Multiple Regression Analysis on Melbourne Housing Price

Sana Amreen

Dufang Qu

Nida Rasool,

Paripon Thangthong

Wilson Wu

(Alphabetize by last name)

Contents

Abstract

Being able to accurately estimate the value of a real estate property is very important for both the housing authorities in charge of urban planning and real estate agencies. In this project, the Great Melbourne area housing data from the online public source are being used to predict the house selling price within this area. Original Melbourne Housing Market dataset has 34858 observations. Therefore, we narrow down by using a random sample, and the new dataset will be our data to build a new model. In data exploration step, Histogram, boxplot, and transformation make our model more accurate and decide how the method would use for building model. Multiple regression method is the major methodology used in this project, so by Multiple using regression model can help us predict the future property price or calculate how the variables influence on property price. In model validation, Data set is split into training and testing, and use a training set to build our accurate model. Testing R-square, Adj R-square, and RMSE to measure how many presents of variation fall independent variable by the model with every independent variation. Moreover, an alternative method, 5- fold cross validation can test and build our final model by using forward, backward, stepwise. While as group members have various hypotheses about most influential factors among the 20 variables, and dealt with the variables differently during data preprocessing phase, for instance, some may drop Suburb and chose Region, while others may have done the opposite, the final models and findings appear quite diverse. Some final models indicate BuildingArea is the most influential factor for housing price, while others suggest that coded Suburb variable is the most significant determinant for housing price. Because of the sample dataset different might, we have some different result. Finally, Based on our analysis, we can use our final model to study the influence of Melbourne Housing Market, analyze price changing in different time, and predict future property price. In future work, if we could get an inside or an average people salary and the growth rate of the business in each area, the prediction can answer the right question for both agent and buyers. Moreover, we might apply more independent variables and modify them into the accurate regression model so that we can find the important role for the property price and how each variable effect and relate to each other.

Introduction

In the modern era, the status of the real estate market is one of the best indicators for the economy of a society. There are so many aspects in term of real estate market our team is longing to discover, such as prediction of rental rate, estimation of land cost, optimal property investing strategies and so on, but in this report, we are only focusing on a fundamental topic: the prediction of housing price from most relevant features. As various features of an urban or suburban residential property may directly affect its price, we believe that the study of the association between them is the foundation of all other real estate market related studies. Besides that, a good model or models eventually coming up with the study which are able to accurately estimate the housing price, would be also a feasible practice in which the housing authorities, real estate companies and real estate brokerages may be interested.

In our project, we chose real estate market open data (2016 - 2018) of the greater area of Melbourne. As the beautiful prosperous capital city of Australia, Melbourne is located by the coast of the Australian state of Victoria and is one of the most populous cities in Oceania. Its housing market has long been a focus of attention. The data our team studied on is original from open real estate data from domain.com.au, which is Australia's second largest real-estate marketing business. Data has been web-scraped and cleaned by a Kaggle.com user and posted in kaggel.com continuously.

This data contents one dependent variable Price, which is number, and 21 independent variables, of which Rooms, Distance, Bedroom2, Bathroom, Car, Landsize, BuildingArea, Latitude, Longtitude and PropertyCount are number type, Suburb, Address, Type, Method, SellerG, PostCode CouncilArea and RegionName are qualitative data, Date and YearBuilt are year/date type. Many of those could have influence on the housing price. For instance, location data will be useful in determining what locations in the Melbourne area have more expensive housing. And time-based data such as year built helps analyze how the age of a house affects its overall value. More factors like landsize and building area including the region where house is located played a signuficant role in deciding the price for the house.This is all valuable information to people in the housing industry and buyers and sellers alike.

In our project, the first goal is to explore each predictor which relation or association between each other. The second goal is investigated what are the real influential predictors among the various features, for the real estate property price in this area as well as to discover the quantitative relationship between these predictors and the property price. The third goal is to build an accurate final model that can predict or calculate the property price.

To reach these goals, we use SAS program to analyze Melbourne Housing Market dataset and the methodologies we used for the project are mainly multiple regression method. Dummy variables and interactive term are created in an attempt to make the most use of the features as well as to achieve a more accurate predictive result. We want to know each predictor association with the dependent variable by using G plots, scatterplot, and Pearson Correlation Coefficients. These approaches allow us to find out if there is any linear between predictor and the dependent variable. Moreover, Pearson Correlation Coefficients VIF and Tolerance can help us fix the

problem of multicollinearity. Testing P-Value and compare the standardized estimate can identify the most influential and significant features of housing prices. We want to have the most accurate model to predict our dependent variable by using model selection and model validation to test. These approaches allow us to know how is the performance of the model by checking R-Square training and R-Square testing. Moreover, we can compare 5- fold cross validation R-square, Adj R-square with the training model to select our final model. Two models at the final stage show respective strengths, while the model which provides the most intuitive interpretations to the effect all features have on the property price is chosen as the best model of this study. Finally, Based on our analysis, we can use our final model to study the influence of Melbourne Housing Market and predict future property price.

Section A (Paripon Thanthong)

Analysis report

The goal of this project is to predict the price of house, apartment and/ or condominium in Melbourne, Australia which shows an interesting trend. The model that are implement in this project point out to see what prospect will have effect in to this dataset. Even though the house selling price are drastically increasing from a recent report in Australia, this model could be used as an support to make a final decision in buying home. Thus, I strongly believe this model can be benefitting for the developer and people who are interested in buying place to live in Melbourne.

Exploratory Analysis

At the beginning, the descriptive statistic of 1499 sample observation from the full model including 5 points summary analysis of twenty variable that suggest to provide adequate information to including in the model. It shows an important information specifically for dependent variable which is price. After computing the histogram of price, the graph show significantly skewed to the right which mean the data of price is very condense around the first quartile. Because of that skewness, price variable is considered to be transformed. Moreover, some original independent variable cannot be use directly, such as regionname and method. Both of them need to be created dummy variable which can use the value 0 or 1 to include in the model. For the exploration on the scatter plot between dependent variable (price) against all independent variable, the result show that some plot show non-predicting shape and some other show violating trend and not linear which need to be adjust before select final model.

From my research on real estate industry, some variable show meaningful value when combining them together like distance(distance from central business),landsize(meter) and buildingarea(meter) which are strongly related to the price. It is because many people tend to stay close to the area that provide them comfort. In this model, distance , landsize and buildingarea variable are used as an interaction term. These interaction term are assigned to in_1(distance * landsize) and in_2(distance_buildingarea).

To check the strength and direction of all independent variable, pearson correlation coefficient table is a graph that provide a result between two variable which can see on the matrix part by using proc corr to compute. It shows some weekness in many of these variable such as, -0.09411 on distance and 0.21796on landsize.

Methodology

To begin with, the dataset of our group selected to do the prediction is Melbourne housing Market which has about 34,000 observations and 13 columns from Kaggle.com.

First, I study the dataset to see what approach I have to go with and see how the dataset look like. Here is the process for my analysis.

Cleaning Process (pick random observation for 3000)

Starting with random the original dataset by using function '=rand()' in Microsoft excel in empty column to compute random number for every row. Then, I sorted '=rand()' column to by number from low to high to re-arrange the whole data set. After that, I pick 3000 sample observations and put it in a new sheet and remove insignificant columns. Next, I import the 3000 observation file in to Oracle developer for eliminating the missing value by writing sql query. The query I wrote was to select all the row that did not have any nulls value (missing value in excel) in every column. The result came out as I collect 1499 observations. So, I exported that query to .csv file name 'Mel66'.

Oracle Developer query part:

```
select *
from melbourne2 m
where (bathroom is not null)
        and
        (price is not null)
        and
        (car is not null)
        and
        (landsize is not null)
        and
        (buildingarea is not null)
       and
       (yearbuilt is not null)
       and
        (councilarea is not null)
        and
        (lattitude is not null)
        and
        (longtitude is not null)
       and
        (regionname is not null)
       and
        (propertycount is not null)
        and
        (method is not null)
```

```
        ;

select *
from mel3 m
where ( price is not null);
```

Pre-Process 1

Importing file in to SAS, I used the procedure import to bring csv file in and assign the data into 'Mel' dataset. The dataset has both qualitative variable which I select 3 qualitative columns which is region name, method and type ,and create 10 dummy variables out of its. Moreover, I also create 2 set of interaction term between distance variable with landsize variable and distance variable with buildingarea.

Exploration Process

For the exploratory process, I compute basic descriptive statistic by using procedure means indicates observation number, minimum, maximum, mean,median, standard deviation, and percentile at 25, 50 and 75. As I set to predict the selling price, my dependent variable is price ,and I compute histogram graph of price variable to see its distribution. The next step is I compute scatter plot and matrix plot with price variable against all other independent to see the trend of each plot.

Analysis Process 1

The analysis process starts with I computing procedure corr to see the correlation between variables in the dataset. Next, compute first full model by using procedure reg to see R square , Adj R square, RMSE(error), F-Value, P-Value and parameter estimate for each variable. After I have full model, I check the diagnostic of the full model by looking at outliers, influential points and check the model assumption. The model assumption I analyze is Constance variance, Independence, Linearity and Normal Probability. Furthermore, I also compute procedure reg with VIF value to see if there any multicollinearity.

Pre-Process 2

After first analysis, I find some obvious pair of influential points and outlier, which I remove it and set another data set to 'new' dataset. With the histogram graph in the first preprocess, the dependent need to be transformed. So, I decide to assign log transformation to dependent variable and create a new dependent variable 'ln_price'. Then, I compute another histogram grap with 'ln_price' to see if the distribute is improved.

Exploration Process 2

Then, I compute another histogram grap with 'ln_price' to see if the distribute is improved.

Analysis Process 2

At this stage, I compute procedure reg again with ln_price variable to see overall improvement and model assumption.

Splitting Data in to Training set and Test Set

For selecting the final model, I split data set into 60% training dataset and 40 % test dataset buy using procedure surveyselect and set in to dataset 'new2'. With 'new2' dataset I use this training dataset to create new variable that relate to the selected value(60% training dataset)

Selection of Model from Training Dataset

Next, I use 3 model selection methods to identify for my final model. The methods are stepwise, backward and CP. From each model selected, I compute another model assumption to check each model. I test performance only two model selected which is from stepwise and backward by using 40% of test dataset, this is because CP and stepwise select the same variable. Then, I multiple ways or cross check test dataset and training dataset. I compare them by looking at RMSE,R square, Adjusted R square, F value, Model assumption number of variable in each model.

Another method, I use 5-Fold cross validated from the full model without splitting data to see the different output in stepwise method and backward method.

Select Final Model base on Training dataset

Selecting final model for prediction which is backward from training dataset then check the following list again.

- Check Model assumption for final model
- R, Influential Points
- Multicollinearity if needed

Check standardized Estimate for the independent variable for final Model

- Computing the standardized estimates to see the association between all independent variable and dependent variable

Compute a prediction by using Datalines combine with 'new4' dataset.

Analysis

First, compute regression full model by using procedure REG that provides important value to analyze. The table shows that R value is about 70.34% and Adj R.sq is about 69.94% which is seem to be a strong model. Furthermore, the F- value , no cut off value, has 175.28 which is a solid number in this case. For the confidence interval at 95%, some of these variable may not suitable for the model which it has to be removed by comparing the Pr>t column to 0.05. In this full model regression,Bedroom2, Propertycount, methodsp,methodsa , rgname_w and rgname_se are shown the p-value that above the 0.05. These insignificant variables are needed to be remove one by one to see improvement of overall model. Although, the full model seem to be in a good shape, the error is 382412 which is very high and may not be good enough to use this model for prediction. It could be from the skewed right of the dependent variable or insignificant variable.

Full model equation at first analysis stage

Price = 244417 + 92009(ROOMs) + 15603(DISTANCE)-237609(BEDROOM2) + 117827(BATHROOM) + 39326(CAR) + 32.65 (LANDSIZE) + 8642.97 (BUILDINGAREA) - 1.2(PROPERTYCOUNT) – 385164(typeu) - 283898 (typet) – 49586(methodsp) - 130742(methodpi) - 134912(methodvb) – 228267(methodsa) - 45573(rgname_w) + 256475(rgname_e) + 453627(rgname_s) + 289061(rgname_se) +8.6(in_d1) -581.31(in_2)+e

Second, checking diagnostic for full model by focusing on outlier, influential points and multicollearity that make an effect to the everall model. By perfroming proc reg, there are obvious sixteen pair of outlier and influential points that are removed from the model. Furthermore, to get more evidence that the full model need adjustment, the model assumption need to be satisfy. By analyzing model assumption of full model, the plot of studentized residuals against all other independent variable show some plot has a pattern and not randomly scatter around zero line which is violated. For all over model assumption, constant variance assumption has shown some funnel shape at studentized residual against x variables (buildingarea, landsize) and the predicted value plot shows some pattern of ‘U’ shape. The linearity assumption is also violated for some independent variable such as some plot show an unexplainable pattern and funnel shape which are needed to be transform. The Normal Probability plot show an obvious ‘S’ shape at both end.

With this full model, the VIF and TOL ,which indicate the multicollinearity, show some good trend which there is no variable that violated this bars. However, with Pearson Correlation Coefficient graph has some pairs of variable with value over 0.9 or close to 0.9. Thus, the model need to be transformed.

With transformation on dependent variable, the log procedure is being implemented to Price variable and set to another name which is ln_price. After computing distribution of ln_price, the histogram plot shows significantly improve from skewed right to be highly normal distribution for full model. Moreover, the regression procedure of ln_price also shows significant improve of R square at 78.54% and adjusted R.sq at 78.24%. Also, the Root MSE (error term) has also decrease to about 0.25.

After transformation process, the overall model assumption has significantly improved. For example, studentized residual against all dependent variables are more randomly scattered around zero line and the predicted Value plot is improved. The normal probability plot is significantly change from 'S' shape to highly linear at 45 degrees.

Third, before starting the model selection for final model, all possible model or method need to be validated. By doing the validated process, the dataset need to be split into 2 parts which Training Set and Test Set and check each model selection with training set to see the difference between model. In this project the training dataset and test dataset has been divided in to 60/40 by using proc surveyselect, and three model selection has been selected which are stepwise, backward and CP. For stepwise and CP, both methods select 13 variables which either one of this method will provide the same result at the regression procedure. The R.sq for stepwise and CP is 77.89% with F-value at 237.32. In contrast, backward method has selected 12 variables which show slightly difference of R.square at 77.83% and F-Value at 256.50. For the outlier that are removed previously, variable 'methodsp' is not in any model because of the low number of observation which has been removed.

Next computing model assumption for all three model selection in training dataset, the model from stepwise selection shows studentized residual against selected variables with more randomly scatter and less outlier. But, some plot are still violate the the rule of model assumption which can be accepted in the model. For example, the plot of residual studentized against landsize variable is still show some funnel shape and show obvious influential points. The predicted value plot has improved to be more randomly scattered from full model,but it still has some outlier which can be accepted because all of them are just slightly over the line and I want to keep as much observations as I can. Because CP and stepwise are selected the same variables, so the result will come out the exact same thing as the analysis of selection from stepwise. Also, the normal probability plot show a strong 45 degrees angle line. However, the selection from backward provides another way for the final model. This selection does not include the variable landsize because the P-value number is over the threshold at 0.05. All the residual plots are the same as I previously analyzed except some value in regression procedure are different such as R.square value and F-value. Outlier and influential point have obviously shown just a pair which I will not remove it from the training dataset that has been already split.

To be precise on selecting the final model, the result from training dataset will be used to validating with test dataset each model. The test dataset is at 40 percent of total observation after removed the outlier and influential points which is 593 observations. The test dataset for stepwise show the two error terms ,which are root mean square error(RMSE) and mean square error(MAE), the RMSE is 0.26851 and the MAE is 0.20433. With Pearson Correlation Coefficients of dependent ,and yhat's value (for computing R for test dataset and CV R.square) at 0.88185. So, R. square this test dataset of stepwise method will be 0.7776 and CV R.square is 0.0013(R square training dataset – $(yhat)^2$). For the model selected from backward method in test dataset, RMSE is 0.27008 and MAE is 0.20526 which generating the R square value at 0.7751 and CV R.square is 0.0031.

To validating the all possible final model, first I will compare Training dataset of stepwise selection model and backward. The result of comparing Rmse, R.square and Adjusted R.square from both model is that the stepwise model overall number is a little better except the F-Value. Second, comparing test dataset for both model shows the result of RMSE, MAE and R square from stepwise method is just slightly better , but the CV R square from the backward method is higher.

Comparing across training dataset and test dataset for each model selection, the result for the stepwise model in training dataset is all error values lower and R.square is slightly better. Also, the result on cross check between training dataset and test dataset in backward method is in the same way as stepwise method.

The another process of validating from full model is 5-Fold Cross Validated. With this process, I compute it with stepwise and backward process to compare the validation and see which one perform better between Cross Validated and previous method. The 5-Fold cross validated process show the ASE test value from backward method is very slightly lower at 0.06922 and ASE Train is 0.06141 with selecting 17 variable to be in the final model. In contrast, ASE test for stepwise method is 0.06954 and ASE train is 0.6288 with 11 variable are selected in the final model. Both of these method show very close R square which is around 79% but backward perform very slightly better. Because of 5-fold cross validation cannot be set when running the process, the result will change every times I re-run the procedure. I decide not to use this validation method to validate final model.

From the information that has been collected from all three models in training dataset, I decide to select model from backward method to be one of my final model. This is because the overall mathematic computation from cross check validation method by splitting dataset show a slightly better in backward than CP and stepwise. Moreover, the number of variable is less which mean it may be a better fit to predict the price of real estate.

Before using the final model for predicting the price of house in Melbourne, Australia, I compute the diagnostic test and model assumption for final model to see if there any thing that can improve the overall predicting. Two pair of outlier and influential points have been found and it is removed, and there is violated value in 2 variable in the final model that show multicollinearity. It is Buildingarea and In_d2. Both of these variable will be ignore to remove out of the final model, because Buildingarea variable is an variable of interest which is the area of buiding in meter that can determine the price in some certain area. In_d2 variable will be ignored to remove out of the model because it is an interaction term that combine 2 variable that related to predict the price. For the model assumption of final model, studentized residual against all other selected independent variable in the final model is randomly scattered without obvious pattern. The normal probability plot is 45 degrees angle.

Therefore the final model equation is :
Log(New_y) = 13.24+0.089(rooms) – 0.022(distance) +0.058(bathroom) +0.0044(buildingarea) -0.51(typeu) – 0.25(typet) – 0.064(methodsp) -0.097(methodpi) – 0.13(methodvb) + 0.26(rgname_e) + 0.36(rgname_s) – 0.00019(in_d2)

For the overall Goodness of Fit Test in this final model,
$H_0 = 0$
$H_a \neq 0$
F-Value is 262.83
P-Value is less than 0.0001
The conclusion is that P-Value in this final model is very small than the threshold at 0.05, therefore we can reject H_0. Moreover there is at least one predictor that is significantly associated with dependent variable and there is a strong support for the final model.

With this final model, the association between selling price and other variable are both numerical and dummies show Buildingarea variable has the most strongest relation against selling price at 57.18%. The second strongest is unit(typeU) which has the value at 42.71%. Other variable show some weak trend from around less than 10% to slightly over 30%. In addition, to interpret the final model from the model equation. For example, we can say that by increasing 1 square meter of building size(buildingarea), the price of house selling in Melbourne, Australia will increase by 0.49 AUS dollars.

Prediction

For this prediction I will predict the price , confidence interval and prediction interval of house selling in Melbourne, Australia by using final model. The given values is distance from central business is 6.4 kilometers, 400 square meters,number of room is 4 ,region is in Southern Metropolitan , type is unit, method is property sold and the relation of distance and buildingarea is 5.

In Southern Metropolitan of Melbourne, Australia, a census data has been provided to predict house selling price in the future. The predicted of house selling price is around 2,850,485.00 AUS dollars with 95% confidence interval equal to (2,557,651.00 AUS dollars, 3,176,530.00 AUS dollars) and 95% prediction interval is equal to (1,734,968.00 AUS dollars 4,682,768.00 AUS dollars).

For the second prediction I will also predict the price , confidence interval and prediction interval of house selling in Melbourne, Australia by using final model. The given values is townhouse(methodt) that has 3 rooms , 2 kilometers from the central business, 1 bathroom, 200 square meters, sell prior before it close and it is in Eastern Metropolitan.

In Eastern Metropolitan of Melbourne, Australia, a census data has been provided to predict house selling price in the future. The predicted of house selling price is around 1,700,614.00 AUS dollars with 95% confidence interval equal to (1,591,519.00AUS dollars, 1,817,368.00AUS dollars) and 95% prediction interval is equal to (1,042,987.00 AUS dollars, 2,773165.00 AUS dollars)

Noted: The predicted value in both of these prediction did not perform the exact same formula the professor suggest due to the result that has shown unlikely number after minus 1 from the exponent value than multiply by hundred. The values go over hundred million AUS dollar which go far beyond the maximum value.

In my opinion, overall performance in this model is strong but it still needs an improvement in some part. For example, some of variable need to clarify more on detail such as landsize and building area which I cannot tell the difference between these two variables whether it is a condominium space or a house. However, I'm satisfy with this model because of the relationship between variable that can defy the selling price not accurately but rationally.

Future Work

In this model, the variables are collected by the mathematical method in SAS which cannot answers all the question if being ask by buyer for a values that out of this model. For further work, I would like to pay attention to the agent which has experience in their area. This experience can give more inside detail about people behavior in some specific area and help predict the price more accurate. Moreover, if I could get an inside or an average people salary and the growth rate of the business in each area, the prediction can answer the right question for both agent and buyers. However, there are some problem with the prediction value that I want to do insight research with it. It is the logarithm of the price that predict a highly value of the price

when I compute with formula my professor gave. I'll try to solve this over value in the further work.

Reference

Cheusheva, Svetlana. August,2018. How to select random sample in Excel. https://www.ablebits.com/office-addins-blog/2018/01/31/excel-random-selection-random-sample/ Accessed at March 2nd ,2019.

CHARTIO , How to SELECT Records With No NULL Values in MySQL. https://chartio.com/resources/tutorials/how-to-select-records-with-no-null-values-in-mysql/ Accessed at March 2nd 2019.

Hall, James. Rentvesting: Sydney and Melbourne property investors looking over the border. 'News.com.au'. https://www.news.com.au/finance/real-estate/buying/rentvesting-sydney-and-melbourne-property-investors-looking-over-the-border/news-story/30744d4a0107990c66ac256eaef26a84 Accessed at March 10, 2019.

Section B (Dufang Qu)

Methodology

Data information, sampling and cleaning

The dataset contains the prices and features of residential properties sold from 2016 to 2018 in Melbourne, web-scraped by an individual user of Kaggle.com from weekly-released open data on Domain.com.au, the second largest real-estate marketing business, and published on Kaggle.com by him (https://www.kaggle.com/anthonypino/melbourne-housing-market).

The full dataset consists of 21 features, of which Price is the dependent variable, and 34,857 sold residential properties, including units, townhouses or houses. It had many missing values in YearBuilt, LandSize, BuildingArea, Car and Price, and many noisy data or inconsistent values in LandSize, BuildingArea and Date.

1) Originally, I attempted to build the model on only the latest data of 2018, but when I was cleaning the data, I noticed there are 63% missing values in variable YearBuilt. I finally chose a subset of the data from July, 2017 to June, 2018. This sample is still timeliness, meantime it contains enough usable samples.

The values of the variable Date (the data at which the house was sold) is inconsistent, mixed with texts and dates. I extract year and month separately so that I could properly sort and select the target subset.

2) All rows with missing values in Price, YearBuilt and Landsize which I considered as MCAR were deleted. Some problematic rows were removed as well. They contain some unreasonable values, such as BuildingArea is greater than Landsize.

3) I used conditional mean imputation technique[1]to impute missing values in Car with the rounded mean value of Car values under the same conditions (same Postcode, similar Distance and LandSize).

4) Replaced around 100 problematic values (such as units have huge land size) in LandSize and BuildingArea with the median of the variable under the same conditions respectively.

5) After data cleaning, there are 3248 observations left, which is a perfect size for this project and became the final data set mbhouse.cvs for my study.

Data preprocessing

1) New variables Years and Ndis were created when importing the cleaned data:

*Years:*obtained from 2019 minus YearBuilt, represents how many years the residential property was built till 2019. This variable was created to make YearBuilt usable.

*Ndis:*obtained from 50 minus distance, represents how many kilometers the residential property is away from a boundary which is centered at CBD and has a radius of 50 kilometers. 50 kilometers is the distance the farthest residential property has in this data set. The purpose of creating Ndis is to better use its potential interactive effect with other variables).

2) Qualitative values of Suburb are coded into new four-level ordinal variable Sublevel.

There are 390 different suburb names in Suburb. Considering that community is usually a big concern when most people purchase real estate property and none of other variables could provide information about the community, I needed to make information in Suburb usable. For this purpose, I discovered Melbourne's most livable suburbs ranking2 (321 suburbs of Melbourne are ranked, released in Nov. 2015 by Domain.com.au) along with some supplementary information for other 69 suburbs, then, divided them into four levels in a top-down order:

1- Highly-ranked suburbs (1-80, top 25%)
2- Upper-middle-ranked suburbs (81- 160, 26% - 50%)
3- Lower-middle-ranked suburbs (161 - 240, 51% - 75%)
4- Low-ranked suburbs (241- 321, 76% -100%)

Based on this classification, I created a new variable SubLevel:

SubLevel: the level of the suburb in which the property is located, has 1, 2, 3, 4 four values

3) I generated histograms for continuous variables and discrete variables. The histogram of Price (see Appendix B.1) shows the dependent variable Price is highly skewed. Other variables, Distance, BuildingArea, LandSize and Yeas, also have skewness (see Appendix B.2 - B.6). I decided to first only apply logarithm transformation on Price and created new variable lnPrice, but kept other skewed variables for further observation.

4) Assuming that different categories in categorical variables have contrasting effects on Price, I generated boxplots of dependent variable price vs discrete variables, for the purpose of discovering which of the variables could be made of better use if corresponding dummy variables are created as well as how to create the dummy variables.

Boxplot for Price vs Sublevel (see Appendix B.7) illustrates that the four levels in Sublevel has very distinct middle 50% ranges, Type "u" has very different middle 50% ranges as Type "t" and Type "h", while the numbers of Rooms, Bathroom and the years of YearSold don't seem make big difference in term of price (see Appendix B.8 - B.14).

Therefore, I decided to create dummy variables below:
*Dsub:*1 if the suburb belongs to Lower-middle-ranked suburbs, corresponding SubLevel=3, otherwise 0;

*Dsub1:*1 if the suburb belongs to higher-middle-ranked suburbs, corresponding SubLevel=2, otherwise 0;
*Dsub2:*1 if the suburb belongs to highly-ranked suburbs, corresponding SubLevel=1, otherwise 0;
Dtype: non-unit or unit, 0 - unit, 1 - townhouse or house.

5) To avoid very small number in parameter estimates, I created three new variable Land, and Tyears:
Land: the land size of the preporty, in 100 square meters
Area:the building area of the preporty, in 100 square meters
Tyears: how many years the residential property was built till 2019, I unit represents 10 years

6) Hypothezing there are interaction effect between Ndis and Dtype, I generated gplot for them. From Scatterplot for LnPrice vs Ndis*Dtype (see Appendix B.15), we can see there is an interaction effect between Ndis and Dtype. When Ndis becomes greater, the LnPrice also increases more rapidly for Dtype =1 than that for Dtype = 0. Therefore, I created Ndis_Dtype:

Ndis_Dtype: interaction term created from Ndis*Dtype

Methodologies in Data Exploration, Analysis and Prediction

After importing cleaned data, I first explored all the variables by generating frequency tables, histograms, boxplots, scatterplots and Pearson correlation coefficient tables. Then, I created new variables, dummy variables and interaction term according to my findings in data exploration.

Next is to start the analysis. A full multiple regression model was built on all usable independent variables. Then, I checked R^2, Adj-R^2 for overall performance, significance test to see if all predictors are significant, VIF values to identify multicollinearity problem, and residual plots for potential issues, such as any suggestion about x or y transformation, other analysis methods, or outliers and influential points. If needed, corresponding actions were taken, including deleting rows, transforming x and y, etc.

After all issues were taken care of, data set was divided into training set and test set. Stepwise and CP model selection methods were used to select important predictors. The two selected models were applied on training data and test data separated, and above-mentioned related metrics and residual plots were generated as well for the comparison and problem identification of the two models and finally came up with a final model.

After I chose a better model from the former two, I don't feel the interpretation of a few estimate coefficients could well explain the relationship between the corresponding predictor and the housing price, so I used untransformed variable build a full model, and repeat the foregoing model selection and comparison processes, finally came up with a satisfactory model Model S3.

The final step is using Model S3 to predict the price of two houses, which have the same condition but in different area.

Analysis, Results and Findings

Data Exploration

For exploring the patterns of each variable, I generated histogram for newly-created variables Lnprice, Area, Land and Tyears. Besides that, I produced Scatterplot matrix and Correlation matrix for non-dummy variables to observe their relationships.

Histogram for LnPrice (see Appendix B.16) shows Lnprice has an approximate normal distribution after the log transformation. As their original variables, Area, Land and Tyears still have right tails on the positive sides (see Appendix B.17 - B.19). Since the dependent variable has been transformed, I kept those variables so far to see if there will be any problems appear in residual plots later.

As for the outliers on the positive direction in Area and Land. After checking the data, I decided also keep them because they look like reasonable data. Most of the observations having extreme Land values are with farther distances from CDB, bigger land size and comparatively higher price than other residuals in the same area with smaller land.

According to the scatterplot matrix (see Appendix B.20) and correlation matrix (see Appendix B.21), some of the variables appear to have a moderate correlation relationship with sale price. In addition, the Pearson Correlation Coefficient table shows Rooms and Bedroom2 are highly correlated (r=0.996). After investigating the information in these two variables, I found them mostly identical, so I chose to drop Bedroom2 and only use Rooms to build the model.

Full model building and diagnotics

Since there are some variables appear correlation to the dependent variable LnPrice, I started by implementing a linear regression model with all 12 variables which I identified usable through data exploration on 3138 samples. I obtained a full model Model F1 (see Appendix B.23):

LnPrize = 13.20 + 0.03Rooms + 0.06Bathroom + 0.01Car - 0.03Ndis + 0.24Dsub
+ 0.45Dsub1 + 0.75Dsub2 - 1.01Dtype + 0.26Area + 0.03Land
+ 0.03Tyears + 0.03Ndis_dtype

Where Dsub = 1 when Sublevel = “2”, (otherwise Dsub = 0) and
Dsub1 = 1 when sublevel = “3”, (otherwise Dsub1 = 0) and
Dsub2 = 1 when Sublevel = “4”, (otherwise Dsub2 = 0) and
Dtype = 1 when Type = “t” or Type = “h”, (otherwise Dtype = 0)

1) Assumption check and problem diagnosis

There are multiple aspects of the model need to be checked to see if the model is a good representative of this dataset.

First of all, the Adj-$R^2$of this model is 0.7996, which indicates its overall performance is good.

Then, we need to perform a Goodness-of-Fit Test:

H_0: $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0$

H_a: At least one coefficient $\beta_j \neq 0$

Since the test statistic F = 1043.86 and the P-value of F is less than 0.001, the null hypothesis of no association between dependent variable LnPrice and x variables is rejected, and there is at least one coefficient $\beta_j$doesn't equal to 0.

The Q-Q plot (see Appendix B.24) illustrates a straight line, therefore normality assumption is satisfied.

In most residual plots (see Appendix B.24), points scatter within a band centered around the horizontal line, linearity assumption is satisfied. However, points in most plots scatter randomly, except plots of Ndis and Ndis_Dtype, which appear a likely funnel shape. Independent assumption and linearity assumption are slightly violated. Ndis need to be checked see if log-transformation is needed.

Other than that, there are a few outliers exceeds ±3 in all residual plots needing to be checked. Through producing Cook's D plot (Appendix B.25), I identified four observations are both outliers and influential points (Appendix B.26): observation 2440, 2547, 2722 and 2825. There are also around 10 influential points of which Cook's Distance are greater than 4/3138(=0.001), especially observation 7, its D value is 0.077.

2) Apply log-transformation and deal with influential points.

To solve funnel pattern in the residual plot of studentized residual vs Ndis, I applied log-transformation on Ndis and created LnNdis and LnNdis_Dtype, then reran the model using LnNdis. The Adj-R2 of the new model decrease to 0.7899, and the residual plots didn't improve either. Considering the performance of this new model is worse off, Log-transformation on Nids seems totally unnecessary, the funnel pattern in residual vs Ndis plot is not very obvious, and the former model performs well, I decided to keep Ndis and Ndis_Dtype.

To solve problem of influential points and outliers, I printed out five most influential observations and investigated them carefully, then I decided that the values of observation 7 are reasonable, could be kept, and other four were problematic values (for instance, observation 2440 has a building area greater than land size) and needed to be removed.

3) The variance inflation factor statistics of three variables Ndis, Ndis_Dtype and Dtype respectively are 31, 71and 57 (greater than the cutoff value of 10), this suggests there are multicollinearity among those variables.

After investigating the variable, I noticed that the VIF value for Dtype is high probably is because Dtype is a dummy variable, and it distinguishes townhouse type and house type from unit type, and the portion of units is very small. This will not affect the regression, so I decided to ignore it.

As for Ndis and Ndis_Dtype, the high VIF values highly likely is due to the fact that Ndis_Dtype is the product of Ndis multiplies Dtype. Nevertheless, I centered the Ndis with its mean and created MNdis and interaction term MNids_Dtype. Reran the regression, VIF values for MNdis, MNids_Dtype and Dtype became much lower (31, 46 and 27), but were still way more above the cutoff value 10. The model didn't change fundamentally.

Considering that the correlation coefficient r between the two variable is actually only 0.48, along with the reasons mentioned above, I chose to ignore it as well.

4) Reran the linear regression with all 12 features on the processed data with 3134 observations, I got the new full model F2:

LnPrize =13.19 + 0.03Rooms + 0.06Bathroom + 0.005Car - 0.02Ndis + 0.24Dsub

+ 0.45Dsub1 + 0.74Dsub2 - 1.00Dtype + 0.26Area + 0.03Land

+ 0.03Tyears + 0.03Ndis_dtype

Where Dsub = 1 when Sublevel = "2", (otherwise Dsub = 0) and

Dsub1 = 1 when sublevel = "3", (otherwise Dsub1 = 0) and

Dsub2 = 1 when Sublevel = "4", (otherwise Dsub2 = 0) and

Dtype = 1 when Type = "t" or Type = "h", (otherwise Dtype = 0)

Up to now, constant variance assumption, independence assumption, linearity assumption and normality assumption are all satisfied. Overall performance is improved (Adj-R2= 0.8013 vs 0.7996). F values is bigger (1053.92 vs 1043.86), and P-value of it is less than 0.001. RMSE is smaller (0.2217 vs 0.2232). No outliers could be observed in residual plots. The P value of Car for T statistic is 0.39, Car is insignificant in this full model, this will be solve during model selection process.

Model Selection and Validation

In order to test the performances of the model, see how well they can predict new data from processed housing features, I used PROC SURVEYSELECT to randomly split the original 3134 samples into training set (2069) and test set (1025) with a ratio of 0.66 vs 0.34 (Appendix B.27).

1) Given the data set has been divided, I used Forward model selection method and CP model selection method to implement model selection procedure on the training set simultaneously.

The Forward model selection procedure performed 12 steps and finally included all 12 predictors, including an insignificant variable Car (Appendix B.28).

The CP model selection procedure computed CP values and R2values (Appendix B.29) for different combination of predictors, in which I considered a 11-predictor model as the best combination, which dropped Car, with a CP value of 11.74 and a R2 value of 0.8111.

I reran the linear regression with the two predictor combinations suggested by the two model selection procedures on the training data, obtained these two models (Appendix B.30, B.31):

Model S1 which is selected through Forward method:

LnPrice = 13.152 + 0.032Rooms + 0.059Bathroom + 0.005Car - 0.024Ndis
+ 0.251Dsub + 0.458Dsub1 + 0.754Dsub2 - 0.962Dtype + 0.276Area
+ 0.033Land + 0.026Tyears + 0.031Ndis_Dtype

Where Dsub = 1 when Sublevel = "2" and
Dsub1 = 1 when sublevel = "3" and
Dsub2 = 1 when Sublevel = "4" and
Dtype = 1 when Type = "t" or Type = "h"

Model S2 which is selected through Forward method:

LnPrice = 13.162 + 0.032Rooms + 0.060Bathroom - 0.024Ndis + 0.251Dsub
+ 0.457Dsub1 + 0.753Dsub2 - 0.969Dtype + 0.277Area
+ 0.034Land + 0.025Tyears + 0.031Ndis_Dtype

Where Dsub = 1 when Sublevel = "2" and
Dsub1 = 1 when sublevel = "3" and
Dsub2 = 1 when Sublevel = "4" and
Dtype = 1 when Type = "t" or Type = "h"

2) Performance comparison of Model S1 and Model S2

I computed performance metrics either in SAS or by hand, all results are summarized in the following Table 1: Comparison of Model S1 and Model S2(also see Appendix B.30 - B.35).

Looking at their performances on training set, the two models have the same adj-r2, RMSE values are almost the same, both Good-of-Fit tests reject the null hypotheses; residual plots suggest both of them satisfy all model assumptions; both models appear multicollinearity issue due to dummy variable or interaction term reason and can be ignored.

The biggest difference is that Model S1 has one more predictor than Model S2 (12 vs 11) and it is not significant. In addition, the F value of Model S1 is greater than the one of Model S2 (802.86 vs 735.92). From this aspect, Model S2 is better than Model S1.

Looking at their performances on test set, except that Adj-R2of Model S2 is slight higher than the one of Model S1, the other four values, RMSE, MAE, R2and CV R2, of Model S2, are all slightly lower. However, the differences are very small, basically negligible.

Finally, compare each model's performances on training set and test set, all differences of two models pertinent to Adj-R2and RMSE are very small, which is good. Besides that, the differences of two models are so small and negligible.

Overall, the two models have very close performances on training data and test set, meanwhile they do not have issue that affects the regression adversely. Nevertheless, Model S2 has fewer predictors and all of them are significant, so Model S2 is a better model and the final model for this section.

Table 1: Comparison of Model S1 and Model S2

Model Name	**Model S1**	**Model S2**
Performance on training set		
Number of Predictors	12 Predictors	11 Predictors
Predictor significance	Predictor Car is not significant	All predictors are significant
Adj-R2	0.8101	0.8101
F value	735.92	802.86
P value for F statistic	< .0001	< .0001
RMSE	0.21657	0.21656

Multicollinearity	VIF values of Ndis, Dtype and Nids_Dtype are high, but can be ignored	VIF values of Ndis, Dtype and Nids_Dtype are high, but can be ignored
Constant variance assumption	Satisfied	Satisfied
Independence assumption	Satisfied	Satisfied
Linearity assumption	Satisfied	Satisfied
Normality assumption	Satisfied	Satisfied
Outliers and influential points	There are a few outliers around ±3 but not a big concern.	There are a few outliers around ±3 but not a big concern.
Performance on test set		
RMSE	0.23179	0.23198
MAE	0.18124	0.18124
R2	0.885492 = 0.7840	0.885302 = 0.7838
Adj-R2	0.7814	0.7815

CV R2	0.0261 < 0.3	0.0263 < 0.3
Difference in Performance on training set and test set		
Difference in Adj-R2	0.8112-0.7840 = 0.0272	0.8111-0.7838 = 0.0273
Difference in RMSE	0.0152	0.0154

3) Rebuilt a model without interaction term

Although the overall performance of Model S2 are good, and no concerning issues are identified, there are still two problems which I considered as unsatisfactory. First, the model has 12 predictors, which might make the model computationally expensive. Second, in this model, the estimated coefficient of Ndis is negative (-0.024). Even though computing it along with the estimated coefficient (0.031) of its product, Ndis_Dtype, the result is still negative. To be more specific, that means 1 kilometer increase in distance between the property and the defined boundary which is centered at CDB and has a radius of 50 kilometers, results in 2.37% decrease in the price of townhouses and houses. which is totally counterintuitive.

Zietz, N. Zietz and Sirmans (2007) argue that "Results differ across (OLS regression) studies, not only in terms of size of OLS coefficients and statistical significance, but sometimes in direction of effect."[2]Their study "suggests that some of the observed variation in the estimated prices of housing characteristics may reflect the fact that characteristics are not priced the same across a given distribution of house prices."[2]This could be one of the possible reasons behind the model.

For the purpose of solving those problems, I decided to rebuild a model base on the original variable Distance without interaction term and Car, which I thought less important. First, I rebuilt a full model and generated residual plots to observe potential problem. The new full Model F3 is as shown below (see Appendix B.37):

LnPrice = 12.189 + 0.040Rooms + 0.061Bathroom - 0.005Distance + 0.245Dsub

+ 0.457Dsub1 + 0.731Dsub2 + 0.318Dtype + 0.270Area

+ 0.030Land + 0.029Tyears

Where Dsub = 1 when Sublevel = "2" and

Dsub1 = 1 when sublevel = "3" and

Dsub2 = 1 when Sublevel = “4” and

Dtype = 1 when Type = “t” or Type = “h”

Check the metrics and residual plots (see Appendix B.38), all four assumptions are all satisfied; overall performance is 0.7930; F values is 1201.44, and P-value of it is less than 0.001; RMSE is 0.2262; no outliers were observed; all predictors are significant in this full model.

After that, I applied CP model selection procedure on the training set and picked the best combination in the suggested list, and reran the regression on the training set. Below is the new model S3 (see Appendix B.39):

LnPrice = 12.179 + 0.037Rooms + 0.057Bathroom - 0.005Distance + 0.254Dsub

+ 0.467Dsub1 + 0.744Dsub2 + 0.319Dtype + 0.284Area

+ 0.029Land + 0.028Tyears

Where Dsub = 1 when Sublevel = “2” and

Dsub1 = 1 when sublevel = “3” and

Dsub2 = 1 when Sublevel = “4” and

Dtype = 1 when Type = “t” or Type = “h”

Check all metrics and residual plots, all assumptions are satisfied, no issues are identified.

4) Compare Model S3 and Model S2, decide the final model

From *Table 2: Comparison of Model S2 and Model S3*below, we can see that both of the models perform well, no issues are identified (ALSO see Appendix B.31, B.33, B.35, B.39, B.40).

Model S2 does slightly better in term of most of the metrics, including $R^2$and Adj-R^2 on both training set and test set, RMSE on both sets, MAE and CV R^2. Model S3 has smaller F value and one less predictors. Overall, the performances of the two models are very close.

However, when looking at the equation of Model S3, the estimated coefficient of Distance is negative, which means the distance between the property and CDB and the property price are negatively correlated, which makes much more sense. Therefore, I decided to choose Model S3 to be the final model of this study.

Table 2: Comparison of Model S2 and Model S3

Model Name	**Model S2**	**Model S3**
Performance on training set		

Number of Predictors	11 Predictors	10 Predictors
Predictor significance	All predictors are significant	All predictors are significant
Adj-R^2	0.8101	0.8022
F value	802.86	839.91
P value for F statistic	< .0001	< .0001
RMSE	0.21656	0.22098
Multicollinearity	VIF values of Ndis, Dtype and Nids_Dtype are high, but can be ignored	None
Constant variance assumption	Satisfied	Satisfied
Independence assumption	Satisfied	Satisfied
Linearity assumption	Satisfied	Satisfied
Normality assumption	Satisfied	Satisfied
Outliers and influential points	There are a few outliers but not a big concern.	There are a few outliers but not a big concern.

Performance on test set		
RMSE	0.23198	0.23664
MAE	0.18124	0.18505
R^2	$0.88530^2 = 0.7838$	$0.88031^2 = 0.7749$
Adj-R^2	0.7815	0.7728
CV R^2	$0.0263 < 0.3$	$0.0283 < 0.3$
Difference in Performance on training set and test set		
Difference in Adj-R^2	0.8111-0.7838 = 0.0273	0.8022-0.7749 = 0.0273
Difference in RMSE	0.0154	0.0157

Final Model and Interpretation

The final regression model is:

LnPrice = 12.179 + 0.037Rooms + 0.057Bathroom - 0.005Distance + 0.254Dsub
+ 0.467Dsub1 + 0.744Dsub2 + 0.319Dtype + 0.284Area
+ 0.029Land + 0.028Tyears

Where Dsub = 1 when Sublevel = “2” and

Dsub1 = 1 when sublevel = “3” and

Dsub2 = 1 when Sublevel = “4” and

Dtype = 1 when Type = “t” or Type = “h”

1) According to standardized estimate values (see Appendix B.39), for this model, the most important predictor is Dsub2 (standardized estimates is 0.70). The other predictors sorted on the level of importance are: Dsub1 (0.41), Area (0.37), Dsub (0.20), Tyears (0.20), Dtype (0.16), Land (0.14), Bathroom (0.08), Distance (-0.08), Rooms (0.06).

