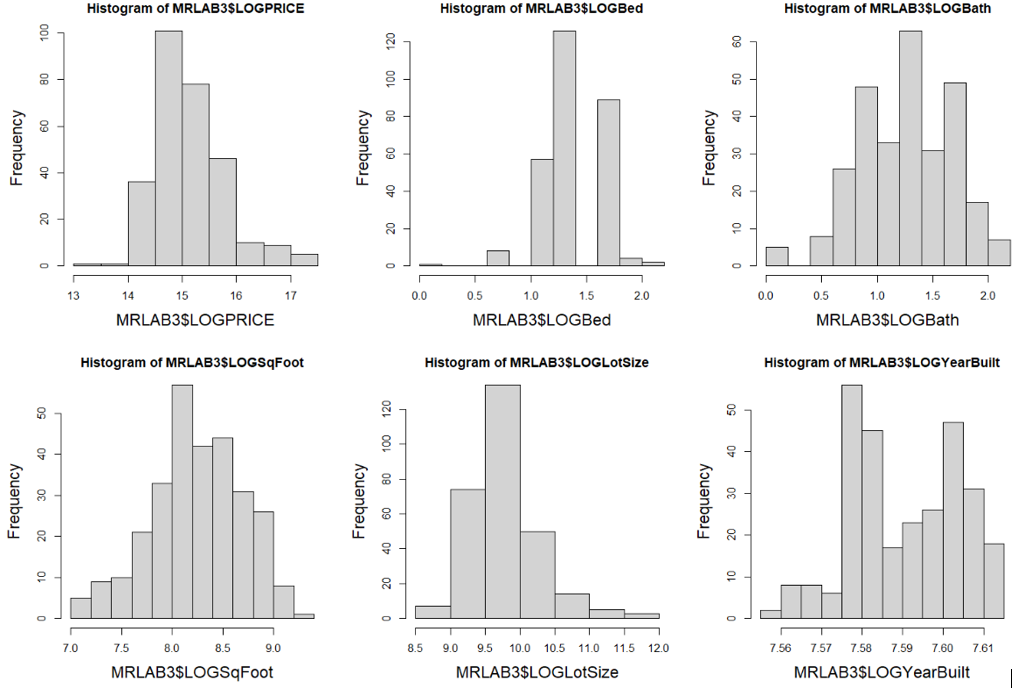


Figure 3‑6: Univariate Properties – With Log Transformation

# REGRESSION ANALYSIS/ OUTPUT

The initial linear model was applied with all variables included and the results are shown in Figure 4‑1. The most significant variables are found to be Beds, LotSize and Square Feet. The R-squared value is 0.703 which is quite high. Next, the model is run with only the significant variables and that has improved the R-Squared to 0.705. The results for that are presented in Figure 4‑2. Finally the model is run with only the most significant variable as such the R-squared value decreased to 0.62 presented in Figure 4‑3. TO check if the huge values are making any changes to the model, log transformation was applied to all variables and the model was run with all variables with log transformation which resulted in an R-Squared value of 0.649 presented in Figure 4‑4. From all these models the best fit was for the model with only significant variables included which has an R-squared value of 0.705.

