Estimation of Unit Area House Prices in a Specific Area Using Multiple Regression Analysis and Correction of Residual Assumptions Using Appropriate Method

Vika Valencia Susanto

Computer Science Department

Bina Nusantara University

Jakarta, Indonesia

[vika.susanto@binus.ac.id](mailto:vika.susanto@binus.ac.id)

Audrey Tabitha Ariani

Computer Science Department

Bina Nusantara University

Jakarta, Indonesia

[audrey.ariani@binus.ac.id](mailto:audrey.ariani@binus.ac.id)

Aditya Chandi Nugroho

Computer Science Department

Bina Nusantara University

Jakarta, Indonesia

[aditya.nugroho003@binus.ac.id](mailto:aditya.nugroho003@binus.ac.id)

# *Abstract-* Many data scientists and engineers have always found it difficult to predict real estate value. Analyzing real estate price projections is the goal of this study. To build a regression model that will precisely predict real estate prices for this study, multiple regression analysis is required. Several presumptions need to be true in order to get a good regression model, such as normality, multicollinearity, heteroscedasticity, and autocorrelation assumptions. The first model, which employs the Ordinary Least Squares technique, deviates from the normality assumption. The created regression model can thus satisfy the normality condition by employing the Robust Regression method, which was selected after comparing it to the original model by computing the residual standard error value. As a result, this study may be used as a reference to estimate the cost of a unit-area house, is anticipated to serve as a basis for similar future research, and can offer suggestions for real estate business players when implementing marketing strategies for real estate agents.

***Keywords- Multiple Linear Regression, Robust Regression***

# INTRODUCTION

One of a person’s primary needs that must be satisfied is a home. Thus, business transactions between people and real estate agencies frequently happen nowadays. Engineers will use this chance to construct real estate, whereas people will purchase houses and land lots to live in or as an investment, and real estate agencies will buy it to run their business. This makes the valuation of real estate increase on a daily basis with public interest soaring high. The unstable, uncertain and unpredictable price makes investors and buyers need a system to predict house prices and valuation based on several housing aspects.

Predicting real estate value has always been a challenge for many data scientists and engineers. The study titled “Building real estate valuation models with comparative approach through case-based reasoning” written by I-Cheng Yeh, Tzu-Kuang Hsu (Yeh & Hsu, 2018) uses five regressions models, namely Linear Model, Quadratic Model, Logarithmic Model, Exponential Model, dan Exponential Growth Model. The result is concluded using a quantitative comparative approach with error hit rate 20%, specifically 74.3 with Rating Scale Mental Effort (RSME) score of 8.65.

Based on the description above, the purpose of this study is to analyze real estate price predictions. For this research, the use of Multiple Regression Analysis is needed to create a regression model that will accurately forecast the real estate prices. To obtain a good regression model, several assumptions must be fulfilled. When one of the assumptions is violated, a method is needed to overcome it. Therefore, in this paper/research, the use of Multiple Linear Regression using several existing variables based on the data set “Real Estate Valuation Data” and using some methods to overcome the violated assumptions will conduct more accurate and wider predictions for real estate prices.

In addition, the benefits obtained from the results of this study are to be able to find out the range of house prices and help us to be someone's consideration in buying and selling houses. This research is also expected to be a reference material for further research that has similarities and to provide input in carrying out marketing strategies for real estate agents and can be considered for real estate business actors.

# METHOD

## Multiple Linear Regression

Multiple linear regression is an extension of linear regression that refers to a statistical approach to predict the outcome variable based on the two or more variables. The dependent variable is the variable we aim to predict, and the factors we use to forecast the value of the dependent variable are known as independent or explanatory variables. Analysts use this technique to determine the model's variation and the proportionate contribution of each independent variable to the total variance.

The equation model of multiple linear regression is:

Where:

* Y = Dependent variable
* X = Independent variable
* = Constants referred to as the model partial regression coefficients (or simply as the regression coefficients)
* = Random disturbance or Error

## Ordinary Least Squares

Ordinary Least Squares regression (OLS) is a widely used method for calculating the coefficients of linear regression equations that describe the connection between one or more independent quantitative variables and a dependent variable (simple or multiple linear regression). The goal of the OLS approach is to reduce the sum of square differences between observed and predicted values as much as possible.