The estimated coefficient of Distance is negative, which indicated Distance negatively associated with the housing price.

2) Interpretation of all parameter estimates:

Intercept:β_1=12.179, meaningless without the meaningful values of other variables in this model, because there would never be a case in which all predictors are 0 at the same time.

Dsub2: the estimated coefficient of which is β_7=0.753, means that comparing to residential properties located in low-ranked suburbs, the prices of a residential property located in highly-ranked suburbs increase by 110.43%.

Dsub1: the estimated coefficient of which is β_6=0.457, means that comparing to residential properties located in low-ranked suburbs, the price of a residential property located in higher-middle-ranked suburbs increased by 59.52%.

Area: the estimated coefficient of which is β_9=0.277, means that 100 m^2increase in building area of the property result in 32.84% increase in the price of the property.

Dsub: the estimated coefficient of which is β_5=0.251, means that comparing to residential properties located in low-ranked suburbs, the price of a residential property located in lower-middle-ranked suburbs increased by 28.92%.

Tyears: the estimated coefficient of which is β_{11}=0.025, means that 10 years increase in the years from the year in which the property was built to 2019 results in 2.84% increase in the price of the property.

Dtype: the estimated coefficient of which is β_8=0.969, means that comparing to the price of unit, the price of townhouse and house increase 37.58%.

Land: the estimated coefficient of which is β_{10}=0.034, means that 100 m^2 increase in the land size of the property results in 2.94% increase in the price of the property.

Bathroom: the estimated coefficient of which is β_3=0.006, means that 1 bathroom increase results in 5.87% increase in the price of the property.

Distance: the estimated coefficient of which is β_4=-0.024, means that 1 km increase on the distance between the property and the boundary which is centered at CDB and has a radius of 50 kms results in 0.50% decrease in the price of the property for townhouses and houses.

Rooms: the estimated coefficient of which is β_2=0.032, means that 1 room increase results in 3.77% increase in the price of the property.

Prediction with C.I. and P.I.

Prediction Case 1:

Assume that House A has 3 rooms and 2 bathrooms, was built in 1979 and located in a highly-ranked suburb. The land size of it is 300 m^2, building area is 150 m^2, and the distance between the house and CBD is 8 kilometers.

After applied prediction procedure in SAS with Model S3 (see Appendix B.41, B.42), I obtained these results:

The predicted LnPrice is 14.0487, with a 95% confidence interval of (14.0293 ~ 14.0681) and a 95% prediction interval of (13.6065 ~ 14.4928).

Transform those numbers back to dollars, the results are:

The predicted price of House A is 1,262,621 AUD, the approximate 95% confidence interval for which is 1,238,362 AUD ~ 1,287,355 AUD, and the 95% prediction interval for which is 811,354 AUD ~ 1,968,535 AUD.

Prediction Case 2:

Assume that House B has 3 rooms and 2 bathrooms, was built in 1979 and located in a lower-ranked suburb. The land size of it is 300 m^2, building area is 150 m^2, and the distance between the house and CBD is 8 kilometers.

The results obtained from prediction procedure in SAS are (see Appendix B.43):

The predicted LnPrice is 13.3178, with a 95% confident interval of (13.2892 ~ 13.3463) and a 95% prediction interval of (12.8732 ~ 13.7623).

Transform those numbers back to Price in dollars, the results are:

The predicted price of this house is 607,921 AUD, the approximate 95% confidence interval for which is 590,781 AUD ~ 625,496 AUD, and the 95% prediction interval for which is 389,726 AUD ~ 948,180 AUD.

We can see that both of the predicted values fall right at the middle the confidence interval, which also proves the model is good.

Findings

From the two cases above, we can see that for houses with exactly the same condition, the prices could have a huge difference (1,262,621 AUD - 607,921 AUD = 654,700 AUD) only because they are located in different suburbs. House A is located in highly-ranked suburb while House B is located in lower-ranked suburb, but the difference between the predicted prices are every significant.

Future Work

This Melbourne housing data set has high quality. It contents the latest data and includes most important housing features commonly related to housing price. The only drawback is that it does not have information about the latest renovation, which I think could've explained why some properties are more expensive than properties under better conditions. This could somehow disturb the effectiveness of the regression model.

Other than that, a fact about housing pricing is that the same housing characteristics may not have the same weight for properties priced at different levels. This aspect is hard to be represented in ordinary least squares regression, and Quantile regression seems to be a more appropriate method to better solve this problem, which is what I am interested next.

Therefore, although multiple regression is still a traditional and classic method to study housing price, and can achieve quite high estimation accuracy and prediction accuracy, there are many more other methodologies, such as Hedonic regression method, Neutral Network method, Random Forest regression and so on, are intriguing to be study with as well.

Reference

1.Soley-Bori, Marina. 2013. Dealing with missing data: Key assumptions and methods for applied analysis. *Technical Report*#4. www.bu.edu/sph/files/2014/05/Marina-tech-report.pdf. Accessed May 6, 2013

2.Sirmans, G. Stacy. Zietz, Emily N. Zietz, Joachim. Determinants of House Prices: A Quantile Regression Approach. *The Journal of Real Estate Finance and Economics*. February, 2008

Section C (Nida Rasool)

Methodology

The first step taken before any analysis could be done was the pre-processing and cleaning of the data. The original dataset was a csv file containing around 65,000 observations and 21 columns. Much of the preprocessing was done using excel tools, beginning with deleting those columns which presented no useful data. For example, the address column which does not present useful data for the analysis. And columns pertaining to the name of the real estate agent who sold the property, or the neighborhood the property was in were deleted because they presented too many unique values to recode into dummy variables. The next step was cleaning the data by deleting those observations with empty values, or impossible values (ex: 0 for landsize) for any of the possible predictors. This all brought the number of observations down to around 8500. After this some columns presented what could possibly be useful data but what wasn't useful data yet. The date the property was sold was given in one column and the year the property was built was given in another column. This was enough to give the age of the property when sold which seemed like a useful predictor. Thus using a python script the csv file was scraped for just the year from the entire date given. Then using excel the year built year sold was subtracted from the year sold in order to determine HouseAge which is the age of the property at time of sale. The final step was making the sample size more manageable to work with for analysis. This was done using excel to choose around 1500 (actual: 1499) observations randomly. This final dataset could then be used for analysis in SAS. This data was then explored, transformed, and validated in order to determine the final model.

Analysis

The data was initially explored in SAS, to set the data up to be analyzed the housing type variable was recoded into a dummy variable. Three values were present in the data 'h' for house, 't' for townhouse, and 'u' for unit. Because over 70% of the data was houses (h) and very few were 't' the dummy variable was coded so that if house type was 'h' then then d_type = 1 otherwise d_type = 0. After this a histogram was created for the data. As can be seen in C.1 the histogram is very right skewed. This was to be expected as property prices in a big city usually center around an average but have high outliers as there are properties which can be very expensive but there aren't many properties which are sold for much below the average price. Essentially property values cannot go as low as they can high leading to an expected right skew in the data. The REG procedure was then performed on this data, where Price was set as the dependent variable, to determine the p-values. As can be seen in C.2 Rooms, Bedroom2, Landsize, and PropertyCount had p-values of >.05 these predictors were removed from the model as they could not be determined to be having a significant effect on the price. The R-Square is .5721The initial model after this is:

Price(AUS $) = $125,058(d_type) - $29,474(distance) + $332,681(Bathroom) + $39,219(Car) + $3,211.33(BuildingArea) + $6,273.77(HouseAge) - $53,449

The R-square of this model is .5701, The change is small even as 4 predictors were deleted. The residual plots (seen in C.3) show clear issues with the data. The quantile vs residual plot displays a clear s-curve rather than a straight 45 degree pattern as does the normal probability plot. The residual plots also display gathering of points around the center line for Building area, HouseAge, and Property Count while no clear shape can be determined the scattering is not completely random. The Distance residual graph also shows a funnel shape. The reg procedure also displayed those points in the data which were influential or outliers. A new dataset was then created where these points were deleted.

Because of all the skew in the data it was determined a transformation was need using the log function. A log transformation was applied to the dependent variable price in order to create an adjusted variable "lnprice" and a new dataset. The histogram generated with this new adjusted variable was much more normal in distribution. Another REG procedure was then performed on the new model with the new dataset without outliers, those predictors previously determined insignificant eliminated, and a new Y variable, lnprice. This resulted in the residual plots being much more satisfactory, the normal probability plot now had no S curve and had a straight 45 degree linear pattern. All predictors had a low p value of < .05 with all except for Car (number of parking spots available) having p value of of < .0001. The R-square value also increased from .57 in the initial model to now .63 in this newer model, and the RMSE decreased to around .3 (Figure C.4). The model after this step is:

lnprice = .169(d_type) - .0246(distance) + .181(bathroom) + .01654(Car)
+ .00295(BuildingArea) + .00496(HouseAge) + 12.912

The next step was to perform a model validation. The dataset was split into two parts, the training set and the test set. Every model selection was tested with the training set to visualize differences between each model. Proc surveyselect was used in order to split the dataset into the training and test set with a 75%, 25% division. Four model selections were chosen, adjrsq, stepwise, backward, and CP. Then the assumptions were computed for all four model selections in the training set. The model from the stepwise selection and the CP have studentized residual plots with the given predictors which display a higher level of random scatter. The normal probability plot has a 45 degree angle pattern. Though there are still plots with clearly visible influential points. While there are influential points visible in the probability plot they were not eliminated so as to preserve the size of the dataset. The rest of the residual plots are mostly similar with slight variations in the F values.

Both the stepwise selection method and the CP selection method suggest that Car, is not significantly effective as a predictor. While its p-val is <.05 it does not make a significant enough change to the R-sq value to keep it. Reference figures C.5 and C.6 to see how in both the stepwise selection method and the CP selection method the display that the Car predictor only adds .01 to the R-sq for the model. Thus car was eliminated for the final model which now has 5 predictors and is determined to be:

lnprice=.18(d_type)-.024(distance)+.181(Bathroom)+.003(BuildingArea)+.005(HouseAge)+12.930

For the overall Goodness of Fit Test in this final model, the null hypothesis is rejected. The F-Value is 348.55, the P-Value is < 0.0001, and the final R-Square is .6150. (Figure C.6)
As the P-value of the final model is much smaller than the .05 we are able to reject the null hypothesis. There is good support for the final model and we can conclude that at least one of the predictors has a statistically significant effect on price.

For the first prediction price will be the y variable and both the confidence interval and prediction interval of properties in Melbourne will be determined using the final model. The prediction will be for a property that is a house (d_type=1), 2 km from the central business district, has 2 bathrooms, a building area of 150 sq m and a house age of 50. The predicted value for the price of the house (figure C.7) is exp(14.1110) or around 1,343,783 AUS dollars with a 95% confidence interval between to (1,297,694 AUS dollars and 1,391,510 AUS dollars) and a 95% prediction interval between (742,516 AUS dollars 2,431,940 AUS dollars). The confidence interval has a relatively narrow range which is a good sign for the final model (Std Error Mean Predict = .0178).

The second prediction will be the same as the first with different data fed to the model. The prediction will be for a property that. The prediction will be for a property that is not a house (unit or townhouse) (d_type=0), .2 km from the central business district, has 1 bathroom, a building area of 80 sq m and a house age of 10. The predicted value for the price of the house (figure C.8) is exp(13.3904) or around 653,697 AUS dollars with a 95% confidence interval between to (620,822 AUS dollars and 688,382 AUS dollars) and a 95% prediction interval between (360,771 AUS dollars 1,184,581 AUS dollars). The confidence interval again has a relatively narrow range which is a good sign but it can be see the prediction interval has a larger range(Std Error Mean Predict = .0263).

Future Work

In the future this model could be greatly improved by better understanding the technicalities of certain variables. To expand on this look to house age, intuitively one would think that the older

a property the less expensive it would be, but this wasn't necessarily true in this data. That may be because of the evolution of property development in Melbourne or it may be because many of the much older properties are more likely to be rehabbed. Whether a property had undergone renovation or not was not present in this data. Also price per sq m would be a useful parameter for data like this. Price per sq ft is often seen on real estate websites (zillow.com, redfin.com, etc.). An apartment unit downtown may be very small and cost 300,000 AUS dollars but a much larger house in a suburb of Melbourne may be 500,000. Even though the house in the suburb is more expensive and houses in the area may be consistently more expensive than an apartment unit downtown doesn't mean that the suburb is a more expensive area to live. The interaction between size of the living area, location, and price should be explored more in a future model in order to develop a more comprehensive understanding of property prices. Property in big cities like Melbourne have many factors at play and being able to explore how those factors are affected by the particular city and with each other would work to build a better predictive model.

References

Luo, Z. Q., Liu, C. and Picken, D. (2007) Hous- ing price diffusion pattern of Australia's state capital cities, *International Journal of Strategic Property Management*, 11, pp. 227–242. Accessed [online] March 7, 2019.

Dataset compiled by Anthony Pino - https://www.kaggle.com/anthonypino/melbourne-housing-market

Section D (Sana Amreen)

METHODOLOGY

Data Cleaning

There are a lot of missing and junk values in the data set which we need to cleanse. For Eg., Price, Bedroom2 and Bathroom have quite a lot of "NA" values (missing) and Landsize has a lot junk values and so on, which are obviously too less to build a house on.So, to cleanse our dataset, we do the following step Filter out the NA values in the columns - Price, Bedroom2, Bathroom and Car.

Number of observations: 1090.

The URL to the site:https://www.kaggle.com/anthonypino/melbourne-housing-market

Data exploration

After the data was cleaned 1090 observations was prepared for analysis.the descriptive analysis was done that may suggest a possible model that is adequate for fitting the data.i nteraction terms were introduced into the import statement, four sets of dummy variables were created for type of housing, regionname, method in which the house is sold, quarter in which the house was sold q1-q4-2016 quarters q5 2017 quarter. Import statement was created and the data was imported with interaction terms and dummy variables. scatter plots were developed which showed nonlinear relationship, the plots were not independent, not normal and dose not have covariance. histogram was developed which was right skewed even it had outliers. multicollinearity was performed and checked for outliers and influential points. Correlation analysis was run and results were taken Correlation procedure was performed and the output was analyzed was any correlation. The values need to be >0.9 to be any collinearity problem as per the output none of the values met that criteria so all the values are strongly related.

Analysis

Linear regression analysis was performed after the full model significant predictors were taken and final model was obtained . residual plots obtained were not linear. One method was implemented to do the analysis that is adjrsq method. The significant predictors were taken and checked for the influential points. three influential points were removed and the adjrsq was obtained too be high. Final model equation was obtained after running the final model. Effect of most important predictor on price was noted and mentioned. prediction was done using two set of numeric data given and predicted was obtained interms of clm and cli.

Since the histogram was right skewed and residual plots obtained from the studentized residuals were not normal log transformation was done.log(price) was created. then regression analysis was run with the ln_price initially full model was taken, significant predictors were obtained then final model was run.the data was then checked for outliers and influential points. Two sets of influential points were removed and again the model was rerun. The f value and adjrsq value were found to be more after removing the two influential points and outliers. Final model equation was fitted. Effect of most significant predictor was noted and exp value was obtained. the histogram was normal after log transformation. The predited plot and studentized residuals were found to be normal.

Predictions

prediction was done using two set of numeric data introduced into the given dataset. We will predict the value of price using clm and cli.

Model validation

Data validation was the next step. In this the data set was divided into test and training sets.sample rate and seed value was given. new_y variable was created which had values for training set and periods(.) value for the test set. Adjrsq method was performed and most important predictors were obtained and regression analysis was performed for the training sets. Predicted value called the yhat was obtained using new_y by the test set.

ANALYSIS

The data was collected, and cleaning was done to prepare data for analysis. Missing values were looked up and data with 1090 observations was prepared for analysis. Since, price is a dependent variable which is a quantitive variable we need to do linear regression to do the analysis.

At the beginning of the analysis dummy variables were created to define the qualitative variables so that they can be included in the analysis. The dummy variables were initially coded in the import statement of the process along with-it interaction terms were also created during importing of the data. Interaction terms proved to be efficient in the entire analysis process.The interaction terms were used in descriptive stage and showed a significant effect. Thus, they were included in the analysis process.

* dummy variables for type where d_type= (type='h') is used as a base;
d_type1= (type='t'); d_type2= (type='u');
* dummy variables for method where d_m= (method='S') is used as a base;
d_m1= (method='PI'); d_m2= (method='SP');d_m3= (method='VB');
* dummy variables for quarter where d_q=(quarter='Q1') is used as a base;
d_q1=(quarter='Q2'); d_q2=(quarter='Q3'); d_q3=(quarter='Q4'); d_q4=(quarter='Q5');
* dummy variables for regionname where d_r=(regionname='N') is used as a base;

d_r1=(regionname='S'); d_r2=(regionname='E'); d_r3=(regionname='W');

DistancePostcode=Distance*Postcode;
RoomsLandsize=Rooms*Landsize;
DistanceLandsize=Distance*Landsize;
PostcodeLandsize=Postcode*Landsize;
Bedroom2Landsize=Bedroom2*Landsize;
CarLandsize=Car*Landsize;
RoomsBuildingArea=Rooms*BuildingArea;
DistanceBuildingArea=Distance*BuildingArea;
LandsizeBuildingArea=Landsize*BuildingArea;
RoomsPropertycount=Rooms*Propertycount;
DistancePropertycount=Distance*Propertycount;
PostcodePropertycount=Postcode*Propertycount;
Bedroom2Propertycount=Bedroom2*Propertycount;
LandsizePropertycount=Landsize*Propertycount;
BuildingAreaPropertycount=BuildingArea*Propertycount;

Data exploration

After importing the data the next step was data exploration stage that includes descriptive that may suggest a possible model that is adequate for fitting the data. The descriptive analysis is shown in the appendix as D2. The histogram was run which is a part of data exploration stage which showed right skewed as shown in D3 in appendix.

Scatter plots were plotted each independent variable against price which are shown in D4 in appendix. The scatter plots are non-linear non-independence the spread is not even along the straight line. The scatter plots were not normal as the points do not spread constant across the line. Thus, the y variable (price) needs to be transformed. The transformation will be helpful. We need to do log transformation.

Thus, it shows that the transformation is needed. Correlation procedure was performed and the output was analyzed was any correlation. The values need to be >0.9 to be any collinearity problem as per the output none of the values met that criteria so all the values are strongly related. D5 appendix.The data was checked for multicollinearity. Tolerance<=0.1 indicates collinearity as per the output result some values are <=0.1 thus it proves that collinearity occurs in the data.D6 appendix.

The cause of collinearity might be that the data is not sufficient, and more data is needed to overcome the collinearity problem.

Analysis

In the next step regression analysis was run assumptions were made by the residual plots. The residual plots were not linear and not normal. The plots were non independent. The method used to run the regression analysis was linear regression after the full model the significant predictors were taken and again the linear regression model was run the result was noted.refer D7,8 in appendix for adjrsquare value for final and full regression model.

A variable selection method was choosen that is adjusted rsquare method which was used and significant predictors were noted from the method adopted that is adjusted rsquare method. The adjusted rsquare method was performed using all the predictors including the interaction terms. Significant predictors were noted and included in the final model.The significant predictors included in the final model are Distance Rooms Bedroom2 Car Landsize BuildingArea d_type1 d_type2 d_q1 d_q2 d_r1 d_r2 d_r3 d_m2 d_m3 DistancePostcode RoomsLandsize PostcodeLandsize Bedroom2Landsize RoomsBuildingArea DistanceBuildingArea LandsizeBuildingArea RoomsPropertycount Bedroom2Propertycount BuildingAreaPropertycount. The output generated f value of 136.91 and adjrsq value of 0.7607 so 76.07% of variation falls in price is explained by this model. Refer D9 in appendix.

Since f value is 136.91 and $p<0.001$ the model meets the criteria of the best fit model and thus depending on the value null hypothesis can be rejected and at least one predictor in the data is significant.

Diagnostics: The skewness in the histogram and the non-linearity, non-independence can be due to outliers and the influential points. Thus, influential points need to be taken care. D10
The next step is removing the outliers and the influential points that is 37,43,63,123,130. and rerun the model. Influential points need to be removed one by one or more than one at a time. In the first stage one influential point was removed that is 37 then the model is rerun for the result adjrsquare was recorded to be 0.7644 so 76.44% of variation in price is explained by this model. refer D11 in appendix. Later three points were removed 43,130,172 and again the model was rerun.

Later three points were removed 43,130,172 and again the model was rerun the adjrsquare was . The residual plot is not normal and studentized residual was not normal,non linear ,non independent.

When we try to fit in the model the model equation is as follows

Predicted value(price)= 303943+(-585789)*distance+(-128419)*rooms+198611*bedroom2+35583*car+(-9710.203)*landsize+9516.824*building area+(-239707)*d_type1+(-387953)*d_type2+(-121114)*d_q1+(-

86313)*d_q2+198659*d_r1+110815*d_r2+70218*d_r3+(-164457)*d_m3+195.5813*distancepostcode+255.704*roomslandsize+3.036*postcodelandsize+(-236.16)*bedroom2landsize+(-362.48)*roomsbuildingarea+(-652.116)*distancebuildingarea+4.104*landsizebuildingarea+13.79*roomspropertycount+(-0.1418)*buildingareapropertcount.

The strongest predictor in the model is d_r1 that is south region. that is when d_r1 is 1 the price of the house increases by $198659.this predictor has a significant effect on the price of the house.
Log transformation of y variable
By doing assumption and diagnostics as mentioned above log transformation needs to be done.a new variable was generated with log option log(price)=ln_price. Histogram was developed after developing the log(price) variable the histogram was found to be normal. D12

Gplots were constructed the scatter plots were not normal not independent.D13
Linear regression analysis was run using the full model the predictors were distance rooms postcode bedroom2 car landsize buildingarea propertycount d_type1 d_type2 d_q1 d_q2 d_q3 d_q4 d_r1 d_r2 d_r3 d_m1 d_m2 d_m3 DistancePostcode RoomsLandsize DistanceLandsize PostcodeLandsize Bedroom2Landsize CarLandsize RoomsBuildingArea DistanceBuildingArea LandsizeBuildingArea RoomsPropertycount DistancePropertycount PostcodePropertycount Bedroom2Propertycount LandsizePropertycount BuildingAreaPropertycount. D14

After the regression analysis significan predictors were taken into consideration and final model was rerun.Distance Rooms Car Landsize BuildingArea d_type1 d_type2 d_q1 d_q2 d_r1 d_r2 d_r3 d_m3 DistancePostcode RoomsLandsize PostcodeLandsize Bedroom2Landsize RoomsBuildingArea DistanceBuildingArea LandsizeBuildingArea BuildingAreaPropertycount.
After the final model for the log transformation the f value was recorded to be 151.62 and adjrsq Was found to be 0.7762 so 77.62% of variation in price is explained by this model. the plot of studentized residuals vs predicted value did not show any pattern. The normal distribution pattern graph was found to be normal after log transformation. refer D15 in appendix. All the predictors in the model were found to be significant.

The logtransformation was then checked for the influential points. A full model was run for the influential points.123 and 172 were regarded as the influential points and the model was rerun after removing the influential points. refer D16 in appendix.

After removing the two influential points the adjrsq value increased to 0.7921 so 79.21% of variation in price is explained by this model the value increased from 0.7762 f value increased to 194.64. thus, removing two outliers and influential points was a good decision and this model is a good fit model as f value is more and p value is <0.001. refer D17 in appendix.

This model is the best model and yes model is satisfied as it normal residual plot was a straight line. model is a good fit model as f value is more and p value is <0.001. the model dosent need further improvement.

Modelequation:
log(y)=13.034+(-0.8267)*distance+0.20662*rooms+0.02164*car+(-0.00372)*landsize+0.008*buildingarea+(-0.2393)*d_type1+(-0.48676)*d_type2+(-0.09152)*d_q1+(-0.06172)*d_q2+0.1575*d_r1+0.13545*d_r2+0.09281*d_r3+(-0.122)*d_m3+0.0002*distancepostcode+0.00008677*roomslandsize+0.00000116*postcodelandsize+(-0.00092818)*roomsbuildingarea+(-00032662)*distancebuildingarea+0.00000155*landsizebuildingarea+(-6.17547e-8)*buildingareapropertycount.
Room is the strongest predictor which has the maximum effect on the ln_price. when room increases by one number ln_price increases by exp(0.2066)=(1.23-1)*100=$23.(australian dollars)

predicted values

prediction was done using two set of numeric data introduced into the given dataset. We will predict the value of price using clm and cli.refere D18 in appendix

Predicted value was calculate for 3 rooms and 1 car and 2 rooms and 1 car.log transformation was done for the predicted value.

Predicted price value for house with 3 rooms and 1 car exp(13.5422)-1*100= 76085546% or 76 million(australian dollars)

Where cl mean=$66917340 or 67 million(australian dollars) to

$86501216% or 86.5 million(australian dollars)

Cl predict = $4754091 or 4.7million(australian dollars) to $126060107% or 126 million (australian dollars)

Predicted price for house with 2 rooms and 1 car

exp(13.3654)-1*100= $63764592 or 63.8 million(australian dollars)

Where cl mean=$565954794 or 566milliob(australian dollars) to $71368716 or 71.4million(australian dollars)

Cl predict = $38623350 or 38.6million(australian dollars) to $105252126%or 105.25million(australian dollars)

Model validation

The data set was divided into training and testing sample rate was 0.8.the seed value was 2519. The data got selected 1 for test 0 for training.refer fig:19 New variable new_y was created for training set where the test values were just periods that is (.) refer D19 in appendix .regression analysis was performed using adjrsquare method for training set the following variables distance rooms postcode bedroom2 car landsize buildingarea propertycount d_type1 d_type2 d_q1 d_q2 d_q3 d_q4 d_r1 d_r2 d_r3 d_m1 d_m2 d_m3 DistancePostcode RoomsLandsize DistanceLandsize PostcodeLandsize Bedroom2Landsize CarLandsize RoomsBuildingArea DistanceBuildingArea LandsizeBuildingArea RoomsPropertycount DistancePropertycount PostcodePropertycount Bedroom2Propertycount LandsizePropertycount BuildingAreaPropertycount.adjrsquare was found to be 0.7668 so 76.68% of variation in price is explained by this model
Significant parameters were noted as Distance Bedroom2 Car Landsize BuildingArea d_type1 d_type2 d_q1 d_q2 d_r1 d_r2 d_r3 d_m3 DistancePostcode RoomsLandsize PostcodeLandsize Bedroom2Landsize RoomsBuildingArea DistanceBuildingArea LandsizeBuildingArea
and final model was run using the regression analysis using price_tt dataset which was created while creating new_y and adjrsq was noted for training set as 0.7545 so 75.45% of variation in price is explained by this model.in the later step regression analysis was done for the test set and yhat value was predicted the adjrsq for test set was found to be 0.7666 so 76.66% of variation in price is explained by this model Difference between Observed and Predicted in Test Set,rmse and mae were calculated prediction values were noted for testing method as yhat. Wqhich is 0.83076.

final model equation for new_y for training set
=320150+(-584167)*distance+117174*bedroom2+32744*car+(-9485.867)*landsize+77241.36*buildingarea+(-233607)*d_type1+(-401627)*d_type2+(-122576)*d_q1+89332*d_q2+216722*d_r1+1476900*d_r2+93690*d_r3+(-160075)*d_m3+194.756*distancepostcode+(-0241.040)*bedroom2landsize+(-269.2399)*roomsbuildingarea+(-648.18)*distancebuildingarea+4.541*landsizebuildingarea.
The most significant predictor in the model was d_r2 that is east region.

future works

the current work done in the analysis helps in identifying that even adding the interaction term to the data which add significance to the analysis. intercation terms that were introduced had a significant effect in exploring, transforming and data validation as many interaction terms were included in the final model at every stage. Thus, they prove to have significant effect on price that is y variable. The x variables that were omitted from the dataset can be included in the future to get a different value of price for melbourne housing. The predicted value of price changes as x variable changes.The number of observations can be increased with more x variables in the dataset to predict changes in the predicted value for the prices of houses in Melbourne.

References

Dataset compiled by Anthony Pino - https://www.kaggle.com/anthonypino/melbourne-housing-market

Residential_Research_Melbourne_Q3-2017.pdf - A research paper of JLL Research

Chris Kohler .Five property market trends to expect in 2018 , www.domain.com.au

Section E (Wilson Wu)

Methodology

This report presents the finding that the variations of the Melbourne Housing Market that influence the property price and using regression to analyze and build the model to predict future property price.

Data Searching, Cleaning and Sampling

We find out dataset Melbourne Housing Market on Kaggle.com[1] because our team is interested in what is the influence in the Housing price. In the dataset, Melbourne Housing Market contains house sale price which is our dependent variable and other 21 independent variables from the Melbourne area in Australia. The total number of observations is 34858. Basically, Melbourne Housing Market dataset has considerable past behavior for us to predict what is the influence in the Housing price in Melbourne. During the cleaning and dealing with missing data, we use listwise deletion[2] in excel to remove entire observations containing one or more unknown, because our dataset which has 34858 observations has sufficient power to drop cases. In the sampling part, we use randomly select in excel and sample[3] 2000 observation into sample dataset.

Data exploration

We determine the quantitative and qualitative in the dataset and figure out the type of dependent value so that we can choose the analysis should be linear regression, logistic regression or polynomial regression. Using the histogram to analyze the distribution of dependent value and independent value, if not normal symmetric might need to apply a transformation on it. Some variables might not available us to use because of time series or location so we remove or irrelative variables. Some quantitative and qualitative have to create dummy variables so one variable will be the base for others. To create an Interaction term, we will have to check whether there is any interaction between the independent variables so that we can create new predictor into the model. Gplots, scatterplot, and Pearson Correlation Coefficients will be our tool to indicate how the association between dependent value and independent value and whether there is any linear correlation. Moreover, Pearson Correlation Coefficients, VIF, and Tolerance can find each independent value have any Multicollinearity so that we can decide to keep or to drop the independent. After Multicollinearity, we might have multiple full models, so we have to use parameter estimates to analyze how significant independent value is in the model. Testing P-Value and compare the standardized estimate which can identify statistically significant and the strongest predictor and using the goodness of fit test to maintain that whether y-variable has least one coefficient x-variable or not. Moreover, R-Square, Adj R-Sq, RMSE can measure how data concentration around the regression model line. Finally, create a normal probability plot to analyze whether the model is linear or curves. If the model has curves, the model should rebuild

to polynomial regression. Furthermore, standardized residuals vs predicted can indicate whether the model is spread constant and independence or not.

Model Validation

Splitting the sample dataset into the training set with 80% observation to estimate the final model so the sample size is 1600 observation. Testing set with 20% to test the predictive performance so the size is 400 observation. Using the training set to do model section, backward selection, and Adjusted R2 Selection, we compare each model by the value of R-Square and Adj R-Sq and select the most precise independent value for the model. If we have multiple models, we will also compare R-Square, Adj R-Sq, and RMSE to build our final model. Creating a normal probability plot to analyze whether the model is linear or curves and standardized residuals vs predicted can indicate whether the model is spread constant and independence or not. Moreover, if there any outlier or influence point in standardized residuals vs predicted. Import the outlier and influence graphic shows, to find which observation is outlier and influence and take further action or remove them. After the action, we would diagnosing the final model, and testing P-Value, which can identify statistically significant, and using the goodness of fit test to maintain that whether y-variable has least one coefficient x-variable or not. Moreover, checking whether R-Square, Adj R-Sq, and RMSE have any improvements. Applying the final model into the testing set to measure predictive performance. We print out the predictive value and the absolute difference value, absd, between observation value and predictive value so that we can analyze how good is the model present. Moreover, we will test how the correlation in the observation variable and predictive variable, and the square of the correlation value is our R-Square so that we can calculate the difference between R-Square training and R-Square testing. We also can use 5- fold cross validation and backward selection to drop insignificant variation in the original sample data set and using AIC, AICC, and ASE to evaluate the performance of the model. Moreover, we compare its R-Square and Adj R-Sq to final model or different model so that we measure the best model for predicting the price.

Estimation and prediction

Since we build our final model for Melbourne Housing Market, we try to merge some predictions into the original data set and apply them into the final model. we will predict the property price and calculate the confidence interval to analyze model performance.

Analysis, Results and Findings

The exploratory analysis of the data

In the first step of exploratory data, displaying the sample dataset to determine which variables are quantitative and which are qualitative (Appendix E.1 Description of Variables). Price is dependent value and quantitative value, so the linear regression would be the method to compute predictive analytics.

The second step, using a histogram to check each variable distributional from. In the dependent value, price, we find that the histogram of the distribution (Appendix E.2 Distribution of Price) shows skewed right, so we apply a transformation log () on the dependent value, ln_Price, to improve the distribution and linearity. (Appendix E.3: Distribution of ln_Price) Moreover, another histogram of the distribution in Landsize (Appendix E.4: Distribution of Landsize) which is our independent shows extremely skewed right, so we also transformation log () on Landsize to improve the distribution. (Appendix E.5: Distribution of ln_Landsize) The other independents show normal or not extreme skewed, so we do not apply the transformation on them.

Third Step, the sample dataset has two independent variables that need to create dummy variables. Type variable has six different qualitative value, so we choose br - bedroom to be baseline in type dummy and create 5 dummy variables.

Where d_Type_h = 1 for type = house, cottage, villa, semi, terrace and d_Type_ br = 0 for bedroom.
Where d_Type_u = 1 for type = unit, duplex and d_Type_ br = 0 for bedroom.
Where d_Type_ t = 1 for type = townhouse and d_Type_ br = 0 for bedroom.
Where d_Type_ dev site = 1 for type = development site and d_Type_ br = 0 for bedroom.
Where d_Type_ o res = 1 for type = other residential and d_Type_ br = 0 for bedroom.

Similarly, in method variable has 11 different qualitative value, so we choose S - property sold to be baseline in type dummy and create 10 dummy variables.
Where d_Method_SP = 1 for method = SP - property sold prior and d_Method_S = 0 for S - property sold.
Where d_Method_PI = 1 for method = PI - property passed in and d_Method_ S = 0 for S - property sold.
Where d_Method_PN = 1 for method = PN - sold prior not disclosed and d_Method_ S = 0 for S -property sold.
Where d_Method_SN = 1 for method = SN - sold not disclosed and d_Method_ S = 0 for S - property sold.
Where d_Method_NB = 1 for method = NB - no bid and d_Method_ S = 0 for S -property sold.
Where d_Method_VB = 1 for method = VB - vendor bid and d_Method_ S = 0 for S -property sold.
Where d_Method_W = 1 for method = W - withdrawn prior to auction and d_Method_ S = 0 for S -property sold.
Where d_Method_SA = 1 for method = SA - sold after auction and d_Method_ S = 0 for S - property sold.

Where d_Method_SS = 1 for method = SS - sold after auction price not disclosed and d_Method_ S = 0 for S -property sold.
Where d_Method_NA= 1 for method = N/A - price or highest bid not available and d_Method_ S = 0 for S -property sold.

Moreover, Because of Bedroom usually have effect on the output of Bathroom[4], so we create an Interaction term Bedroom2_ Bathroom.

In gplots and scatterplot (Appendix E.6 Scatterplot Matrix data) we can find that between ln_price and other dependent doesn't show significant association to be linear. However, in gplots, ln_price with Distance, Building Area and Room_bathrom (Appendix E.7 GPlot), has a slight association. We discover a little association. Similarly, we apply in Pearson Correlation Coefficients. The Pearson Correlation Coefficients value in each value doesn't show significantly. Correlation values of ln_price with BuildingArea are 0.51646 show that there is a medium association. Correlation values of ln_price with Rooms, Bedrooms, Bathrooms, and Rooms_Bathroom are around 0.42 so that there also show a medium association. On Car, ln_landsize, d_Type_t, d_Method_SP, d_Method_PI, d_Method_VB and d_Method_SA show almost around 0, so basically all these independent variables are no correlation or not strong relationship with the ln_price. (Appendix E.8 Pearson Correlation Coefficients)

Being insight into the Pearson Correlation Coefficients, we discover Rooms and Bedroom2 are highly correlated with each other. Their Correlation Coefficients are 0.95372. (Appendix A.8 Pearson Correlation Coefficients) Moreover, when we check the VIF and Tolerance of full regression model for ln_Price. (Appendix E.9 Full Regression Model Parameter Estimates) Rooms VIF value is 12.85571 which is over the normal value 10 and Bedroom2 VIF value is 17.91567. Independent variables of Rooms and Bedroom2 have Multicollinearity. When we go to the data sheet and discover that Rooms and Bedroom2 are the same and also in the description the number of rooms is similar. Based on these reasons, we decide to drop Independent variable Rooms.

In the new Pearson Correlation Coefficients, Multicollinearity problem improves and now only the Bathroom and Bedroom2_ Bathroom highly related to each other. Their Correlation Coefficients are 0.91366. (Appendix E.10 New Pearson Correlation Coefficients (without Rooms)) Furthermore, when we check the VIF and Tolerance of new full regression model for ln_Price. (Appendix E.11 New Regression Model Parameter Estimates (without Rooms)) we discover Bedroom2_Bathroom VIF value is 13.70834. Independent variables of Bathroom and Bedroom2_ Bathroom have Multicollinearity, but we decide not to drop each of them because Bedroom2_ Bathroom is Interaction term of Bedroom2 and Bathroom. Moreover, the Correlation Coefficient is 0.91366 just above 0.9 so we maintain that it is still useful in the full model. In addition, we create two more models which one has Bathroom but do not have

Bedroom2_ Bathroom (Appendix E.12 New1 Regression Model Parameter Estimates (without Rooms, Bedroom2_ Bathroom)) and another one has Bedroom2_ Bathroom but does not have Bathroom. (Appendix E.13 New2 Regression Model Parameter Estimates (without Rooms, Bathroom)) We can use these three models to figure out the best model for Analysis on Property Sales Price of Melbourne Housing Market.

Before doing the model selection and validation to get the final regression model, we use the t-test on every parameter to check the full regression model on each. Basically, they're not much different. The p-value for the t-test on the coefficient, for all three models, Car, d_Method_PI, d_Method_VB, and d_Method_SA values are over 0.05. Therefore, we cannot reject the hypothesis that Car, d_Method_PI, d_Method_VB, and d_Method_SA have no effect on the ln_price. And on d_Type_t, d_Type_devsite and d_Type_ores, the value did not present so that we cannot reject the hypothesis that d_Type_t, d_Type_devsite and d_Type_ores have no effect on the ln_price.

New Regression Model Parameter Estimates (without Rooms) (Appendix E.11 New Regression Model Parameter Estimates (without Rooms))

For this model, R-Square is 0.5631, so 56.31% of variation fall in ln_price explained by the model with every independent variation. Adj R-Sq is 0.5600, so 56% of variation fall in ln_price explained by the model with significant independent variations. Both R-Square and Adj R-Sq present this model is a medium model. Moreover, the RMSE is 0.34341which is close to 0, present the data is well concentrated around the linear regression line. Test on Goodness-of-Fit, H_0: $b_j = 0$ which suppose the hypothesis that the model is completely wrong, and no any x-variable are significant. However, parameter estimates of the model present that there are ten independent variations are significant, so H_a: $b_j \neq 0$. Furthermore, F- value is 182.73 and P-value is lower than 0.0001. Therefore, we reject H0 and conclude model has at least one independent variable that has a significant effect on the dependent.

Full Model 1: ln_Price =12.77640-0.03675*Distance +0.12306* Bedroom2+0.23433* Bathroom +0.06993*ln_Landsize+0.00185*BuildingArea-0.00000392*Propertycount+0.23901 *d_Type_h - 0.15267* d_Type_u - 0.06562 * d_Method_SP- 0.03304 *Bedroom2_Bathroom

Where d_Type_h = 1 for type = house, cottage, villa, semi, terrace and d_Type_br = 0 for bedroom.
Where d_Type_u = 1 for type = unit, duplex and d_Type_br = 0 for bedroom.
Where d_Method_SP = 1 for method = SP - property sold prior and d_Method_S = 0 for S - property sold.

On the normal probability plot (Appendix E.14 New full Regression Model Normal probability plot (without Rooms)) shows almost 45 degrees line without any significant curve, so it's linear and normal distributed. Moreover, the normal probability plot indicates that this model is multiple linear regression.

New1 Regression Model Parameter Estimates (without Rooms, Bedroom2_ Bathroom) (Appendix E.12 New1 Regression Model Parameter Estimates (without Rooms, Bedroom2_ Bathroom))

For this model, R-Square is 0. 5575, so 55.75% of variation fall in ln_price explained by the model with every independent variation. Adj R-Sq is 0.5547, so 55.47% of variation fall in ln_price explained by the model with significant independent variations. Both R-Square and Adj R-Sq present this model is a medium model. Moreover, the RMSE is 0.34549 which is close to 0, present the data is well concentrated around the linear regression line. Test on Goodness-of-Fit, H_0: $b_j = 0$ which suppose the hypothesis that the model is completely wrong, and no any x-variable are significant. However, parameter estimates of the model present that there are ten independent variations are significant, so H_a: $b_j \neq 0$. Furthermore, F- value is 192.51 and P-value is lower than 0.0001. Therefore, we reject H0 and conclude model has at least one independent variable that has a significant effect on the dependent.

Full Model 2: ln_Price=13.00940-0.03611*Distance + 0.06024* Bedroom2+0.10553*Bathroom +0.06998*ln_Landsize+0.00185*BuildingArea-0.00000415*Propertycount+0.23179*d_Type_h -0.18627* d_Type_u-0.18627* d_Method_SP

Where d_Type_h = 1 for type = house, cottage, villa, semi, terrace and d_Type_br = 0 for bedroom.
Where d_Type_u = 1 for type = unit, duplex and d_Type_br = 0 for bedroom.
Where d_Method_SP = 1 for method = SP - property sold prior and d_Method_S = 0 for S - property sold.

On the normal probability plot (Appendix E.15 New full Regression Model Normal probability plot (without Rooms, Bedroom2_ Bathroom)) shows almost 45 degrees line without any significant curve, so it's a linear and normal distribution. Moreover, the normal probability plot indicates that this model is multiple linear regression.

New2 Regression Model Parameter Estimates (without Rooms, Bathroom) (Appendix E.13 New2 Regression Model Parameter Estimates (without Rooms, Bathroom))

For this model, R-Square is 0.5492, so 54.92% of variation fall in ln_price explained by the model with every independent variation. Adj R-Sq is 0.5462, so 54.62% of variation fall in

ln_price explained by the model with significant independent variations. Both R-Square and Adj R-Sq present this model is a medium model. Moreover, the RMSE is 0. 34875 which is close to 0, present the data is well concentrated around the linear regression line. Test on Goodness-of-Fit, H_0: $b_j = 0$ which suppose the hypothesis that the model is completely wrong, and no any x-variable are significant. However, parameter estimates of the model present that there are ten independent variations are significant, so H_a: $b_j \neq 0$. Furthermore, F- value is 186.08 and P-value is lower than 0.0001. Therefore, we reject H0 and conclude model has at least one independent variable that has a significant effect on the dependent.

Full Model 3: ln_Price=13.13428-0.03591*Distance+0.05746*Bedroom2+0.06632* ln_Landsize +0.00203*BuildingArea-0.00000442 *Propertycount + 0.21083*d_Type_h -0.20637 * d_Type_u - 0.06522* d_Method_SP-0.01261*Bedroom2_Bathroom

Where d_Type_h = 1 for type = house, cottage, villa, semi, terrace and d_Type_br = 0 for bedroom.
Where d_Type_u = 1 for type = unit, duplex and d_Type_br = 0 for bedroom.
Where d_Method_SP = 1 for method = SP - property sold prior and d_Method_S = 0 for S - property sold.
On the normal probability plot (Appendix E.16 New full Regression Model Normal probability plot (without Rooms, Bathroom)) shows almost 45 degrees line without any significant curve, so it's linear and normal distributed. Moreover, the normal probability plot indicates that this model is multiple linear regression.

Model Validation
Using model validation to evaluate our model, because we want to test the model's prediction accuracy. We split the sample dataset into a training set with 80% observation to estimate the final model so the sample size is 1600 observation. Testing set with 20% to test the predictive performance. (Appendix E.17 Training and Test set) Therefore, in the training set, the observations that selected use new y to be our training set dependent value. (Appendix E.18 New training and test set in selected) In the model selection, to build a good model, we use backward elimination and Adjusted R2 Selection to compare and select. Basically, we compare by R-Square and Adj R-Sq, and the most precise independent value.