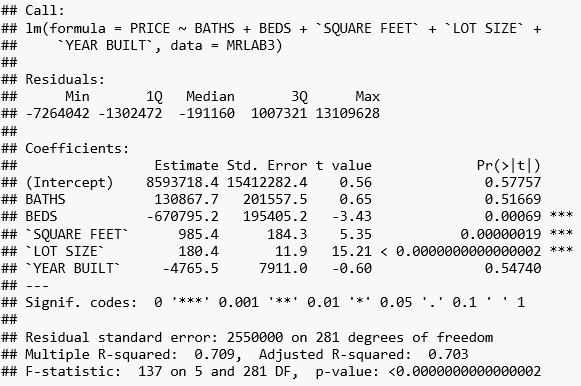


Figure 4‑1: Regression Output – All Variables Included

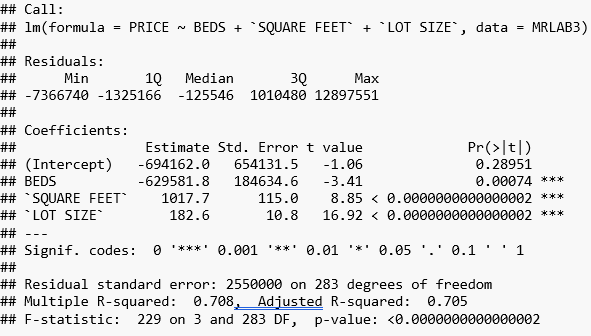


Figure 4‑2: Regression Output – Only Significant Variables Included

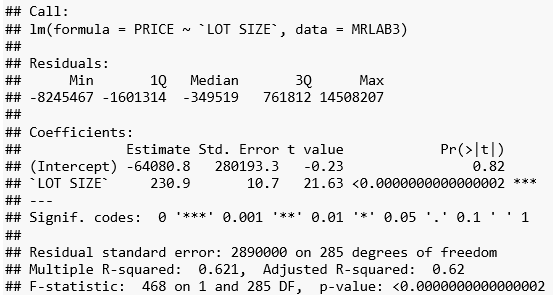


Figure 4‑3: Regression Output – Only Most Significant Variable Included

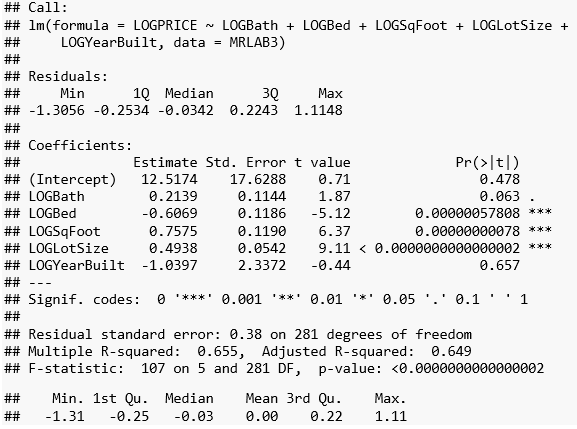


Figure 4‑4: Regression Output – Only All Variable Log Transformation Included

# POST REGRESSION ANALYSIS

The initial linear model was applied with all variables included and the residuals are shown in Figure 5‑1 and Figure 5‑2. The most significant variables are found to be Beds, LotSize and Square Feet. The R-squared value is 0.703 which is quite high. Next, the model is run with only the significant variables and that has improved the R-Squared to 0.705. The residuals for that are presented in Figure 5‑3 and Figure 5‑4. Finally the model is run with only the most significant variable as such the R-squared value decreased to 0.62 and the residual values are presented in Figure 5‑5 and Figure 5‑6. T0 check if the huge values are making any changes to the model, log transformation was applied to all variables and the model was run with all variables with log transformation which resulted in an R-Squared value of 0.649. The residuals for the log transformation model are presented in Figure 5‑7 and Figure 5‑8. AS we can see from graphs some of the values are very far and seems to influence the results a lot.