In this approach, the coefficients' vector can be estimated using the formula:

Where:

* X = matrix contains of the value of the predictors variable (the value of is usually 1)
* = the transpose of matrix X
* Y = vector of the dependent variable

## Assumptions of Multiple Linear Regression

To conduct a good regression models, several assumptions must be fulfilled, which are:  
1. The errors εi where i = 1, 2, …, n, have a normal distribution or *normality* assumption;

2. The attributes xi where i = 1, 2, …, n, have no multicollinearity;

3. The errors εi where i = 1, 2, …, n, have homoscedasticity;

4. The errors εi where i = 1, 2, …, n, have no autocorrelation.

### Normality test

Datasets are preferred to be normally distributed, unless the sample size is deemed relatively large, in order to build an unbiased prediction model. The normality test serves to assess the likelihood that the residuals are normally distributed. When the distribution of a cumulative distribution is not normal, it can usually be explained by the presence of extreme values and Insufficient data discrimination. This normality test could be carried out by doing the Kolmogorov-Smirnov test for normality, and Lilliefors Corrected K-S test. The normality test hypothesis is:

* H0: The errors are normally distributed
* H1: The errors are not normally distributed

The Kolmogorov-Smirnov test statistic for a given cumulative distribution function *F(x)* is:

Where:

* *n* is the total number of data
* is the supremum of the set of distances
* is the ordered sample
* is the Kolmogorov-Smirnov test statistic

The value of will then be compared to obtained from the Kolmogorov-Smirnov table.

### Heteroscedasticity Test

The Breusch-Pagan test is one of the most popular methods to test for heteroscedasticity, which describes if the residuals are homoscedastic (independently identically distributed) or not (Klein et al., 2016). The hypothesis of the heteroscedasticity test is:

* H0: The errors are homoscedastic
* H1: The errors are heteroscedastic

The Breusch-Pagan test statistic is denoted as such:

Where:

* SSR\* is the regression sum of squares from the regression between and *X*
* SSE is the error sum of squares from the regression between *X* and *Y*

The test statistic will be compared to the from the Chi-Square table. If is larger than then it can be decided that the residuals are heteroscedastic.

### Autocorrelation Test

The degree of correlation between two consecutive time intervals is referred to as autocorrelation. It assesses the relationship between the lagged version of a variable's value and the original version in a time series. Durbin Watson test is used to test autocorrelation in data.

The hypothesis for Durbin Watson Test:  
 no autocorrelation exists

autocorrelation exists

The formula for Durbin Watson test:

Where:

* = residuals from OLS Regression

The test statistic will be compared to the critical value from the Durbin Watson table.

### Multicollinearity Test

In regression analysis, the Variance Inflation Factor (VIF) evaluates the severity of multicollinearity. It's a statistical notion that describes how collinearity increases the variance of a regression coefficient.

VIF can be calculated using the formula:

Where is the uncorrected coefficient of determination for regressing the i-th independent variables on the remaining ones.

If the value of VIF is more than 10, the observation can be concluded as multicollinearity.

## Robust Regression

Robust Regression is a form of regression analysis designed to overcome some limitations of traditional parametric and non-parametric methods. Regression analysis seeks to find the relationship between one or more independent variables and a dependent variable.

The finite sample breakdown of an estimator/procedure is the smallest fraction α of data points such that if [nα] points → ∞ then the estimator/procedure also becomes infinite. The sample mean of is

=

=

and so if is large enough then can be made as large as desired regardless of the other n − 1 values.

Linear model:

For the *i*th of n observations. Given an estimator b for , the fitted model is

And the residuals are given by

With M-estimation, the estimates b are determined by minimizing a particular objective function over all b

where the function ρ gives the contribution of each residual to the objective function. A reasonable ρ should have the following properties:

* Always nonnegative, ρ(e) ≥ 0
* Equal to zero when its argument is zero, ρ(0) = 0
* Symmetric, ρ(e) = ρ(−e)
* Monotone in |ei |, ρ(ei) ≥ ρ(ei 0) for |ei | > |ei 0|

The estimating equations may be written as :

## Dataset

Our paper conducts a study based on the dataset obtained from Kaggle with a total of 414 observations, and 6 attributes from 2018. Kaggle is a platform that provides plenty of datasets in extensive topics, which are often used for machine learning competitions and practices. The six attributes are: (1) Transaction Date; (2) House Age; (3) Distance to the Nearest MRT Station; (4) Number of Convenience Stores; (5) Latitude; (6) Longitude. The target variable is the House Price of Unit Area.