In New Regression Model Parameter Estimates in the Training set (without Rooms) (Appendix E.21 Regression Model Parameter Estimates md1 in training)

On the backward selection (Appendix E.19 Selection Method: backward Selection Method md1), it selects 10 independent variations which are Distance, Bedroom2, Bathroom, ln_Landsize, BuildingArea, Propertycount, d_Type_h, d_Type_u, d_Method_SP, and Bedroom2_Bathroom. However, the p-value for the t-test on the Propertycount is 0.0818 which are over 0.05. We

decide to keep it, because, in the full model, Propertycount is 0.038 which is lower than the 0.05. Basically, Propertycount is a significant independent. In the Adjusted R2 Selection (Appendix A.20 Selection Method: Adj-R2 Selection Method md1), there are 11, 12 ,10 variations in different model and all shows 0.5670 in Adj R-Sq. The difference between them just add d_Method_SA and Car, both of them didn't indicate significant in full model p-value, so we choose the 10 variations one. Moreover, we find one has the same model with backward selection. Therefore, we use the same one for the model.

For this model, R-Square is 0.5697, so 56.97% of variation fall in new_y explained by the model with every independent variation. Adj R-Sq is 0.5670, so 56.70% of variations fall in new_y explained by the model with significant independent variations. Both R-Square and Adj R-Sq present this model is a medium model. Moreover, the RMSE is 0.34329 which is close to 0, present the data is well concentrated around the linear regression line. Test on Goodness-of-Fit, H0: $b_j = 0$ which suppose the hypothesis that the model is completely wrong, and no any x-variable are significant. However, parameter estimates of the model present that there are ten independent variations are significant, so Ha: $b_j \neq 0$. Furthermore, F- value is 210.35 and P-value is lower than 0.0001. Therefore, we reject H0 and conclude model has at least one independent variable that has a significant effect on the dependent.

Full Model 1: new_y = 12.69342 -0.03776*Distance+0.14178*Bedroom2+0.25948* Bathroom +0.07537*ln_Landsize+0.00175*BuildingArea-0.00000355*Propertycount+ 0.26018*d_Type_h -0.13285* d_Type_u -0.08071* d_Method_SP - 0.03840*Bedroom2_Bathroom

Where d_Type_h = 1 for type = house, cottage, villa, semi, terrace and d_Type_br = 0 for bedroom.
Where d_Type_u = 1 for type = unit, duplex and d_Type_br = 0 for bedroom.
Where d_Method_SP = 1 for method = SP - property sold prior and d_Method_S = 0 for S - property sold.

In New Regression Model Parameter Estimates in the Training set (without Rooms, Bedroom2_ Bathroom) (Appendix E.24 Regression Model Parameter Estimates md2 in training)

On the backward selection (Appendix E.22 Selection Method: backward Selection Method md2), it selects 9 independent variations which are Distance, Bedroom2, Bathroom, ln_Landsize, BuildingArea, Propertycount, d_Type_h, d_Type_u, and d_Method_SP. However, the p-value for the t-test on the Propertycount is 0.0525 which are slightly over 0.05. We decide to keep it, because, in the full model, Propertycount is 0.038 which is lower than the 0.05. Basically, Propertycount is a significant independent. Furthermore, it just over 0.0025 which is so close to 0 so it is acceptable. In the Adjusted R2 Selection (Appendix E.23 Selection Method: Adj-R2

Selection Method md2), there are 11, 10, 9 variations in a different model. 11, and 10 variations models show 0.5590 in Adj R-Sq and 9 variations models indicate 0.5588 which the difference only has 0.0002, so 9 variations model presents an almost same model with 10 or 11 variations, which means 9 variations model are more efficient for predict. Moreover, we find one has the same model with backward selection. Therefore, we use the same one for the model.

For this model, R-Square is 0.5613, so 56.13% of variation fall in new_y explained by the model with every independent variation. Adj R-Sq is 0.5588, so 55.88% of variation fall in new_y explained by the model with significant independent variations. Both R-Square and Adj R-Sq present this model is a medium model. Moreover, the RMSE is 0.34650 which is close to 0, present the data is well concentrated around the linear regression line. Test on Goodness-of-Fit, H0: $b_j = 0$ which suppose the hypothesis that the model is completely wrong, and no any x-variable are significant. However, parameter estimates of the model present that there are ten independent variations are significant, so Ha: $b_j \neq 0$. Furthermore, F- value is 226.06 and P-value is lower than 0.0001. Therefore, we reject H0 and conclude model has at least one independent variable that has a significant effect on the dependent.

Full Model 2: new_y = 12.96775-0.03707*Distance+ 0.06965*Bedroom2+0.10563* Bathroom +0.07510*ln_Landsize+0.00176*BuildingArea-0.00000399*Propertycount+ 0.25055*d_Type_h -0.17187*d_Type_u -0.08150* d_Method_SP

Where d_Type_h = 1 for type = house, cottage, villa, semi, terrace and d_Type_br = 0 for bedroom.
Where d_Type_u = 1 for type = unit, duplex and d_Type_br = 0 for bedroom.
Where d_Method_SP = 1 for method = SP - property sold prior and d_Method_S = 0 for S - property sold.

In New Regression Model Parameter Estimates in the Training set (without Rooms, Bathroom) (Appendix E.27 Regression Model Parameter Estimates md3 in training)

On the backward selection (Appendix E.25 Selection Method: backward Selection Method md3), it selects 9 independent variations which are Distance, Bedroom2, ln_Landsize, BuildingArea, Propertycount, d_Type_h, d_Type_u, d_Method_SP, and Bedroom2_Bathroom. the p-value for the t-test on each independent is lower than 0.05 so that these 9 variations are significant. In the Adjusted R2 Selection (Appendix E.26 Selection Method: Adj-R2 Selection Method md3), there are 12, 11, 10, 9 variations in a different model. 11, 12, 10 variations models show around 0.5496 to 0.5494 in Adj R-Sq and 9 variations models indicate 0.5494 which the difference only has 0.0002, so 9 variations model presents an almost same model with 10, 11 or 12 variations, which means 9 variations model are more efficient for predict. Moreover, we find one has the same model with backward selection. Therefore, we use the same one for the model.

For this model, R-Square is 0.5519, so 55.19% of variation fall in new_y explained by the model with every independent variation. Adj R-Sq is 0.5494, so 55.88% of variation fall in new_y explained by the model with significant independent variations. Both R-Square and Adj R-Sq present this model is a medium model. Moreover, the RMSE is 0.35019 which is close to 0, present the data is well concentrated around the linear regression line. Test on Goodness-of-Fit, H0: $b_j = 0$ which suppose the hypothesis that the model is completely wrong, and no any x-variable are significant. However, parameter estimates of the model present that there are ten independent variations are significant, so Ha: $b_j \neq 0$. Furthermore, F- value is 217.62 and P-value is lower than 0.0001. Therefore, we reject H0 and conclude model has at least one independent variable that has a significant effect on the dependent

Full Model 3: new_y= 13.07309-0.03706*Distance+0.07761*Bedroom2+0.07269*ln_Landsize +0.00195*BuildingArea-0.00000428*Propertycount+0.22401*d_Type_h-0.19197*d_Type_u-0.08233* d_Method_SP+ 0.01002*Bedroom2_Bathroom

Where d_Type_h = 1 for type = house, cottage, villa, semi, terrace and d_Type_br = 0 for bedroom.
Where d_Type_u = 1 for type = unit, duplex and d_Type_br = 0 for bedroom.
Where d_Method_SP = 1 for method = SP - property sold prior and d_Method_S = 0 for S - property sold.

Comparing three Model, the Full Model 1 only without rooms has highest Adj R-Sq which is 0.5670 and R-Square which is 0.5697. Moreover, Full Model 1 indicates the lowest RMSE which is 0.34329. Furthermore, on the normal probability plot (Appendix E.28 Final Regression Model Normal probability plot (without Rooms)) shows almost 45 degrees line without any significant curve, so it's a linear and normal distribution and multiple linear regression. In standardized residuals vs predicted, it indicated that points are randomly scattered around the zero line. we can find that there are outliers over the +3 and -3 of the standardized residuals. (Appendix E.29 final model standardized residuals vs predicted)

The outlier and influence graphic shows that observations 103, 756, 982, 1232, 1385 and 1844 are simultaneously outlier and influence point, so we first remove them and re-run the model. Adj R-Sq becomes higher to 0.5819 and also R-Square improves to 0.5845. (Appendix E.30 Final Regression Model 1 in training (First remove)) Then, new outlier and Influence graphic indicate that observations 452 and 1617are in same outlier and influence point, so we remove them. Therefore, Adj R-Sq gains to 0.5862 and also R-Square improves to 0.5888. (Appendix E.31 Final Regression Model 1 in training (Second remove)) After re-running the model, finally, there only left one outlier point. It's acceptable to have one outlier in dataset since we already remove 8 observations.

The final Model Parameter Estimates in the Training set (Appendix E.32 Final Regression Model Parameter Estimates in training)

For this model, R-Square is 0.5888, so 58.88% of variation fall in new_y explained by the model with every independent variation. Adj R-Sq is 0.5862 so 58.62% of variations fall in new_y explained by the model with significant independent variations. Both R-Square and Adj R-Sq present this model is a medium model. Moreover, the RMSE is 0.33317 which is close to 0, present the data is well concentrated around the linear regression line. Test on Goodness-of-Fit, H0: $b_j = 0$ which suppose the hypothesis that the model is completely wrong, and no any x-variable are significant. However, parameter estimates of the model present that there is ten independent variations are significant, so Ha: $b_j \neq 0$. Furthermore, F- value is 226.42 and P-value is lower than 0.0001. Therefore, we reject H0 and conclude model has at least one independent variable that has a significant effect on the dependent.

Final Model: new_y = 12.77372-0.03723*Distance+ 0.12740*Bedroom2+ 0.24974* Bathroom +0.06141*ln_Landsize+0.00213*BuildingArea-0.00000428*Propertycount+0.26817 *d_Type_h -0.12032 * d_Type_u -0.08264* d_Method_SP-0.03834 *Bedroom2_Bathroom

Where d_Type_h = 1 for type = house, cottage, villa, semi, terrace and d_Type_br = 0 for bedroom.
Where d_Type_u = 1 for type = unit, duplex and d_Type_br = 0 for bedroom.
Where d_Method_SP = 1 for method = SP - property sold prior and d_Method_S = 0 for S - property sold.

The parameter for margin indicates that assuming every independent fixed, for any kilometers increase in distance, new_y or ln_price will increase by 3.7%. Similarly, we can say that for any bedroom increase in Bedroom2, new_y or ln_price will increase by 13.58%. For any bathroom increase in Bathroom, new_y or ln_price will increase by 28.36%. For any size increase in ln_Landsize, new_y or ln_price will increase by 6.33%. For any size increase in BuildingArea, new_y or ln_price will increase by 0.21%. For any number increase Propertycount, new_y or ln_price will increase by 0.0004%. For any type for a house is d_Type_h, new_y or ln_price will increase by 30.75%. For any type for a house is d_Type_u, new_y or ln_price will increase by 12.78%. For any property sold type for a house is d_Method_SP, new_y or ln_price will increase by 3.91%.

Variable distance has the largest absolute value of the standardized coefficient which is 0.48868 that has the greatest influence on ln_price, and the second variable is bathroom which standardized coefficient value is 0.36079. (Appendix E.32 Final Regression Model Parameter Estimates in training) On the normal probability plot (Appendix E.33 Final Regression Model Normal probability plot) shows almost 45 degrees line without any significant curve, so it's

linear and normal distributed. Moreover, the normal probability plot indicates that this model is multiple linear regression. In standardized residuals vs predicted, it indicated that points are randomly scattered around the zero line which spread constant and independence, so the final model is acceptable. (Appendix E.34 final model standardized residuals vs predicted)

Applying the final model into testing set to measure predictive performance. In the validation-test set (Appendix A.35 Difference between Observed and Predicted in Test Set), the absolute difference value, absd, between ln_price and yhat which are observed value and predict value has a mean of $ 0.276313 and a median of $ 0.250581. The minimum difference is $ 0.00134298. The Maximum is $ 1.40131241. Furthermore, In Normal distribution, Q1 = 0.11687964, Q3 = 0.39719138 The interquartile range is IQR = 0.39719138-0.1168796 = 0.28031178. (Appendix E.36 Descriptive of abs(d) in testing set)

For the final model apply to the testing set R-Square is 0.73236^2 = 0.5364, so 53.64% of variation fall in new_y explained by the model with every independent variation. |modelR2 - R2CV| = 0.5888- 0.5364 = 0.0524. Moreover, the RMSE is 0.34338 and Mae is 0.2763, both are low and close to zero so that it indicates that is well concentrated around the linear regression line. (Appendix E.37 Final model rmse, mae and Pearson Correlation Coefficients in testing set)

Five-fold cross validation + 25% testing set

On 5- fold cross validation, we use backward selection to drop insignificant variation in the original sample data set. Similarly, it selects same 10 independent variations which are Distance, Bedroom2, Bathroom, ln_Landsize, BuildingArea, Propertycount, d_Type_h, d_Type_u, d_Method_SP, and Bedroom2_Bathroom as the final model in 80% training set, but the Parameter Estimates value is slightly different. When model drop d_Method_PI has the highest Adj R-Sq This model shows AIC is -1736.30455 and also when model drop Car has the lowest AIC and AICC. (Appendix E.38 Fit Criteria for ln_Price) However, in the progression of ASE, when modeling drop d_Method_SA, the testing set square error decrease. (Appendix E.39 Progression of Average Squared Error ln_Price) Moreover, we compare its R-Square and Adj R-Sq to the final model. This model R-Square is 0.5595 and Adj R-Sq is0.5566 which both lower than the Final Model. Moreover, the RMSE is 0.34154 is bigger than the Final Model which is 0.33317. (Appendix E.40 5-fold cross validation Parameter Estimates)

Test on Goodness-of-Fit, H0: $b_j = 0$ which suppose the hypothesis that the model is completely wrong, and no any x-variable are significant. However, parameter estimates of the model present that there are ten independent variations are significant, so Ha:$b_j \neq 0$. Furthermore, F- value is 192.03 and P-value is lower than 0.0001. Therefore, we reject H0 and conclude model has at least one independent variable that has a significant effect on the dependent.

5-fold cross validation Model: ln_ price = 12.806859-0.037498*Distance + 0.134504*Bedroom2 + 0.218868*Bathroom + 0.066786* ln_Landsize + 0.001838* BuildingArea - 0.000004274* Propertycount +0.243519* d_Type_h-0.151383* d_Type_u - 0.057892 *d_Method_SP-0.03344* Bedroom2_Bathroom

Where d_Type_h = 1 for type = house, cottage, villa, semi, terrace and d_Type_br = 0 for bedroom.
Where d_Type_u = 1 for type = unit, duplex and d_Type_br = 0 for bedroom.
Where d_Method_SP = 1 for method = SP - property sold prior and d_Method_S = 0 for S - property sold.

Estimation and prediction

Merging two predictions into the original data set, we create observation one to predicted probability of 10 Kilometers distance from the major central business district, 3 Bedroom and 1 Bathroom in the house. ln_Landsize is 5.34 meters and building Area is 209 meters. 5682 of properties that exist in the suburb. The property type is house and is property sold prior. Observation two is 20 kilometers distance from the major central business district, 5 Bedroom and 2 Bathroom in the house. ln_Landsize is 6.87 meters and building Area is 280 meters. 8888 of properties that exist in the suburb. The property type is unit and is property sold. (Appendix E.41 Merge prediction dataset with original dataset)

In observation one, the predicted value for ln_price is 13.8351 so the property price is $1019782.57, and with a 95% C.L Mean of 13.7807 to 13.8895, which is $965788.36 to $ 107679. 44 and with 95% C.L Predict of 13.1596 to 14.5105, which is $518969.3 to $2003687.92. Therefore, there is a 95% chance that the confidence interval means contains the property price from $965788.36 to $ 107679. 44 and there is a 95% chance that the confidence interval predict contains the property price from $518969.3 to $2003687.92. In observation two the predicted value for ln_price is 13.6259 so the property price is $827281.25, and with a 95% CL Mean of 13.5462 to 13.7056, which is $763905.98 to $895914.26 and with 95% CL Predict of 12.9479 to 14.3038, which is $419953.8 to $1629526.44. Therefore, there is 95% chance that the confidence interval means contains the property price from $763905.98 to $895914.26 and there is 95% chance that the confidence interval predict contains the property price from $419953.8 to $1629526.44. (Appendix E.42 Output Predicted Value and C.L, P.L)

Future Work

As multiple regression analysis on Melbourne Housing Market shows, the final model results validate the independent variables in the dataset has a correlation to the dependent variables. Moreover, the analysis indicates how the distance from major central business district, the

number of bedrooms and bathrooms, and the size of the building area influence the property price and predict the future value. However, some independent variables, such as the dated house sold, the year house built and general region name, are not applied in this analysis so that the Adj R-Sq of the model and RMSE of the model still not perform well yet. In future work, for a deeper analysis of Melbourne Housing Market, we will apply and modify them into the accurate regression model so that we can find the important role for the property price and how each variable effect and relate to each other. Finally, an accurate regression model helps us know the changing of property price in time and predict the property price in the future.

Reference

1. Tony Pino. 2018. kaggle.com. *Melbourne Housing Market dataset* , [online] Available at: https://www.kaggle.com/anthonypino/melbourne-housing-market Accessed February 25, 2019.

2. Karen Grace-Martin. , The Analysis Factor. *When Listwise Deletion works for Missing Data* [online] Available at: https://www.theanalysisfactor.com/when-listwise-deletion-works/ Accessed March 10, 2019.

3. Svetlana Cheusheva. April 2, 2018. Ablebit.com *How to select random sample in Excel* [online] Available at: https://www.ablebits.com/office-addins-blog/2018/01/31/excel-random-selection-random-sample/ Accessed March 10, 2019.

4. NYS Department of Taxation & Finance Office of Real Property Tax Services. (2018).*How to Estimate the Market Value of Your Home.* [online] Available at: https://www.tax.ny.gov/pdf/publications/orpts/mv_estimates.pdf Accessed February 25, 2019.

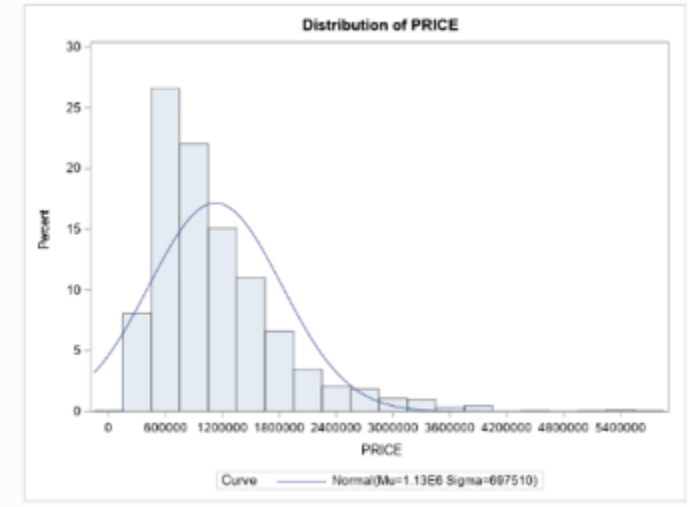
Appendix

Appendix - A (Paripon Thangthong)

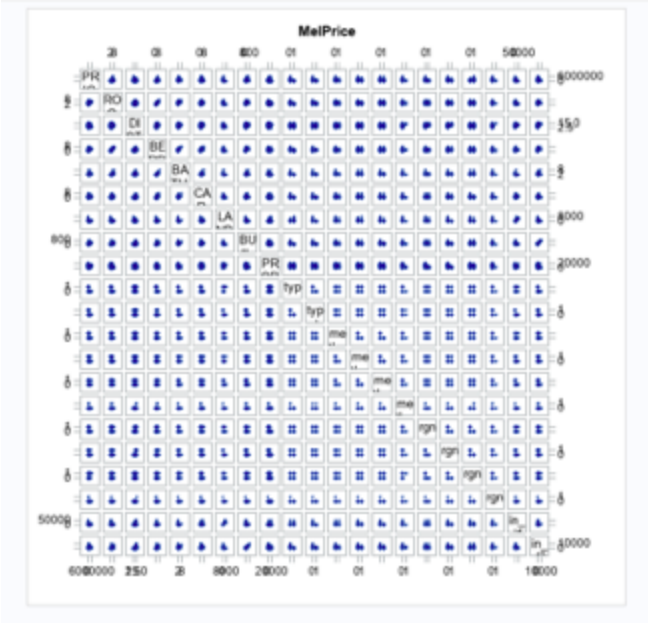
A.1 Descritive 5 points statistic chart

Variable	N	Minimum	Maximum	Mean	Median	Std Dev	25th Pctl	50th Pctl	75th Pctl
ROOMS	1499	1.0000000	8.0000000	2.9412942	3.0000000	0.9759573	2.0000000	3.0000000	4.0000000
PRICE	1499	131000.00	5700000.00	1128016.30	930000.00	697509.91	650000.00	930000.00	1415000.00
DISTANCE	1499	1.2000000	15.0000000	8.4965310	8.7000000	3.6090213	5.6000000	8.7000000	11.2000000
BEDROOM2	1499	0	9.0000000	2.8879253	3.0000000	0.9753885	2.0000000	3.0000000	3.0000000
BATHROOM	1499	1.0000000	8.0000000	1.5983989	1.0000000	0.7435688	1.0000000	1.0000000	2.0000000
CAR	1499	0	8.0000000	1.5416945	1.0000000	0.8987991	1.0000000	1.0000000	2.0000000
LANDSIZE	1499	0	7455.00	410.9306204	327.0000000	499.2038821	127.0000000	327.0000000	606.0000000
BUILDINGAREA	1499	0	792.0000000	142.0667111	125.0000000	79.5432276	90.0000000	125.0000000	173.0000000
PROPERTYCOUNT	1499	438.0000000	21650.00	7751.24	6795.00	4439.25	4675.00	6795.00	10412.00
typeu	1499	0	1.0000000	0.2461641	0	0.4309190	0	0	0
typet	1499	0	1.0000000	0.0940627	0	0.2920134	0	0	0
methodsp	1499	0	1.0000000	0.1247498	0	0.3305453	0	0	0
methodpi	1499	0	1.0000000	0.1260841	0	0.3320548	0	0	0
methodvb	1499	0	1.0000000	0.0933956	0	0.2910831	0	0	0
methodsa	1499	0	1.0000000	0.0020013	0	0.0447064	0	0	0
rgname_w	1499	0	1.0000000	0.2234823	0	0.4167179	0	0	0
rgname_e	1499	0	1.0000000	0.0687125	0	0.2530490	0	0	0
rgname_s	1499	0	1.0000000	0.3875917	0	0.4873631	0	0	1.0000000
rgname_se	1499	0	1.0000000	0.0013342	0	0.0365148	0	0	0
in_d1	1499	0	52404.80	3777.67	2756.00	4301.07	572.0000000	2756.00	6118.00
in_d2	1499	0	9662.40	1257.23	1064.80	955.8297669	530.4000000	1064.80	1703.00

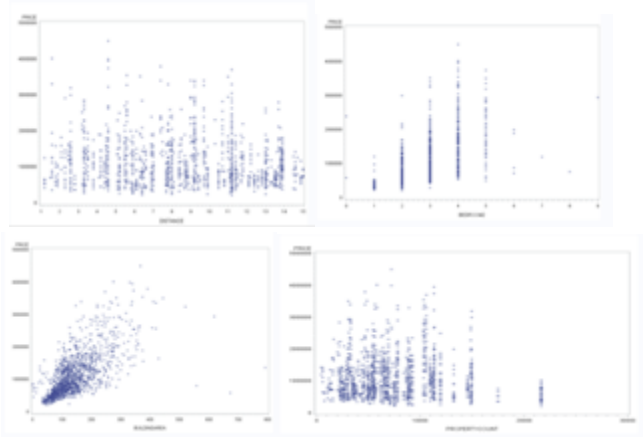
Price variable histogram



A.2 Scatter plot Matrix



A.3 G-plot example



A.4 Procedure Full model

Number of Observations Read	1499
Number of Observations Used	1499

Analysis of Variance

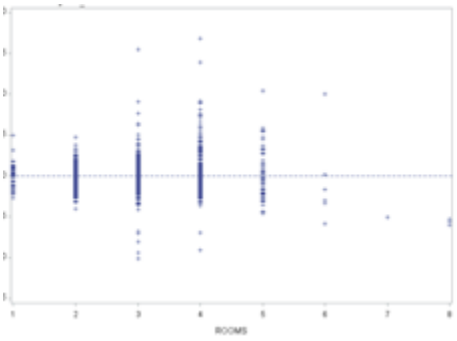
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	20	5.126663E14	2.563331E13	175.28	<.0001
Error	1478	2.161408E14	1.462387E11		
Corrected Total	1498	7.288071E14			

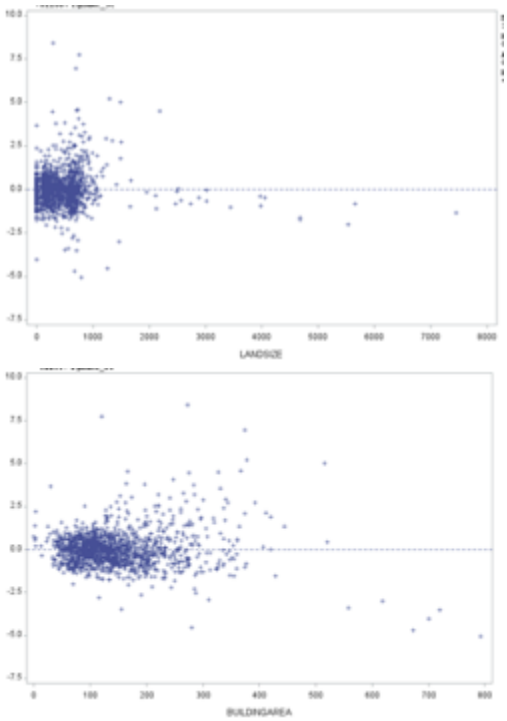
Root MSE	382412	R-Square	0.7034
Dependent Mean	1128016	Adj R-Sq	0.6994
Coeff Var	33.90126		

Parameter Estimates

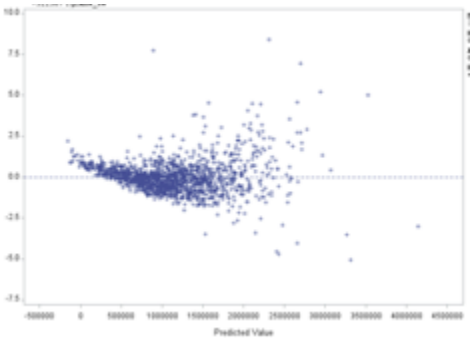
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	24417	69468	0.35	0.7253
ROOMS	1	92009	28805	3.19	0.0014
DISTANCE	1	15603	6462.68290	2.41	0.0159
BEDROOM2	1	-37609	27959	-1.35	0.1788
BATHROOM	1	117827	18643	6.32	<.0001
CAR	1	39326	12969	3.03	0.0025
LANDSIZE	1	32.65107	44.42779	0.73	0.4625
BUILDINGAREA	1	8642.97538	442.83368	19.52	<.0001
PROPERTYCOUNT	1	-1.20958	2.47341	-0.49	0.6249
typeu	1	-385164	30817	-12.50	<.0001
typet	1	-283898	36561	-7.77	<.0001
methodsp	1	-49586	30859	-1.61	0.1083
methodpi	1	-130742	30834	-4.24	<.0001
methodvb	1	-134912	35244	-3.83	0.0001
methodsa	1	-228267	222403	-1.03	0.3049
rgname_w	1	-45573	30291	-1.50	0.1327
rgname_e	1	256475	45967	5.58	<.0001
rgname_s	1	453627	25633	17.70	<.0001
rgname_se	1	289061	274133	1.05	0.2918
in_d1	1	8.62151	6.04481	1.43	0.1540
in_d2	1	-581.30715	42.30290	-13.74	<.0001

A.5 Studentized Residual vs (all X variable) some of the violation plot

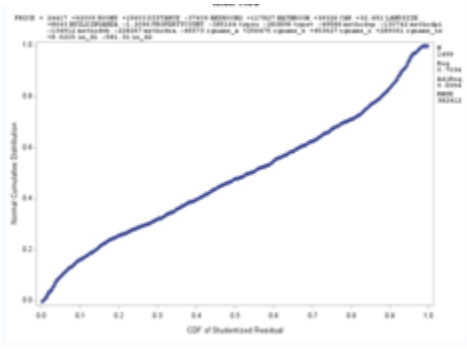




A.6 Studentized vs Predicted

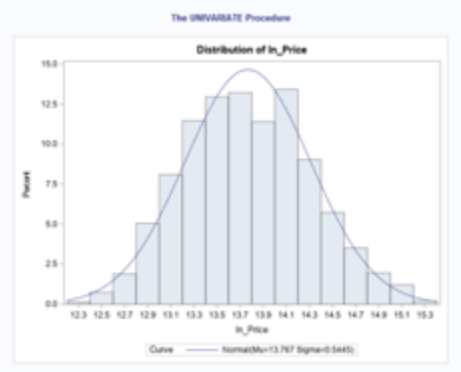


Npp Plot



A.7 Transformation with Log () method

Histogram of Log(ln_price)



A.8 Descriptive Statistic after log transformation

The UNIVARIATE Procedure
Variable: ln_Price

Moments			
N	1483	Sum Weights	1483
Mean	13.7666312	Sum Observations	20415.9141
Std Deviation	0.54454013	Variance	0.29652395
Skewness	0.11699053	Kurtosis	-0.415533
Uncorrected SS	281497.808	Corrected SS	439.448493
Coeff Variation	3.95550748	Std Error Mean	0.01414032

Basic Statistical Measures			
Location		Variability	
Mean	13.76663	Std Deviation	0.54454
Median	13.73863	Variance	0.29652
Mode	14.07787	Range	3.03655
		Interquartile Range	0.77498

A.9 Full model with Transformation proc reg (ln_price)

The REG Procedure
Model: MODEL1
Dependent Variable: ln_Price

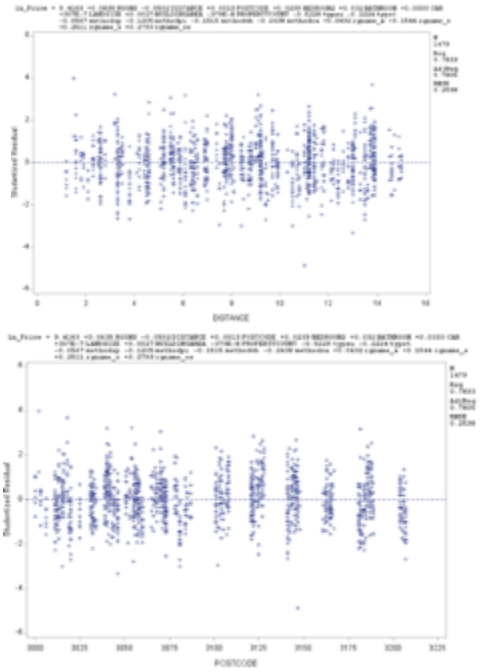
Number of Observations Read	1483
Number of Observations Used	1483

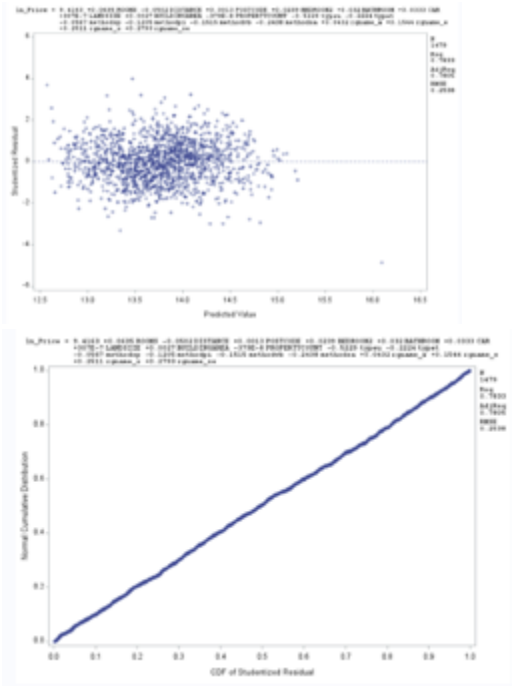
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	20	345.13113	17.25656	267.49	<.0001
Error	1462	94.31736	0.06451		
Corrected Total	1482	439.44849			

Root MSE	0.25399	R-Square	0.7854
Dependent Mean	13.76663	Adj R-Sq	0.7824
Coeff Var	1.84499		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	13.19991	0.04712	280.12	<.0001
ROOMS	1	0.07059	0.02043	3.45	0.0006
DISTANCE	1	-0.01401	0.00438	-3.20	0.0014
BEDROOM2	1	0.01803	0.01970	0.92	0.3602
BATHROOM	1	0.06340	0.01295	4.90	<.0001
CAR	1	0.02690	0.00880	3.06	0.0023
LANDSIZE	1	0.00001442	0.00002963	0.49	0.6265
BUILDINGAREA	1	0.00471	0.00031568	14.94	<.0001
PROPERTYCOUNT	1	-0.00000281	0.00000165	-1.70	0.0887
typeu	1	-0.51876	0.02059	-25.20	<.0001
typet	1	-0.22464	0.02434	-9.23	<.0001
methodsp	1	-0.06596	0.02060	-3.20	0.0014
methodpi	1	-0.12176	0.02066	-5.89	<.0001
methodvb	1	-0.16200	0.02364	-6.85	<.0001
methodsa	1	-0.16408	0.14774	-1.11	0.2669
rgname_w	1	-0.00917	0.02019	-0.45	0.6498
rgname_e	1	0.22667	0.03072	7.38	<.0001
rgname_s	1	0.38449	0.01711	22.46	<.0001
rgname_se	1	0.36832	0.18222	2.02	0.0434
in_d1	1	0.00000412	0.00000405	1.02	0.3085
in_d2	1	-0.00027976	0.00002904	-9.63	<.0001

A.10 Some example of studentized plot, npp plot and predicted value after transformation





A.11 Split Data for Validation

Number of Observations Read	1483
Number of Observations Used	890
Number of Observations with Missing Values	593

A.12 Stepwise and CP select (13) variables in Training Dataset

Stepwise Selection: Step 13

Variable LANDSIZE Entered: R-Square = 0.7789 and C(p) = 12.0303

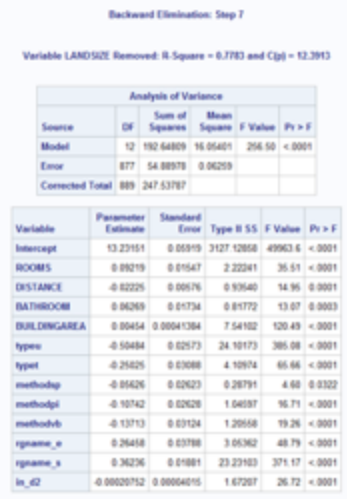
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	192.79696	14.83046	237.32	<.0001
Error	876	54.74191	0.06249		
Corrected Total	889	247.53787			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	13.22646	0.05924	3115.12483	49849.4	<.0001
ROOMS	0.09065	0.01549	2.13941	34.24	<.0001
DISTANCE	-0.02217	0.00575	0.92839	14.86	0.0001
BATHROOM	0.06362	0.01734	0.84111	13.46	0.0003
LANDSIZE	0.00002795	0.00001817	0.14787	2.37	0.1243
BUILDINGAREA	0.00455	0.00041353	7.55623	120.92	<.0001
typeu	-0.50256	0.02575	23.80494	380.94	<.0001
typet	-0.24390	0.03113	3.83477	61.37	<.0001
methodsp	-0.05963	0.02630	0.32120	5.14	0.0236
methodpi	-0.10928	0.02628	1.08021	17.29	<.0001
methodvb	-0.13903	0.03124	1.23727	19.80	<.0001
rgname_e	0.26232	0.03788	2.99731	47.96	<.0001
rgname_s	0.36313	0.01880	23.31336	373.07	<.0001
in_d2	-0.00021167	0.00004021	1.73178	27.71	<.0001

Number in Model	C(p)	R-Square	Variables in Model
13	12.0303	0.7789	ROOMS DISTANCE BATHROOM LANDSIZE BUILDINGAREA typeu typet methodsp methodpi methodvb rgname_e rgname_s in_d2

A.13 Backward select (12) variables in Training Dataset



Backward Elimination: Step 7

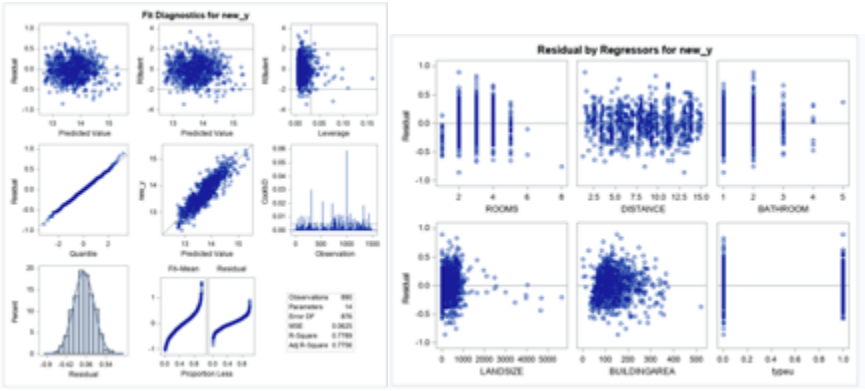
Variable LANDSIZE Removed: R-Square = 0.7783 and C(p) = 12.3913

Analysis of Variance

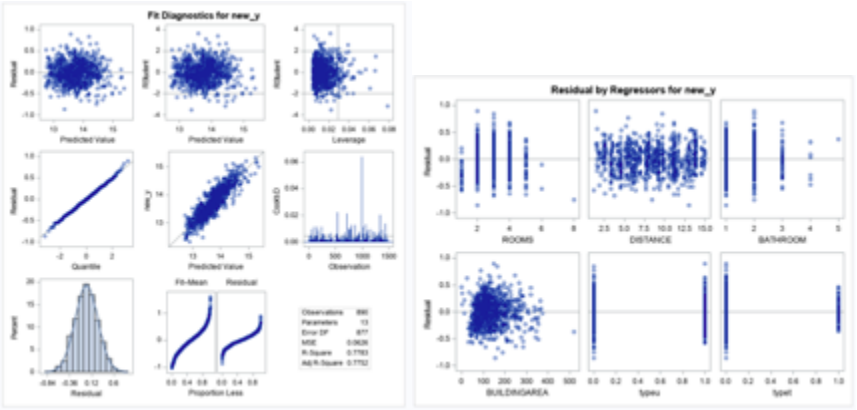
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	192.64809	16.05401	256.50	<.0001
Error	877	54.88978	0.06259		
Corrected Total	889	247.53787			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	13.23151	0.05919	3127.12858	49963.6	<.0001
ROOMS	0.09219	0.01547	2.22241	35.51	<.0001
DISTANCE	-0.02225	0.00576	0.93640	14.96	0.0001
BATHROOM	0.06269	0.01734	0.81772	13.07	0.0003
BUILDINGAREA	0.00454	0.00041384	7.54102	120.49	<.0001
typeu	-0.50484	0.02573	24.10173	385.08	<.0001
typet	-0.25825	0.03088	4.10974	65.66	<.0001
methodsp	-0.05626	0.02623	0.28791	4.60	0.0322
methodpi	-0.10742	0.02628	1.04597	16.71	<.0001
methodvb	-0.13713	0.03124	1.20558	19.26	<.0001
rgnaeme_e	0.26458	0.03788	3.05362	48.79	<.0001
rgnaeme_s	0.36236	0.01881	23.23103	371.17	<.0001
ln_d2	-0.00020752	0.00004015	1.67207	26.72	<.0001

A.14 Model assumption for CP and Stepwise in Training Dataset



A.15 Model assumption for Backward in Training Dataset



A.16 Stepwise and Cp Test dataset

Validation statistics for model

Obs	_TYPE_	_FREQ_	rmse	mae
1	0	593	0.26851	0.20433

Validation statistics for model

The CORR Procedure

2 Variables: ln_Price yhat

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
ln_Price	593	13.79350	0.56830	8180	12.28303	15.20180	
yhat	593	13.79127	0.51818	8178	12.68052	15.92867	Predicted Value of new_y

Pearson Correlation Coefficients, N = 593
Prob > |r| under H0: Rho=0

	ln_Price	yhat
ln_Price	1.00000	0.88185 <.0001
yhat Predicted Value of new_y	0.88186 <.0001	1.00000

A.17 Backward Test Dataset

Validation statistics for model

Obs	_TYPE_	_FREQ_	rmse	mae
1	0	593	0.27008	0.20526

Validation statistics for model

The CORR Procedure

2 Variables: ln_Price yhat

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
ln_Price	593	13.79350	0.56830	8180	12.28303	15.20180	
yhat	593	13.79137	0.51847	8178	12.68596	15.92207	Predicted Value of new_y

Pearson Correlation Coefficients, N = 593
Prob > |r| under H0: Rho=0

	ln_Price	yhat
ln_Price	1.00000	0.88044 <.0001
yhat Predicted Value of new_y	0.88044 <.0001	1.00000

A.18 5 Fold Cross validate for Stepwise

Validation statistics for model

The GLMSELECT Procedure

Data Set	WORK.NEW1
Dependent Variable	ln_Price
Selection Method	Stepwise
Select Criterion	SBC
Stop Criterion	Cross Validation
Cross Validation Method	Split
Cross Validation Fold	5
Effect Hierarchy Enforced	None
Random Number Seed	154060001

Number of Observations Read	1483
Number of Observations Used	1483
Number of Observations Used for Training	919
Number of Observations Used for Testing	564

Dimensions	
Number of Effects	21
Number of Parameters	21

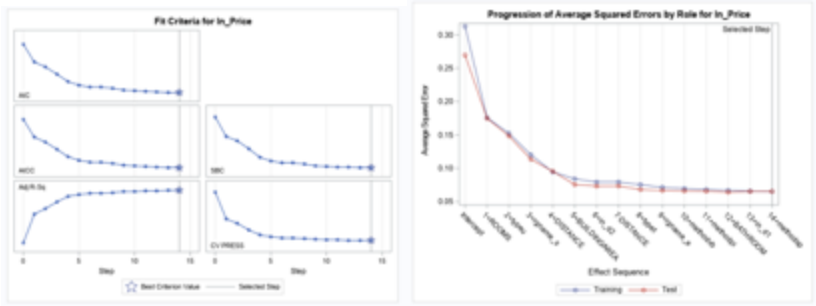
Validation statistics for model

The GLMSELECT Procedure

Stepwise Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	SBC	ASE	Test ASE	CV PRESS
0	Intercept		1	-1061.4818	0.3127	0.2696	288.1418
1	ROOMS		2	-1582.7857	0.1760	0.1758	161.9633
2	typeu		3	-1797.3520	0.1526	0.1486	140.6772
3	rgnname_s		4	-1917.9146	0.1294	0.1126	111.8492
4	DISTANCE		5	-2132.7759	0.0946	0.0955	87.7412
5	BUILDINGAREA		6	-2232.1811	0.0843	0.0757	78.3606
6	ln_d2		7	-2279.0576	0.0796	0.0726	74.2295
7		DISTANCE	6	-2284.7966	0.0796	0.0737	73.9542
8	typet		7	-2326.6329	0.0755	0.0683	70.7186
9	rgnname_e		8	-2368.7000	0.0716	0.0669	66.6364
10	methodvb		9	-2394.6401	0.0698	0.0662	65.6756
11	methodpi		10	-2396.3537	0.0694	0.0658	64.9157
12	BATHROOM		11	-2406.8630	0.0672	0.0643	63.8019
13	ln_d1		12	-2410.7537	0.0660	0.0654	61.9884
14	methodsp		13	-2417.9400*	0.0654	0.0655	61.5462*

* Optimal Value of Criterion



A.19 Cross validated For Backward

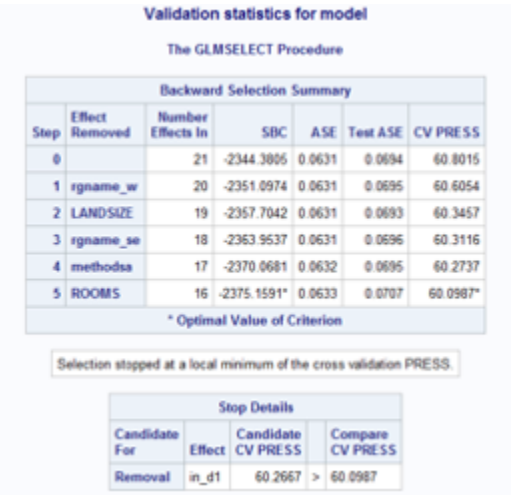
Validation statistics for model

The GLMSELECT Procedure

Data Set	WORK.NEW1
Dependent Variable	ln_Price
Selection Method	Backward
Select Criterion	SBC
Stop Criterion	Cross Validation
Cross Validation Method	Split
Cross Validation Fold	5
Effect Hierarchy Enforced	None
Random Number Seed	38836700

Number of Observations Read	1483
Number of Observations Used	1483
Number of Observations Used for Training	900
Number of Observations Used for Testing	583

Dimensions	
Number of Effects	21
Number of Parameters	21



Validation statistics for model

The GLMSELECT Procedure

Backward Selection Summary

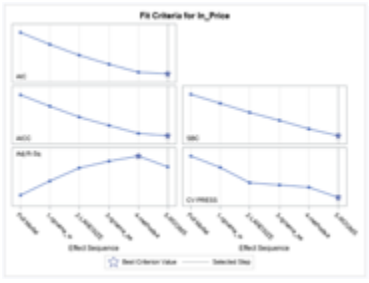
Step	Effect Removed	Number Effects In	SBC	ASE	Test ASE	CV PRESS
0		21	-2344.3806	0.0631	0.0694	60.8015
1	rgname_w	20	-2351.0974	0.0631	0.0695	60.6054
2	LANDSIZE	19	-2357.7042	0.0631	0.0693	60.3457
3	rgname_se	18	-2363.9537	0.0631	0.0696	60.3116
4	methodsa	17	-2370.0681	0.0632	0.0695	60.2737
5	ROOMS	16	-2375.1591*	0.0633	0.0707	60.0987*

* Optimal Value of Criterion

Selection stopped at a local minimum of the cross validation PRESS.