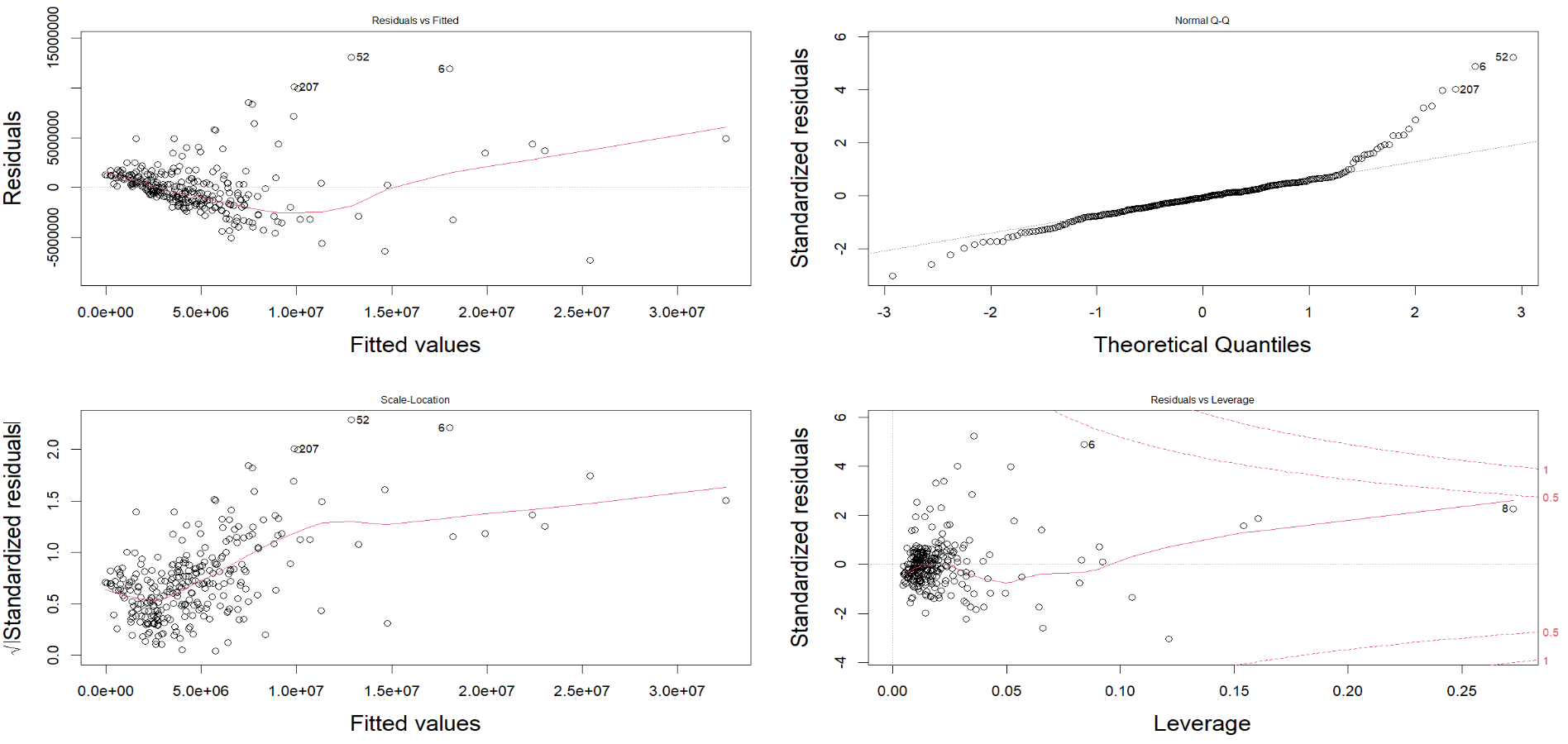


Figure 5‑1: Residual Plots – All Variable Included

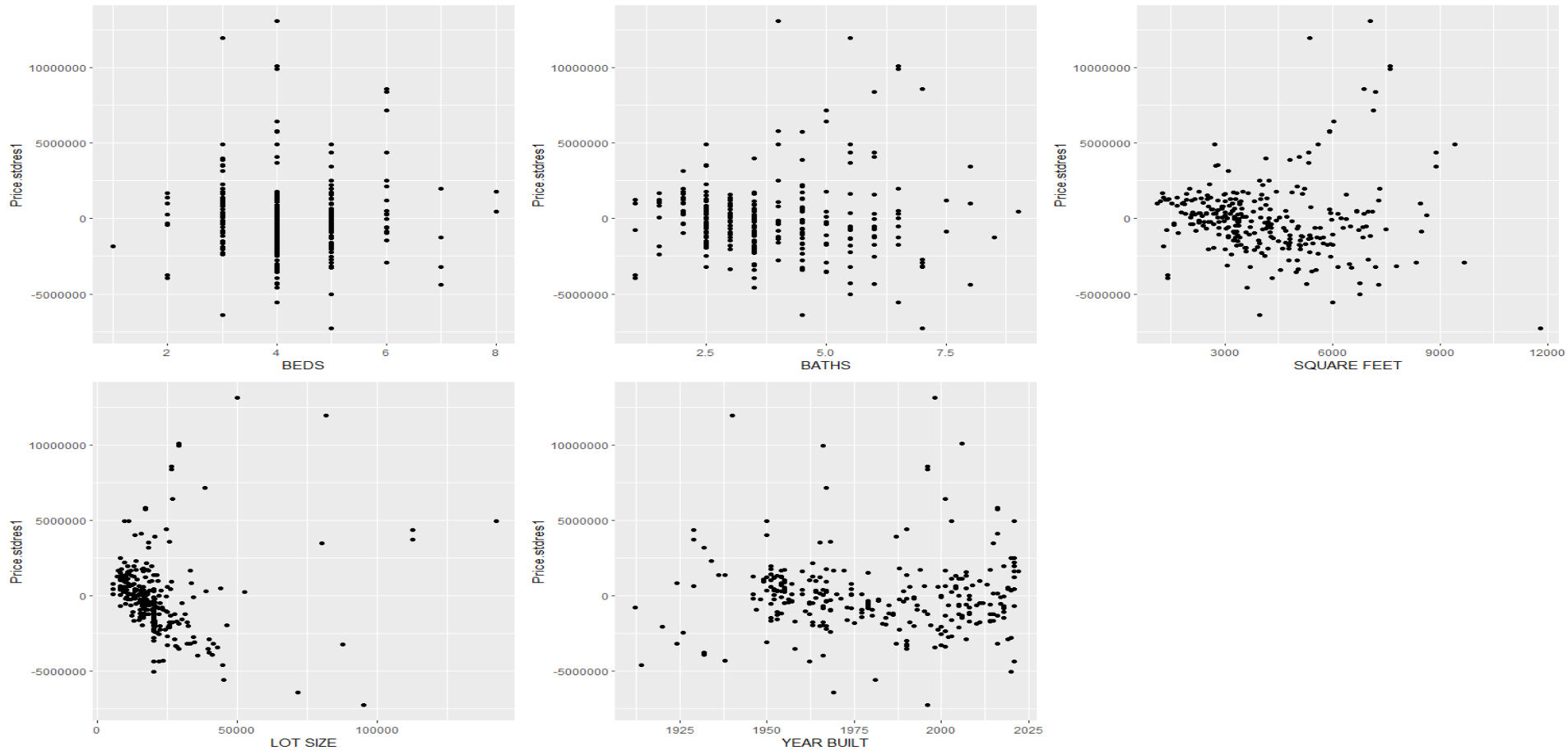


Figure 5‑2: Residual Plots – All Variable Included-2

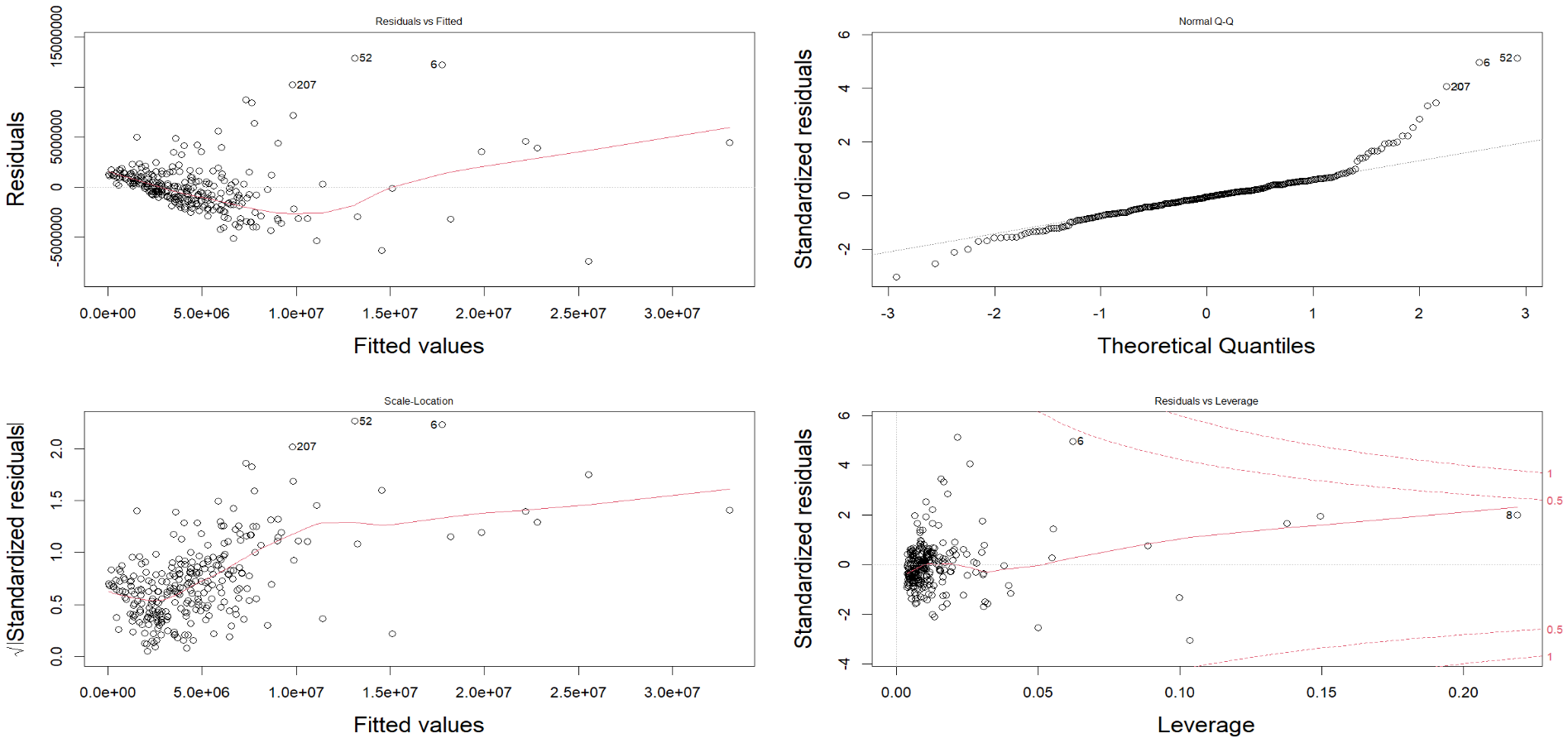


Figure 5‑3: Residual Plots – Only Significant Variables Included