The data has limited description about each attribute, however the best guess of explanation is the following:

1. Transaction Date: -
2. House Age: Age of the house
3. Distance to the Nearest MRT: The distance to the nearest MRT station
4. Number of Convenience Stores: The number of convenience stores per unit of radius
5. Latitude: Latitude point of the house
6. Longitude: Longitude point of the house
7. House Price Unit Area: House price per unit area

Using the summary() function in R software, a simple Exploratory Data Analysis (EDA) of the data is obtained.

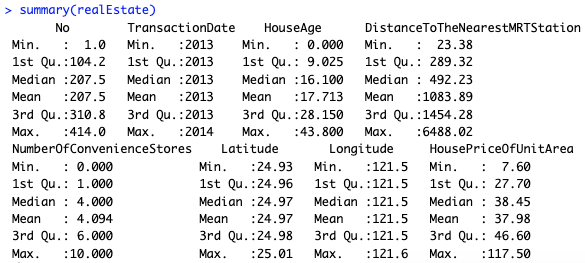


Fig. 1: Summary of the dataset

The large difference between the third quartile and maximum values in the attribute DistanceToTheNearestMRTStation suggests that it contains at least 1 outlier. If the attribute is visualized with a boxplot using the ggplot2 library, it would be clear to see that some outliers exist within this particular attribute.

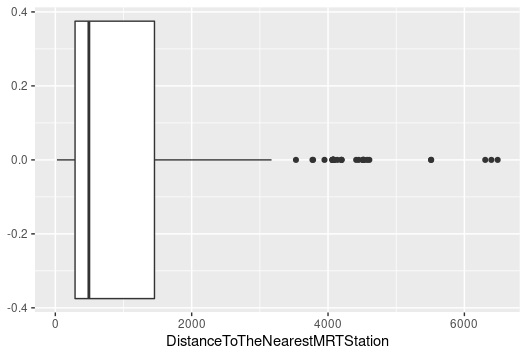


Fig. 2: Boxplot of DistanceToTheNearestMRTStation shows the existence of outliers

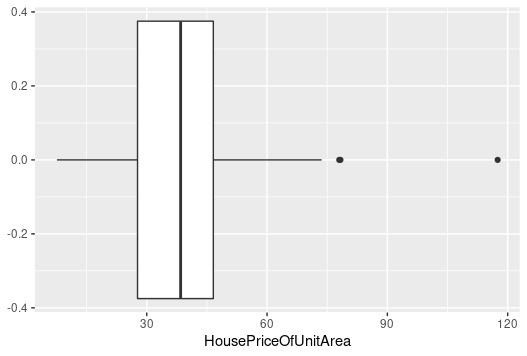


Fig. 3: Boxplot of HousePriceOfUnitArea shows the existence of outliers

# RESULT AND DISCUSSION

The use of R software is needed to calculate further computation results. For the multiple linear regression model, the predictor variables that will be used are house age, distance to the nearest MRT, number of convenience stores and the predicted variable is house price unit area. The variables longitude and latitude can not be used directly, as it is unlikely there is a true linear relationship. Those variables will be applied if they have already been converted into variable like geodesic distance.

For the variable selection, the p-value for each predictor variable in the t-test is shown in the table below. Predictor variable with p-value greater than α or the significance level will be eliminated, in this case the α is 0.05.

|  |  |
| --- | --- |
| **Predictor Variable** | **P-value** |
| House age |  |
| Distance to the nearest MRT station |  |
| Number of convenience stores |  |

Table 1: Table of t-test p-values for three attributes (House age, Distance to the nearest MRT station, Number of convenience stores)

Since there is no such p-value that exceeds α (0.05), then no variables are removed. This means that those variables are used to estimate the response variable, “house price of unit area”, and using the OLS method of Multiple Linear Regression, the equation provided is:

To carry out a good regression model, check this regression equation's residual assumption next. As mentioned in the Methods section, the tests utilized are the Kolmogorov-Smirnov, Durbin-Watson, and Breusch-Pagan tests. The p-value for the regression model in each test is displayed in the table below. The desired outcome is bigger than α (0.05), which suggests that all of the null-hypotheses fail to be rejected.