Stop Details

Candidate For	Effect	Candidate CV PRESS		Compare CV PRESS
Removal	in_d1	60.2667	>	60.0987



A.20 Prediction 1

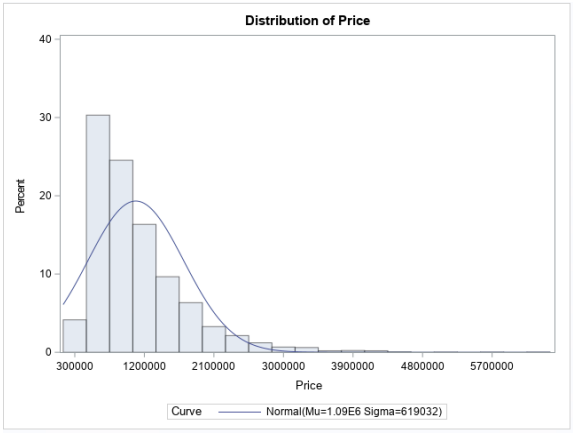
Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	14.8630	0.0552	14.7546	14.9713	14.3665	15.3594	.

A.21 Prediction 2

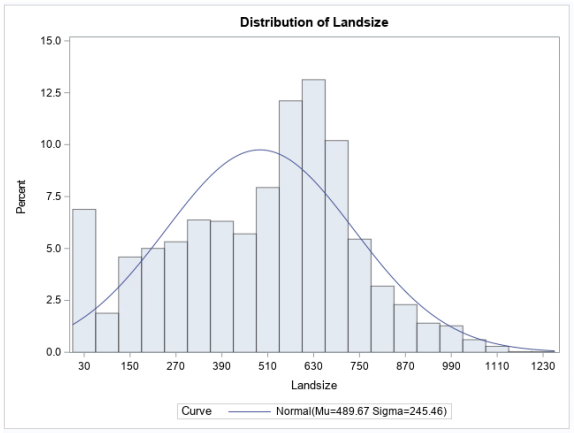
Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	14.3465	0.0338	14.2802	14.4129	13.8576	14.8355	.

Appendix-B (Dufang Qu)

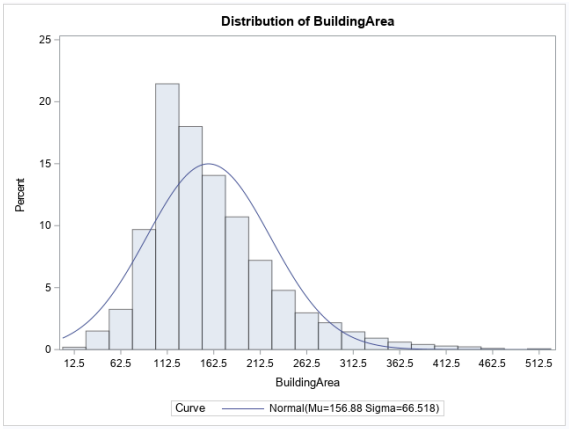
B.1 Histogram of Price



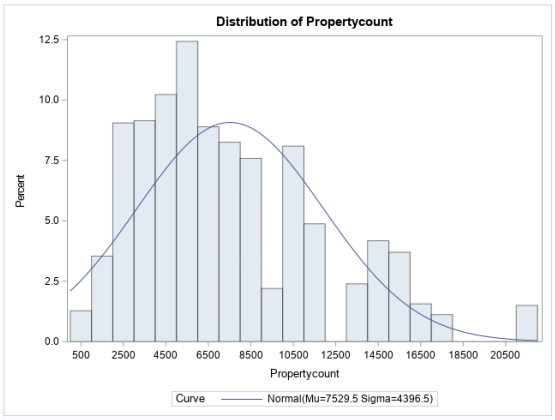
B.2 Histogram of LandSize



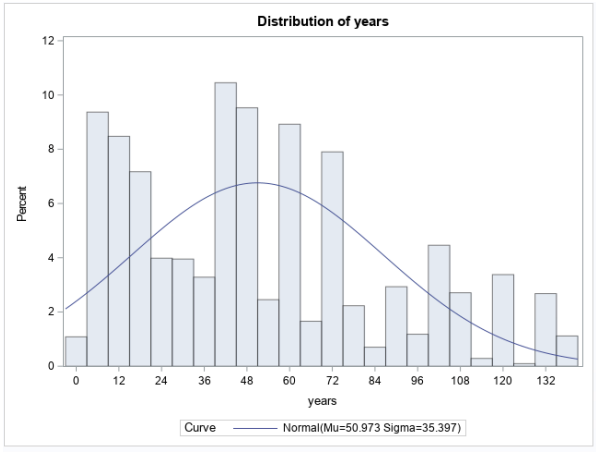
B.3 Histogram of BuildingArea



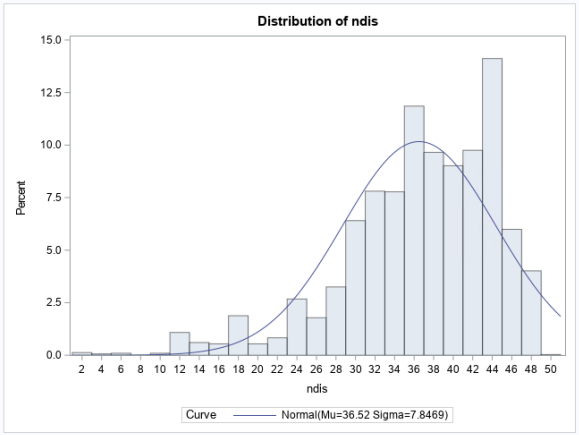
B.4 Histogram of Propertycount



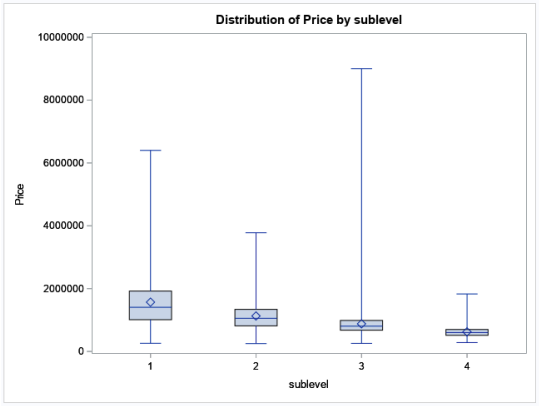
B.5 Histogram of Years



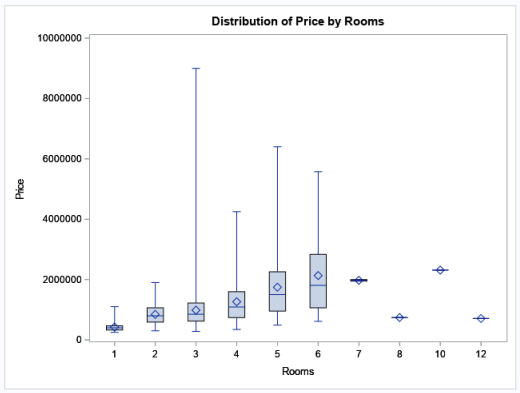
B.6 Histogram of Ndis



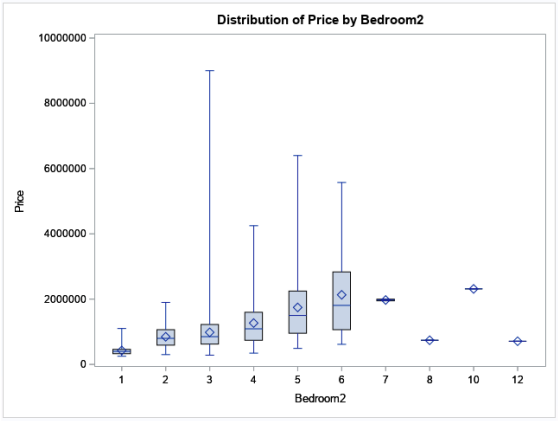
B.7 Boxplot for Price vs Sublevel



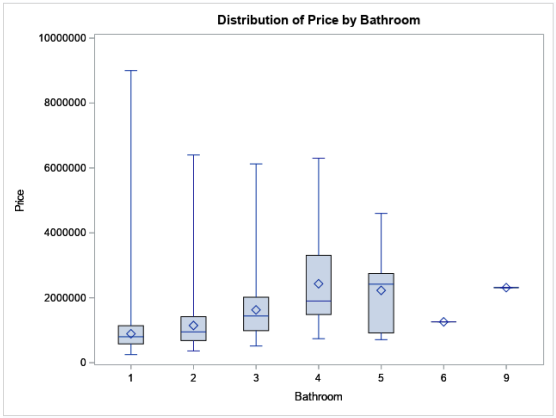
B.8 Boxplot for Price vs Rooms



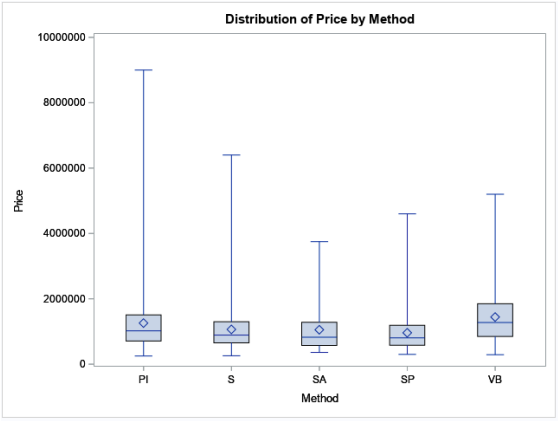
B.9 Boxplot for Price vs Bedroom2



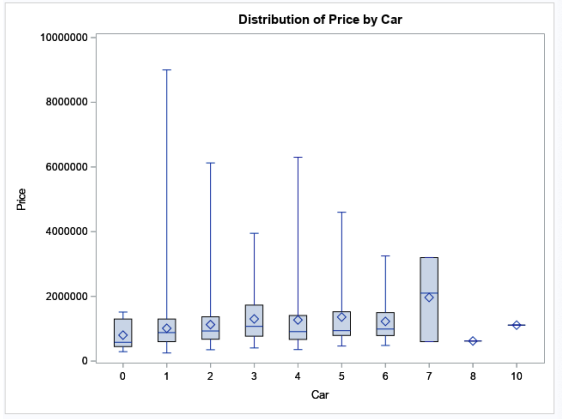
B.10 Boxplot for Price vs Bathroom



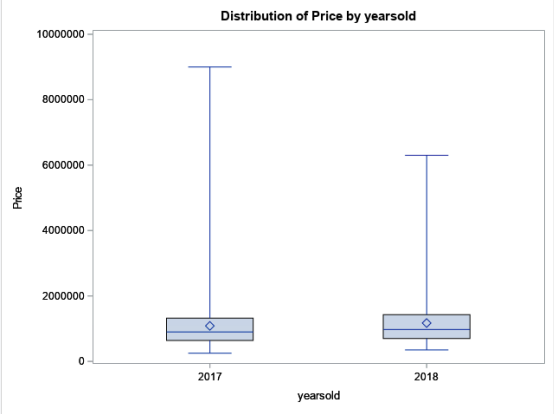
B.11 Boxplot for Price vs Method



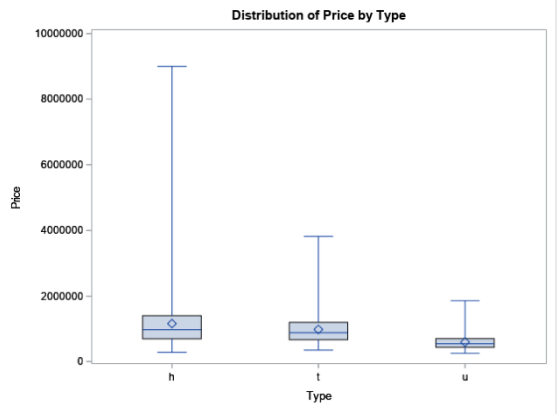
B.12 Boxplot for Price vs Car



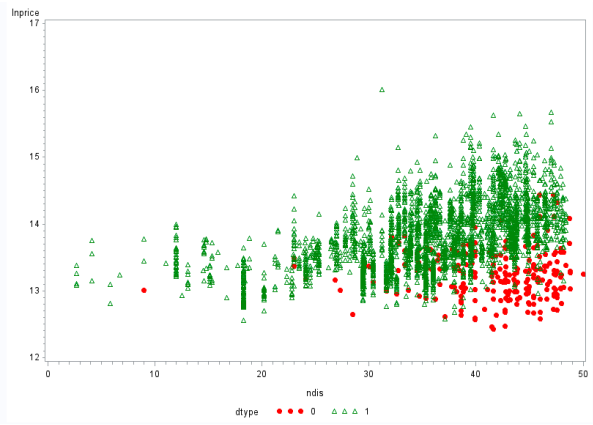
B.13 Boxplot for Price vs YearSold



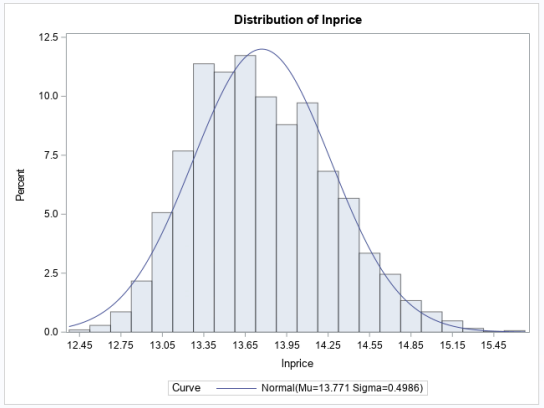
B.14 Boxplot for Price vs Type



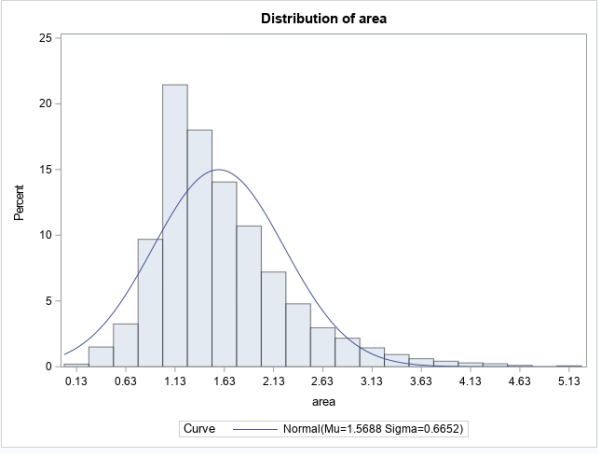
B.15 Scatterplot for LnPrice vs Ndis*Dtype



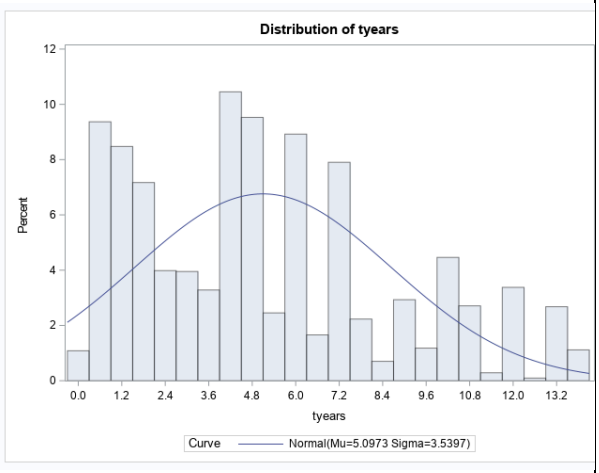
B.16 Histogram for LnPrice



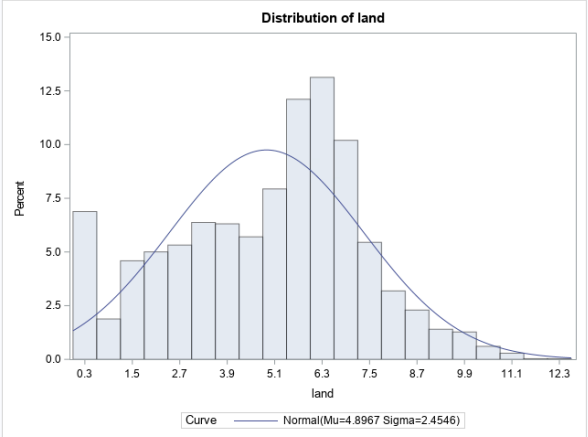
B.17 Histogram for Area



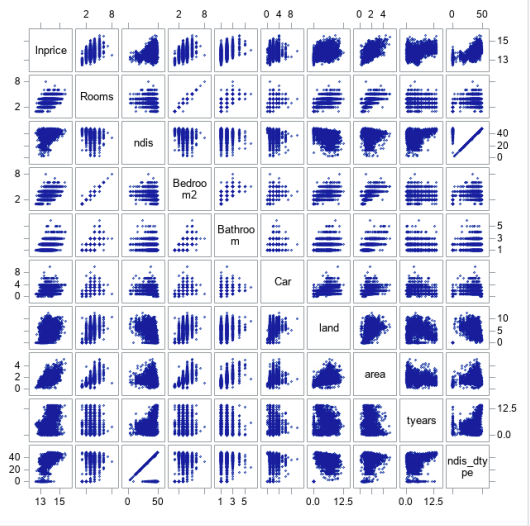
B.18 Histogram for Tyears



B.19 Histogram for Land



B.20 Scatterplot Matrix



B.21 Pearson Correlation Coefficients

Pearson Correlation Coefficients, N = 3138
Prob > |r| under H0: Rho=0

	lnprice	Rooms	ndis	Bedroom2	Bathroom	Car	land	area	tyears	ndis_dtype
lnprice	1.00000	0.39646 <.0001	0.43247 <.0001	0.39686 <.0001	0.36414 <.0001	0.11851 <.0001	0.22132 <.0001	0.51280 <.0001	0.42557 <.0001	0.56026 <.0001
Rooms	0.39646 <.0001	1.00000	-0.20458 <.0001	0.99603 <.0001	0.61628 <.0001	0.33466 <.0001	0.54291 <.0001	0.70222 <.0001	-0.11357 <.0001	0.26225 <.0001
ndis	0.43247 <.0001	-0.20458 <.0001	1.00000	-0.20364 <.0001	-0.08884 <.0001	-0.19139 <.0001	-0.43118 <.0001	-0.10110 <.0001	0.42500 <.0001	0.48015 <.0001
Bedroom2	0.39686 <.0001	0.99603 <.0001	-0.20364 <.0001	1.00000	0.61833 <.0001	0.33898 <.0001	0.54141 <.0001	0.70279 <.0001	-0.11414 <.0001	0.26173 <.0001
Bathroom	0.36414 <.0001	0.61628 <.0001	-0.08884 <.0001	0.61833 <.0001	1.00000	0.24884 <.0001	0.26791 <.0001	0.67028 <.0001	-0.25815 <.0001	0.13657 <.0001
Car	0.11851 <.0001	0.33466 <.0001	-0.19139 <.0001	0.33898 <.0001	0.24884 <.0001	1.00000	0.40141 <.0001	0.30396 <.0001	-0.15320 <.0001	0.06775 0.0001
land	0.22132 <.0001	0.54291 <.0001	-0.43118 <.0001	0.54141 <.0001	0.26791 <.0001	0.40141 <.0001	1.00000	0.42684 <.0001	-0.05285 0.0031	0.19193 <.0001
area	0.51280 <.0001	0.70222 <.0001	-0.10110 <.0001	0.70279 <.0001	0.67028 <.0001	0.30396 <.0001	0.42684 <.0001	1.00000	-0.19917 <.0001	0.21850 <.0001
tyears	0.42557 <.0001	-0.11357 <.0001	0.42500 <.0001	-0.11414 <.0001	-0.25815 <.0001	-0.15320 <.0001	-0.05285 0.0031	-0.19917 <.0001	1.00000	0.37354 <.0001
ndis_dtype	0.56026 <.0001	0.26225 <.0001	0.48015 <.0001	0.26173 <.0001	0.13657 <.0001	0.06775 0.0001	0.19193 <.0001	0.21850 <.0001	0.37354 <.0001	1.00000

B.22 Printed First Row

The SAS System

Obs	Suburb	Rooms	Type	Price	YearBuilt	Method	Distance	Bedroom2	Bathroom	Car	Landsize	BuildingArea	Propertycou
1	Chelsea	2	u	639000	1980	S	27	2	1	2	0	104	39

yearsold	sublevel	years	ndis	lnprice	dsub	dsub1	dsub2	dtype	area	land	tyears	ndis_dtype
2017	4	39	23	13.3677	0	0	0	0	1.04	0	3.9	0

B.23 Full Model F1

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: lnprice

Number of Observations Read	3138
Number of Observations Used	3138

Analysis of Variance

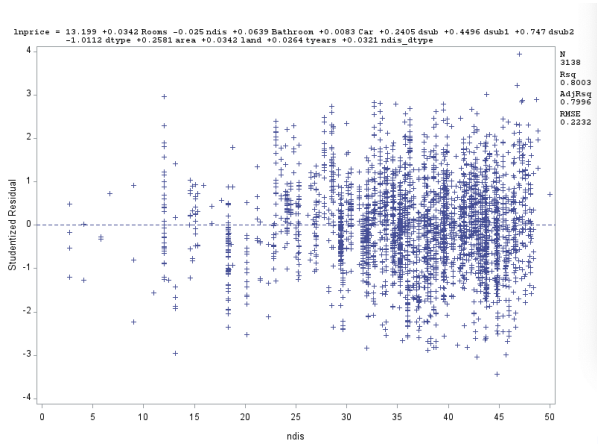
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	624.24337	52.02028	1043.86	<.0001
Error	3125	155.73306	0.04983		
Corrected Total	3137	779.97644			

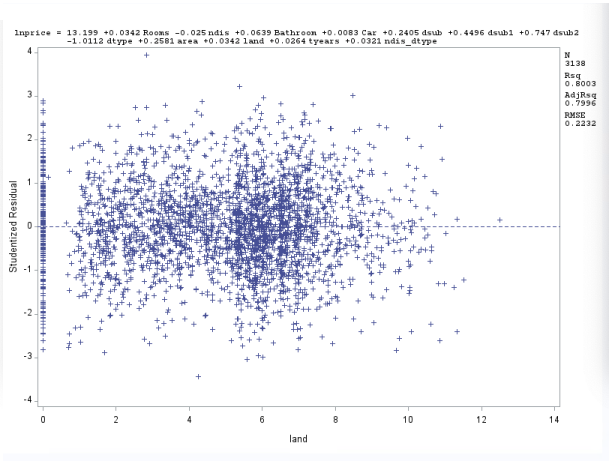
Root MSE	0.22324	R-Square	0.8003
Dependent Mean	13.77090	Adj R-Sq	0.7996
Coeff Var	1.62107		

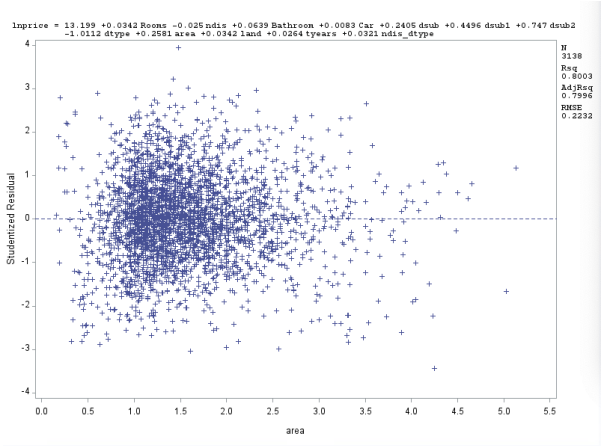
Parameter Estimates

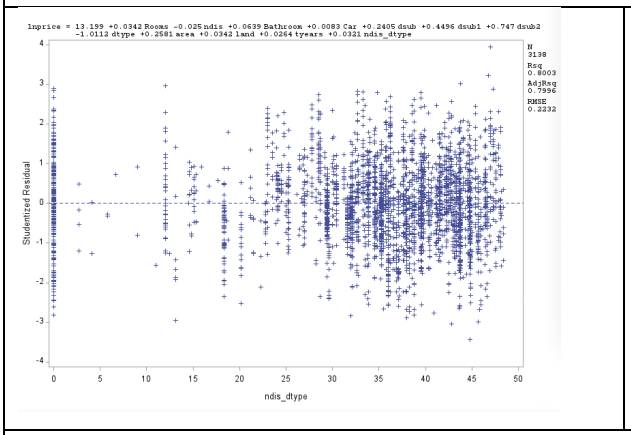
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	13.19102	0.11820	111.60	<.0001	0
Rooms	1	0.03331	0.00744	4.48	<.0001	2.69360
Bathroom	1	0.06295	0.00826	7.62	<.0001	2.11929
Car	1	0.00854	0.00480	1.78	0.0757	1.25591
ndis	1	-0.02487	0.00280	-8.88	<.0001	30.76638
dsub	1	0.24085	0.01294	18.61	<.0001	1.75983
dsub1	1	0.44983	0.01409	31.93	<.0001	2.37239
dsub2	1	0.74437	0.01741	42.75	<.0001	4.21434
dtype	1	-0.99904	0.11898	-8.40	<.0001	57.17553
area	1	0.26239	0.00970	27.05	<.0001	2.64435
land	1	0.03352	0.00249	13.46	<.0001	2.38150
tyears	1	0.02671	0.00150	17.78	<.0001	1.79845
ndis_dtype	1	0.03175	0.00283	11.24	<.0001	71.31113

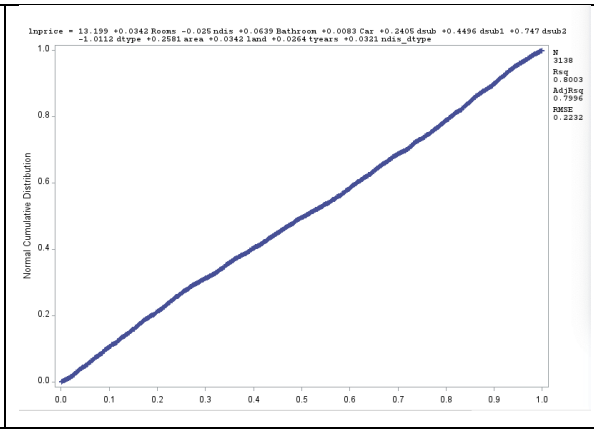
B.24 Residual Plots for Model F1



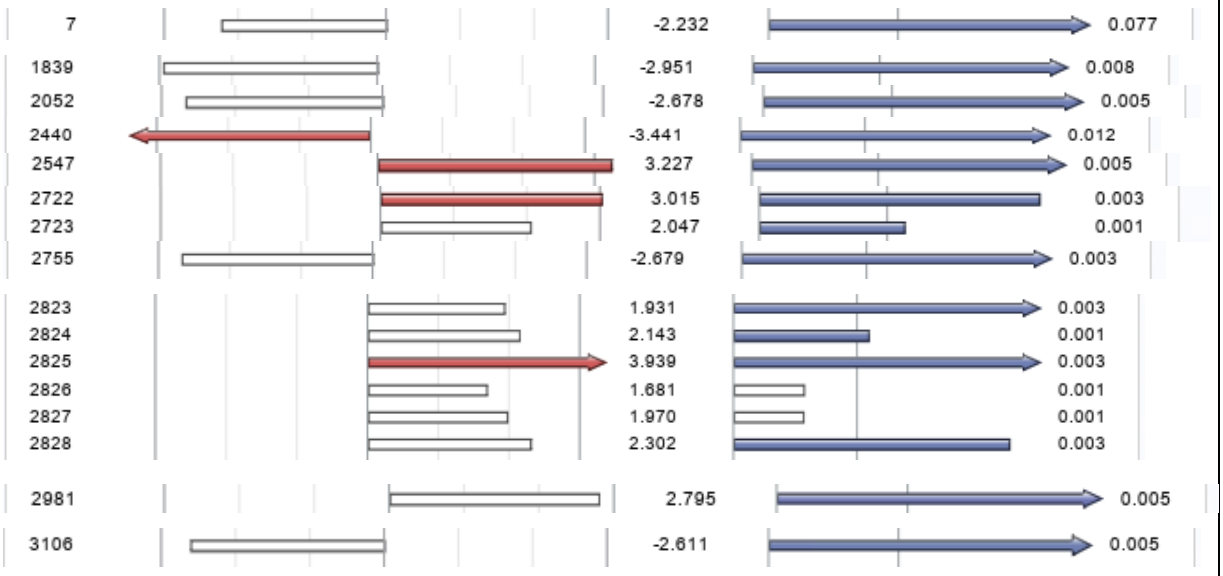








B.25 Part of Cook's D for Model F1



B.26 Most influential observations (5)

Obs	Suburb	Rooms	Type	Price	YearBuilt	Method	Distance	Bedroom2	Bathroom	Car	Landsize	BuildingArea	Propertycount
2440	Brunswick	4	h	1585000	1915	S	5.2	4	2	1	426	425	11918

The SAS System

Obs	Suburb	Rooms	Type	Price	YearBuilt	Method	Distance	Bedroom2	Bathroom	Car	Landsize	BuildingArea	Propertycount
2547	Carlton Nort	4	h	2718000	2015	S	3.2	4	2	1	538	142	3106

The SAS System

Obs	Suburb	Rooms	Type	Price	YearBuilt	Method	Distance	Bedroom2	Bathroom	Car	Landsize	BuildingArea	Propertycount
2722	Hawthorn Eas	3	h	3600000	1929	S	6.2	3	2	2	848	158	6482

The SAS System

Obs	Suburb	Rooms	Type	Price	YearBuilt	Method	Distance	Bedroom2	Bathroom	Car	Landsize	BuildingArea	Propertycount
2825	Middle Park	3	h	3750000	1890	SA	3	3	1	1	284	148	2019

yearsold	sublevel	years	ndis	lnprice	dsub	dsub1	dsub2	dtype	area	land	tyears	ndis_dtype	lnndis	lnndis_dtype
2018	1	104	44.8	14.2761	0	0	1	1	4.25	4.26	10.4	44.8	3.80221	3.80221

earsold	sublevel	years	ndis	lnprice	dsub	dsub1	dsub2	dtype	area	land	tyears	ndis_dtype	lnndis	lnndis_dtype
2017	1	4	46.8	14.8154	0	0	1	1	1.42	5.38	0.4	46.8	3.84588	3.84588

yearsold	sublevel	years	ndis	lnprice	dsub	dsub1	dsub2	dtype	area	land	tyears	ndis_dtype	lnndis	lnndis_dtype
2017	1	90	43.8	15.0964	0	0	1	1	1.58	8.48	9	43.8	3.77963	3.77963

earsold	sublevel	years	ndis	lnprice	dsub	dsub1	dsub2	dtype	area	land	tyears	ndis_dtype	lnndis	lnndis_dtype
2018	1	129	47	15.1373	0	0	1	1	1.48	2.84	12.9	47	3.85015	3.85015

The SAS System

Obs	Suburb	Rooms	Type	Price	YearBuilt	Method	Distance	Bedroom2	Bathroom	Car	Landsize	BuildingArea	Propertycount	yearsold
7	Langwarrin	3	u	447000	2002	S	41	3	1	2	0	102	8743	2017

sublevel	years	ndis	lnprice	dsub	dsub1	dsub2	dtype	area	land	tyears	ndis_dtype	lnndis	lnndis_dtype
4	17	9	13.0103	0	0	0	0	1.02	0	1.7	0	2.19722	0

B.27 Validation Output

The SAS System

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling

Input Data Set	MBHOUSE
Random Number Seed	298369
Sampling Rate	0.66
Sample Size	2069
Selection Probability	0.660179
Sampling Weight	0
Output Data Set	TATSET

B.28 Stepwise Procedure for Model S1

Forward Selection: Step 12

Variable Car Entered: R-Square = 0.8112 and C(p) = 13.0000

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	414.21338	34.51778	735.92	<.0001
Error	2056	96.43468	0.04690		
Corrected Total	2068	510.64807			

Summary of Forward Selection

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	ndis_dtype	1	0.3215	0.3215	5322.33	979.24	<.0001
2	area	2	0.1637	0.4852	3541.83	657.05	<.0001
3	dsub2	3	0.1718	0.6569	1673.86	1033.90	<.0001
4	dsub1	4	0.0627	0.7197	993.079	461.74	<.0001
5	tyears	5	0.0375	0.7572	586.361	318.98	<.0001
6	dsub	6	0.0235	0.7807	332.302	221.17	<.0001
7	land	7	0.0186	0.7994	131.347	191.49	<.0001
8	Bathroom	8	0.0054	0.8047	75.0873	56.45	<.0001
9	Rooms	9	0.0015	0.8062	61.0070	15.69	<.0001
10	ndis	10	0.0006	0.8068	56.9114	5.96	0.0147
11	dtype	11	0.0043	0.8111	11.7448	47.17	<.0001
12	Car	12	0.0001	0.8112	13.0000	0.74	0.3882

B.29 CP Procedure for Model S2

Model B

The REG Procedure
Model: MODEL2
Dependent Variable: new_lp

C(p) Selection Method

Number of Observations Read	3134
Number of Observations Used	2069
Number of Observations with Missing Values	1065

Number in Model	C(p)	R-Square	Variables in Model
11	11.7448	0.8111	Rooms Bathroom ndis dsub dsub1 dsub2 dtype area land tyears ndis_dtype
12	13.0000	0.8112	Rooms Bathroom Car ndis dsub dsub1 dsub2 dtype area land tyears ndis_dtype
10	22.4760	0.8099	Bathroom ndis dsub dsub1 dsub2 dtype area land tyears ndis_dtype
11	23.3596	0.8100	Bathroom Car ndis dsub dsub1 dsub2 dtype area land tyears ndis_dtype
10	46.1007	0.8077	Rooms ndis dsub dsub1 dsub2 dtype area land tyears ndis_dtype
11	47.0465	0.8078	Rooms Car ndis dsub dsub1 dsub2 dtype area land tyears ndis_dtype
10	56.9114	0.8068	Rooms Bathroom ndis dsub dsub1 dsub2 area land tyears ndis_dtype
11	57.3499	0.8069	Rooms Bathroom Car ndis dsub dsub1 dsub2 area land tyears ndis_dtype
9	61.0070	0.8062	Rooms Bathroom dsub dsub1 dsub2 area land tyears ndis_dtype
10	61.6601	0.8063	Rooms Bathroom Car dsub dsub1 dsub2 area land tyears ndis_dtype

B.30 Model S1 on Training Set

B.31 Model S2 on Training Set

The REG Procedure
Model: MODEL1
Dependent Variable: new_lp

Number of Observations Read	3134
Number of Observations Used	2069
Number of Observations with Missing Values	1065

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	414.21338	34.51778	735.92	<.0001
Error	2056	96.43468	0.04690		
Corrected Total	2068	510.64807			

Root MSE	0.21657	R-Square	0.8112
Dependent Mean	13.77114	Adj R-Sq	0.8101
Coeff Var	1.57266		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	13.15179	0.13988	94.02	<.0001	0
Rooms	1	0.03169	0.00901	3.52	0.0004	2.74084
Bathroom	1	0.05942	0.00990	6.00	<.0001	2.13048
Car	1	0.00486	0.00563	0.86	0.3882	1.23794
ndis	1	-0.02413	0.00333	-7.26	<.0001	29.45819
dsub	1	0.25085	0.01565	16.03	<.0001	1.74806
dsub1	1	0.45760	0.01705	26.84	<.0001	2.40985
dsub2	1	0.75366	0.02109	35.73	<.0001	4.28974
dtype	1	-0.96187	0.14128	-6.81	<.0001	53.70319
area	1	0.27605	0.01185	23.30	<.0001	2.67127
land	1	0.03299	0.00298	11.08	<.0001	2.36516
tyears	1	0.02554	0.00182	14.02	<.0001	1.81997
ndis_dtype	1	0.03099	0.00337	9.21	<.0001	68.07490

B.32 Validation Stats for Model S1

Validation stats for model S1

Obs	_TYPE_	_FREQ_	rmse	mae
1	0	1065	0.23179	0.18124

Validation stats for model S1

The CORR Procedure

2 Variables:	Inprice yhat

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
Inprice	1065	13.76646	0.49843	14661	12.42922	15.65606	
yhat	1065	13.76001	0.45087	14654	12.66585	15.38641	Predicted Value of new_lp

Pearson Correlation Coefficients, N = 1065
Prob > |r| under H0: Rho=0

	Inprice	yhat
Inprice	1.00000	0.88549 <.0001
yhat Predicted Value of new_lp	0.88549 <.0001	1.00000

The REG Procedure
Model: MODEL2
Dependent Variable: new_lp

Number of Observations Read	3134
Number of Observations Used	2069
Number of Observations with Missing Values	1065

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	414.17845	37.65259	802.86	<.0001
Error	2057	96.46962	0.04690		
Corrected Total	2068	510.64807			

Root MSE	0.21656	R-Square	0.8111
Dependent Mean	13.77114	Adj R-Sq	0.8101
Coeff Var	1.57256		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	13.16176	0.13939	94.42	<.0001	0	0
Rooms	1	0.03211	0.00900	3.57	0.0004	0.05653	2.73266
Bathroom	1	0.05965	0.00989	6.03	<.0001	0.08432	2.12891
ndis	1	-0.02425	0.00332	-7.30	<.0001	-0.37929	29.41008
dsub	1	0.25088	0.01565	16.03	<.0001	0.20312	1.74805
dsub1	1	0.45734	0.01705	26.83	<.0001	0.39907	2.40907
dsub2	1	0.75323	0.02108	35.73	<.0001	0.70892	4.28735
dtype	1	-0.96875	0.14105	-6.87	<.0001	-0.48158	53.53179
area	1	0.27652	0.01184	23.36	<.0001	0.36553	2.66565
land	1	0.03361	0.00289	11.63	<.0001	0.16632	2.22702
tyears	1	0.02541	0.00182	14.00	<.0001	0.18037	1.80785
ndis_dtype	1	0.03114	0.00336	9.27	<.0001	0.73175	67.88680

B.33 Validation Stats for Model S2

Validation stats for Model S2

Obs	_TYPE_	_FREQ_	rmse	mae
1	0	1065	0.23198	0.18124

Validation stats for Model S2

The CORR Procedure

2 Variables:	Inprice yhat

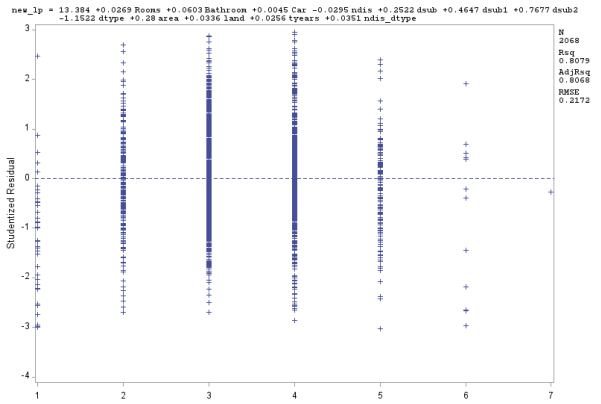
Simple Statistics

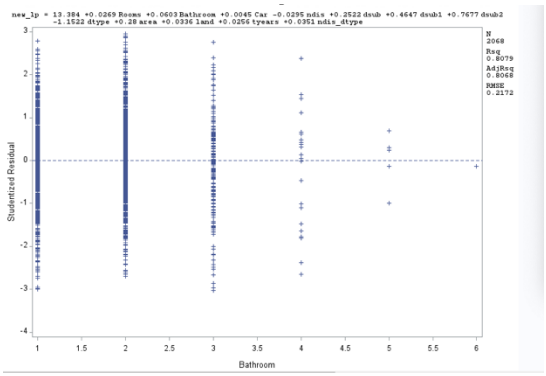
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
Inprice	1065	13.76646	0.49843	14661	12.42922	15.65606	
yhat	1065	13.76008	0.45086	14654	12.66755	15.38080	Predicted Value of new_lp

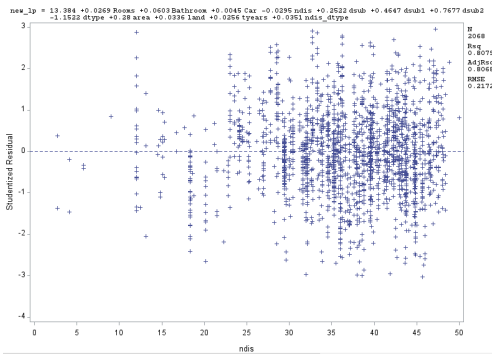
Pearson Correlation Coefficients, N = 1065
Prob > |r| under H0: Rho=0

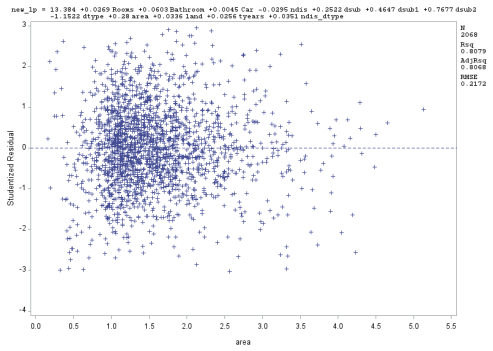
	Inprice	yhat
Inprice	1.00000	0.88530 <.0001
yhat Predicted Value of new_lp	0.88530 <.0001	1.00000