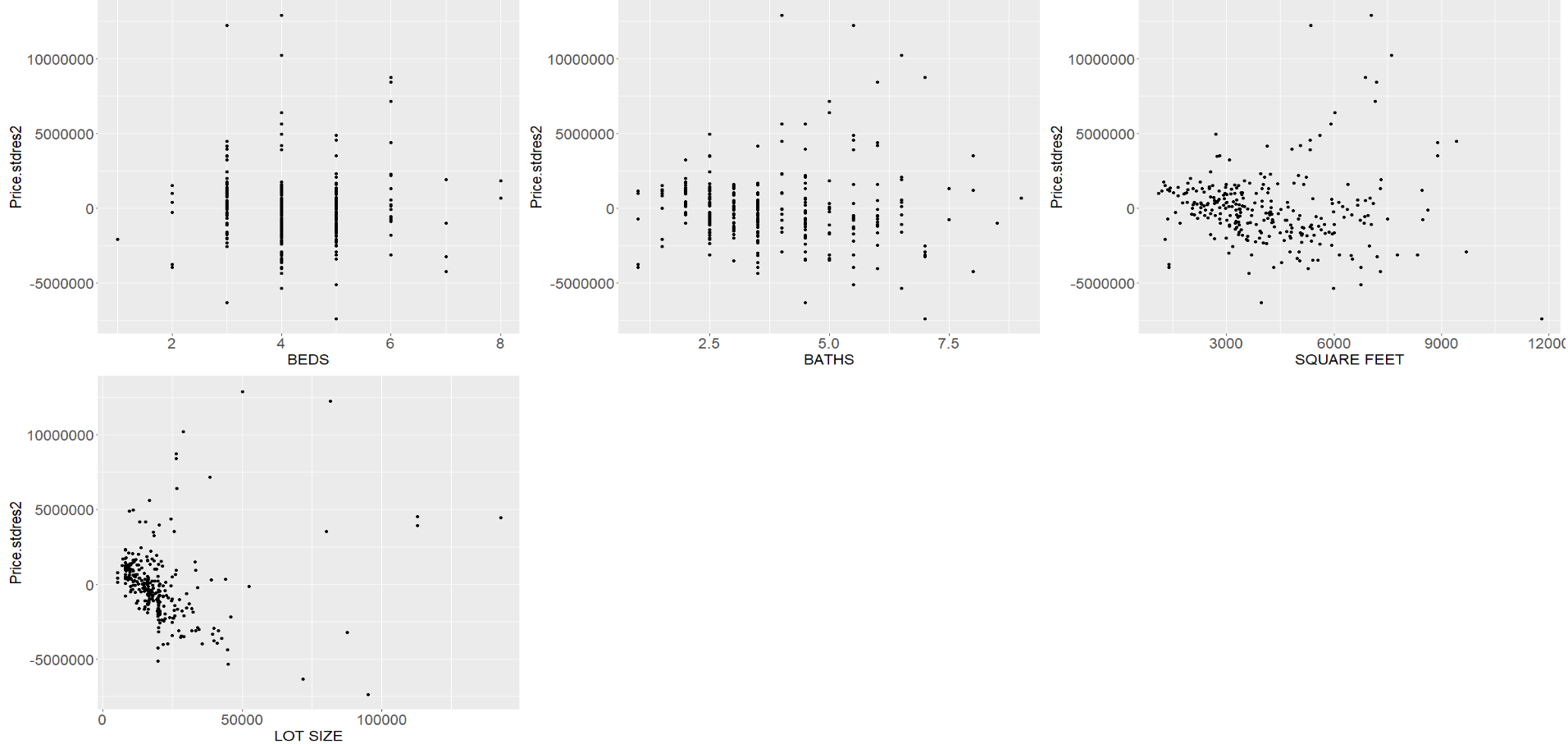


Figure 5‑4: Residual Plots – Only Significant Variables Included-2

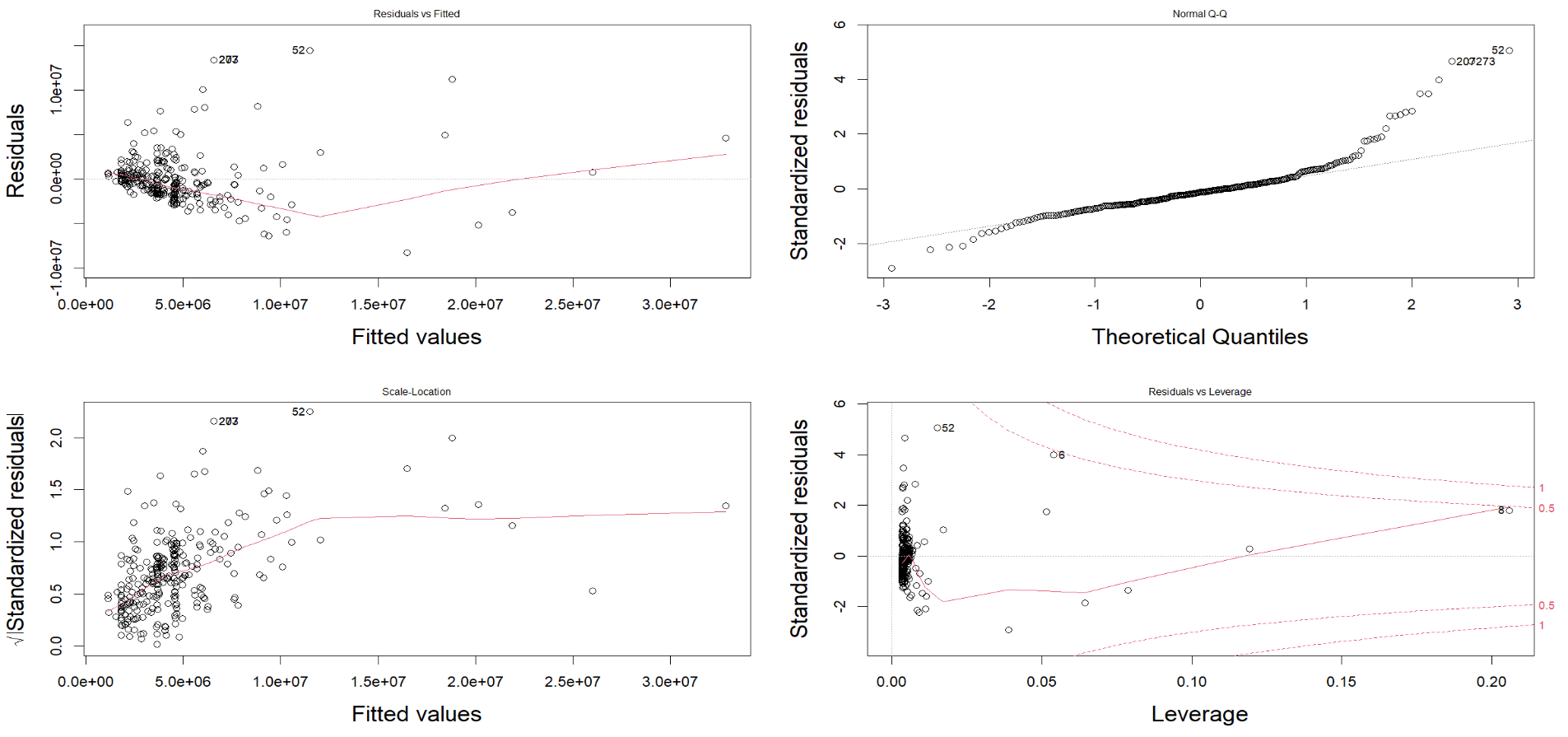


Figure 5‑5: Residual Plots – Only Most Significant Variable Included

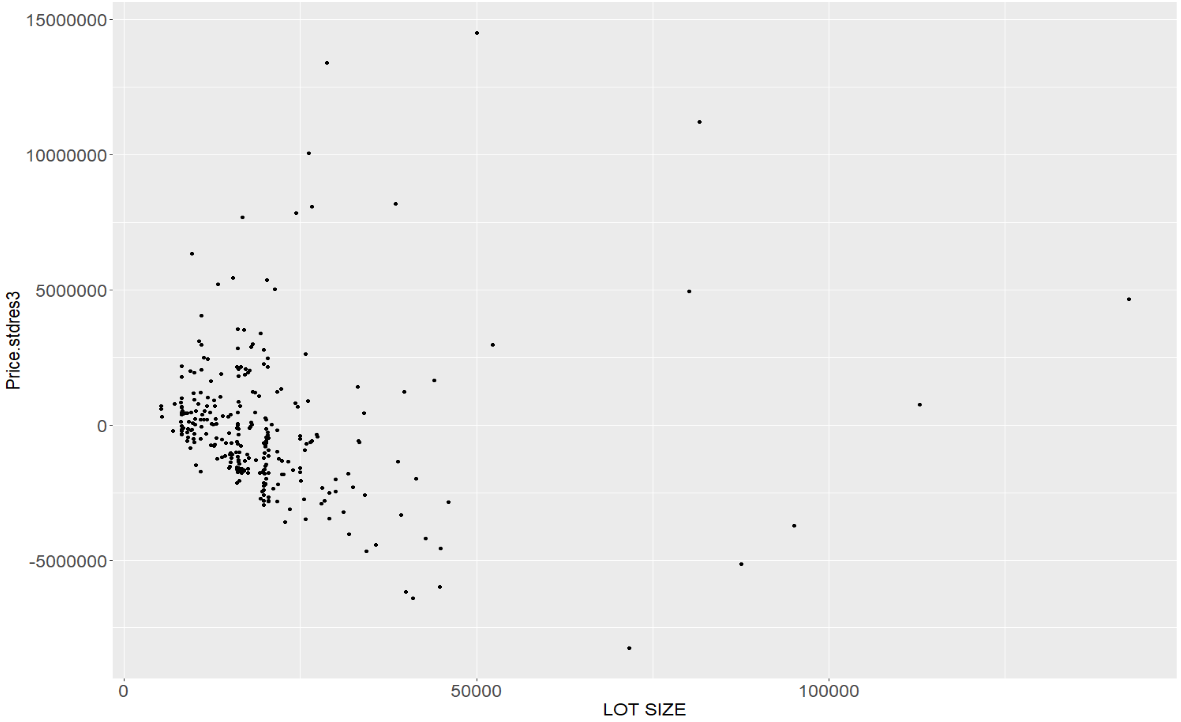