|  |  |  |
| --- | --- | --- |
| **Assumption** | **Test** | **P-value** |
| Normality | Kolmogorov-Smirnov |  |
| Autocorrelation | Durbin-Watson |  |
| Heteroscedascity | Breusch-Pagan |  |

Table 2: Table of p-values for three regression assumption tests

Using the multicollinearity test, determine whether each variable's VIF score is greater than 10 or less than 10. The variable has multicollinearity if the VIF score is larger than 10, according to the definition.

|  |  |
| --- | --- |
| **Variable** | **VIF** |
| House age | 1.007349 |
| Distance to the nearest MRT station | 1.577579 |
| Number of convenience stores | 1.580431 |

Table 3: Table of VIF values for three attributes (House age, Distance to the nearest MRT station, Number of convenience stores)

According to the outcome shown in the table above, the model only does not satisfy the residual assumption in the Kolmogorov-Smirnov test for normality, because the P-value exceeds the significance level (0.05), which shows that the model is not normally distributed. Regarding this violation, the Robust Regression approach is the best regression model to satisfy the normality assumption. Thus, the table below shows the regressor coefficients for each developed regression model.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Regression Model** | |
| Ordinary Least Squares | Robust Regression |
| Intercept | 42.977286 | 42.0485 |
| House Age | -0.252856 | -0.2760 |
| Distance to the Nearest MRT Station | -0.005379 | -0.0052 |
| Number of Convenience Stores | 1.297443 | 1.3860 |

Table 4: Table of regressor coefficient for two models (Ordinary Least Squares, Robust Regression)

The residual standard error (RSE) was chosen in this study as the indicator of how well our created regression models performed. The better a regression model fits our data, the lower its RSE value should be. The RSE for both models are shown below.

|  |  |
| --- | --- |
| **Regression Model** | **RSE Value** |
| Ordinary Least Squares (OLS) | 9.251359 |
| Robust Regression | 7.080152 |

Table 5: Table of RSE values for two models (Ordinary Least Squares, Robust Regression)

The regression model created using the Robust Regression method appears to be the most accurate model for our data when comparing the RSE value of each regression model. Consequently, the Robust Regression method-developed regression model,

# CONCLUSION

By creating a regression model using the Robust Regression approach and using the house age, distance to the nearest MRT, and number of convenience stores as the predictor variables, it is feasible to generate an estimate of the house price in a unit area. The Multiple Linear Regression Model broke the requirement of normality, and this method is employed to make up for it because the Robust Regression Model provided a superior fit for the data.

Thus, this study could be used as a reference to determine the price of a unit-area house, also is anticipated to be a starting point for related future research, and can be utilized to inform marketing plans for real estate brokers and other players in the industry. In order to potentially identify a more accurate regression model, it may be necessary to construct regression models in future research or studies employing larger versions of the methodology.

# REFERENCES

Fernandez-Duran, L., Llorca, A., Ruiz, N.,Valero, S., Botti, V. (2011). The impact of location on housing prices: applying the Artificial Neural Network Model as an analytical tool. *ERSA conference papers*.

Yeh, I. C., & Hsu, T. K. (2018). Building real estate valuation models with comparative approach through casebased reasoning. Applied Soft Computing Journal, 65, 260–271. https://doi.org/10.1016/j.asoc.2018.01. 029

Ghasemi, A., Zahediasl, S. (2012). Normality Tests for Statistical Analysis: A Guide for Non-Statisticians. International Journal of Endocrinology and Metabolism, 10(2):486-489. <https://www.researchgate.net/publication/248398138_Normality_Tests_for_Statistical_Analysis_A_Guide_for_Non-Statisticians>

Singh, A. S., Masuku, M. B. (2014). Normality And Data Transformation for Applied Statistical Analysis. International Journal of Economics, Commerce and Management, Vol. II, Issue 7.

Fox, J., Weisberg, S. (2012). Robust

Regression.

[https://www.bauer.uh.edu/rsusmel/phd/fwro bustreg.pdf](https://www.bauer.uh.edu/rsusmel/phd/fwrobustreg.pdf)

Schmidt, Amand F., and Chris Finan.“Linear

Regression and the Normality Assumption.”

*Journal of Clinical Epidemiology*, vol. 98,

June 2018, pp. 146–151,

10.1016/j.jclinepi.2017.12.006.

P. Huber, "Robust Estimation of a Location

Parameter", *The Annals of Mathematical*

*Statistics*, vol. 35, no. 1, 1964.