B.34 Residual plots for Model S1

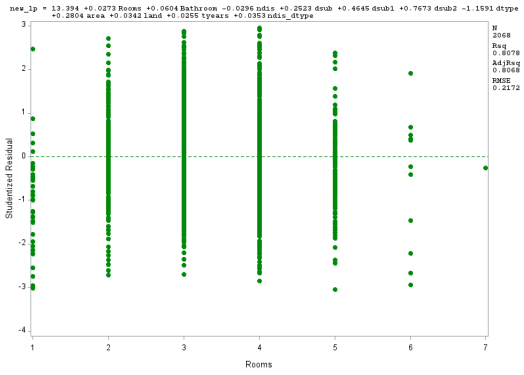


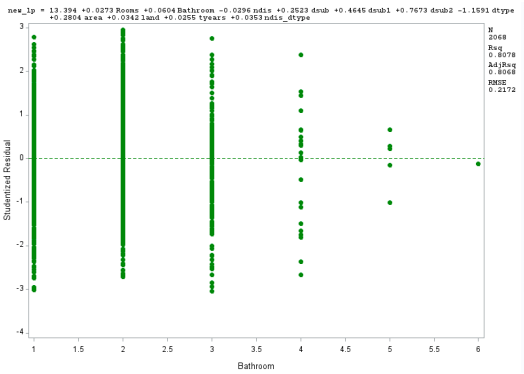


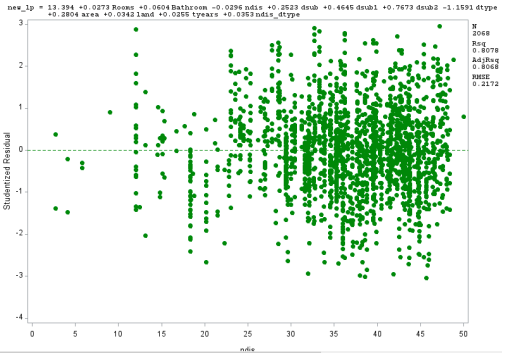


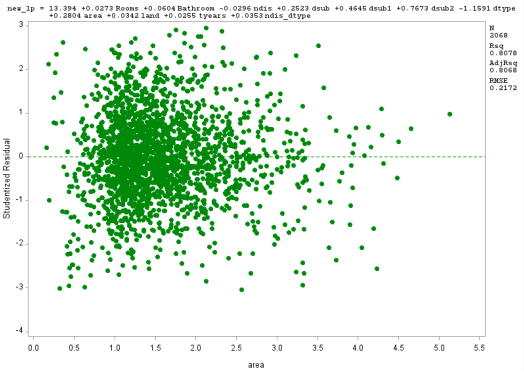


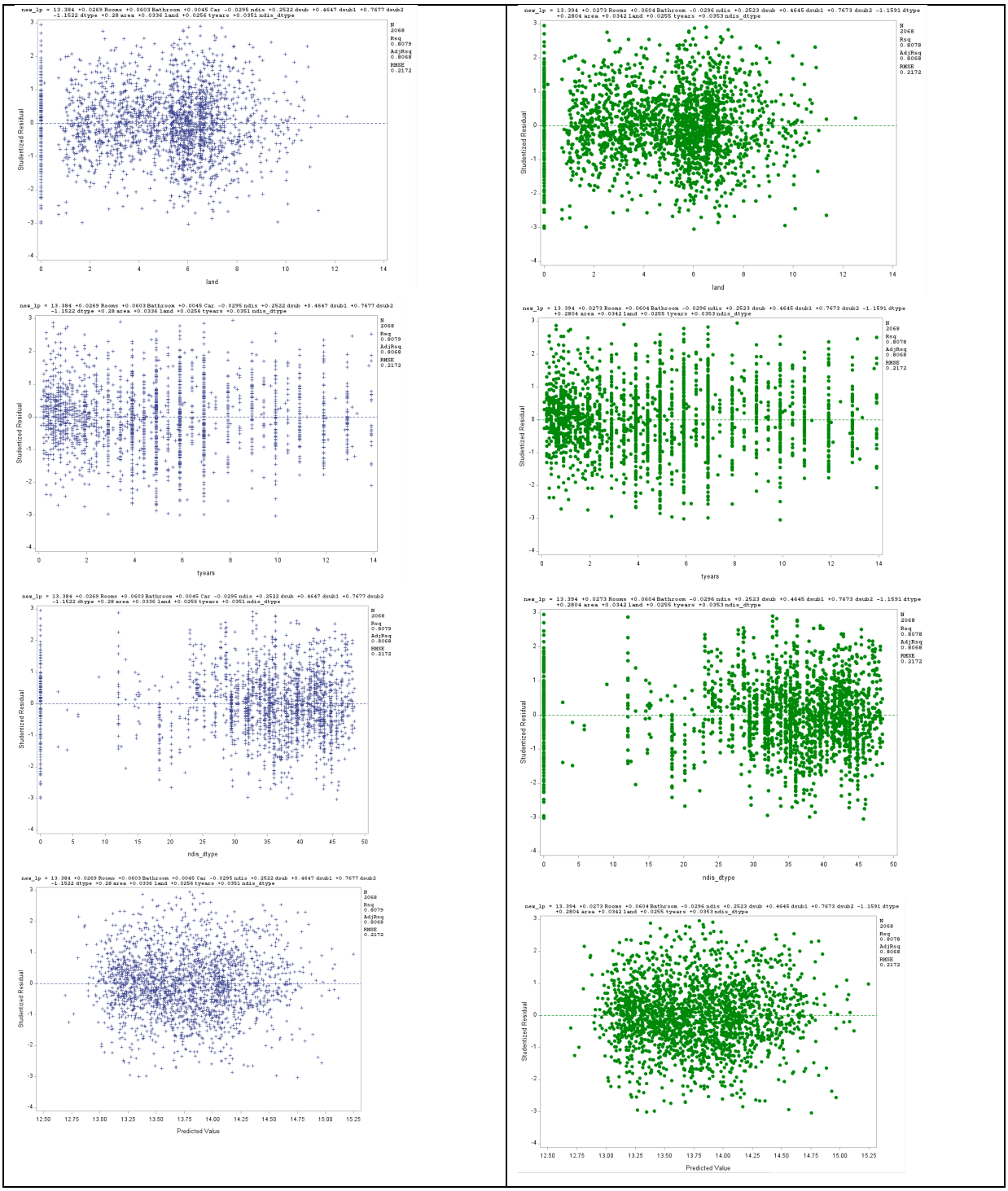
B.35 Residual plots for Model S1

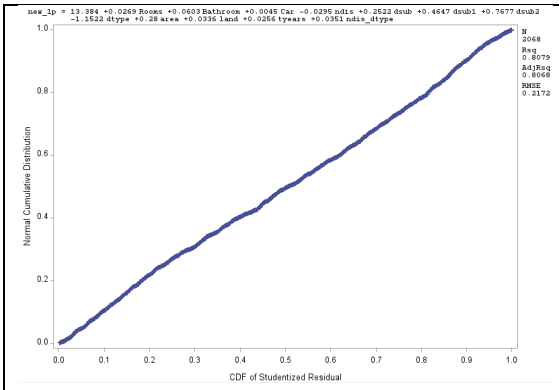


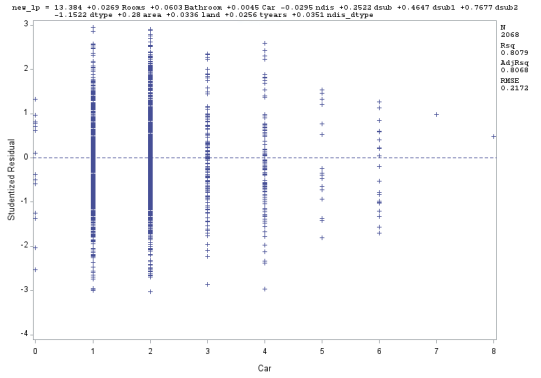


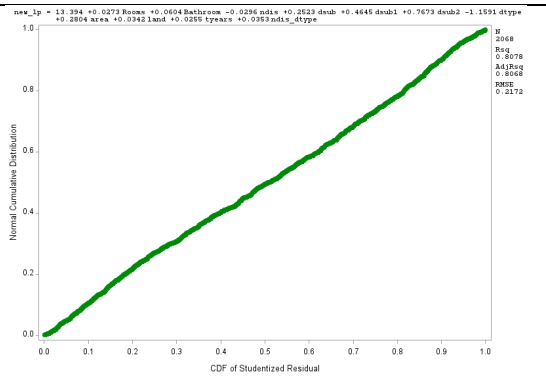





Studentized Residual
land
tyears
ndis_dtype
Predicted Value







B.36 CP Procedure for Model F3

Number in Model	C(p)	R-Square	Variables in Model
10	11.0000	0.7937	Rooms Bathroom Distance dsub dsub1 dsub2 dtype area land tyears
9	36.4647	0.7919	Bathroom Distance dsub dsub1 dsub2 dtype area land tyears
9	51.5342	0.7909	Rooms Bathroom dsub dsub1 dsub2 dtype area land tyears

B.37 Reg Procedure for F3

Full Model F3 PROC REG DATA=mbhouse

The REG Procedure
Model: MODEL2
Dependent Variable: lnprice

Number of Observations Read	3134
Number of Observations Used	3134

Analysis of Variance

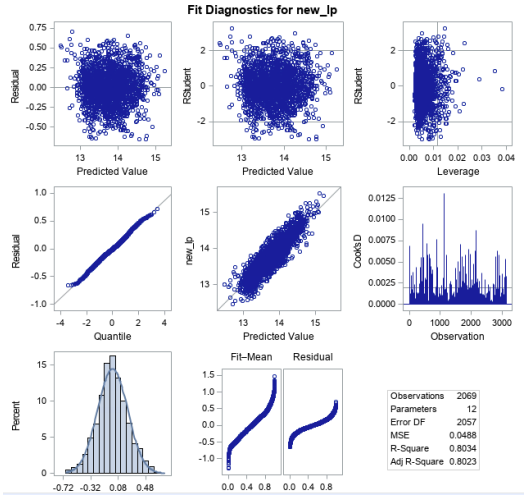
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	615.10947	61.51095	1201.44	<.0001
Error	3123	159.89101	0.05120		
Corrected Total	3133	775.00048			

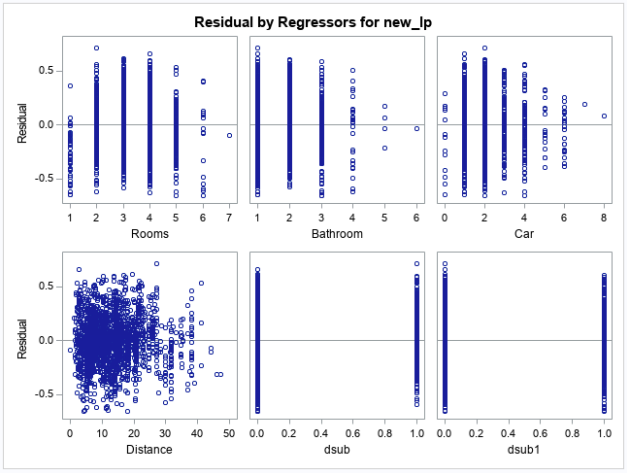
Root MSE	0.22627	R-Square	0.7937
Dependent Mean	13.76955	Adj R-Sq	0.7930
Coeff Var	1.64326		

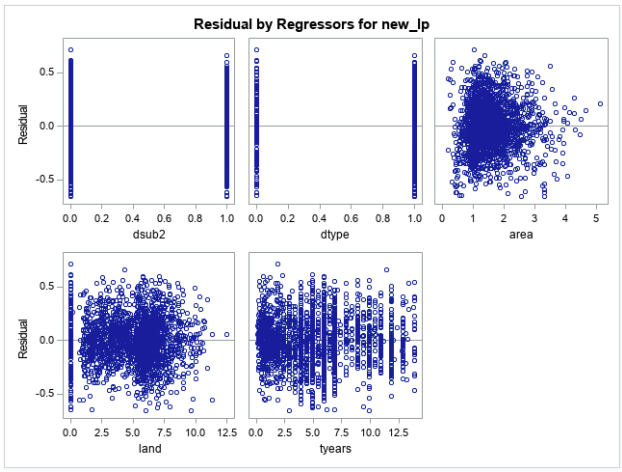
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	12.18917	0.02709	449.94	<.0001	0
Rooms	1	0.03965	0.00757	5.24	<.0001	2.67425
Bathroom	1	0.06114	0.00843	7.26	<.0001	2.11470
Distance	1	-0.00534	0.00081897	-6.52	<.0001	2.52598
dsub	1	0.24461	0.01320	18.52	<.0001	1.75861
dsub1	1	0.45701	0.01436	31.82	<.0001	2.36702
dsub2	1	0.73094	0.01773	41.22	<.0001	4.19472
dtype	1	0.31849	0.02008	15.86	<.0001	1.56414
area	1	0.27031	0.00987	27.38	<.0001	2.62802
land	1	0.03013	0.00243	12.39	<.0001	2.18233
tyears	1	0.02894	0.00151	19.15	<.0001	1.74764

B.38 Residual Plots for F3







B.39 Model S3 on Training Set

Model S3

The REG Procedure
Model: MODEL1
Dependent Variable: new_lp

Number of Observations Read	3134
Number of Observations Used	2069
Number of Observations with Missing Values	1065

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	410.15066	41.01507	839.91	<.0001
Error	2058	100.49741	0.04883		
Corrected Total	2068	510.64807			

Root MSE	0.22098	R-Square	0.8032
Dependent Mean	13.77114	Adj R-Sq	0.8022
Coeff Var	1.60467		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	12.17883	0.03301	368.99	<.0001	0	0
Rooms	1	0.03737	0.00917	4.08	<.0001	0.06578	2.72182
Bathroom	1	0.05742	0.01009	5.69	<.0001	0.08117	2.12765
Distance	1	-0.00514	0.00101	-5.09	<.0001	-0.08047	2.61223
dsub	1	0.25443	0.01596	15.94	<.0001	0.20600	1.74700
dsub1	1	0.46690	0.01736	26.89	<.0001	0.40741	2.40023
dsub2	1	0.74410	0.02149	34.62	<.0001	0.70032	4.27799
dtype	1	0.31939	0.02446	13.06	<.0001	0.15877	1.54546
area	1	0.28364	0.01205	23.53	<.0001	0.37494	2.65443
land	1	0.02921	0.00291	10.04	<.0001	0.14453	2.16681
tyears	1	0.02781	0.00183	15.17	<.0001	0.19737	1.77121

B.40 Validation Stats for Model S3

Validation stats for Model S3

Obs	_TYPE_	_FREQ_	rmse	mae
1	0	1065	0.23664	0.18505

Validation stats for Model S3

The CORR Procedure

2 Variables:	lnprice yhat

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
lnprice	1065	13.76646	0.49843	14661	12.42922	15.65606	
yhat	1065	13.76205	0.44825	14657	12.54027	15.37268	Predicted Value of new_lp

Pearson Correlation Coefficients, N = 1065
Prob > |r| under H0: Rho=0

	lnprice	yhat
lnprice	1.00000	0.88031 <.0001
yhat Predicted Value of new_lp	0.88031 <.0001	1.00000

B.41 Case 1

The SAS System

Obs	lnprice	rooms	bathroom	distance	dsub	dsub1	dsub2	dtype	area	land	tyears	Suburb	Type
1	.	3	2	8	0	0	1	1	1.50	3	4.0		
2	13.0919	2	1	3.3	0	0	1	0	0.73	0	4.9	South Yarra	u

B.42 Predicted Results 1 by Model S3

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	14.0487	0.009887	14.0293	14.0681	13.6046	14.4928	.
2	13.1	13.3821	0.0161	13.3504	13.4137	12.9373	13.8268	-0.2902

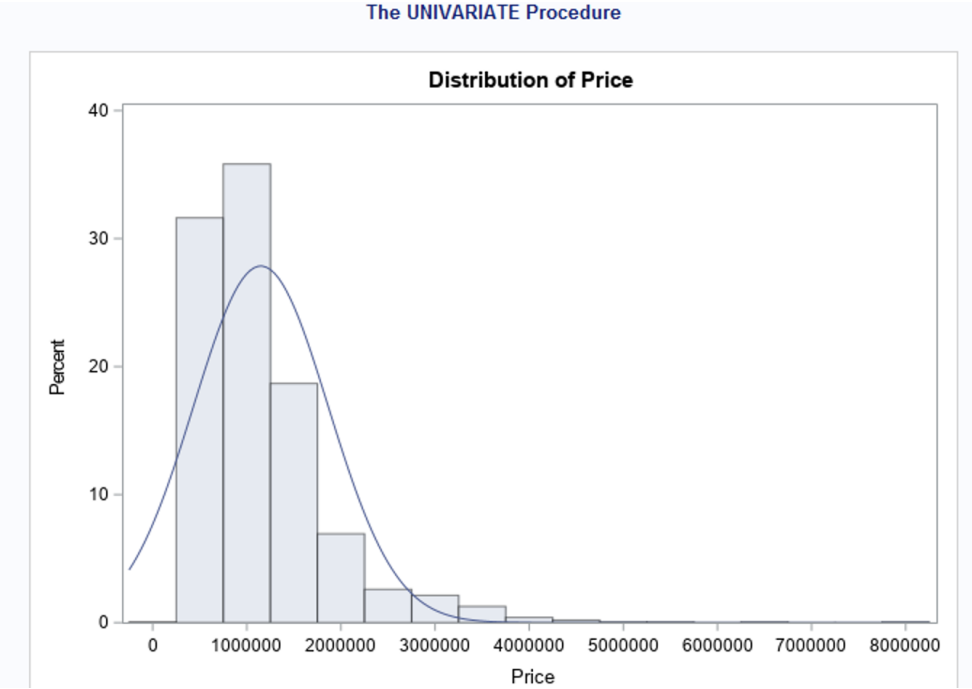
B.43 Predicted Results 2 by Model S3

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	13.3178	0.0145	13.2892	13.3463	12.8732	13.7623	.
2	13.1	13.3821	0.0161	13.3504	13.4137	12.9373	13.8268	-0.2902

Appendix - C (Nida Rasool)

C.1



C.2

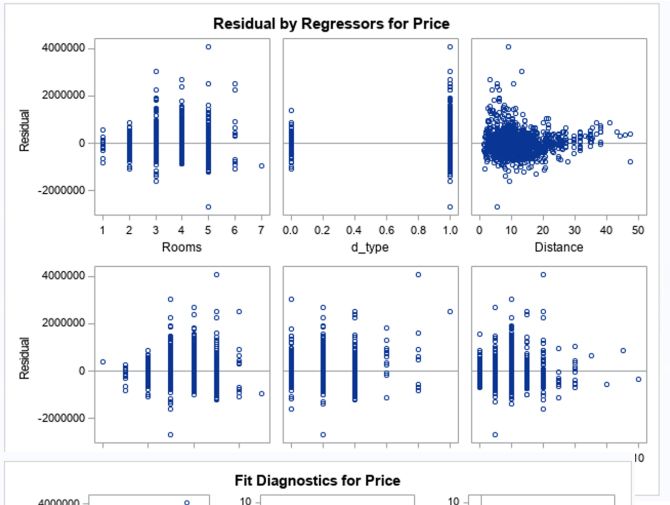
Model	10	4.392362E14	4.392362E13	198.95	<.0001
Error	1488	3.285104E14	2.207731E11		
Corrected Total	1498	7.677466E14			

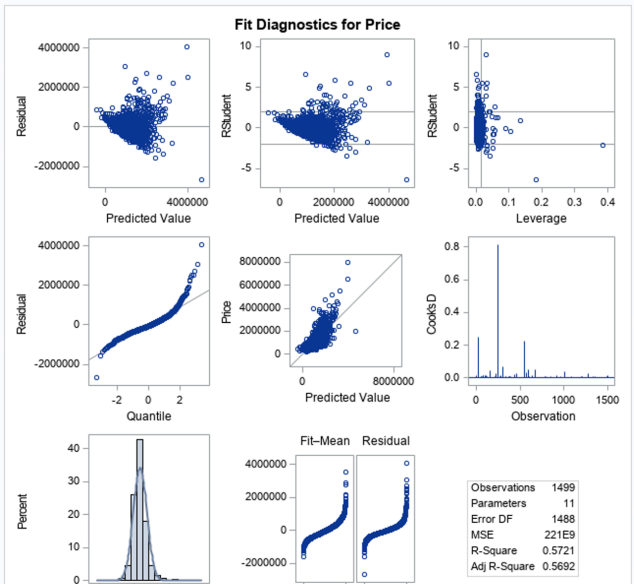
Root MSE	469865	R-Square	0.5721
Dependent Mean	1149861	Adj R-Sq	0.5692
Coeff Var	40.86276		

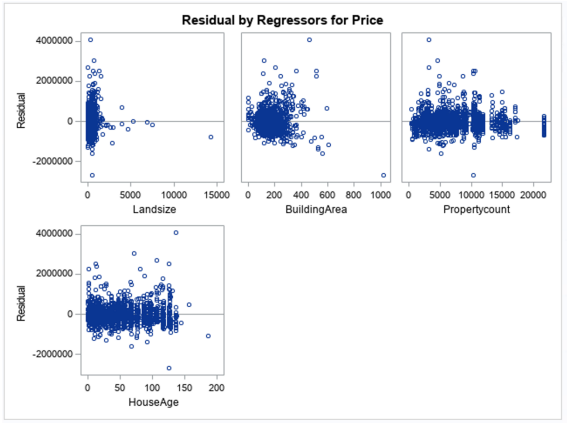
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-91879	60013	-1.53	0.1260
Rooms	1	61046	57540	1.06	0.2889
d_type	1	107207	40266	2.66	0.0078
Distance	1	-29957	2032.75606	-14.74	<.0001
Bedroom2	1	-23602	56634	-0.42	0.6769
Bathroom	1	314235	23251	13.51	<.0001
Car	1	35784	12761	2.80	0.0051
Landsize	1	35.59427	20.95592	1.70	0.0896
BuildingArea	1	3064.10533	215.61508	14.21	<.0001
Propertycount	1	-2.17795	2.70679	-0.80	0.4212
HouseAge	1	6236.62309	412.76085	15.11	<.0001

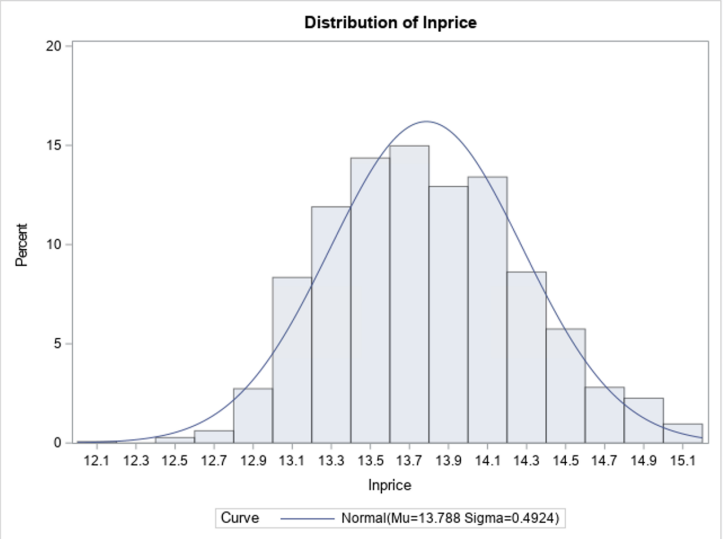
C.3


Residual by Regressors for Price
Residual
Rooms
d_type
Distance


Fit Diagnostics for Price
Residual
RStudent
Predicted Value
Leverage
Quantile
Price
CooksD
Observation
Percent
Fit–Mean
Residual
Observations 1499
Parameters 11
Error DF 1488
MSE 221E9
R-Square 0.5721
Adj R-Square 0.5692


Residual by Regressors for Price
Residual
Landsize
BuildingArea
Propertycount
HouseAge

C.4



Number of Observations Read	1462
Number of Observations Used	1462

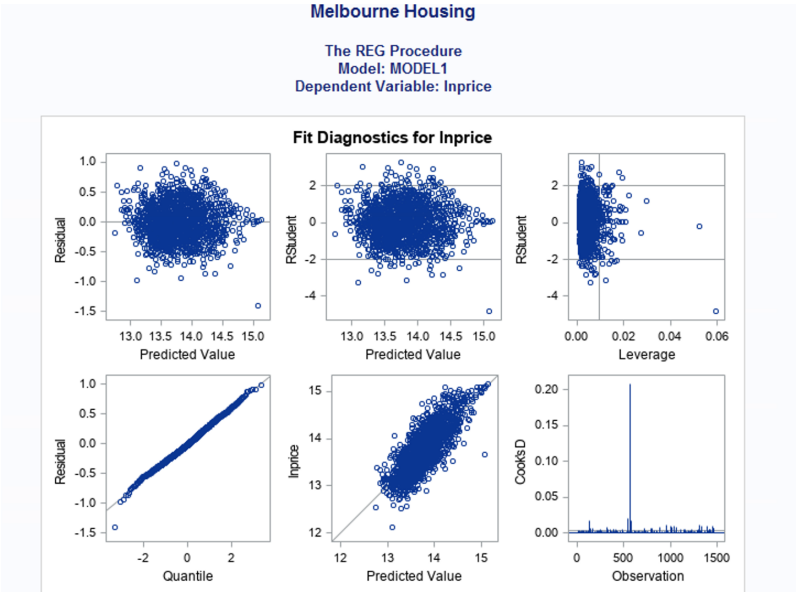
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	221.85767	36.97628	406.53	<.0001
Error	1455	132.34234	0.09096		
Corrected Total	1461	354.20001			

Root MSE	0.30159	R-Square	0.6264
Dependent Mean	13.78750	Adj R-Sq	0.6248
Coeff Var	2.18742		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	12.91222	0.03453	373.99	<.0001
d_type	1	0.16907	0.02581	6.55	<.0001
Distance	1	-0.02455	0.00133	-18.49	<.0001
Bathroom	1	0.18072	0.01521	11.88	<.0001
Car	1	0.01654	0.00848	1.95	0.0514
BuildingArea	1	0.00295	0.00016387	17.98	<.0001
HouseAge	1	0.00496	0.00027634	17.95	<.0001



C.5

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	BuildingArea		1	0.2495	0.2495	1036.08	364.11	<.0001
2	HouseAge		2	0.2315	0.4811	381.214	488.11	<.0001
3	Distance		3	0.0855	0.5666	140.521	215.74	<.0001
4	Bathroom		4	0.0353	0.6020	42.2564	96.96	<.0001
5	d_type		5	0.0130	0.6150	7.2655	36.95	<.0001
6	Car		6	0.0008	0.6158	7.0000	2.27	0.1326

C.6

Number in Model	C(p)	R-Square	Variables in Model
6	7.0000	0.6158	d_type Distance Bathroom Car BuildingArea HouseAge
5	7.2655	0.6150	d_type Distance Bathroom BuildingArea HouseAge
5	38.9706	0.6038	Distance Bathroom Car BuildingArea HouseAge
4	42.2564	0.6020	Distance Bathroom BuildingArea HouseAge
5	105.9605	0.5802	d_type Distance Car BuildingArea HouseAge
4	108.1787	0.5787	d_type Distance BuildingArea HouseAge
4	134.6474	0.5694	Distance Car BuildingArea HouseAge

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	165.78090	33.15618	348.55	<.0001
Error	1091	103.78257	0.09513		
Corrected Total	1096	269.56346			

Root MSE	0.30843	R-Square	0.6150
Dependent Mean	13.78057	Adj R-Sq	0.6132
Coeff Var	2.23812		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	12.92955	0.04005	322.84	<.0001
d_type	1	0.18001	0.02961	6.08	<.0001
Distance	1	-0.02414	0.00151	-15.97	<.0001
Bathroom	1	0.18062	0.01781	10.14	<.0001
BuildingArea	1	0.00295	0.00019050	15.46	<.0001
HouseAge	1	0.00490	0.00032447	15.09	<.0001

C.7

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	14.1110	0.0178	14.0761	14.1459	13.5178	14.7042	.
2	13.3	13.4759	0.0282	13.4206	13.5312	12.8812	14.0707	-0.1344
3	13.0	13.3363	0.0257	13.2859	13.3867	12.7420	13.9306	-0.3083

C.8

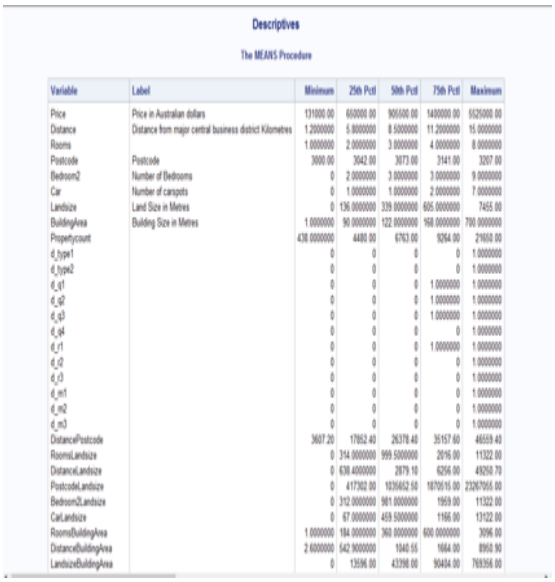
Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	13.3904	0.0263	13.3388	13.4421	12.7960	13.9849	.
2	13.3	13.4759	0.0282	13.4206	13.5312	12.8812	14.0707	-0.1344
3	13.0	13.3363	0.0257	13.2859	13.3867	12.7420	13.9306	-0.3083

Appendix - D (Sana Amreen)

D1-Variables

#	Variable	Description	Type
1	Suburb	Melbourne Suburb	Qualitative
3	Room	Number of rooms	Number
4	Type	h - house,cottage,villa, semi,terrace; u - unit, duplex; t - townhouse;	Qualitative
5	Price	Price in Australian dollars	Number
6	Method	S - property sold; SP - property sold prior; PI - property passed in; VB - vendor bid;	Qualitative
7	quarter	quarter house was sold(q1-q4=2016,q5=2017)	Date
8	Distance	Distance from major central business district Kilometres	Number
9	Postcode	Postcode	Qualitative
10	Bedroom2	Number of Bedrooms	Number
11	Car	Number of carspots	Number
12	Landsize	Land Size in Metres	Number
13	BuildingArea	Building Size in Metres	Number
14	CouncilArea	Governing council for the area	Qualitative
15	Regionname	General Region (West, North West, North, North east ...etc)	Qualitative
16	Propertycount	Number of properties that exist in the suburb.	Number

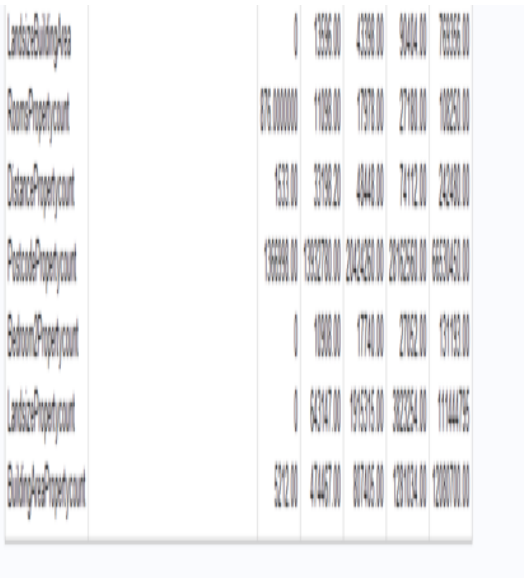
D2-descriptives



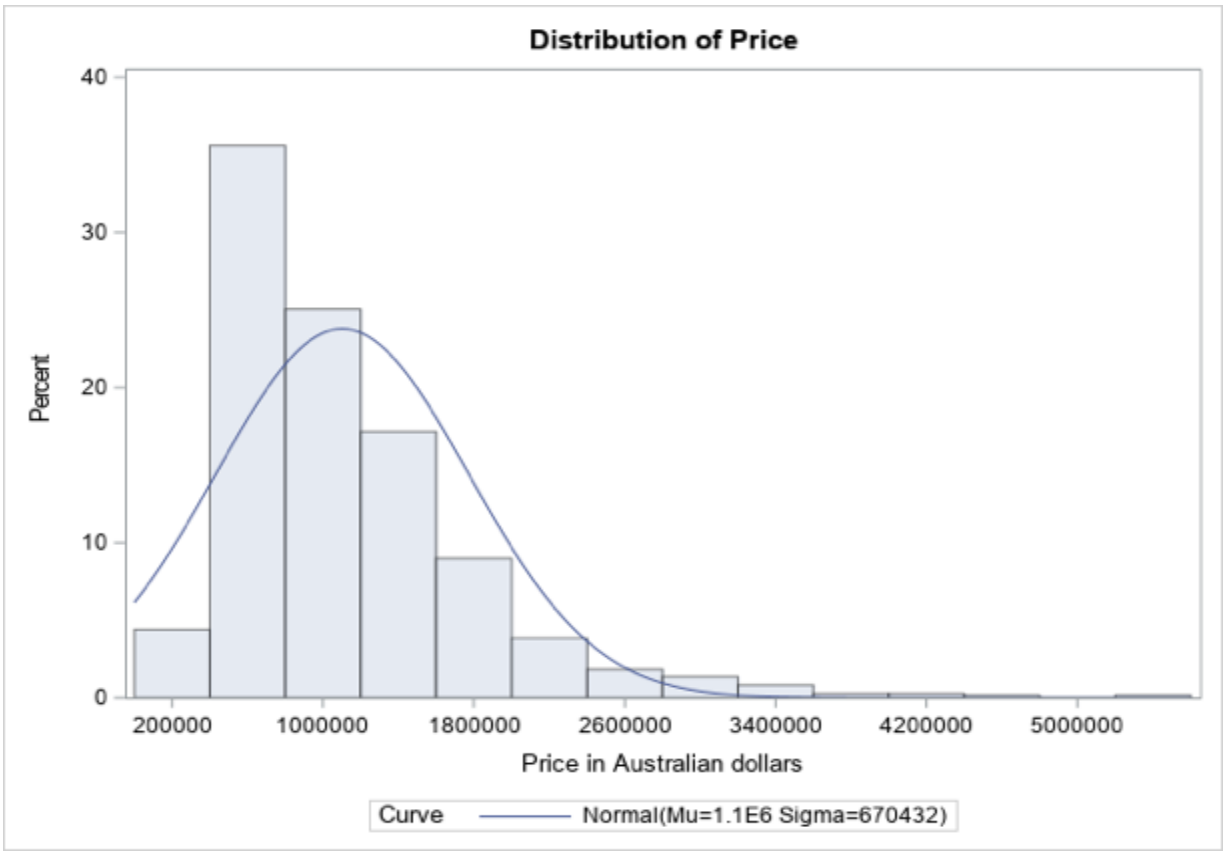
Descriptives

The MEANS Procedure

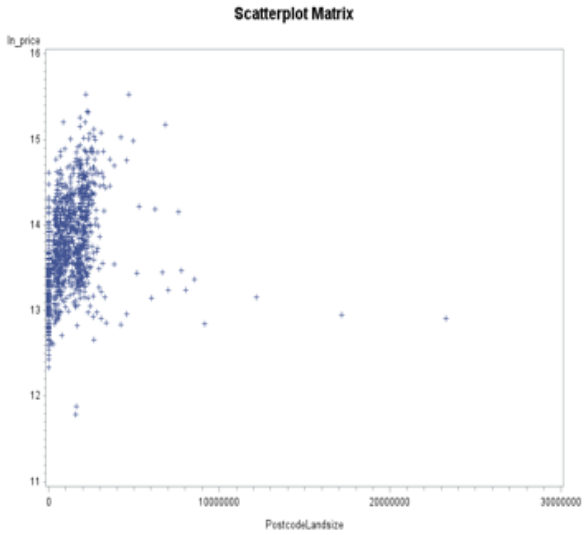
Variable	Label	Minimum	25th Pctl	50th Pctl	75th Pctl	Maximum
Price	Price in Australian dollars	131000.00	650000.00	905500.00	1400000.00	5525000.00
Distance	Distance from major central business district Kilometres	1.2000000	5.8000000	8.5000000	11.2000000	15.0000000
Rooms		1.0000000	2.0000000	3.0000000	4.0000000	8.0000000
Postcode	Postcode	3000.00	3042.00	3073.00	3141.00	3207.00
Bedroom2	Number of Bedrooms	0	2.0000000	3.0000000	3.0000000	9.0000000
Car	Number of carspots	0	1.0000000	1.0000000	2.0000000	7.0000000
Landsize	Land Size in Metres	0	136.0000000	339.0000000	605.0000000	7455.00
BuildingArea	Building Size in Metres	1.0000000	90.0000000	122.0000000	168.0000000	780.0000000
Propertycount		438.0000000	4480.00	6763.00	9264.00	21650.00
d_type1		0	0	0	0	1.0000000
d_type2		0	0	0	0	1.0000000
d_s1		0	0	0	1.0000000	1.0000000
d_s2		0	0	0	1.0000000	1.0000000
d_s3		0	0	0	1.0000000	1.0000000
d_s4		0	0	0	0	1.0000000
d_r1		0	0	0	1.0000000	1.0000000
d_r2		0	0	0	0	1.0000000
d_r3		0	0	0	0	1.0000000
d_m1		0	0	0	0	1.0000000
d_m2		0	0	0	0	1.0000000
d_m3		0	0	0	0	1.0000000
DistancePostcode		3607.20	17862.40	26378.40	35157.60	46559.40
RoomsLandsize		0	314.0000000	999.5000000	2016.00	11322.00
DistanceLandsize		0	638.4000000	2879.10	6256.00	43259.70
PostcodeLandsize		0	417302.00	1035652.50	1670515.00	23267055.00
Bedroom2Landsize		0	312.0000000	981.0000000	1959.00	11322.00
CarLandsize		0	67.0000000	459.5000000	1166.00	13122.00
RoomsBuildingArea		1.0000000	184.0000000	360.0000000	600.0000000	3096.00
DistanceBuildingArea		2.6000000	542.9000000	1040.55	1664.00	8960.90
LandsizeBuildingArea		0	13596.00	43298.00	90404.00	769356.00

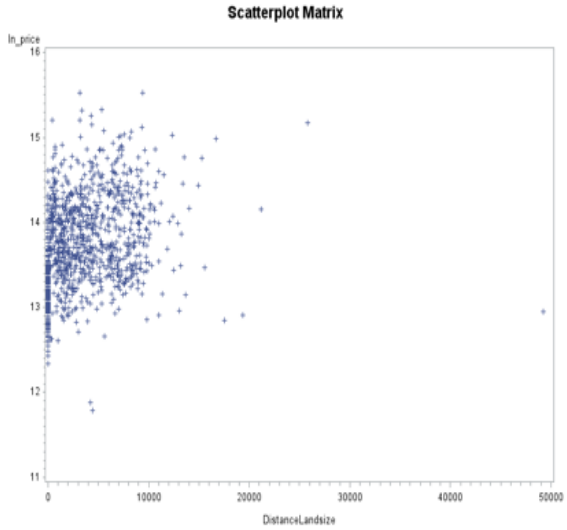


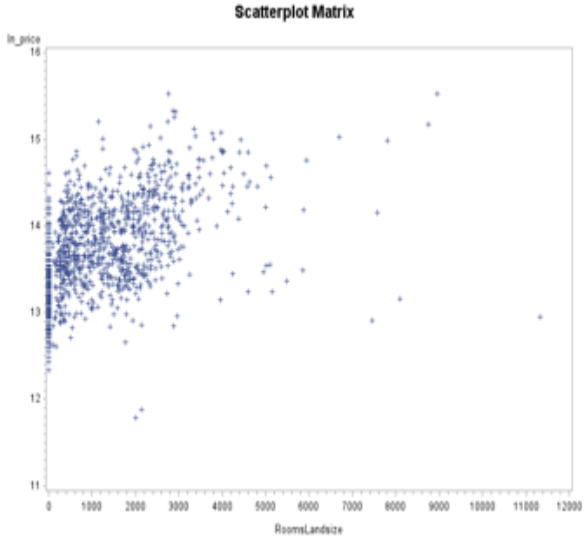
D3 -histogram

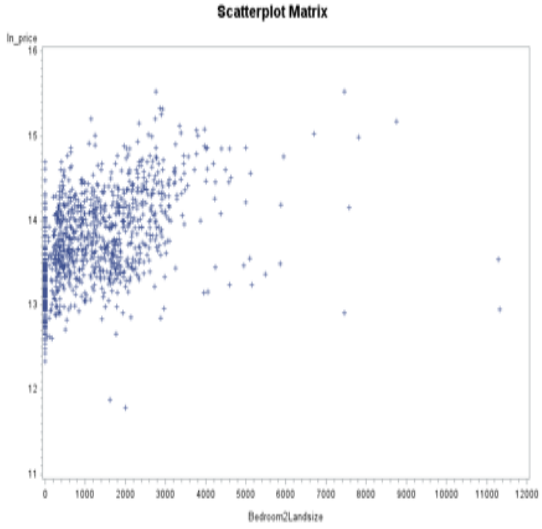


D4-scatter plots









D5-correlation procedure

	Price	Distance	Rooms	Postcode	Bedroom2	Car	Landsize	BuildingArea	Propertycount	d_type1	d_type2	d_q1	d_q2	d_q3	d_q4	d_r1	d_r2	d_r3	
Price Price in Australian dollars	1.00000 1090	-0.07207 0.0173 1090	0.55447 <.0001 1090	0.29421 <.0001 1090	0.51894 <.0001 1090	0.32064 <.0001 1090	0.25004 <.0001 1090	0.64497 <.0001 1070	-0.09033 0.0028 1090	-0.12080 <.0001 1090	-0.43937 <.0001 1090	-0.04077 0.1786 1090	-0.04134 0.1726 1090	0.08715 0.0040 1090	-0.00309 0.9189 1090	0.33830 <.0001 1090	0.03561 0.2402 1090	-0.16836 <.0001 1090	0.(0
Distance Distance from major central business district Kilometres	-0.07207 0.0173 1090	1.00000 1090	0.28757 <.0001 1090	0.08028 0.0080 1090	0.28739 <.0001 1090	0.27006 <.0001 1090	0.18690 <.0001 1090	0.19571 <.0001 1070	-0.15227 <.0001 1090	0.06020 0.0469 1090	-0.22936 <.0001 1090	-0.03718 0.2200 1090	-0.00741 0.8068 1090	0.04310 0.1550 1090	0.00078 0.9796 1090	0.01455 0.6313 1090	0.28017 <.0001 1090	0.10612 0.0004 1090	0.(0
Rooms	0.55447 <.0001 1090	0.28757 <.0001 1090	1.00000 1090	-0.00166 0.9563 1090	0.92185 <.0001 1090	0.44134 <.0001 1090	0.27324 <.0001 1090	0.68666 <.0001 1070	-0.14892 <.0001 1090	0.00213 0.9441 1090	-0.55826 <.0001 1090	0.02935 0.3330 1090	-0.04285 0.1575 1090	0.04151 0.1709 1090	-0.02674 0.3778 1090	0.03383 0.2645 1090	0.11934 <.0001 1090	0.06266 0.0386 1090	0.(0
Postcode Postcode	0.29421 <.0001 1090	0.08028 0.0080 1090	-0.00166 0.9563 1090	1.00000 1090	0.00071 0.9812 1090	0.00844 0.7808 1090	0.01560 0.6070 1090	0.05345 0.0806 1070	0.12599 <.0001 1090	-0.03979 0.1893 1090	0.10226 0.0007 1090	-0.02707 0.3719 1090	0.00185 0.9514 1090	-0.00187 0.9509 1090	0.02321 0.4440 1090	0.82512 <.0001 1090	0.05672 0.0612 1090	-0.60490 <.0001 1090	0.(0
Bedroom2 Number of Bedrooms	0.51894 <.0001 1090	0.28739 <.0001 1090	0.92185 <.0001 1090	0.00071 0.9812 1090	1.00000 1090	0.45069 <.0001 1090	0.25869 <.0001 1090	0.64940 <.0001 1070	-0.14453 <.0001 1090	-0.00017 0.9956 1090	-0.53348 <.0001 1090	-0.00449 0.8823 1090	-0.05347 0.0777 1090	0.06815 0.0244 1090	-0.00524 0.8628 1090	0.03185 0.2934 1090	0.11033 0.0003 1090	0.05642 0.0626 1090	0.(0
Car Number of carspots	0.32064 <.0001 1090	0.27006 <.0001 1090	0.44134 <.0001 1090	0.00844 0.7808 1090	0.45069 <.0001 1090	1.00000 1090	0.27760 <.0001 1090	0.40233 <.0001 1070	-0.05950 0.0495 1090	-0.03791 0.2111 1090	-0.26926 <.0001 1090	-0.00054 0.9857 1090	-0.00776 0.7979 1090	0.06154 0.0422 1090	-0.06939 0.0220 1090	0.02508 0.4082 1090	0.12766 <.0001 1090	0.01626 0.5919 1090	0.(0
Landsize Land Size in Metres	0.25004 <.0001 1090	0.18690 <.0001 1090	0.27324 <.0001 1090	0.01560 0.6070 1090	0.25869 <.0001 1090	0.27760 <.0001 1090	1.00000 1090	0.26265 <.0001 1070	-0.01106 0.7152 1090	-0.11831 <.0001 1090	-0.18925 <.0001 1090	0.04119 0.1741 1090	-0.07903 0.0091 1090	0.01900 0.5309 1090	0.03010 0.3207 1090	-0.00428 0.8878 1090	0.13173 <.0001 1090	0.00356 0.9065 1090	0.(0
BuildingArea Building Size in Metres	0.64497 <.0001 1070	0.19571 <.0001 1070	0.68666 <.0001 1070	0.05345 0.0806 1070	0.64940 <.0001 1070	0.40233 <.0001 1070	0.26265 <.0001 1070	1.00000 1070	-0.09366 0.0022 1070	-0.00673 0.8259 1070	-0.45118 <.0001 1070	0.00744 0.8078 1070	-0.00113 0.9705 1070	0.04154 0.1745 1070	-0.05953 0.0516 1070	0.10917 0.0003 1070	0.06948 0.0230 1070	0.03420 0.2636 1070	0.(<
Propertycount	-0.09033 0.0028 1090	-0.15227 <.0001 1090	-0.14892 <.0001 1090	0.12599 <.0001 1090	-0.14453 <.0001 1090	-0.05950 0.0495 1090	-0.01106 0.7152 1090	-0.09366 0.0022 1070	1.00000 1090	-0.03106 0.3055 1090	0.14926 <.0001 1090	-0.00112 0.9704 1090	0.04403 0.1463 1090	-0.06087 0.0445 1090	0.04236 0.1622 1090	0.06647 0.0282 1090	-0.19835 <.0001 1090	-0.25245 <.0001 1090	0.(0
d_type1	-0.12080 <.0001	0.06020 0.0469	0.00213 0.9441	-0.03979 0.1893	-0.00017 0.9956	-0.03791 0.2111	-0.11831 <.0001	-0.00673 0.8259	-0.03106 0.3055	1.00000 	-0.18467 <.0001	-0.04179 0.1680	0.00436 0.8857	0.02400 0.4286	0.01567 0.6052	-0.01561 0.6066	-0.01897 0.5316	0.02274 0.4532	0.(0