Figure 5‑6: Residual Plots – Only Most Significant Variable Included-2

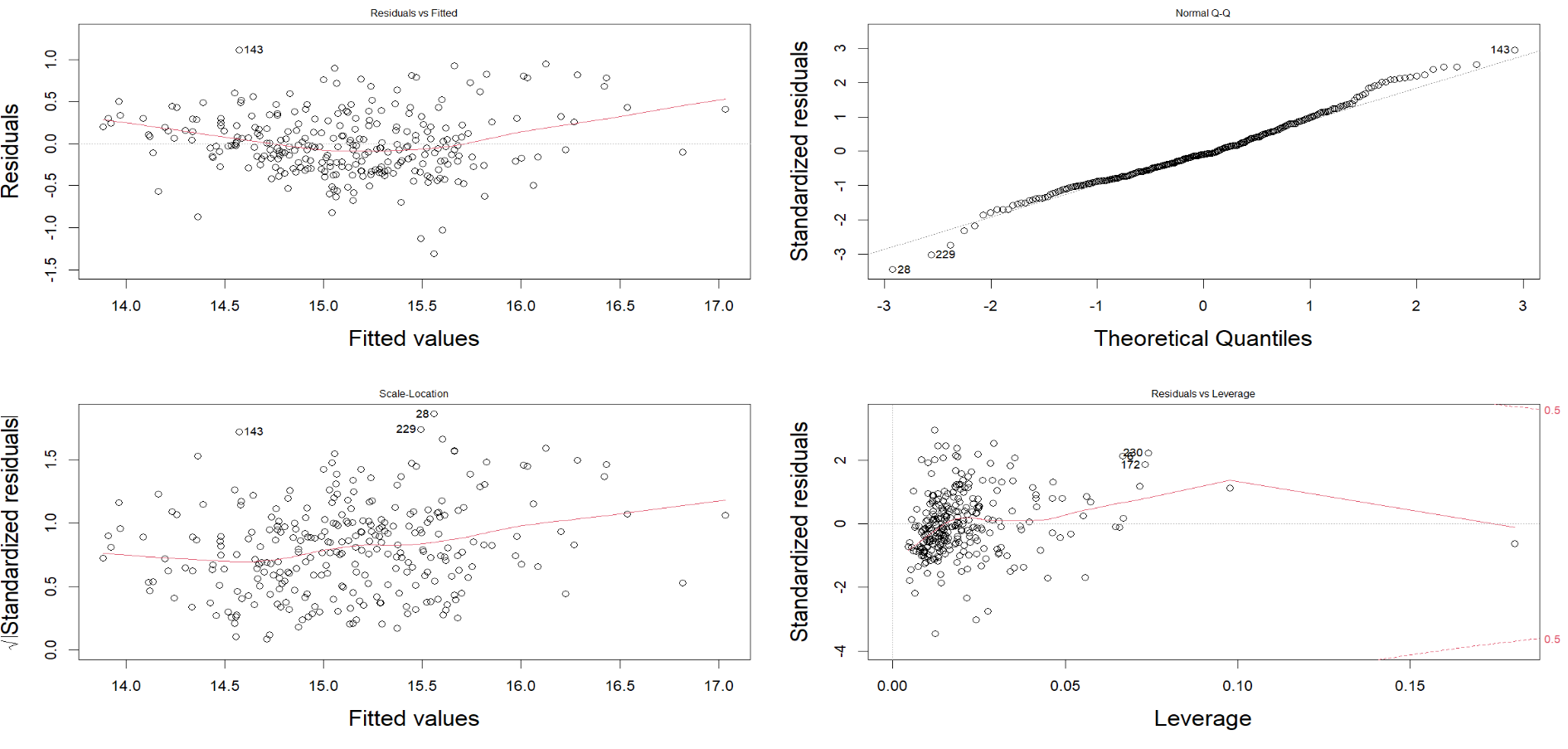


Figure 5‑7: Residual Plots – All Variable Log Transformation Included

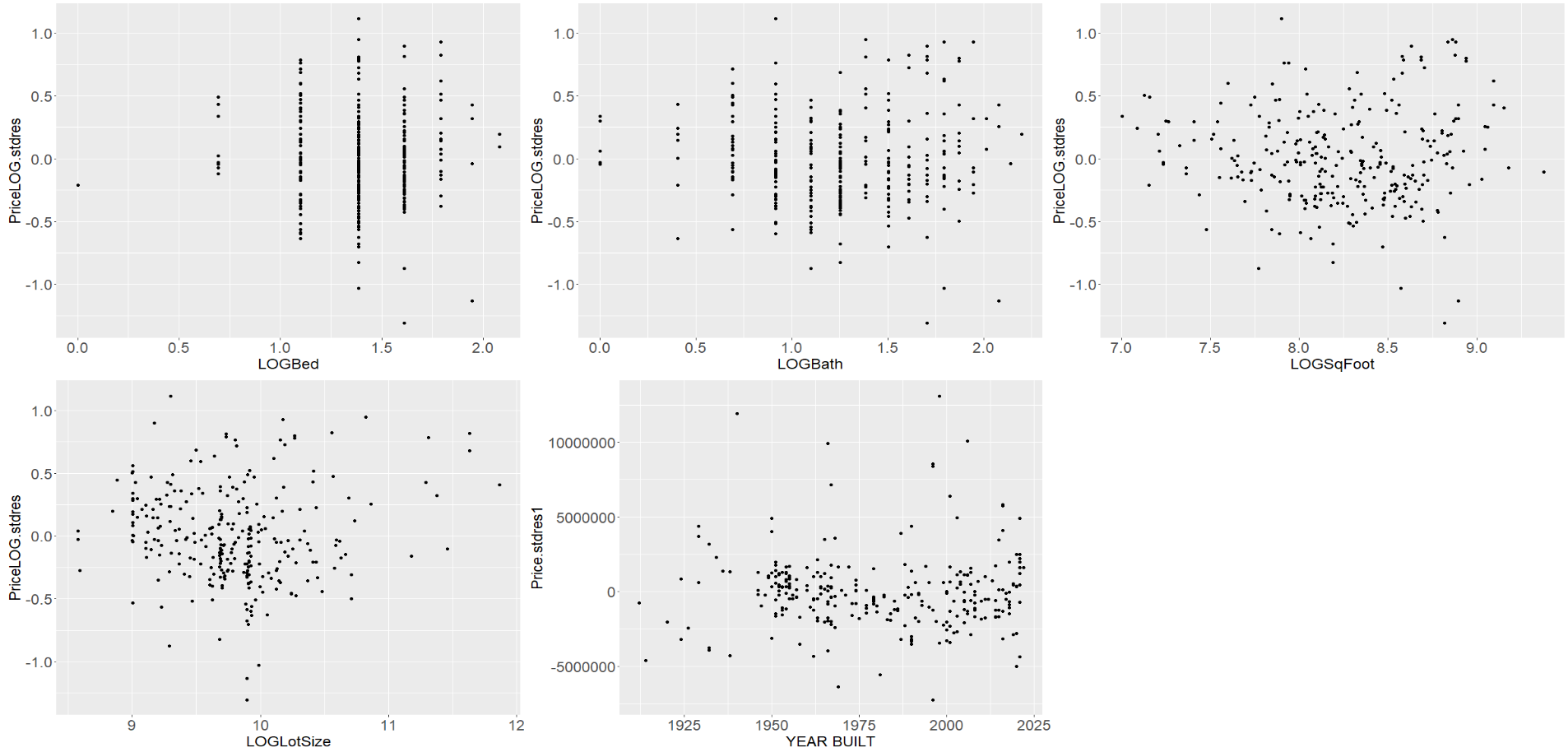


Figure 5‑8: Residual Plots – All Variable Log Transformation Included-2