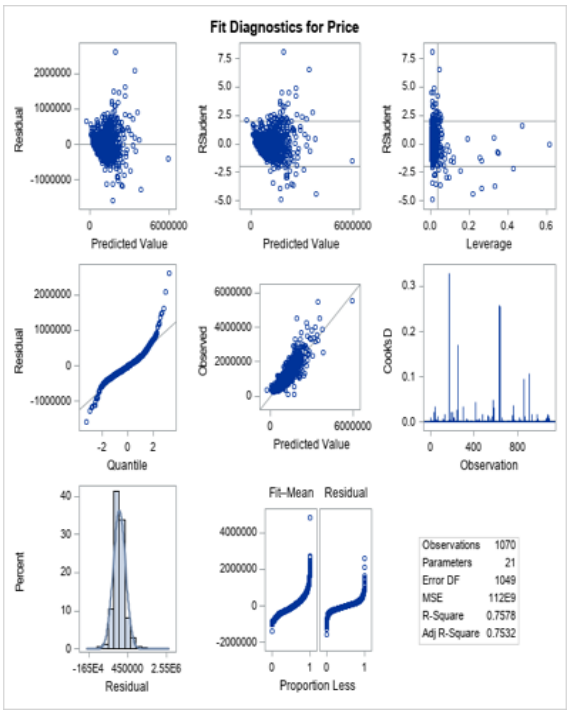
d_type2	-0.43937 <.0001 1090	-0.22936 <.0001 1090	-0.55826 <.0001 1090	0.10226 0.0007 1090	-0.53348 <.0001 1090	-0.26926 <.0001 1090	-0.18925 <.0001 1090	-0.45118 <.0001 1070	0.14926 <.0001 1090	-0.18467 <.0001 1090	1.00000 1090	-0.00966 0.7501 1090	0.01564 0.6059 1090	-0.04443 0.1427 1090	0.04816 0.1120 1090	0.12617 <.0001 1090	-0.08887 0.0033 1090		
d_q1	-0.04077 0.1786 1090	-0.03718 0.2200 1090	0.02935 0.3330 1090	-0.02707 0.3719 1090	-0.00449 0.8823 1090	-0.00054 0.9857 1090	0.04119 0.1741 1090	0.00744 0.8078 1070	-0.00112 0.9704 1090	-0.04179 0.1680 1090	-0.00966 0.7501 1090	1.00000 1090	-0.37960 <.0001 1090	-0.43259 <.0001 1090	-0.23348 <.0001 1090	-0.00782 0.7964 1090	-0.00287 0.9246 1090		
d_q2	-0.04134 0.1726 1090	-0.00741 0.8068 1090	-0.04285 0.1575 1090	0.00185 0.9514 1090	-0.05347 0.0777 1090	-0.00776 0.7979 1090	-0.07903 0.0091 1090	-0.00113 0.9705 1070	0.04403 0.1463 1090	0.00436 0.8857 1090	0.01564 0.6059 1090	-0.37960 <.0001 1090	1.00000 1090	-0.41504 <.0001 1090	-0.22402 <.0001 1090	-0.01897 0.5316 1090	-0.03046 0.3150 1090		
d_q3	0.08715 0.0040 1090	0.04310 0.1550 1090	0.04151 0.1709 1090	-0.00187 0.9509 1090	0.06815 0.0244 1090	0.06154 0.0422 1090	0.01900 0.5309 1090	0.04154 0.1745 1070	-0.06087 0.0445 1090	0.02400 0.4286 1090	-0.04443 0.1427 1090	-0.43259 <.0001 1090	-0.41504 <.0001 1090	1.00000 1090	-0.25528 <.0001 1090	-0.01455 0.6314 1090	0.06841 0.0239 1090		
d_q4	-0.00309 0.9189 1090	0.00078 0.9796 1090	-0.02674 0.3778 1090	0.02321 0.4440 1090	-0.00524 0.8628 1090	-0.06939 0.0220 1090	0.03010 0.3207 1090	-0.05953 0.0516 1070	0.04236 0.1622 1090	0.01567 0.6052 1090	0.04816 0.1120 1090	-0.23348 <.0001 1090	-0.22402 <.0001 1090	-0.25528 <.0001 1090	1.00000 1090	0.05071 0.0942 1090	-0.04605 0.1287 1090		
d_r1	0.33830 <.0001 1090	0.01455 0.6313 1090	0.03383 0.2645 1090	0.82512 <.0001 1090	0.03185 0.2934 1090	0.02508 0.4082 1090	-0.00428 0.8878 1090	0.10917 0.0003 1070	0.06647 0.0282 1090	-0.01561 0.6066 1090	0.12617 <.0001 1090	-0.00782 0.7964 1090	-0.01897 0.5316 1090	-0.01455 0.6314 1090	0.05071 0.0942 1090	1.00000 1090	-0.21889 <.0001 1090		
d_r2	0.03561 0.2402 1090	0.28017 <.0001 1090	0.11934 <.0001 1090	0.05672 0.0612 1090	0.11033 0.0003 1090	0.12766 <.0001 1090	0.13173 <.0001 1090	0.06948 0.0230 1070	-0.19835 <.0001 1090	-0.01897 0.5316 1090	-0.08887 0.0033 1090	-0.00287 0.9246 1090	-0.03046 0.3150 1090	0.06841 0.0239 1090	-0.04605 0.1287 1090	-0.21889 <.0001 1090	1.00000 1090		
d_r3	-0.16836 <.0001 1090	0.10612 0.0004 1090	0.06266 0.0386 1090	-0.60490 <.0001 1090	0.05642 0.0626 1090	0.01626 0.5919 1090	0.00356 0.9065 1090	0.03420 0.2636 1070	-0.25245 <.0001 1090	0.02274 0.4532 1090	-0.07502 0.0132 1090	0.04527 0.1353 1090	-0.04572 0.1314 1090	0.02224 0.4632 1090	-0.01991 0.5113 1090	-0.41437 <.0001 1090	-0.16215 <.0001 1090		
d_m1	0.08241 0.0065 1090	0.01847 0.5424 1090	0.10195 0.0007 1090	0.04661 0.1241 1090	0.10687 0.0004 1090	0.08797 0.0037 1090	0.08391 0.0056 1090	0.12001 <.0001 1070	0.04148 0.1711 1090	0.04808 0.1126 1090	-0.06625 0.0287 1090	0.02981 0.3255 1090	0.01821 0.5480 1090	-0.01893 0.5325 1090	-0.03058 0.3131 1090	0.05992 0.0480 1090	-0.01162 0.7016 1090		
d_m2	-0.15368 <.0001 1090	0.00234 0.9386 1090	-0.12156 <.0001 1090	-0.08880 0.0033 1090	-0.10508 0.0005 1090	-0.03726 0.2190 1090	-0.04007 0.1862 1090	-0.12496 <.0001 1070	0.03321 0.2733 1090	-0.04466 0.1407 1090	0.13442 <.0001 1090	0.02821 0.3521 1090	-0.00470 0.8768 1090	-0.03915 0.1965 1090	0.02420 0.4248 1090	-0.07488 0.0134 1090	-0.06693 0.0271 1090		
d_m3	0.05331 1090	-0.08202 1090	0.10097 1090	0.00467 1090	0.09999 1090	0.07032 1090	0.03989 1070	0.09335 1090	-0.01803 1090	0.02000 1090	0.00623 1090	0.07508 1090	-0.04719 1090	-0.00867 1090	-0.02054 1090	0.05779 1090	0.03470 1090		
DistancePostcode	-0.05726 0.0588 1090	0.99887 <.0001 1090	0.28650 <.0001 1090	0.12354 <.0001 1090	0.28635 <.0001 1090	0.26896 <.0001 1090	0.18666 <.0001 1090	0.19757 <.0001 1070	-0.14699 <.0001 1090	0.05910 0.0511 1090	-0.22353 <.0001 1090	-0.03859 0.2030 1090	-0.00656 0.8288 1090	0.04315 0.1546 1090	0.00100 0.9738 1090	0.05165 0.0883 1090	0.28158 <.0001 1090	0.07608 0.0120 1090	0.(0
RoomsLandsize	0.49368 <.0001 1090	0.29264 <.0001 1090	0.63821 <.0001 1090	0.04299 0.1561 1090	0.59515 <.0001 1090	0.45341 <.0001 1090	0.83769 <.0001 1090	0.53903 <.0001 1070	-0.09134 0.0025 1090	-0.14932 <.0001 1090	-0.35516 <.0001 1090	0.02898 0.3391 1090	-0.06383 0.0351 1090	0.03423 0.2588 1090	0.00681 0.8223 1090	0.04974 0.1007 1090	0.18049 <.0001 1090	0.00915 0.7629 1090	0.(0
DistanceLandsize	0.25390 <.0001 1090	0.55143 <.0001 1090	0.40633 <.0001 1090	0.05479 0.0706 1090	0.39640 <.0001 1090	0.39626 <.0001 1090	0.81165 <.0001 1090	0.36029 <.0001 1070	-0.12916 <.0001 1090	-0.14061 <.0001 1090	-0.32082 <.0001 1090	0.03312 0.2746 1090	-0.06377 0.0353 1090	0.03092 0.3078 1090	-0.00143 0.9623 1090	0.00356 0.9064 1090	0.27900 <.0001 1090	0.02594 0.3922 1090	0.(0
PostcodeLandsize	0.25687 <.0001 1090	0.18824 <.0001 1090	0.27420 <.0001 1090	0.03060 0.3128 1090	0.26000 <.0001 1090	0.27741 <.0001 1090	0.99972 <.0001 1090	0.26466 <.0001 1070	-0.00940 0.7565 1090	-0.11851 <.0001 1090	-0.18824 <.0001 1090	0.04053 0.1812 1090	-0.07871 0.0093 1090	0.01864 0.5387 1090	0.03098 0.3068 1090	0.00841 0.7815 1090	0.13280 <.0001 1090	-0.00663 0.8270 1090	0.(0
Bedroom2Landsize	0.47331 <.0001 1090	0.29626 <.0001 1090	0.61331 <.0001 1090	0.04939 0.1031 1090	0.65371 <.0001 1090	0.47370 <.0001 1090	0.81141 <.0001 1090	0.52673 <.0001 1070	-0.08855 0.0034 1090	-0.14617 <.0001 1090	-0.35268 <.0001 1090	0.02193 0.4696 1090	-0.06909 0.0225 1090	0.04008 0.1860 1090	0.01487 0.6238 1090	0.05179 0.0874 1090	0.16788 <.0001 1090	0.00840 0.7819 1090	0.(0
CarLandsize	0.37183 <.0001 1090	0.23506 <.0001 1090	0.38839 <.0001 1090	0.01754 0.5629 1090	0.40182 <.0001 1090	0.69652 <.0001 1090	0.77068 <.0001 1090	0.38782 <.0001 1070	-0.06478 0.0325 1090	-0.12005 <.0001 1090	-0.25051 <.0001 1090	0.02347 0.4389 1090	-0.05924 0.0505 1090	0.04378 0.1486 1090	-0.00795 0.7931 1090	0.01002 0.7411 1090	0.16070 <.0001 1090	0.01573 0.6039 1090	0.(0
RoomsBuildingArea	0.65995 <.0001 1070	0.20312 <.0001 1070	0.81673 <.0001 1070	0.06067 0.0472 1070	0.75727 <.0001 1070	0.42961 <.0001 1070	0.28534 <.0001 1070	0.95008 <.0001 1070	-0.11061 0.0003 1070	-0.03471 0.2566 1070	-0.42833 <.0001 1070	0.00935 0.7600 1070	-0.01058 0.7296 1070	0.04444 0.1463 1070	-0.05216 0.0881 1070	0.12152 <.0001 1070	0.07351 0.0162 1070	0.02742 0.3703 1070	0.(<
DistanceBuildingArea	0.38704 <.0001 1070	0.63117 <.0001 1070	0.61631 <.0001 1070	0.09662 0.0016 1070	0.58927 <.0001 1070	0.41294 <.0001 1070	0.29575 <.0001 1070	0.84314 <.0001 1070	-0.13752 <.0001 1070	0.01985 0.5166 1070	-0.40653 <.0001 1070	-0.01918 0.5309 1070	0.00079 0.9794 1070	0.04739 0.1213 1070	-0.03993 0.1918 1070	0.08851 0.0038 1070	0.20588 <.0001 1070	0.06593 0.0311 1070	0.(0
LandsizeBuildingArea	0.59211 <.0001 1070	0.24249 <.0001 1070	0.56502 <.0001 1070	0.07097 0.0203 1070	0.53474 <.0001 1070	0.44758 <.0001 1070	0.70571 <.0001 1070	0.75002 <.0001 1070	-0.08369 0.0062 1070	-0.13262 <.0001 1070	-0.31902 <.0001 1070	0.01592 0.6030 1070	-0.03676 0.2296 1070	0.03331 0.2763 1070	-0.01249 0.6832 1070	0.09382 0.0021 1070	0.15524 <.0001 1070	0.00054 0.9858 1070	0.(0
RoomsPropertycount	0.22457 <.0001 1090	0.01311 0.6654 1090	0.39077 <.0001 1090	0.10908 0.0003 1090	0.35414 <.0001 1090	0.18290 <.0001 1090	0.10729 0.0004 1090	0.28174 <.0001 1070	0.80710 <.0001 1090	-0.01126 0.7103 1090	-0.19709 <.0001 1090	0.01074 0.7232 1090	0.01753 0.5632 1090	-0.02864 0.3448 1090	0.02068 0.4953 1090	0.08512 0.0049 1090	-0.13440 <.0001 1090	-0.19739 <.0001 1090	0.(0

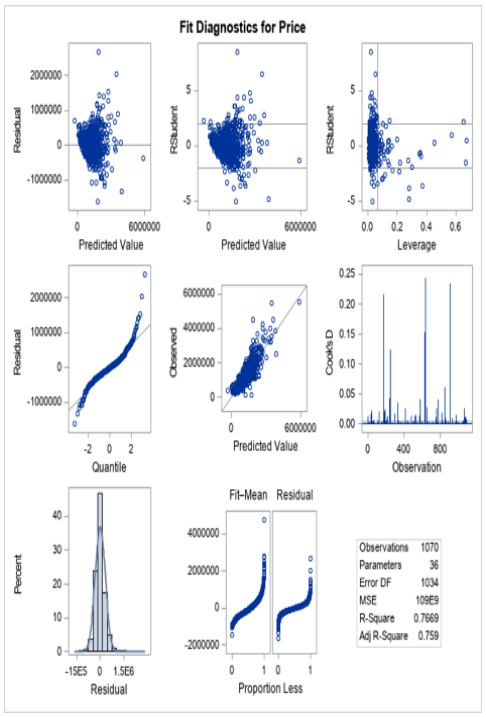
D6-multicollinearity

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	Intercept	1	1821189	2184800	0.83	0.4047		0
Distance	Distance from major central business district Kilometres	1	-636712	162975	-3.91	<.0001	0.00030235	3307.43989
Rooms		1	-137446	67380	-2.04	0.0416	0.02315	43.20215
Postcode	Postcode	1	-508.37656	707.67442	-0.72	0.4727	0.06102	16.38722
Bedroom2	Number of Bedrooms	1	196230	59891	3.28	0.0011	0.02834	35.28085
Car	Number of carspots	1	42018	22947	1.83	0.0674	0.23716	4.21663
Landsize	Land Size in Metres	1	-9629.13488	1335.43334	-7.21	<.0001	0.00027665	3614.66042
BuildingArea	Building Size in Metres	1	9374.90212	830.67682	11.29	<.0001	0.02371	42.18326
Propertycount		1	-131.04268	179.48400	-0.73	0.4655	0.00016516	6054.90129
d_type1		1	-240414	36583	-6.57	<.0001	0.83919	1.19162
d_type2		1	-387307	34671	-11.17	<.0001	0.48303	2.07027
d_q1		1	-22437	120376	-0.19	0.8522	0.03458	28.91826
d_q2		1	11524	120139	0.10	0.9236	0.03590	27.85814
d_q3		1	98958	120018	0.82	0.4098	0.03249	30.78281
d_q4		1	99111	122343	0.81	0.4181	0.06509	15.36234
d_r1		1	205816	46161	4.46	<.0001	0.20884	4.78928
d_r2		1	113131	48810	2.32	0.0207	0.59133	1.69110
d_r3		1	62212	34431	1.81	0.0711	0.47617	2.10007
d_m1		1	-29735	32582	-0.91	0.3616	0.91161	1.09696
d_m2		1	-39551	30776	-1.29	0.1990	0.91319	1.09506
d_m3		1	-170911	38180	-4.48	<.0001	0.89850	1.11296
DistancePostcode		1	211.27502	52.96529	3.99	<.0001	0.00029581	3380.54360
RoomsLandsize		1	255.75001	67.91958	3.77	0.0002	0.01311	76.25854
DistanceLandsize		1	-1.73497	10.94105	-0.16	0.8731	0.86995	14.49720
PostcodeLandsize		1	3.02862	0.43429	6.97	<.0001	0.00027322	3660.10253
Bedroom2Landsize		1	-237.33248	63.99637	-3.72	0.0002	0.01499	66.69210
CarLandsize		1	-11.63868	35.71307	-0.33	0.7446	0.07723	12.94519
RoomsBuildingArea		1	-349.31453	161.88688	-2.16	0.0309	0.05235	19.44307
DistanceBuildingArea		1	-643.89749	49.97031	-12.15	<.0001	0.04792	20.86739
LandsizeBuildingArea		1	4.14963	0.59122	7.02	<.0001	0.04474	22.35317
RoomsPropertycount		1	14.48144	5.91176	2.45	0.0144	0.01596	62.67201
DistancePropertycount		1	0.25548	0.72047	0.35	0.7230	0.09045	11.05634
PostcodePropertycount		1	0.04166	0.05796	0.72	0.4731	0.00016498	6061.34776
BedroomPropertycount		1	-9.64563	5.96666	-1.62	0.1521	0.01593	62.76839
LandsizePropertycount		1	-0.00261	0.00731	-0.36	0.7209	0.06572	15.21552
BuildingAreaPropertycount		1	-0.14203	0.03270	-4.34	<.0001	0.11666	8.57166

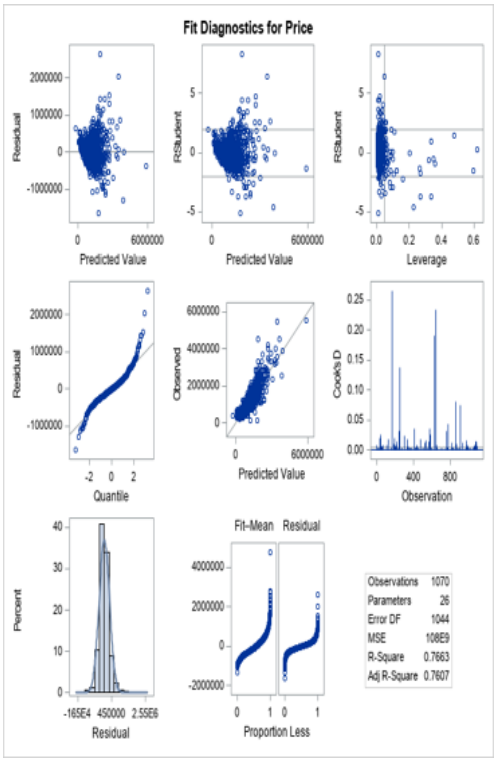
D7 -regression full model

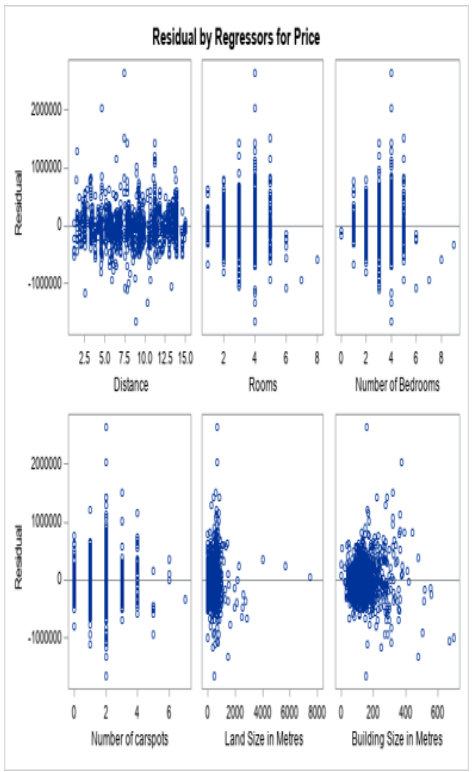
D8 regression final model





D9-adjrsquare method output





Final Model w/ options

The REG Procedure
Model: MODEL1
Dependent Variable: Price Price in Australian dollars

Number of Observations Read	1090
Number of Observations Used	1070
Number of Observations with Missing Values	20

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	25	3.7071E14	1.48284E13	136.91	<.0001
Error	1044	1.130693E14	1.083039E11		
Corrected Total	1069	4.837793E14			

Root MSE	329096	R-Square	0.7663
Dependent Mean	1107083	Adj R-Sq	0.7607
Coeff Var	29.72638		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	302943	94421	3.22	0.0013
Distance	Distance from major central business district Kilometres	1	-585789	124510	-4.70	<.0001
Rooms		1	-128419	63875	-2.01	0.0446
Bedroom2	Number of Bedrooms	1	199611	58663	3.39	0.0007
Car	Number of carspots	1	35583	13405	2.65	0.0081
Landsize	Land Size in Metres	1	-9710.20341	1291.98555	-7.52	<.0001
BuildingArea	Building Size in Metres	1	9516.82410	803.40561	11.85	<.0001
d_type1		1	-239707	35464	-6.76	<.0001
d_type2		1	-387953	33766	-11.49	<.0001
d_q1		1	-121114	24550	-4.93	<.0001
d_q2		1	-86313	24840	-3.47	0.0005
d_r1		1	198659	38906	5.11	<.0001
d_r2		1	110815	46749	2.37	0.0179
d_r3		1	70218	33175	2.12	0.0345
d_m2		1	-33455	30210	-1.11	0.2684
d_m3		1	-164457	37430	-4.39	<.0001
DistancePostcode		1	195.58135	40.24426	4.86	<.0001
RoomsLandsize		1	255.70438	65.59730	3.90	0.0001
PostcodeLandsize		1	3.03664	0.41525	7.31	<.0001
Bedroom2Landsize		1	-236.16684	59.46655	-3.97	<.0001
RoomsBuildingArea		1	-362.48846	153.95777	-2.35	0.0187
DistanceBuildingArea		1	-652.11668	44.75196	-14.57	<.0001
LandsizeBuildingArea		1	4.10464	0.53635	7.65	<.0001
RoomsPropertycount		1	13.79153	5.65889	2.44	0.0150
Bedroom2Propertycount		1	-9.20598	5.58829	-1.65	0.0998
BuildingAreaPropertycount		1	-0.14187	0.03216	-4.41	<.0001

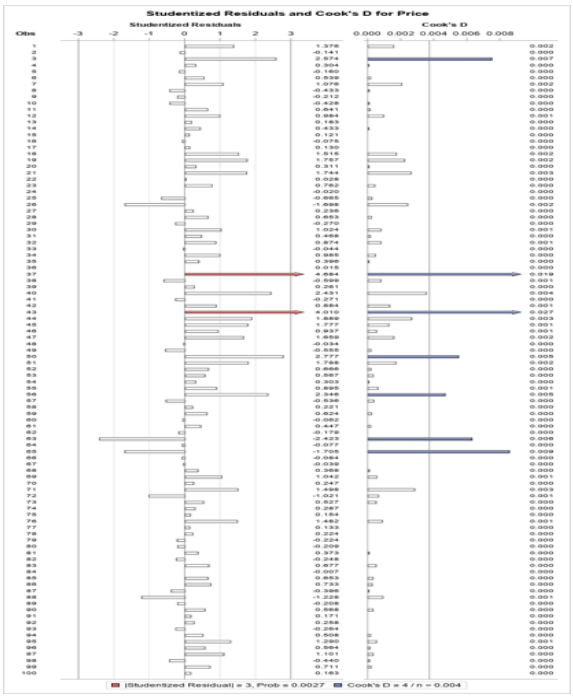
Model: MODEL1
Dependent Variable: Price

Adjusted R-Square Selection Method

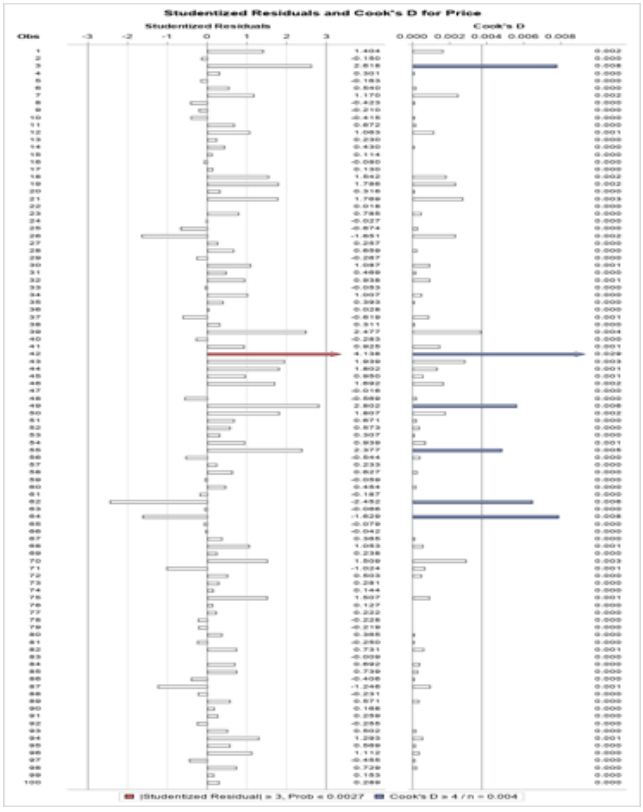
Number of Observations Read	1090
Number of Observations Used	1070
Number of Observations with Missing Values	20

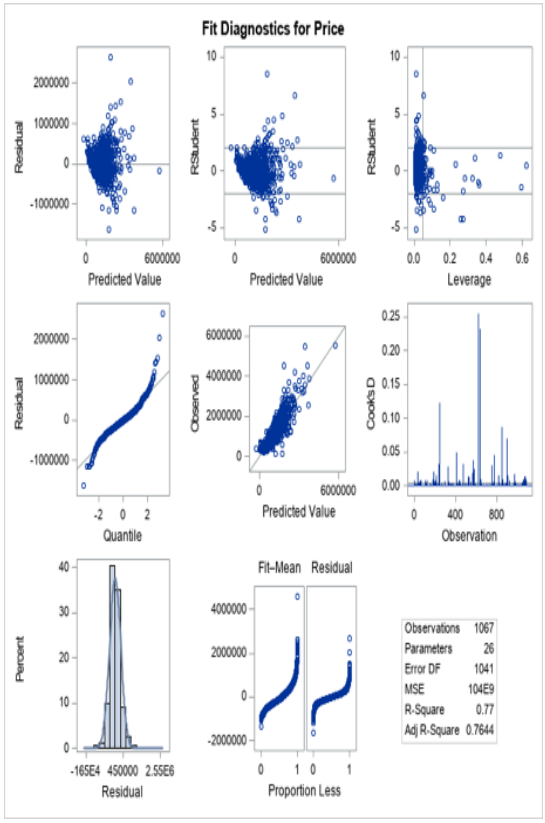
Number in Model	Adjusted R-Square	R-Square	Variables in Model
25	0.7607	0.7663	Distance Rooms Bedroom2 Car Landsize BuildingArea d_type1 d_type2 d_q1 d_q2 d_r1 d_r2 d_r3 d_m2 d_m3 DistancePostcode RoomsLandsize PostcodeLandsize Bedroom2Landsize RoomsBuildingArea DistanceBuildingArea LandsizeBuildingArea RoomsPropertycount Bedroom2Propertycount BuildingAreaPropertycount
26	0.7606	0.7665	Distance Rooms Bedroom2 Car Landsize BuildingArea d_type1 d_type2 d_q1 d_q2 d_r1 d_r2 d_r3 d_m1 d_m2 d_m3 DistancePostcode RoomsLandsize PostcodeLandsize Bedroom2Landsize RoomsBuildingArea DistanceBuildingArea LandsizeBuildingArea RoomsPropertycount Bedroom2Propertycount BuildingAreaPropertycount
24	0.7606	0.7660	Distance Rooms Bedroom2 Car Landsize BuildingArea d_type1 d_type2 d_q1 d_q2 d_r1 d_r2 d_r3 d_m3 DistancePostcode RoomsLandsize PostcodeLandsize Bedroom2Landsize RoomsBuildingArea DistanceBuildingArea LandsizeBuildingArea RoomsPropertycount Bedroom2Propertycount BuildingAreaPropertycount
26	0.7606	0.7664	Distance Rooms Bedroom2 Car Landsize BuildingArea d_type1 d_type2 d_q1 d_q3 d_q4 d_r1 d_r2 d_r3 d_m2 d_m3 DistancePostcode RoomsLandsize PostcodeLandsize Bedroom2Landsize RoomsBuildingArea DistanceBuildingArea LandsizeBuildingArea RoomsPropertycount Bedroom2Propertycount BuildingAreaPropertycount

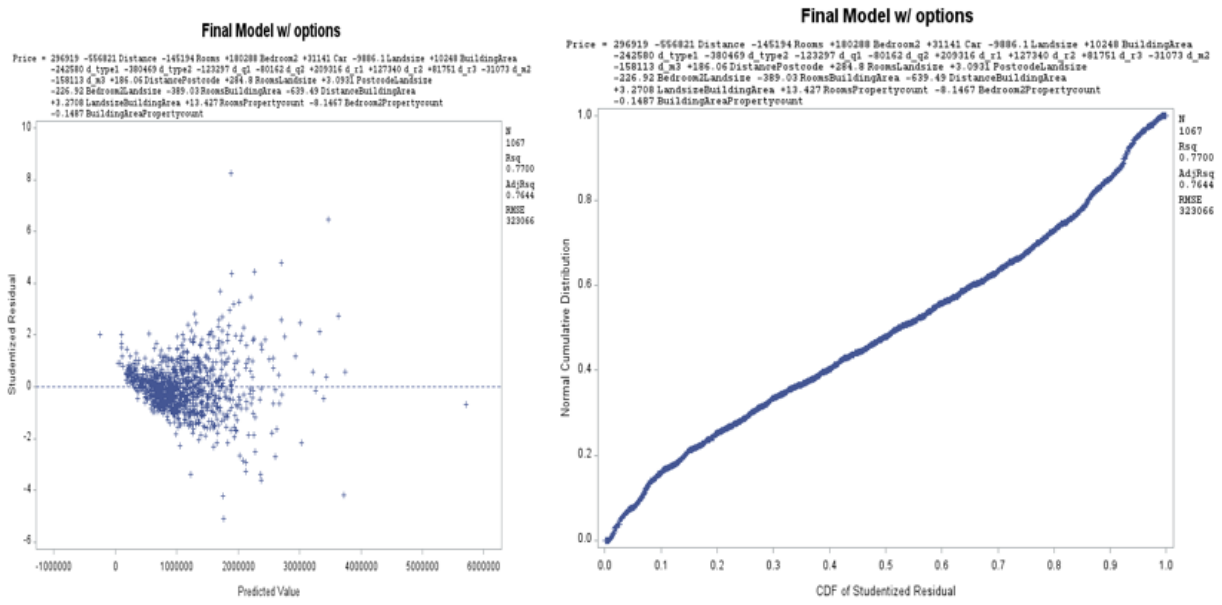
D10-influential points and outliers.



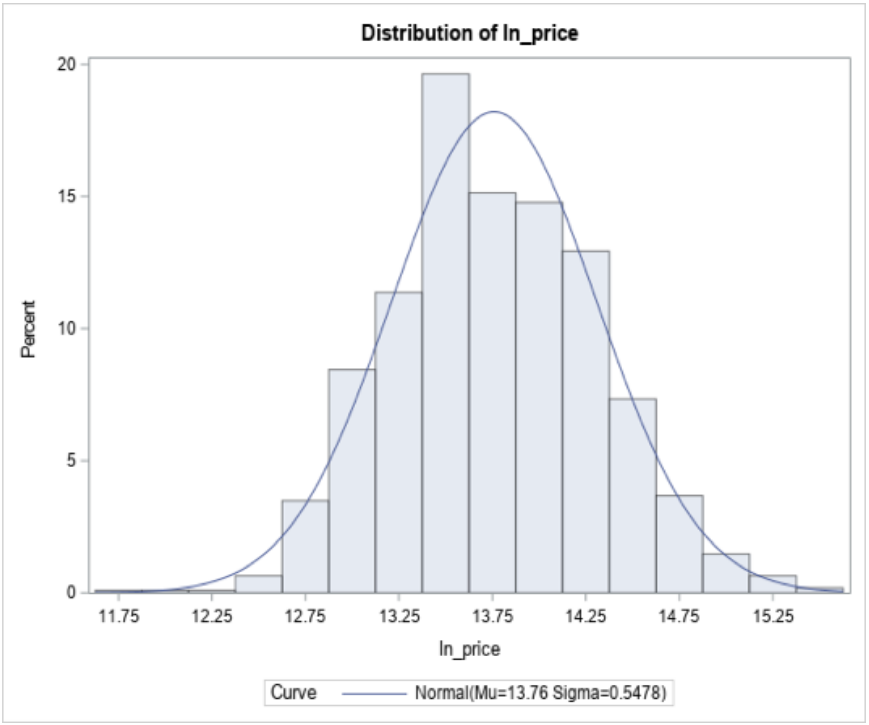
D11-after removing three influential point/outliers.



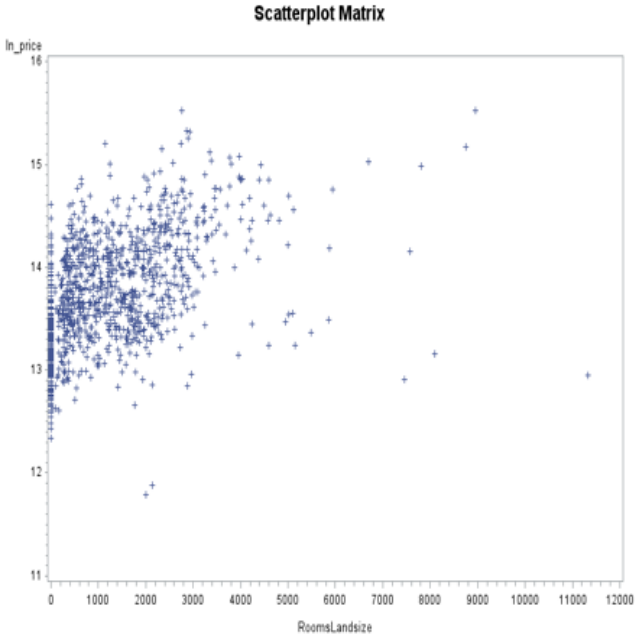


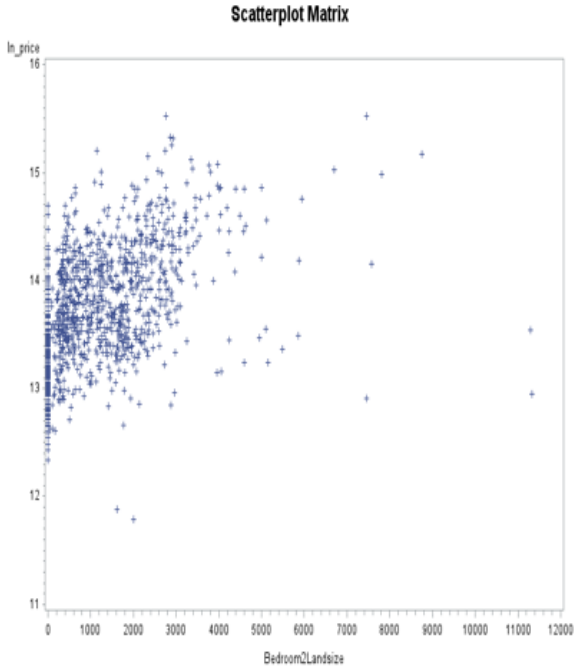


D12-histogram for ln_price

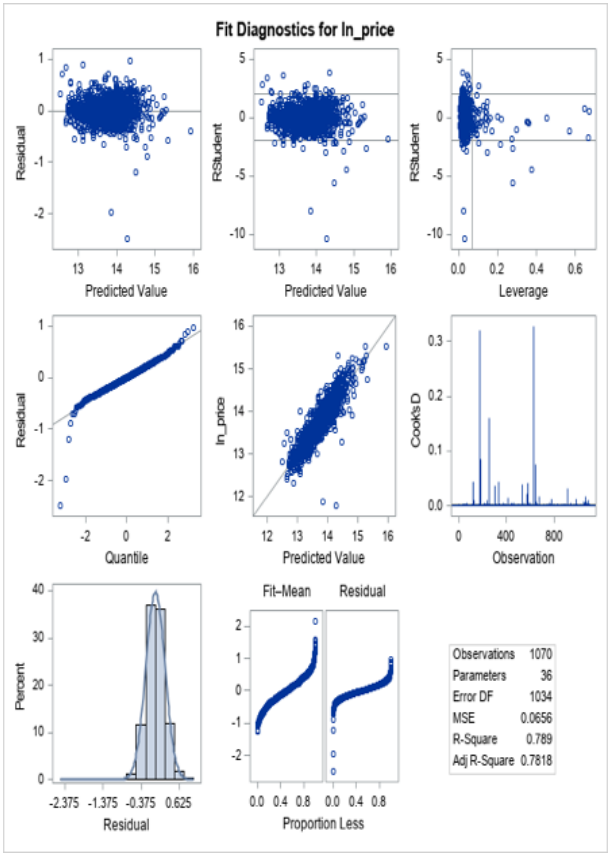


D13-scatter plots for ln_price

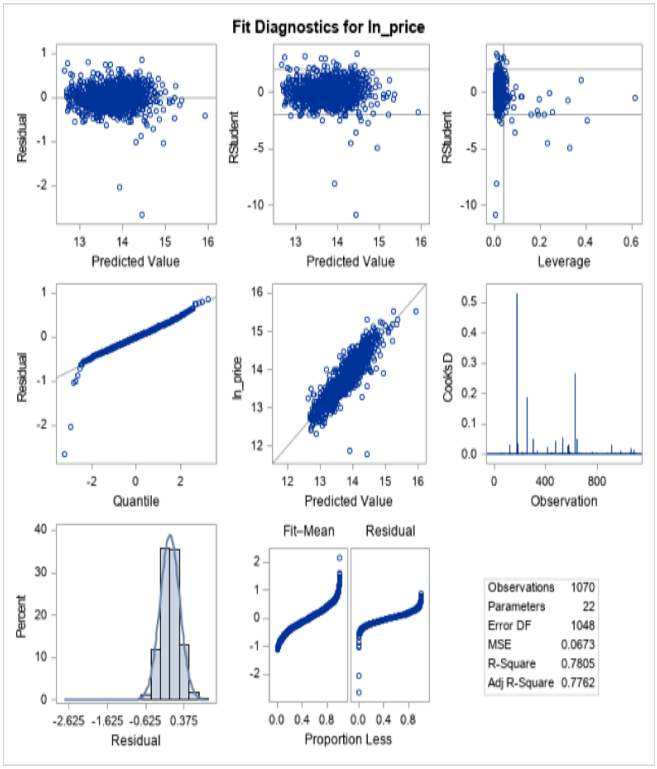


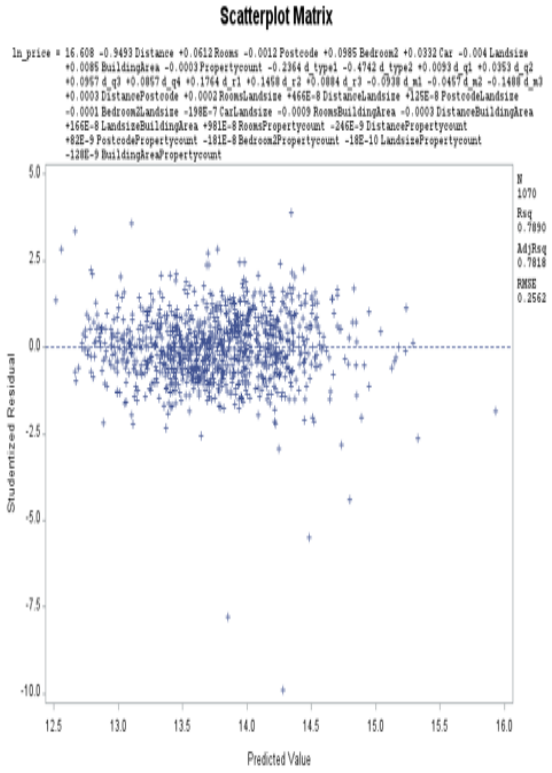


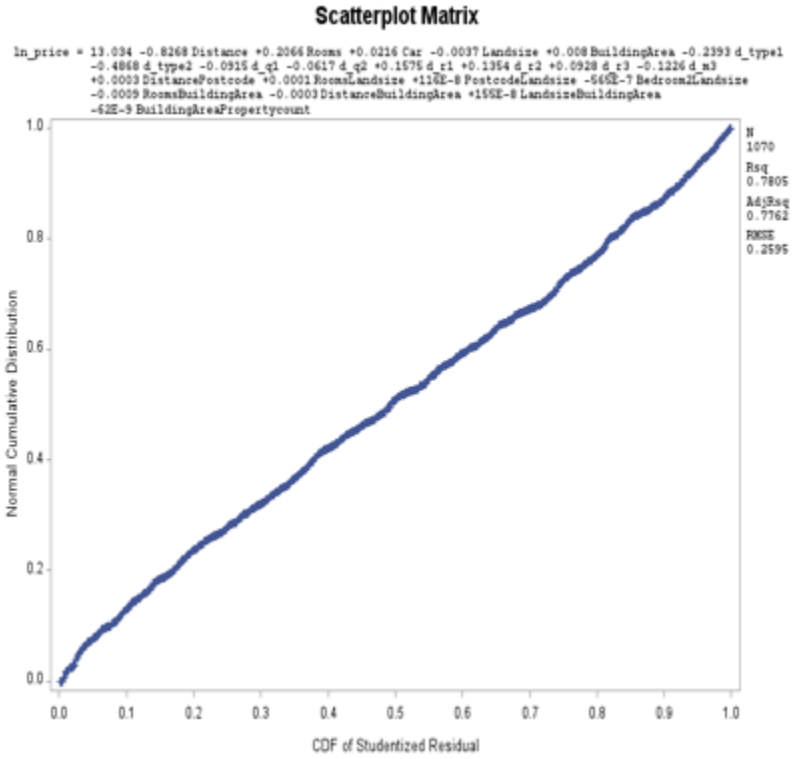
D14



D15 -after log transformation

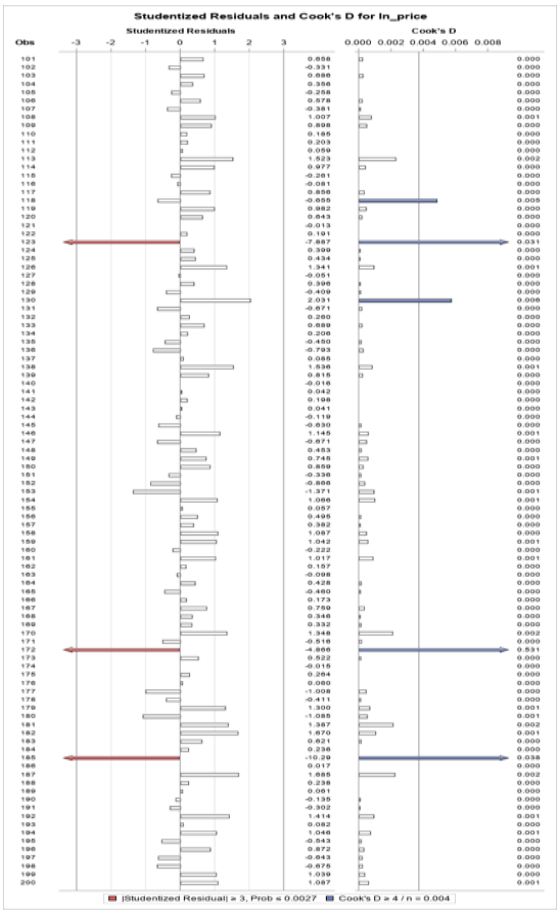


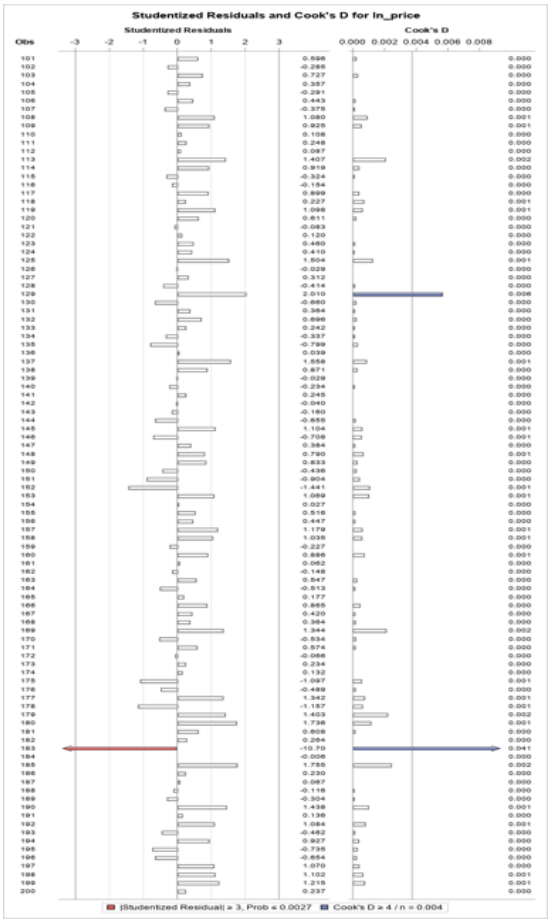




D16 - influential points/outliers

D17- after removing influential points/outliers





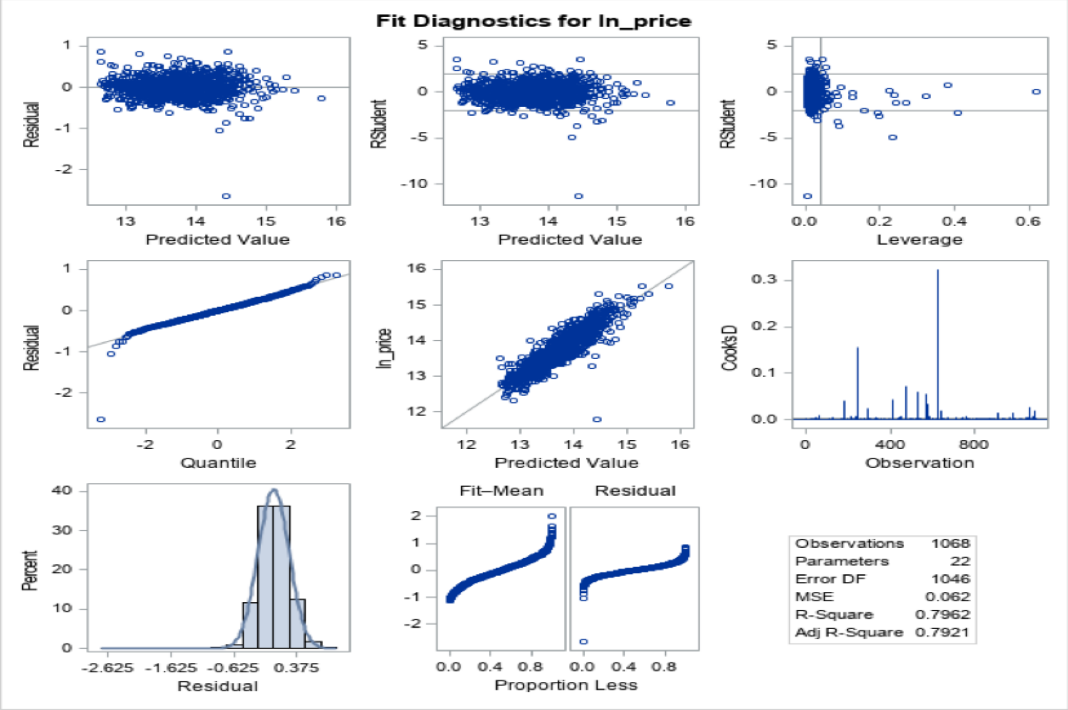
The REG Procedure
Model: MODEL1
Dependent Variable: ln_price