# PREDICTION

The predict function is used to predict the value of house prices. The predicted house price vs the actual house price comparison for each variable is shown in Figure 6‑1. You can see from the graph that some of the lower prices values are over predicted and some of the higher price values are underpredicted.

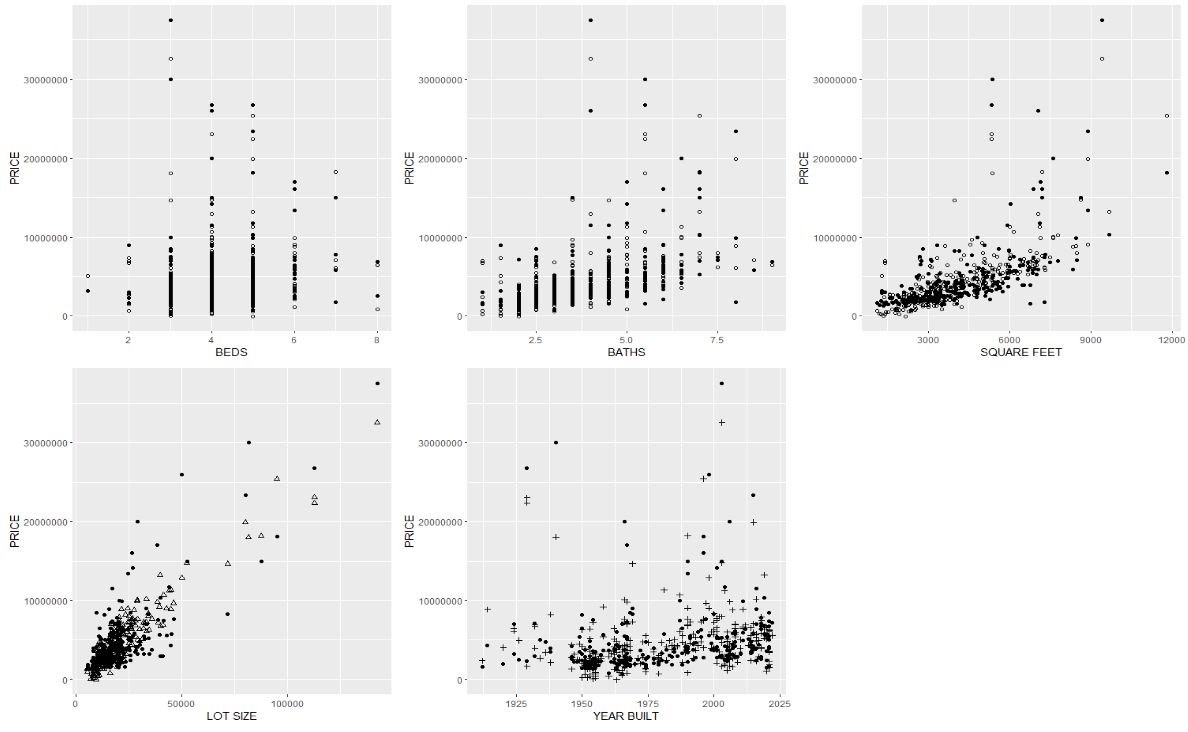


Figure 6‑1: Predicted Vs Actual Values

# CONCLUSION

From the complete analysis, it’s quite evident that the lot size plays a major role in the richest neighborhood in/around Seattle.

It’s also interesting to find that though the log transformations have normalized the variables to a great extent it didn’t help much in improving the R-squared value of the model. Also, the data has some outliers which influence the data/ regression model a lot.