Number of Observations Read	1088
Number of Observations Used	1068
Number of Observations with Missing Values	20

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	21	253.21600	12.05790	194.64	<.0001
Error	1046	64.80010	0.06195		
Corrected Total	1067	318.01611			

Root MSE	0.24890	R-Square	0.7962
Dependent Mean	13.76532	Adj R-Sq	0.7921
Coeff Var	1.80815		



D18-predicted values

Compute Predictions

The REG Procedure
Model: MODEL1
Dependent Variable: ln_price

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	13.5422	0.0654	13.4138	13.6705	13.0372	14.0471	.
2	.	13.3654	0.0575	13.2526	13.4782	12.8642	13.8667	.
3	14.6	14.3817	0.0370	14.3090	14.4543	13.8879	14.8754	0.2223
4	13.7	13.6939	0.0278	13.6394	13.7484	13.2025	14.1853	-0.000539
5	14.8	14.3411	0.0371	14.2683	14.4140	13.8473	14.8349	0.4299
6	13.8	13.6776	0.0318	13.6152	13.7400	13.1852	14.1700	0.1479

The REG Procedure
Model: MODEL1
Dependent Variable: ln_price

Number of Observations Read	1090
Number of Observations Used	1068
Number of Observations with Missing Values	22

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	21	253.21600	12.05790	194.64	<.0001
Error	1046	64.80010	0.06195		
Corrected Total	1067	318.01611			

Root MSE	0.24890	R-Square	0.7962
Dependent Mean	13.76532	Adj R-Sq	0.7921
Coeff Var	1.80815		

D19-data validation

Test and Train Sets for price

Obs	Selected	suburb	Rooms	Type	Price	Method	QUARTER	Distance	Postcode	Bedroom2	Car	Land
1	1	Parkville	2	h	2200000	S	Q4	2.6	3052	2	1	
2	1	North Melbourne	2	h	885000	S	Q3	2.3	3051	2	1	
3	1	South Yarra	4	h	2600000	PI	Q5	3.3	3141	4	2	
4	1	Carlton	2	h	1010000	S	Q5	1.8	3053	2	0	
5	1	Abbotsford	3	h	1200000	S	Q3	2.5	3067	3	1	
6	1	Clifton Hill	3	h	1360000	S	Q2	3.4	3068	3	0	
7	1	North Melbourne	4	h	2920000	S	Q3	2.3	3051	4	4	
8	0	Port Melbourne	3	h	1670000	S	Q2	3.8	3207	3	1	
9	0	Fitzroy North	2	h	1210000	S	Q4	3.5	3068	2	1	
10	1	Kew	3	h	1410000	S	Q4	5.6	3101	3	2	
11	1	Armadale	3	h	1939000	S	Q4	6.3	3143	3	2	
12	1	Brunswick	5	h	2500000	S	Q4	5.2	3056	5	2	
13	1	Fitzroy North	4	h	1950000	S	Q2	3.5	3068	4	2	

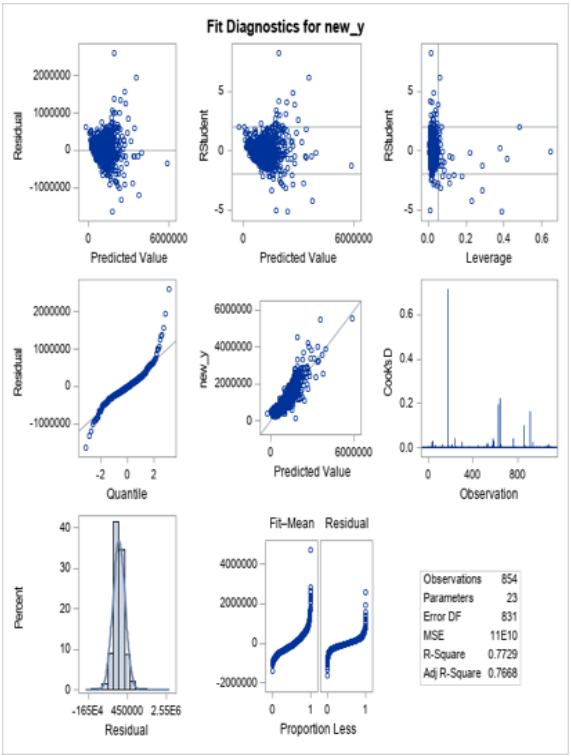
Test and Train Sets for price

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling

Input Data Set	PRICE
Random Number Seed	2519
Sampling Rate	0.8
Sample Size	872
Selection Probability	0.8
Sampling Weight	0
Output Data Set	PRICE_TT

ropertycount	Bedroom2Propertycount	LandsizePropertycount	BuildingAreaPropertycount	ln_price	new_y
7047068	4618	438710	397148	14.6040	2200000
20810871	13642	757131	497933	13.6933	885000
46760067	59548	2099067	1890649	14.7710	2600000
20717658	13572	685386	407160	13.8255	1010000
12326273	12057	454147	442090	13.9978	1200000
9062872	8862	348572	348572	14.1230	1360000
20810871	27284	2114510	1882596	14.8871	2920000
27734136	25944	1210720	1461512	14.3283	.
19156592	12488	1392412	736792	14.0061	.
32036431	30993	3140624	1312037	14.1591	1410000
15199548	14508	1431456	725400	14.4777	1939000
36421408	59590	6614490	3170188	14.7318	2500000
19156592	24976	1454852	1348704	14.4833	1950000



Validation - Test Set

The REG Procedure
Model: MODEL1
Dependent Variable: new_y

Number of Observations Read	1090
Number of Observations Used	854
Number of Observations with Missing Values	236

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	20	3.097656E14	1.548828E13	141.05	<.0001
Error	833	9.146617E13	1.098033E11		
Corrected Total	853	4.012317E14			

Root MSE	331366	R-Square	0.7720
Dependent Mean	1120256	Adj R-Sq	0.7666
Coeff Var	29.57947		

The REG Procedure
Model: MODEL1
Dependent Variable: Price Price in Australian dollars

Number of Observations Read	1090
Number of Observations Used	1070
Number of Observations with Missing Values	20

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	20	3.672502E14	1.836251E13	165.30	<.0001
Error	1049	1.165291E14	1.110859E11		
Corrected Total	1069	4.837793E14			

Root MSE	333295	R-Square	0.7591
Dependent Mean	1107083	Adj R-Sq	0.7545
Coeff Var	30.10574		

Validation statistics for Model

Obs	_TYPE_	_FREQ_	rmse	mae
1	0	218	350901.38	232343.27

The CORR Procedure

2 Variables: Price yhat

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
Price	218	1051769	614943	229285750	145000	4500000	Price in Australian dollars
yhat	216	1099159	584797	237418336	-15586	3732289	Predicted Value of new_y

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

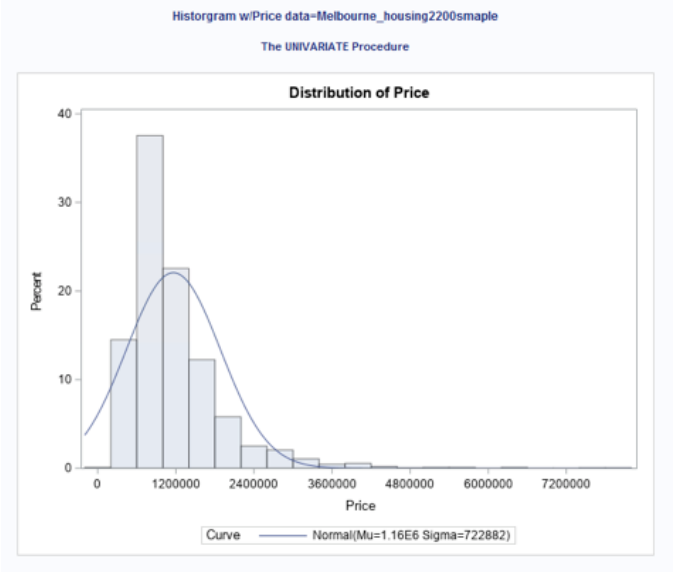
	Price	yhat
Price Price in Australian dollars	1.00000 218	0.83076 <.0001 216
yhat Predicted Value of new_y	0.83076 <.0001 216	1.00000 216

Appendix - E (Wilson Wu)

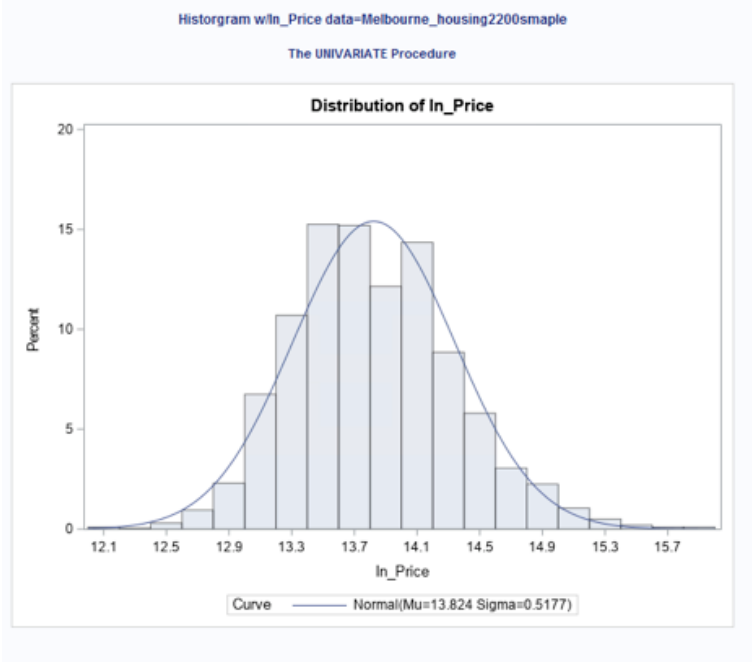
E.1: Description of Variable

#	Variable	Description	Type
1	Suburb	Melbourne Suburb	Qualitative
2	Address	House Address	Qualitative
3	Room	Number of rooms	Number
4	Type	br - bedroom(s); h - house,cottage,villa, semi,terrace; u - unit, duplex; t - townhouse; dev site - development site; o res - other residential.	Qualitative
5	Price	Price in Australian dollars	Number
6	Method	S - property sold; SP - property sold prior; PI - property passed in; PN - sold prior not disclosed; SN - sold not disclosed; NB - no bid; VB - vendor bid; W - withdrawn prior to auction; SA - sold after auction;	Qualitative
		SS - sold after auction price not disclosed. N/A - price or highest bid not available.	
7	SellerG	Real Estate Agent	Qualitative
8	Date	Date house was sold	Date
9	Distance	Distance from major central business district Kilometres	Number
10	Postcode	Postcode	Qualitative
11	Bedroom2	Number of Bedrooms	Number
12	Bathroom	Number of Bathrooms	Number
13	Car	Number of carspots	Number
14	Landsize	Land Size in Metres	Number
15	BuildingArea	Building Size in Metres	Number
16	YearBulit	Year the house was built	Year
17	CouncilArea	Governing council for the area	Qualitative
18	Latitude	Latitude of the property	Number
19	Longtitude	Longitude of the property	Number
20	Regionname	General Region (West, North West, North, North east ...etc)	Qualitative
21	Propertycount	Number of properties that exist in the suburb.	Number

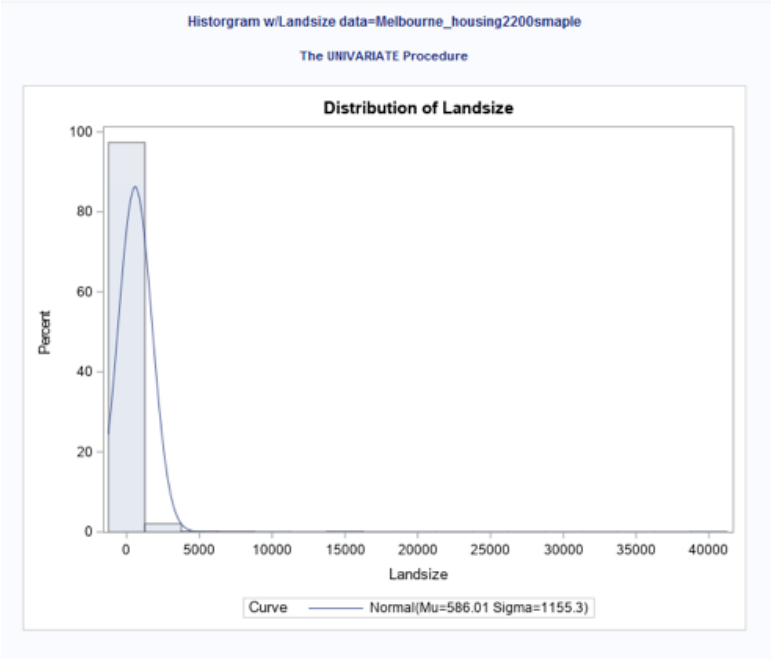
E.2: Distribution of Price



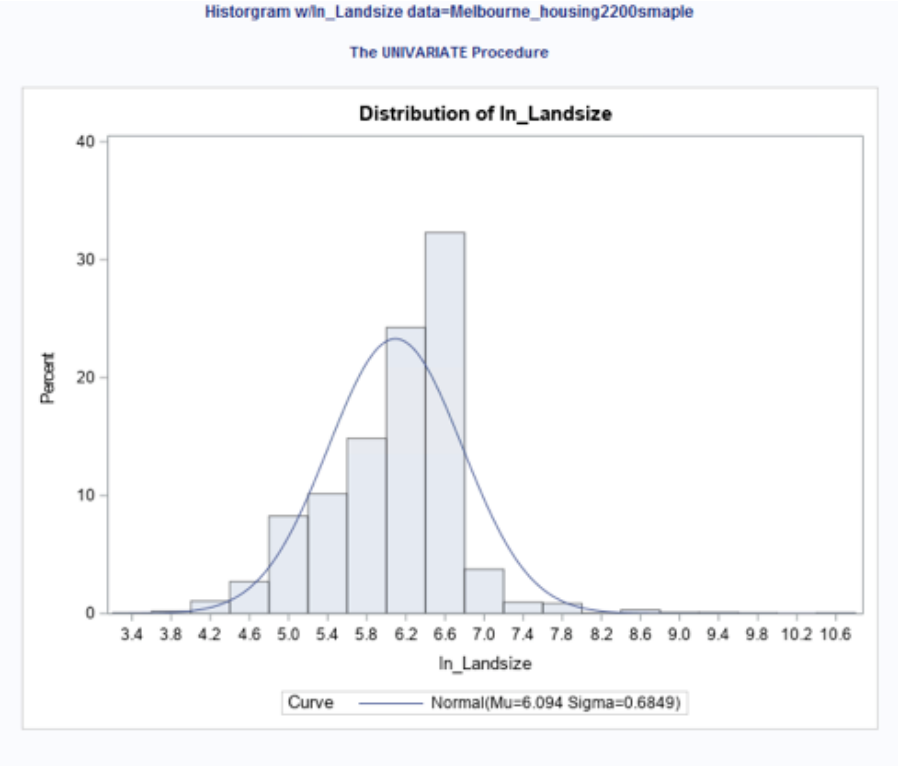
E.3: Distribution of ln_Price



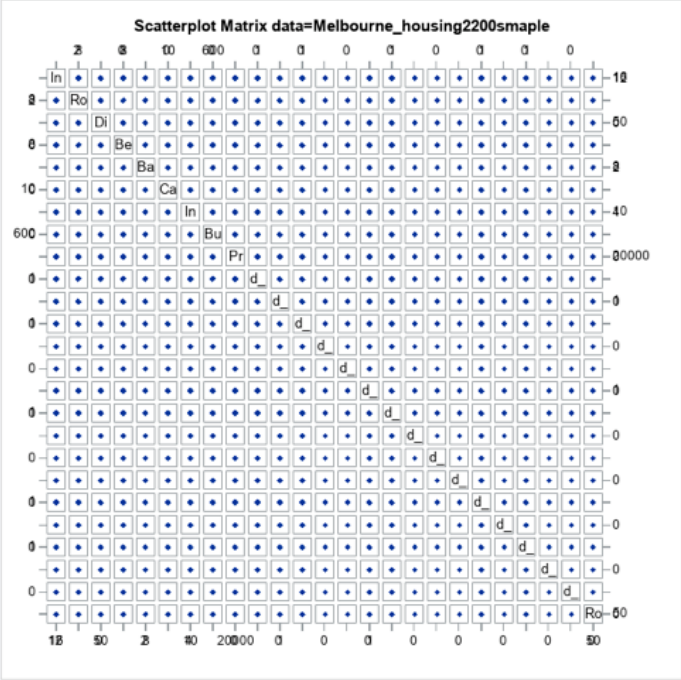
E.4: Distribution of Landsize



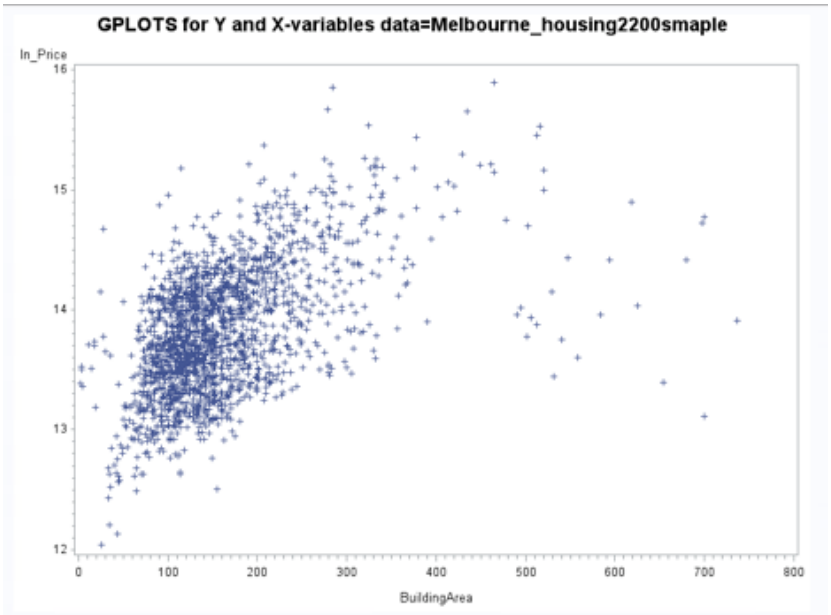
E.5: Distribution of ln_Landsize

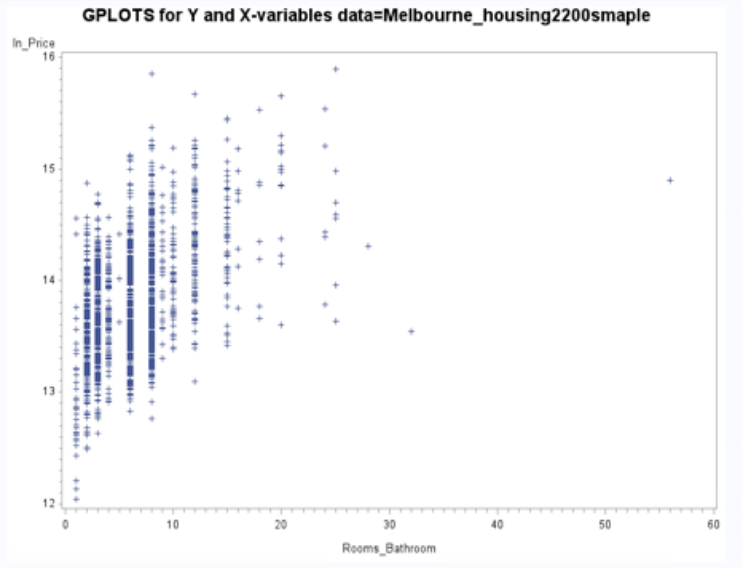


E.6 Scatterplot Matrix data



E.7 Glopt





E.8 Pearson Correlation Coefficients

Pearson Correla
Prob > |r

	ln_Price	Rooms	Distance	Bedroom2	Bathroom	Car	ln_Landsize	BuildingArea	Propertycount	d_Type_h	d_Type_u	d_Type_t	d_Type_devsite
ln_Price	1.00000	0.44276 <.0001	-0.37597 <.0001	0.41940 <.0001	0.41673 <.0001	0.17572 <.0001	0.15082 <.0001	0.51646 <.0001	-0.06733 0.0026	0.34093 <.0001	-0.33654 <.0001	-0.11658 <.0001	.
Rooms	0.44276 <.0001	1.00000	0.19190 <.0001	0.95372 <.0001	0.61617 <.0001	0.37002 <.0001	0.34666 <.0001	0.61202 <.0001	-0.03563 0.1111	0.37595 <.0001	-0.39303 <.0001	-0.10539 <.0001	.
Distance	-0.37597 <.0001	0.19190 <.0001	1.00000	0.19485 <.0001	0.07945 0.0004	0.21770 <.0001	0.30315 <.0001	0.07080 0.0015	0.02059 0.3573	0.12079 <.0001	-0.10144 <.0001	-0.06013 0.0072	.
Bedroom2	0.41940 <.0001	0.95372 <.0001	0.19485 <.0001	1.00000	0.62661 <.0001	0.38868 <.0001	0.34353 <.0001	0.59531 <.0001	-0.03029 0.1757	0.36008 <.0001	-0.37858 <.0001	-0.09868 <.0001	.
Bathroom	0.41673 <.0001	0.61617 <.0001	0.07945 0.0004	0.62661 <.0001	1.00000	0.30845 <.0001	0.17374 <.0001	0.58800 <.0001	-0.04699 0.0356	0.11795 <.0001	-0.21193 <.0001	0.06062 0.0067	.
Car	0.17572 <.0001	0.37002 <.0001	0.21770 <.0001	0.38868 <.0001	0.30845 <.0001	1.00000	0.36484 <.0001	0.32884 <.0001	-0.02074 0.3539	0.18877 <.0001	-0.19336 <.0001	-0.05713 0.0106	.
ln_Landsize	0.15082 <.0001	0.34666 <.0001	0.30315 <.0001	0.34353 <.0001	0.17374 <.0001	0.36484 <.0001	1.00000	0.28731 <.0001	-0.03877 0.0831	0.27635 <.0001	-0.07588 0.0007	-0.30267 <.0001	.
BuildingArea	0.51646 <.0001	0.61202 <.0001	0.07080 0.0015	0.59531 <.0001	0.58800 <.0001	0.32884 <.0001	0.28731 <.0001	1.00000	-0.03259 0.1452	0.26025 <.0001	-0.27883 <.0001	-0.06582 0.0032	.
Propertycount	-0.06733 0.0026	-0.03563 0.1111	0.02059 0.3573	-0.03029 0.1757	-0.04699 0.0356	-0.02074 0.3539	-0.03877 0.0831	-0.03259 0.1452	1.00000	0.02230 0.3188	0.02782 0.2136	-0.06031 0.0070	.
d_Type_h	0.34093 <.0001	0.37595 <.0001	0.12079 <.0001	0.36008 <.0001	0.11795 <.0001	0.18877 <.0001	0.27635 <.0001	0.26025 <.0001	0.02230 0.3188	1.00000	-0.69543 <.0001	-0.65033 <.0001	.
d_Type_u	-0.33654 <.0001	-0.39303 <.0001	-0.10144 <.0001	-0.37858 <.0001	-0.21193 <.0001	-0.19336 <.0001	-0.07588 0.0007	-0.27883 <.0001	0.02782 0.2136	-0.69543 <.0001	1.00000	-0.09362 <.0001	.
d_Type_t	-0.11658 <.0001	-0.10539 <.0001	-0.06013 0.0072	-0.09868 <.0001	0.06062 0.0067	-0.05713 0.0106	-0.30267 <.0001	-0.06582 0.0032	-0.06031 0.0070	-0.65033 <.0001	-0.09362 <.0001	1.00000	.

ıtion Coefficients, N = 2000
r| under H0: Rho=0

d_Type_ores	d_Method_SP	d_Method_PI	d_Method_PN	d_Method_SN	d_Method_NB	d_Method_VB	d_Method_W	d_Method_SA	d_Method_SS	d_Method_NA	Bedroom2_Bathroom
. .	-0.10432 <.0001	0.06865 0.0021	. .	. .	. .	0.11041 <.0001	. .	0.01332 0.5516	. .	. .	0.42522 <.0001
. .	-0.04789 0.0322	0.06399 0.0042	. .	. .	. .	0.03153 0.1586	. .	0.00805 0.7189	. .	. .	0.75389 <.0001
. .	0.06887 0.0021	-0.05853 0.0088	. .	. .	. .	-0.14536 <.0001	. .	0.02970 0.1842	. .	. .	0.10051 <.0001
. .	-0.04066 0.0691	0.06878 0.0021	. .	. .	. .	0.02877 0.1984	. .	0.01520 0.4970	. .	. .	0.80419 <.0001
. .	-0.04395 0.0494	0.08218 0.0002	. .	. .	. .	0.06915 0.0020	. .	0.01251 0.5761	. .	. .	0.91366 <.0001
. .	-0.01467 0.5119	0.05200 0.0200	. .	. .	. .	-0.00932 0.6769	. .	0.01936 0.3868	. .	. .	0.36290 <.0001
. .	0.02552 0.2540	0.01938 0.3864	. .	. .	. .	-0.03443 0.1238	. .	0.05697 0.0108	. .	. .	0.25073 <.0001
. .	-0.04363 0.0511	0.07550 0.0007	. .	. .	. .	0.04587 0.0402	. .	0.03050 0.1727	. .	. .	0.62209 <.0001
. .	-0.01828 0.4140	0.02604 0.2445	. .	. .	. .	-0.02008 0.3696	. .	-0.04432 0.0475	. .	. .	-0.03519 0.1156
. .	-0.01547 0.4894	0.01286 0.5654	. .	. .	. .	-0.02932 0.1899	. .	-0.02687 0.2297	. .	. .	0.19565 <.0001
. .	0.03067 0.1704	-0.00488 0.8273	. .	. .	. .	-0.00437 0.8453	. .	0.02507 0.2624	. .	. .	-0.23994 <.0001
. .	-0.01099 0.6232	-0.01266 0.5715	. .	. .	. .	0.04524 0.0431	. .	0.01072 0.6318	. .	. .	-0.01742 0.4362

E.9 Regression Model Parameter Estimates

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	B	12.80722	0.08789	145.71	<.0001	.	0
Rooms	1	0.09102	0.03062	2.97	0.0030	0.07779	12.85571
Distance	1	-0.03672	0.00120	-30.62	<.0001	0.84103	1.18901
Bedroom2	1	0.03186	0.03546	0.90	0.3690	0.05582	17.91567
Bathroom	1	0.20855	0.03065	6.81	<.0001	0.11257	8.88305
Car	1	0.00860	0.00864	1.00	0.3195	0.75751	1.32011
ln_Landsize	1	0.06778	0.01345	5.04	<.0001	0.69212	1.44483
BuildingArea	1	0.00179	0.00012902	13.88	<.0001	0.53090	1.88361
Propertycount	1	-0.00000385	0.00000181	-2.13	0.0336	0.98633	1.01386
d_Type_h	B	0.23359	0.03043	7.68	<.0001	0.44637	2.24028
d_Type_u	B	-0.15240	0.03984	-3.83	0.0001	0.44726	2.23582
d_Type_t	0	0	.	.	.	.	.
d_Type_devsite	0	0	.	.	.	.	.
d_Type_ores	0	0	.	.	.	.	.
d_Method_SP	1	-0.06336	0.02291	-2.77	0.0057	0.94653	1.05649
d_Method_PI	1	-0.01437	0.02387	-0.60	0.5473	0.93633	1.06801
d_Method_PN	0	0	.	.	.	.	.
d_Method_SN	0	0	.	.	.	.	.
d_Method_NB	0	0	.	.	.	.	.
d_Method_VB	1	0.01937	0.02680	0.72	0.4699	0.92899	1.07644
d_Method_W	0	0	.	.	.	.	.
d_Method_SA	1	0.07667	0.08175	0.94	0.3484	0.98527	1.01495
d_Method_SS	0	0	.	.	.	.	.
d_Method_NA	0	0	.	.	.	.	.
Bedroom2_Bathroom	1	-0.02689	0.00689	-3.90	<.0001	0.06638	15.06570

E.10 New Pearson Correlation Coefficients (without Rooms)

Pearson Co
Pro

	ln_Price	Distance	Bedroom2	Bathroom	Car	ln_Landsize	BuildingArea	Propertycount	d_Type_h	d_Type_u	d_Type_t	d_Type_devsite
ln_Price	1.00000	-0.37597 <.0001	0.41940 <.0001	0.41673 <.0001	0.17572 <.0001	0.15082 <.0001	0.51646 <.0001	-0.06733 0.0026	0.34093 <.0001	-0.33654 <.0001	-0.11658 <.0001	. .
Distance	-0.37597 <.0001	1.00000	0.19485 <.0001	0.07945 0.0004	0.21770 <.0001	0.30315 <.0001	0.07080 0.0015	0.02059 0.3573	0.12079 <.0001	-0.10144 <.0001	-0.06013 0.0072	. .
Bedroom2	0.41940 <.0001	0.19485 <.0001	1.00000	0.62661 <.0001	0.38868 <.0001	0.34353 <.0001	0.59531 <.0001	-0.03029 0.1757	0.36008 <.0001	-0.37858 <.0001	-0.09868 <.0001	. .
Bathroom	0.41673 <.0001	0.07945 0.0004	0.62661 <.0001	1.00000	0.30845 <.0001	0.17374 <.0001	0.58800 <.0001	-0.04699 0.0356	0.11795 <.0001	-0.21193 <.0001	0.06062 0.0067	. .
Car	0.17572 <.0001	0.21770 <.0001	0.38868 <.0001	0.30845 <.0001	1.00000	0.36484 <.0001	0.32884 <.0001	-0.02074 0.3539	0.18877 <.0001	-0.19336 <.0001	-0.05713 0.0106	. .
ln_Landsize	0.15082 <.0001	0.30315 <.0001	0.34353 <.0001	0.17374 <.0001	0.36484 <.0001	1.00000	0.28731 <.0001	-0.03877 0.0831	0.27635 <.0001	-0.07588 0.0007	-0.30267 <.0001	. .
BuildingArea	0.51646 <.0001	0.07080 0.0015	0.59531 <.0001	0.58800 <.0001	0.32884 <.0001	0.28731 <.0001	1.00000	-0.03259 0.1452	0.26025 <.0001	-0.27883 <.0001	-0.06582 0.0032	. .
Propertycount	-0.06733 0.0026	0.02059 0.3573	-0.03029 0.1757	-0.04699 0.0356	-0.02074 0.3539	-0.03877 0.0831	-0.03259 0.1452	1.00000	0.02230 0.3188	0.02782 0.2136	-0.06031 0.0070	. .
d_Type_h	0.34093 <.0001	0.12079 <.0001	0.36008 <.0001	0.11795 <.0001	0.18877 <.0001	0.27635 <.0001	0.26025 <.0001	0.02230 0.3188	1.00000	-0.69543 <.0001	-0.65033 <.0001	. .
d_Type_u	-0.33654 <.0001	-0.10144 <.0001	-0.37858 <.0001	-0.21193 <.0001	-0.19336 <.0001	-0.07588 0.0007	-0.27883 <.0001	0.02782 0.2136	-0.69543 <.0001	1.00000	-0.09362 <.0001	. .
d_Type_t	-0.11658 <.0001	-0.06013 0.0072	-0.09868 <.0001	0.06062 0.0067	-0.05713 0.0106	-0.30267 <.0001	-0.06582 0.0032	-0.06031 0.0070	-0.65033 <.0001	-0.09362 <.0001	1.00000	. .

rrelation Coefficients, N = 2000
ob > |r| under H0: Rho=0

d_Type_ores	d_Method_SP	d_Method_PI	d_Method_PN	d_Method_SN	d_Method_NB	d_Method_VB	d_Method_W	d_Method_SA	d_Method_SS	d_Method_NA	Bedroom2_Bathroom
. .	-0.10432 <.0001	0.06865 0.0021	. .	. .	. .	0.11041 <.0001	. .	0.01332 0.5516	. .	. .	0.42522 <.0001
. .	0.06887 0.0021	-0.05853 0.0088	. .	. .	. .	-0.14536 <.0001	. .	0.02970 0.1842	. .	. .	0.10051 <.0001
. .	-0.04066 0.0691	0.06878 0.0021	. .	. .	. .	0.02877 0.1984	. .	0.01520 0.4970	. .	. .	0.80419 <.0001
. .	-0.04395 0.0494	0.08218 0.0002	. .	. .	. .	0.06915 0.0020	. .	0.01251 0.5761	. .	. .	0.91366 <.0001
. .	-0.01467 0.5119	0.05200 0.0200	. .	. .	. .	-0.00932 0.6769	. .	0.01936 0.3868	. .	. .	0.36290 <.0001
. .	0.02552 0.2540	0.01938 0.3864	. .	. .	. .	-0.03443 0.1238	. .	0.05697 0.0108	. .	. .	0.25073 <.0001
. .	-0.04363 0.0511	0.07550 0.0007	. .	. .	. .	0.04587 0.0402	. .	0.03050 0.1727	. .	. .	0.62209 <.0001
. .	-0.01828 0.4140	0.02604 0.2445	. .	. .	. .	-0.02008 0.3696	. .	-0.04432 0.0475	. .	. .	-0.03519 0.1156
. .	-0.01547 0.4894	0.01286 0.5654	. .	. .	. .	-0.02932 0.1899	. .	-0.02687 0.2297	. .	. .	0.19565 <.0001
. .	0.03067 0.1704	-0.00488 0.8273	. .	. .	. .	-0.00437 0.8453	. .	0.02507 0.2624	. .	. .	-0.23994 <.0001
. .	-0.01099 0.6232	-0.01266 0.5715	. .	. .	. .	0.04524 0.0431	. .	0.01072 0.6318	. .	. .	-0.01742 0.4362

E.11 New Regression Model Parameter Estimates (without Rooms)

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	B	12.77640	0.08745	146.09	<.0001	.	0
Distance	1	-0.03675	0.00120	-30.58	<.0001	0.84110	1.18892
Bedroom2	1	0.12306	0.01781	6.91	<.0001	0.22215	4.50153
Bathroom	1	0.23433	0.02945	7.96	<.0001	0.12237	8.17173
Car	1	0.00763	0.00865	0.88	0.3780	0.75860	1.31821
ln_Landsize	1	0.06993	0.01346	5.20	<.0001	0.69414	1.44064
BuildingArea	1	0.00185	0.00012767	14.50	<.0001	0.54429	1.83726
Propertycount	1	-0.00000392	0.00000182	-2.16	0.0308	0.98650	1.01369
d_Type_h	B	0.23901	0.03044	7.85	<.0001	0.44798	2.23225
d_Type_u	B	-0.15267	0.03992	-3.82	0.0001	0.44727	2.23581
d_Type_t	0	0	.	.	.	.	.
d_Type_devsite	0	0	.	.	.	.	.
d_Type_ores	0	0	.	.	.	.	.
d_Method_SP	1	-0.06562	0.02294	-2.86	0.0043	0.94758	1.05532
d_Method_PI	1	-0.01519	0.02391	-0.64	0.5253	0.93645	1.06786
d_Method_PN	0	0	.	.	.	.	.
d_Method_SN	0	0	.	.	.	.	.
d_Method_NB	0	0	.	.	.	.	.
d_Method_VB	1	0.01912	0.02686	0.71	0.4767	0.92900	1.07643
d_Method_W	0	0	.	.	.	.	.
d_Method_SA	1	0.06824	0.08186	0.83	0.4046	0.98646	1.01373
d_Method_SS	0	0	.	.	.	.	.
d_Method_NA	0	0	.	.	.	.	.
Bedroom2_Bathroom	1	-0.03304	0.00658	-5.02	<.0001	0.07295	13.70834

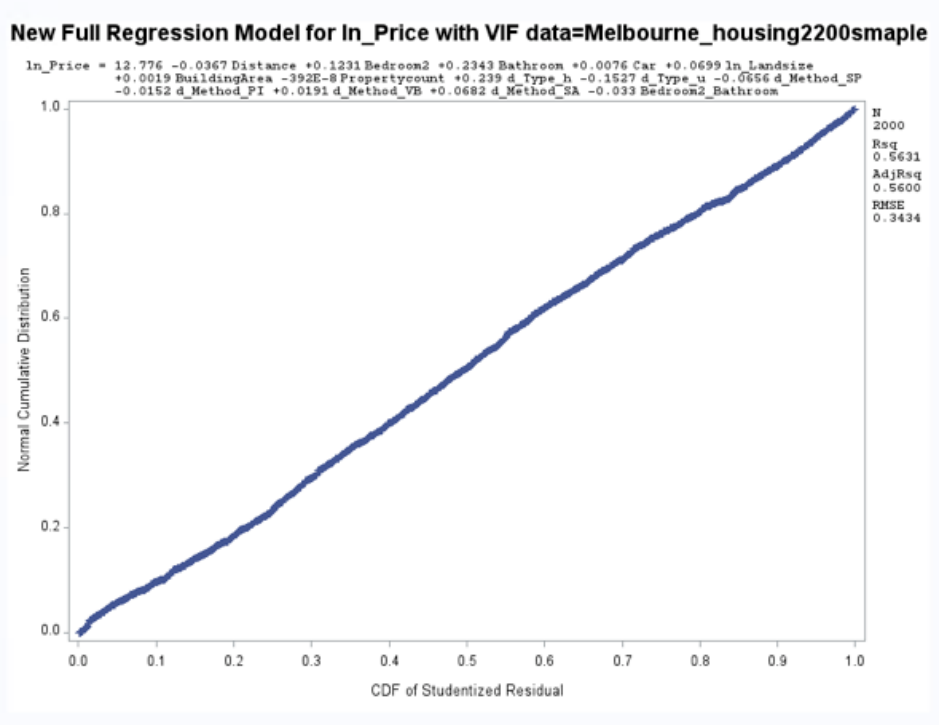
E.12 New1 Regression Model Parameter Estimates (without Rooms, Bedroom2_ Bathroom)

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	B	13.00940	0.07456	174.48	<.0001	.	0
Distance	1	-0.03611	0.00120	-30.04	<.0001	0.85066	1.17555
Bedroom2	1	0.06024	0.01274	4.73	<.0001	0.43904	2.27767
Bathroom	1	0.10553	0.01453	7.26	<.0001	0.50873	1.96567
Car	1	0.00541	0.00869	0.62	0.5335	0.76058	1.31478
ln_Landsize	1	0.06998	0.01354	5.17	<.0001	0.69414	1.44064
BuildingArea	1	0.00185	0.00012845	14.43	<.0001	0.54430	1.83722
Propertycount	1	-0.00000415	0.00000183	-2.27	0.0231	0.98711	1.01306
d_Type_h	B	0.23179	0.03059	7.58	<.0001	0.44898	2.22726
d_Type_u	B	-0.18627	0.03959	-4.70	<.0001	0.46021	2.17292
d_Type_t	0	0	.	.	.	.	.
d_Type_devsite	0	0	.	.	.	.	.
d_Type_ores	0	0	.	.	.	.	.
d_Method_SP	1	-0.06533	0.02308	-2.83	0.0047	0.94758	1.05532
d_Method_PI	1	-0.01589	0.02406	-0.66	0.5090	0.93648	1.06782
d_Method_PN	0	0	.	.	.	.	.
d_Method_SN	0	0	.	.	.	.	.
d_Method_NB	0	0	.	.	.	.	.
d_Method_VB	1	0.02332	0.02701	0.86	0.3880	0.92990	1.07538
d_Method_W	0	0	.	.	.	.	.
d_Method_SA	1	0.07605	0.08235	0.92	0.3559	0.98681	1.01336
d_Method_SS	0	0	.	.	.	.	.
d_Method_NA	0	0	.	.	.	.	.

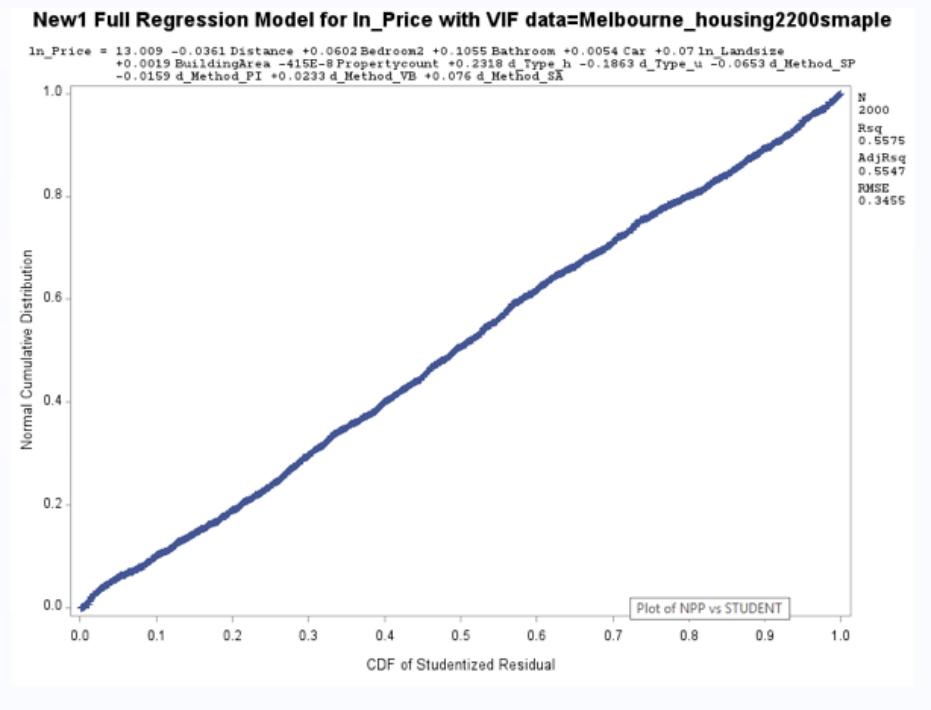
E.13 New2 Regression Model Parameter Estimates (without Rooms, Bathroom)

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	B	13.00940	0.07456	174.48	<.0001	.	0
Distance	1	-0.03611	0.00120	-30.04	<.0001	0.85066	1.17555
Bedroom2	1	0.06024	0.01274	4.73	<.0001	0.43904	2.27767
Bathroom	1	0.10553	0.01453	7.26	<.0001	0.50873	1.96567
Car	1	0.00541	0.00869	0.62	0.5335	0.76058	1.31478
ln_Landsize	1	0.06998	0.01354	5.17	<.0001	0.69414	1.44064
BuildingArea	1	0.00185	0.00012845	14.43	<.0001	0.54430	1.83722
Propertycount	1	-0.00000415	0.00000183	-2.27	0.0231	0.98711	1.01306
d_Type_h	B	0.23179	0.03059	7.58	<.0001	0.44898	2.22726
d_Type_u	B	-0.18627	0.03959	-4.70	<.0001	0.46021	2.17292
d_Type_t	0	0	.	.	.	.	.
d_Type_devsite	0	0	.	.	.	.	.
d_Type_ores	0	0	.	.	.	.	.
d_Method_SP	1	-0.06533	0.02308	-2.83	0.0047	0.94758	1.05532
d_Method_PI	1	-0.01589	0.02406	-0.66	0.5090	0.93648	1.06782
d_Method_PN	0	0	.	.	.	.	.
d_Method_SN	0	0	.	.	.	.	.
d_Method_NB	0	0	.	.	.	.	.
d_Method_VB	1	0.02332	0.02701	0.86	0.3880	0.92990	1.07538
d_Method_W	0	0	.	.	.	.	.
d_Method_SA	1	0.07605	0.08235	0.92	0.3559	0.98681	1.01336
d_Method_SS	0	0	.	.	.	.	.
d_Method_NA	0	0	.	.	.	.	.

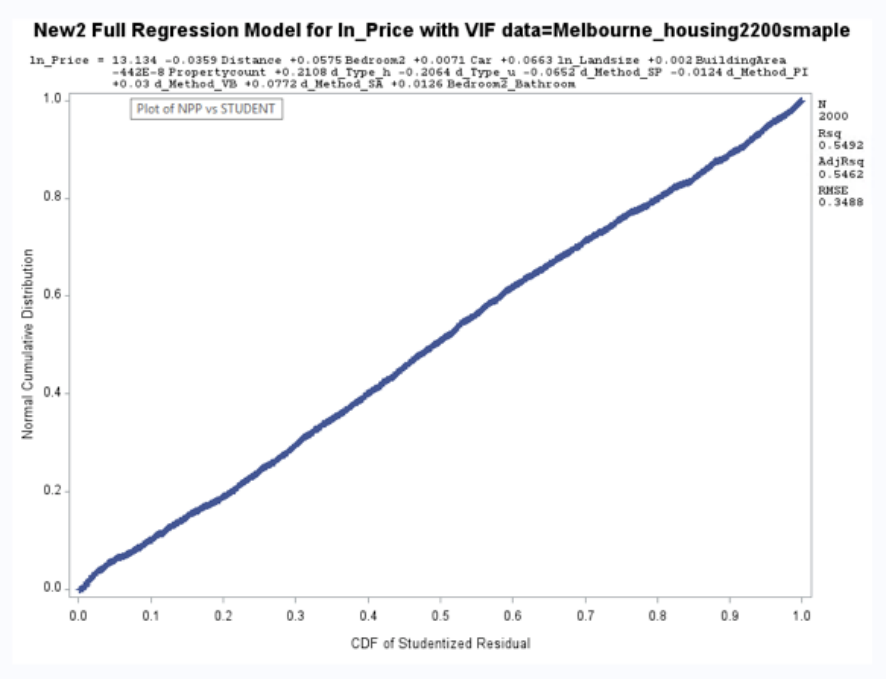
E.14 New full Regression Model Normal probability plot (without Rooms)



E.15 New full Regression Model Normal probability plot (without Rooms, Bedroom2_ Bathroom)



E.16 New full Regression Model Normal probability plot (without Rooms, Bathroom)



E.17 Training and Test set

Training and test sets

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling

Input Data Set	MELBOURNE_HOUSING_2000DATA
Random Number Seed	4378
Sampling Rate	0.8
Sample Size	1600
Selection Probability	0.8
Sampling Weight	0
Output Data Set	MH_TT

E.18New training and test set in selected

Training and test sets in selected

Obs	Selected	Suburb	Address	Rooms	Type	Price	Method	SellerG	Date	Distance	Postcode	Bedroom2	Bathroom	Car	Landsize
1	1	Williams	54 Victo	5	h	2550000	PI	Williams	19/08/20	6.8	3016	5	2	2	840
2	1	Malvern	3/197 Wa	5	h	2325000	S	Marshall	25/02/20	7.4	3144	5	3	2	501
3	1	Fitzroy	431 Geor	3	h	1250000	PI	Nelson	17/09/20	1.6	3065	3	1	1	113
4	1	Yallambi	36 Longa	4	h	953000	S	Fletcher	15/10/20	15.0	3085	4	2	2	551
5	0	Malvern	1 Ferncr	3	h	2200000	VB	Marshall	24/02/20	8.4	3145	3	2	2	735
6	1	Murrumbe	9/14 Wah	3	t	956000	S	Woodards	15/07/20	10.1	3163	3	2	1	137

d_Method_NB	d_Method_VB	d_Method_W	d_Method_SA	d_Method_SS	d_Method_NA	Bedroom2_Bathroom	new_y
0	0	0	0	0	0	10	14.7516
0	0	0	0	0	0	15	14.6592
0	0	0	0	0	0	3	14.0387
0	0	0	0	0	0	8	13.7674
0	1	0	0	0	0	6	.
0	0	0	0	0	0	6	13.7705
0	0	0	0	0	0	4	13.6231

E.19 Selection Method: backward Selection Method md1

Backward Elimination: Step 4

Variable d_Method_SA Removed: R-Square = 0.5697 and C(p) = 10.1509

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	247.89781	24.78978	210.35	<.0001
Error	1589	187.26382	0.11785		
Corrected Total	1599	435.16162			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	12.69342	0.09355	2169.86670	18412.1	<.0001
Distance	-0.03776	0.00135	92.30448	783.24	<.0001
Bedroom2	0.14178	0.01916	6.45282	54.75	<.0001
Bathroom	0.25948	0.03206	7.71924	65.50	<.0001
ln_Landsize	0.07537	0.01433	3.25951	27.66	<.0001
BuildingArea	0.00175	0.00013921	18.56829	157.56	<.0001
Propertycount	-0.00000355	0.00000204	0.35730	3.03	0.0818
d_Type_h	0.26018	0.03331	7.19020	61.01	<.0001
d_Type_u	-0.13285	0.04346	1.10129	9.34	0.0023
d_Method_SP	-0.08071	0.02530	1.19974	10.18	0.0014
Bedroom2_Bathroom	-0.03840	0.00692	3.62985	30.80	<.0001

E.20 Selection Method: Adj-R2 Selection Method md1

Selection Method: Adj-R2 Selection Method md1

The REG Procedure
Model: MODEL1
Dependent Variable: new_y

Adjusted R-Square Selection Method

Number of Observations Read	2000
Number of Observations Used	1600
Number of Observations with Missing Values	400

Number in Model	Adjusted R-Square	R-Square	Variables in Model
11	0.5670	0.5700	Distance Bedroom2 Bathroom ln_Landsize BuildingArea Propertycount d_Type_h d_Type_t d_Method_SP d_Method_SA Bedroom2_Bathroom
11	0.5670	0.5700	Distance Bedroom2 Bathroom ln_Landsize BuildingArea Propertycount d_Type_h d_Type_u d_Method_SP d_Method_SA Bedroom2_Bathroom
11	0.5670	0.5700	Distance Bedroom2 Bathroom ln_Landsize BuildingArea Propertycount d_Type_u d_Type_t d_Method_SP d_Method_SA Bedroom2_Bathroom
12	0.5670	0.5703	Distance Bedroom2 Bathroom Car ln_Landsize BuildingArea Propertycount d_Type_h d_Type_t d_Method_SP d_Method_SA Bedroom2_Bathroom
12	0.5670	0.5703	Distance Bedroom2 Bathroom Car ln_Landsize BuildingArea Propertycount d_Type_h d_Type_u d_Method_SP d_Method_SA Bedroom2_Bathroom
12	0.5670	0.5703	Distance Bedroom2 Bathroom Car ln_Landsize BuildingArea Propertycount d_Type_u d_Type_t d_Method_SP d_Method_SA Bedroom2_Bathroom
10	0.5670	0.5697	Distance Bedroom2 Bathroom ln_Landsize BuildingArea Propertycount d_Type_h d_Type_t d_Method_SP Bedroom2_Bathroom
10	0.5670	0.5697	Distance Bedroom2 Bathroom ln_Landsize BuildingArea Propertycount d_Type_h d_Type_u d_Method_SP Bedroom2_Bathroom

E.21 Regression Model Parameter Estimates md1 in training

modl 1 Full Regression Model for new_y with VIF data=Melbourne_housing2200smaple

The REG Procedure
Model: MODEL1
Dependent Variable: new_y

Number of Observations Read	2000
Number of Observations Used	1600
Number of Observations with Missing Values	400

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	247.89781	24.78978	210.35	<.0001
Error	1589	187.26382	0.11785		
Corrected Total	1599	435.16162			

Root MSE	0.34329	R-Square	0.5697
Dependent Mean	13.82606	Adj R-Sq	0.5670
Coeff Var	2.48294		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	12.69342	0.09355	135.69	<.0001
Distance	1	-0.03776	0.00135	-27.99	<.0001
Bedroom2	1	0.14178	0.01916	7.40	<.0001
Bathroom	1	0.25948	0.03206	8.09	<.0001
ln_Landsize	1	0.07537	0.01433	5.26	<.0001
BuildingArea	1	0.00175	0.00013921	12.55	<.0001
Propertycount	1	-0.00000355	0.00000204	-1.74	0.0818
d_Type_h	1	0.26018	0.03331	7.81	<.0001
d_Type_u	1	-0.13285	0.04346	-3.06	0.0023
d_Method_SP	1	-0.08071	0.02530	-3.19	0.0014
Bedroom2_Bathroom	1	-0.03840	0.00692	-5.55	<.0001

E.22 Selection Method: backward Selection Method md2

Backward Elimination: Step 4

Variable d_Method_SA Removed: R-Square = 0.5613 and C(p) = 9.2353

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	244.26795	27.14088	226.06	<.0001
Error	1590	190.89367	0.12006		
Corrected Total	1599	435.16162			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	12.96775	0.08016	3141.92286	26169.8	<.0001
Distance	-0.03707	0.00136	89.71727	747.28	<.0001
Bedroom2	0.06965	0.01421	2.88449	24.03	<.0001
Bathroom	0.10563	0.01626	5.06808	42.21	<.0001
ln_Landsize	0.07510	0.01446	3.23677	26.96	<.0001
BuildingArea	0.00176	0.00014049	18.79127	156.52	<.0001
Propertycount	-0.00000399	0.00000206	0.45196	3.76	0.0525
d_Type_h	0.25055	0.03357	6.68624	55.69	<.0001
d_Type_u	-0.17187	0.04329	1.89273	15.76	<.0001
d_Method_SP	-0.08150	0.02553	1.22341	10.19	0.0014

E.23 Selection Method: Adj-R2 Selection Method md2

Selection Method: Adj-R2 Selection Method md2

The REG Procedure
Model: MODEL1
Dependent Variable: new_y

Adjusted R-Square Selection Method

Number of Observations Read	2000
Number of Observations Used	1600
Number of Observations with Missing Values	400

Number in Model	Adjusted R-Square	R-Square	Variables in Model
11	0.5590	0.5620	Distance Bedroom2 Bathroom ln_Landsize BuildingArea Propertycount d_Type_h d_Type_t d_Method_SP d_Method_VB d_Method_SA
11	0.5590	0.5620	Distance Bedroom2 Bathroom ln_Landsize BuildingArea Propertycount d_Type_h d_Type_u d_Method_SP d_Method_VB d_Method_SA
11	0.5590	0.5620	Distance Bedroom2 Bathroom ln_Landsize BuildingArea Propertycount d_Type_u d_Type_t d_Method_SP d_Method_VB d_Method_SA
10	0.5590	0.5617	Distance Bedroom2 Bathroom ln_Landsize BuildingArea Propertycount d_Type_h d_Type_t d_Method_SP d_Method_SA
10	0.5590	0.5617	Distance Bedroom2 Bathroom ln_Landsize BuildingArea Propertycount d_Type_h d_Type_u d_Method_SP d_Method_SA
10	0.5590	0.5617	Distance Bedroom2 Bathroom ln_Landsize BuildingArea Propertycount d_Type_u d_Type_t d_Method_SP d_Method_SA
9	0.5588	0.5613	Distance Bedroom2 Bathroom ln_Landsize BuildingArea Propertycount d_Type_h d_Type_t d_Method_SP
9	0.5588	0.5613	Distance Bedroom2 Bathroom ln_Landsize BuildingArea Propertycount d_Type_h d_Type_u d_Method_SP

E.24 Regression Model Parameter Estimates md2 in training

modl 2 Full Regression Model for new_y with VIF data=Melbourne_housing2200smaple

The REG Procedure
Model: MODEL1
Dependent Variable: new_y

Number of Observations Read	2000
Number of Observations Used	1600
Number of Observations with Missing Values	400

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	244.26795	27.14088	226.06	<.0001
Error	1590	190.89367	0.12006		
Corrected Total	1599	435.16162			

Root MSE	0.34650	R-Square	0.5613
Dependent Mean	13.82606	Adj R-Sq	0.5588
Coeff Var	2.50610		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	12.96775	0.08016	161.77	<.0001
Distance	1	-0.03707	0.00136	-27.34	<.0001
Bedroom2	1	0.06965	0.01421	4.90	<.0001
Bathroom	1	0.10563	0.01626	6.50	<.0001
ln_Landsize	1	0.07510	0.01446	5.19	<.0001
BuildingArea	1	0.00176	0.00014049	12.51	<.0001
Propertycount	1	-0.00000399	0.00000206	-1.94	0.0525
d_Type_h	1	0.25055	0.03357	7.46	<.0001
d_Type_u	1	-0.17187	0.04329	-3.97	<.0001
d_Method_SP	1	-0.08150	0.02553	-3.19	0.0014

E.25 Selection Method: backward Selection Method md3

Backward Elimination: Step 4

Variable d_Method_VB Removed: R-Square = 0.5519 and C(p) = 9.7720

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	240.17857	26.68651	217.62	<.0001
Error	1590	194.98306	0.12263		
Corrected Total	1599	435.16162			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	13.07309	0.08256	3074.92196	25074.6	<.0001
Distance	-0.03706	0.00137	89.28255	728.06	<.0001
Bedroom2	0.07761	0.01779	2.33317	19.03	<.0001
ln_Landsize	0.07269	0.01462	3.03336	24.74	<.0001
BuildingArea	0.00195	0.00013965	23.95324	195.33	<.0001
Propertycount	-0.00000428	0.00000208	0.51866	4.23	0.0399
d_Type_h	0.22401	0.03367	5.42782	44.26	<.0001
d_Type_u	-0.19197	0.04370	2.36642	19.30	<.0001
d_Method_SP	-0.08233	0.02580	1.24841	10.18	0.0014
Bedroom2_Bathroom	0.01002	0.00355	0.97869	7.98	0.0048

E.26 Selection Method: Adj-R2 Selection Method md3

Adjusted R-Square Selection Method

Number of Observations Read	2000
Number of Observations Used	1600
Number of Observations with Missing Values	400

Number in Model	Adjusted R-Square	R-Square	Variables in Model
11	0.5496	0.5527	Distance Bedroom2 ln_Landsize BuildingArea Propertycount d_Type_h d_Type_t d_Method_SP d_Method_VB d_Method_SA Bedroom2_Bathroom
11	0.5496	0.5527	Distance Bedroom2 ln_Landsize BuildingArea Propertycount d_Type_h d_Type_u d_Method_SP d_Method_VB d_Method_SA Bedroom2_Bathroom
11	0.5496	0.5527	Distance Bedroom2 ln_Landsize BuildingArea Propertycount d_Type_u d_Type_t d_Method_SP d_Method_VB d_Method_SA Bedroom2_Bathroom
12	0.5496	0.5529	Distance Bedroom2 Car ln_Landsize BuildingArea Propertycount d_Type_h d_Type_t d_Method_SP d_Method_VB d_Method_SA Bedroom2_Bathroom
12	0.5496	0.5529	Distance Bedroom2 Car ln_Landsize BuildingArea Propertycount d_Type_h d_Type_u d_Method_SP d_Method_VB d_Method_SA Bedroom2_Bathroom
12	0.5496	0.5529	Distance Bedroom2 Car ln_Landsize BuildingArea Propertycount d_Type_u d_Type_t d_Method_SP d_Method_VB d_Method_SA Bedroom2_Bathroom
10	0.5495	0.5523	Distance Bedroom2 ln_Landsize BuildingArea Propertycount d_Type_h d_Type_t d_Method_SP d_Method_VB Bedroom2_Bathroom
10	0.5495	0.5523	Distance Bedroom2 ln_Landsize BuildingArea Propertycount d_Type_h d_Type_u d_Method_SP d_Method_VB Bedroom2_Bathroom
10	0.5495	0.5523	Distance Bedroom2 ln_Landsize BuildingArea Propertycount d_Type_u d_Type_t d_Method_SP d_Method_VB Bedroom2_Bathroom
10	0.5495	0.5523	Distance Bedroom2 ln_Landsize BuildingArea Propertycount d_Type_h d_Type_t d_Method_SP d_Method_SA Bedroom2_Bathroom
10	0.5495	0.5523	Distance Bedroom2 ln_Landsize BuildingArea Propertycount d_Type_h d_Type_u d_Method_SP d_Method_SA Bedroom2_Bathroom
10	0.5495	0.5523	Distance Bedroom2 ln_Landsize BuildingArea Propertycount d_Type_u d_Type_t d_Method_SP d_Method_SA Bedroom2_Bathroom
11	0.5495	0.5526	Distance Bedroom2 Car ln_Landsize BuildingArea Propertycount d_Type_h d_Type_t d_Method_SP d_Method_VB Bedroom2_Bathroom
11	0.5495	0.5526	Distance Bedroom2 Car ln_Landsize BuildingArea Propertycount d_Type_h d_Type_u d_Method_SP d_Method_VB Bedroom2_Bathroom
11	0.5495	0.5526	Distance Bedroom2 Car ln_Landsize BuildingArea Propertycount d_Type_u d_Type_t d_Method_SP d_Method_VB Bedroom2_Bathroom
11	0.5494	0.5525	Distance Bedroom2 Car ln_Landsize BuildingArea Propertycount d_Type_h d_Type_t d_Method_SP d_Method_SA Bedroom2_Bathroom
11	0.5494	0.5525	Distance Bedroom2 Car ln_Landsize BuildingArea Propertycount d_Type_h d_Type_u d_Method_SP d_Method_SA Bedroom2_Bathroom
11	0.5494	0.5525	Distance Bedroom2 Car ln_Landsize BuildingArea Propertycount d_Type_u d_Type_t d_Method_SP d_Method_SA Bedroom2_Bathroom
9	0.5494	0.5519	Distance Bedroom2 ln_Landsize BuildingArea Propertycount d_Type_h d_Type_t d_Method_SP Bedroom2_Bathroom
9	0.5494	0.5519	Distance Bedroom2 ln_Landsize BuildingArea Propertycount d_Type_h d_Type_u d_Method_SP Bedroom2_Bathroom

E.27 Regression Model Parameter Estimates md3 in training

modl 3 Full Regression Model for new_y with VIF data=Melbourne_housing2200smaple

The REG Procedure
Model: MODEL1
Dependent Variable: new_y

Number of Observations Read	2000
Number of Observations Used	1600
Number of Observations with Missing Values	400

Analysis of Variance

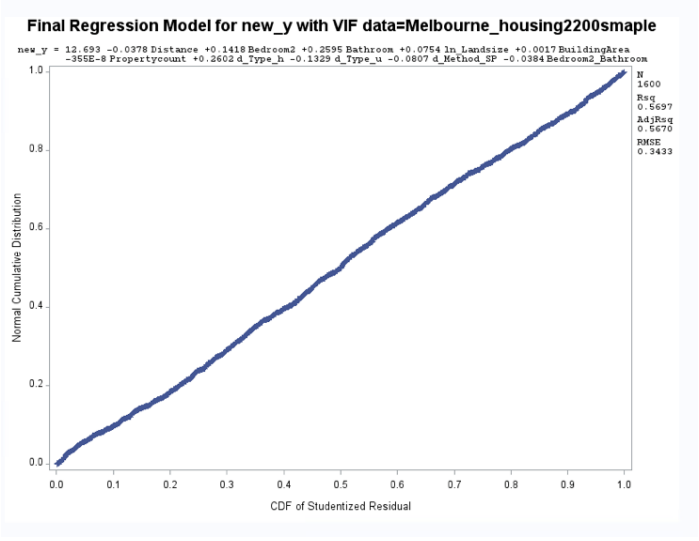
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	240.17857	26.68651	217.62	<.0001
Error	1590	194.98306	0.12263		
Corrected Total	1599	435.16162			

Root MSE	0.35019	R-Square	0.5519
Dependent Mean	13.82606	Adj R-Sq	0.5494
Coeff Var	2.53280		

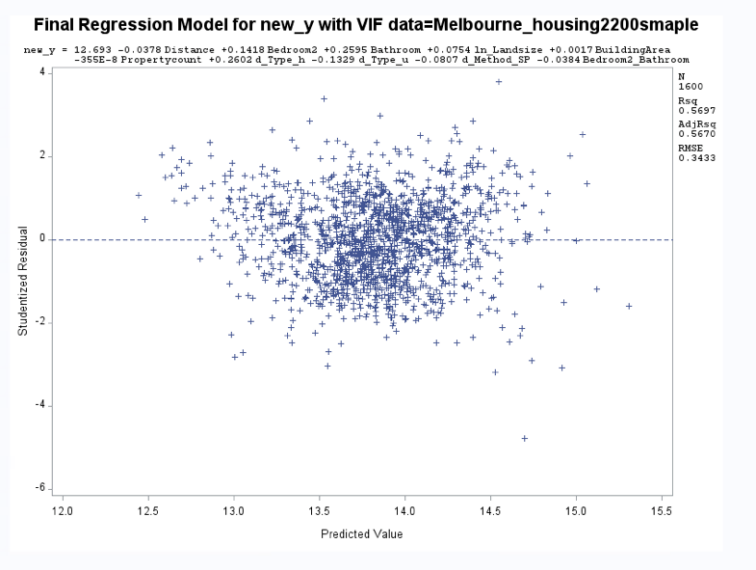
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	13.07309	0.08256	158.35	<.0001
Distance	1	-0.03706	0.00137	-26.98	<.0001
Bedroom2	1	0.07761	0.01779	4.36	<.0001
ln_Landsize	1	0.07269	0.01462	4.97	<.0001
BuildingArea	1	0.00195	0.00013965	13.98	<.0001
Propertycount	1	-0.00000428	0.00000208	-2.06	0.0399
d_Type_h	1	0.22401	0.03367	6.65	<.0001
d_Type_u	1	-0.19197	0.04370	-4.39	<.0001
d_Method_SP	1	-0.08233	0.02580	-3.19	0.0014
Bedroom2_Bathroom	1	0.01002	0.00355	2.83	0.0048

E.28 Final Regression Model Normal probability plot (without Rooms)



E.29 final model standardized residuals vs predicted



E.30 Final Regression Model 1 in training (First remove)

Final Regression Model for new_y with VIF data=Melbourne_housing2200smaple

The REG Procedure
Model: MODEL1
Dependent Variable: new_y

Number of Observations Read	1994
Number of Observations Used	1594
Number of Observations with Missing Values	400

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	250.12217	25.01222	222.68	<.0001
Error	1583	177.80896	0.11232		
Corrected Total	1593	427.93113			

Root MSE	0.33515	R-Square	0.5845
Dependent Mean	13.82573	Adj R-Sq	0.5819
Coeff Var	2.42409		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	12.76532	0.09187	138.95	<.0001	0	0
Distance	1	-0.03735	0.00132	-28.32	<.0001	-0.48978	1.13989
Bedroom2	1	0.12772	0.01881	6.79	<.0001	0.22824	4.30372
Bathroom	1	0.24908	0.03141	7.93	<.0001	0.35949	7.83147
ln_Landsize	1	0.06429	0.01405	4.57	<.0001	0.08656	1.36408
BuildingArea	1	0.00205	0.00014297	14.34	<.0001	0.31825	1.87703
Propertycount	1	-0.00000398	0.00000200	-1.99	0.0463	-0.03248	1.01068
d_Type_h	1	0.26725	0.03253	8.21	<.0001	0.19791	2.21158
d_Type_u	1	-0.12485	0.04244	-2.94	0.0033	-0.07014	2.16589
d_Method_SP	1	-0.08256	0.02470	-3.34	0.0009	-0.05443	1.01043
Bedroom2_Bathroom	1	-0.03764	0.00676	-5.56	<.0001	-0.32028	12.61982

E.31 Final Regression Model 1 in training (Second remove)

Final Regression Model for new_y with VIF data=Melbourne_housing2200smaple

The REG Procedure
Model: MODEL1
Dependent Variable: new_y

Number of Observations Read	1992
Number of Observations Used	1592
Number of Observations with Missing Values	400

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	251.33637	25.13364	226.42	<.0001
Error	1581	175.49521	0.11100		
Corrected Total	1591	426.83158			

Root MSE	0.33317	R-Square	0.5888
Dependent Mean	13.82512	Adj R-Sq	0.5862
Coeff Var	2.40989		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	12.77372	0.09135	139.84	<.0001	0	0
Distance	1	-0.03723	0.00131	-28.38	<.0001	-0.48868	1.14014
Bedroom2	1	0.12740	0.01873	6.80	<.0001	0.22782	4.31370
Bathroom	1	0.24974	0.03124	8.00	<.0001	0.36079	7.82929
ln_Landsize	1	0.06141	0.01399	4.39	<.0001	0.08274	1.36677
BuildingArea	1	0.00213	0.00014393	14.80	<.0001	0.32884	1.89716
Propertycount	1	-0.00000428	0.00000199	-2.15	0.0314	-0.03492	1.01071
d_Type_h	1	0.26817	0.03234	8.29	<.0001	0.19882	2.21108
d_Type_u	1	-0.12032	0.04220	-2.85	0.0044	-0.06768	2.16713
d_Method_SP	1	-0.08264	0.02456	-3.36	0.0008	-0.05455	1.01049
Bedroom2_Bathroom	1	-0.03834	0.00673	-5.70	<.0001	-0.32651	12.62054

E.32 Final Regression Model Parameter Estimates in training

Final Regression Model for new_y with VIF data=Melbourne_housing2200smaple

The REG Procedure
Model: MODEL1
Dependent Variable: new_y

Number of Observations Read	1994
Number of Observations Used	1594
Number of Observations with Missing Values	400

Analysis of Variance

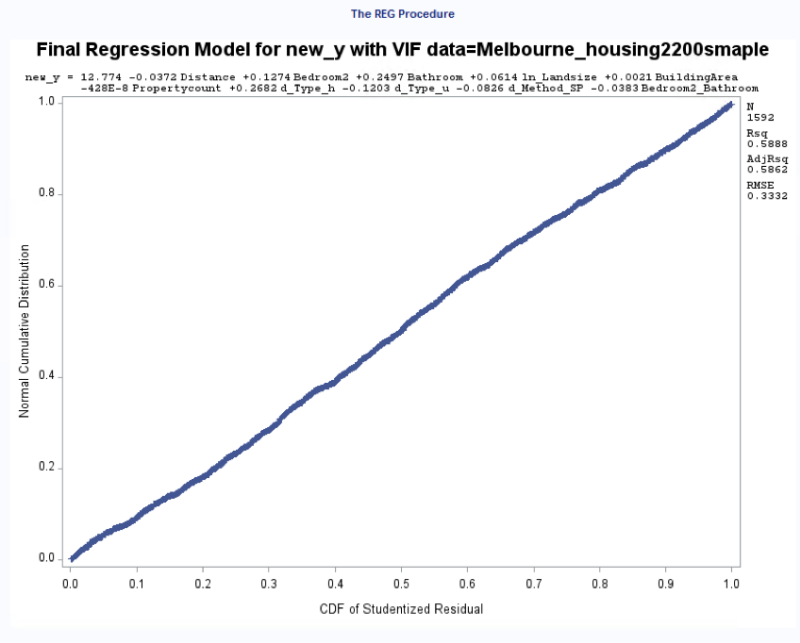
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	250.12217	25.01222	222.68	<.0001
Error	1583	177.80896	0.11232		
Corrected Total	1593	427.93113			

Root MSE	0.33515	R-Square	0.5845
Dependent Mean	13.82573	Adj R-Sq	0.5819
Coeff Var	2.42409		

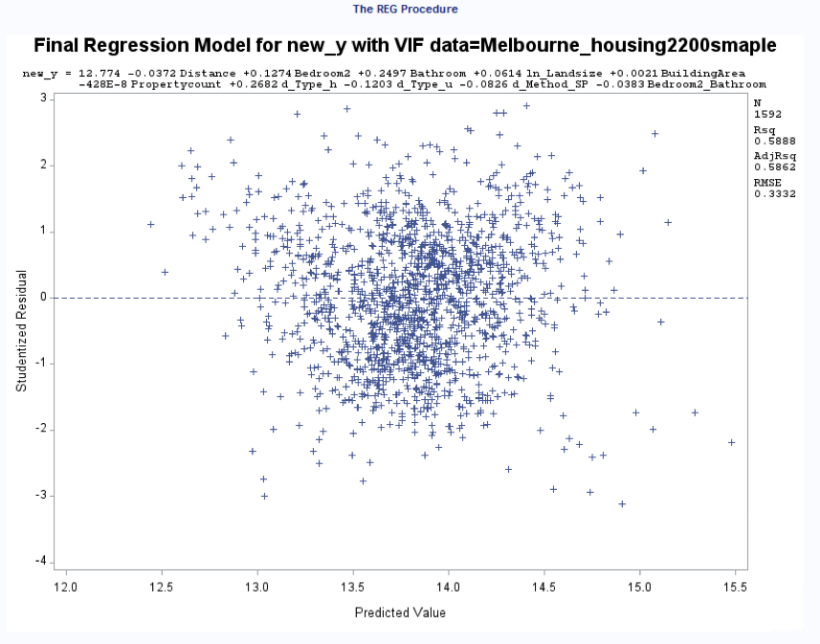
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	12.76532	0.09187	138.95	<.0001	0	0
Distance	1	-0.03735	0.00132	-28.32	<.0001	-0.48978	1.13989
Bedroom2	1	0.12772	0.01881	6.79	<.0001	0.22824	4.30372
Bathroom	1	0.24908	0.03141	7.93	<.0001	0.35949	7.83147
ln_Landsize	1	0.06429	0.01405	4.57	<.0001	0.08656	1.36408
BuildingArea	1	0.00205	0.00014297	14.34	<.0001	0.31825	1.87703
Propertycount	1	-0.00000398	0.00000200	-1.99	0.0463	-0.03248	1.01068
d_Type_h	1	0.26725	0.03253	8.21	<.0001	0.19791	2.21158
d_Type_u	1	-0.12485	0.04244	-2.94	0.0033	-0.07014	2.16589
d_Method_SP	1	-0.08256	0.02470	-3.34	0.0009	-0.05443	1.01043
Bedroom2_Bathroom	1	-0.03764	0.00676	-5.56	<.0001	-0.32028	12.61982

E.33 Final Regression Model Normal probability plot



E.34 final model standardized residuals vs predicted



E.35 Difference between Observed and Predicted in Test Set

Validation - Test Set

Obs	Selected	Suburb	Address	Rooms	Type	Price	Method	SellerG	Date	Distance	Postcode	Bedroom2	Bathroom	Car	Landsize	BuildingArea
1	0	Malvern	1 Ferncr	3	h	2200000	VB	Marshall	24/02/20	8.4	3145	3	2	2	735	224.00
2	0	West Foo	31 Napol	3	h	860000	S	Biggin	26/07/20	8.2	3012	3	1	2	501	147.00
3	0	Prahran	7/72 Wil	2	u	550000	SP	Thomson	13/05/20	4.5	3181	2	1	1	72	69.00
4	0	Brighton	40 Whyte	3	h	1900000	PI	Nick	6/1/18	10.5	3186	3	2	2	386	118.00
5	0	Hoppers	27 Quarr	3	h	473000	S	YPA	17/03/20	18.4	3029	3	2	1	392	108.00
6	0	Carnegie	55 Tranm	3	h	1840000	S	Ray	15/07/20	10.1	3163	3	2	1	483	171.00
7	0	Glen Wav	707 Wave	5	h	1116500	S	Harcourt	15/07/20	16.7	3150	5	2	2	701	182.85

d_Type_devsite	d_Type_ores	d_Method_SP	d_Method_PI	d_Method_PN	d_Method_SN	d_Method_NB	d_Method_VB	d_Method_W	d_Method_SA	d_Method_SS	d_Method_NA	Bedroom2_Bathroom	new_y	yhat	d	absd
0	0	0	0	0	0	0	1	0	0	0	0	6	.	14.2257	0.37827	0.37827
0	0	0	0	0	0	0	0	0	0	0	0	3	.	13.9268	-0.26213	0.26213
0	0	1	0	0	0	0	0	0	0	0	0	2	.	13.2077	0.00999	0.00999
0	0	0	1	0	0	0	0	0	0	0	0	6	.	13.8745	0.58289	0.58289
0	0	0	0	0	0	0	0	0	0	0	0	6	.	13.5461	-0.47921	0.47921
0	0	0	0	0	0	0	0	0	0	0	0	6	.	14.0279	0.39740	0.39740
0	0	0	0	0	0	0	0	0	0	0	0	10	.	13.8996	0.02610	0.02610

E.36 Descriptive of abs(d) in testing set

Descriptives

The MEANS Procedure

Analysis Variable : absd								
Mean	Std Dev	Std Error	Lower 95% CL for Mean	Upper 95% CL for Mean	Minimum	25th Pctl	50th Pctl	75th Pctl	Maximum
0.2763134	0.2034080	0.0101704	0.2563192	0.2963077	0.0013430	0.1168796	0.2505812	0.3971914	1.4013124

E.37 Final model rmse, mae and Pearson Correlation Coefficients in testing set

Difference between Observed and Predicted in Test Set

Obs	_TYPE_	_FREQ_	rmse	mae
1	0	400	0.34338	0.27631

Validation statistics for Model

The CORR Procedure

2 Variables:	ln_Price yhat

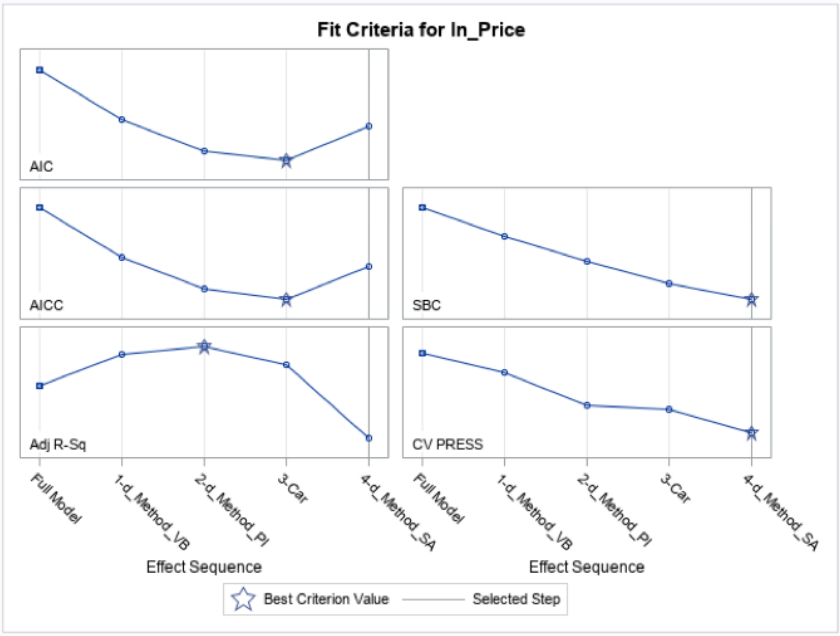
Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
ln_Price	400	13.81441	0.50205	5526	12.62807	15.66711	
yhat	400	13.81195	0.40002	5525	12.30143	15.04478	Predicted Value of new_y

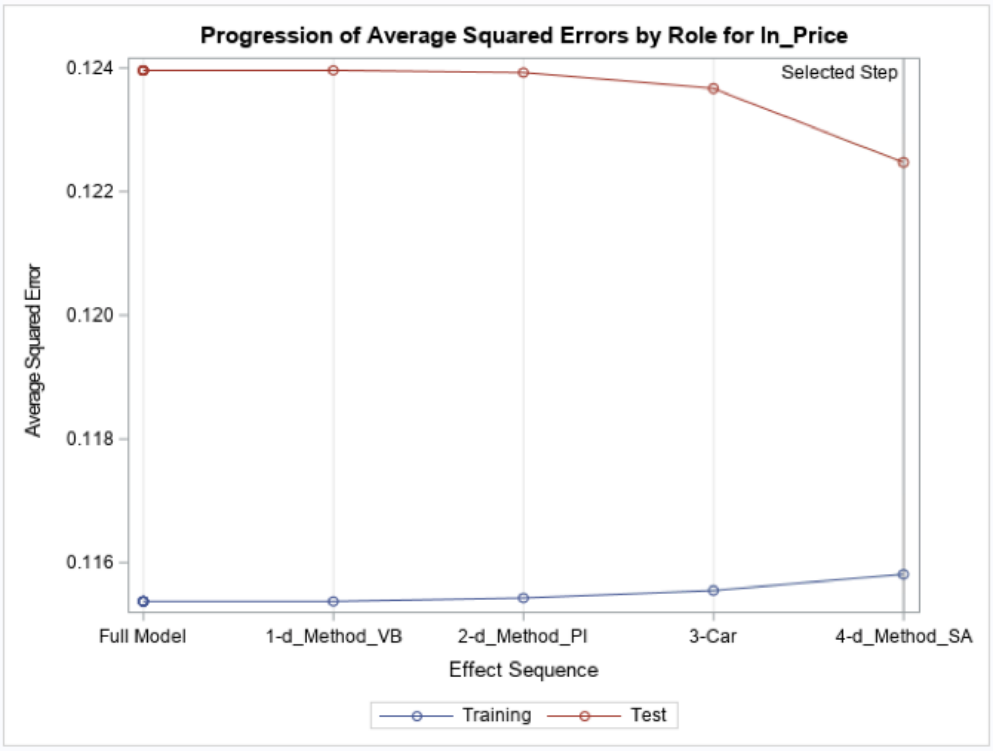
Pearson Correlation Coefficients, N = 400
Prob > |r| under H0: Rho=0

	ln_Price	yhat
ln_Price	1.00000	0.73236 <.0001
yhat Predicted Value of new_y	0.73236 <.0001	1.00000

E.38 Fit Criteria for ln_Price



E.39 Progression of Average Squared Error ln_Price



E.40 5-fold cross validation Parameter Estimates

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	10	224.00182	22.40018	192.03
Error	1512	176.37695	0.11665	
Corrected Total	1522	400.37877		

Root MSE	0.34154
Dependent Mean	13.82651
R-Square	0.5595
Adj R-Sq	0.5566
AIC	-1736.30455
AICC	-1736.09793
SBC	-3202.69174
ASE (Train)	0.11581
ASE (Test)	0.12247
CV PRESS	182.64466

Parameter Estimates									
					Cross Validation Estimates				
Parameter	DF	Estimate	Standard Error	t Value	1	2	3	4	5
Intercept	1	12.806859	0.097646	131.16	1.28E+01	1.27E+01	1.30E+01	1.29E+01	1.27E+01
Distance	1	-0.037498	0.001374	-27.28	-3.65E-02	-3.69E-02	-3.75E-02	-3.81E-02	-3.84E-02
Bedroom2	1	0.134504	0.019649	6.85	1.40E-01	1.58E-01	1.11E-01	1.14E-01	1.34E-01
Bathroom	1	0.218868	0.032086	6.82	2.32E-01	2.22E-01	1.90E-01	2.09E-01	2.14E-01
ln_Landsize	1	0.066786	0.014717	4.54	5.53E-02	7.22E-02	5.03E-02	6.51E-02	8.92E-02
BuildingArea	1	0.001838	0.000142	12.95	1.87E-03	1.70E-03	1.69E-03	2.13E-03	1.82E-03
Propertycount	1	-0.000004274	0.000002111	-2.02	-4.76E-06	-5.28E-06	-3.54E-06	-3.59E-06	-4.18E-06
d_Type_h	1	0.243519	0.034878	6.98	2.62E-01	2.33E-01	2.32E-01	2.56E-01	2.35E-01
d_Type_u	1	-0.151383	0.045599	-3.32	-1.27E-01	-1.71E-01	-1.90E-01	-1.21E-01	-1.49E-01
d_Method_SP	1	-0.057892	0.026310	-2.20	-8.13E-02	-3.43E-02	-4.87E-02	-7.18E-02	-5.81E-02
Bedroom2_Bathroom	1	-0.033440	0.006944	-4.82	-3.64E-02	-3.63E-02	-1.96E-02	-3.33E-02	-3.39E-02

E.41 Merge prediction dataset with original dataset

Merge prediction dataset with original datase

Obs	Distance	Bedroom2	Bathroom	ln_Landsize	BuildingArea	Propertycount	d_Type_h	d_Type_u	d_Method_SP	Bedroom2_Bathroom	Suburb	Address	Rooms
1	10.0	3	1	5.3400	209.00	5682	1	0	1	3			.
2	20.0	5	2	6.8700	280.00	8888	0	1	0	10			.
3	6.8	5	2	6.7334	244.00	6380	1	0	0	10	Williams	54 Victo	5
4	7.4	5	3	6.2166	233.00	4675	1	0	0	15	Malvern	3/197 Wa	5
5	1.6	3	1	4.7274	105.00	5825	1	0	0	3	Fitzroy	431 Geor	3

E.42 Output Predicted Value and C.L, P.L

Final Regression Model for new_y with VIF data=Melbourne_housing2200smaple

The REG Procedure
Model: MODEL1
Dependent Variable: ln_Price

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	13.8351	0.0277	13.7807	13.8895	13.1596	14.5105	.
2	.	13.6259	0.0406	13.5462	13.7056	12.9479	14.3038	.
3	14.8	14.4358	0.0219	14.3928	14.4789	13.7612	15.1104	0.3158
4	14.7	14.4325	0.0209	14.3916	14.4735	13.7581	15.1070	0.2267