Regression Analysis on the Price of Real Estate Property

EMSE 4765/6765

SPRING 2021

Kahang Ngau

EMSE 6765

Apr 27th, 2021

Introduction

The price of housing has always been an interesting and popular topic. Sellers and investors would probably love to see the price of housing increase, while buyers are looking for a cheaper price of housing. There are a number of attributes that can influence the price of a real estate property. And also, linear regression model is one of the most popular and useful machine learning techniques in projecting results and identifying the strength of effect that the independent variables have on a dependent variable, it will be very useful to use it to conduct analytical results. This report aims to address the problems of the following:

1. Conduct a linear regression model base on the house price data, while using *Log (Price)/Log(Y)* as the dependent variable as opposed to *Price/Y*.
2. Find the best estimated linear regression of *Log (Price)/Log(Y)* on an appropriate set of explanatory variables using the properties and interpret the results.
3. Perform diagnostic analysis on the regression analysis of the final selected model.
4. Forecast the median and average of *Price/Y* of a real estate property for the following values of the explanatory variables and provide a 95% prediction interval for *Price/Y* and an approximate 95% confidence interval for *E[Y]*.

There are a total of 10 explanatory variables and 2 dependent variables that will be used in conducting the data analysis. The table below demonstrates the sample subset of the house price data (10 out of a total of 80 properties):

Table 1: Sample of Original Housing Price Data

	Y	Log(Y)	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
Property	PRICE	Log(PRICE)	bedrooms	bathrooms	sqft_living	sqft_lot	floors	Numbers of times viewed	Quality Grade	sqft_above	sqft_basement	Built or Renovated
1	$ 440,000.00	5.64345	3	2.5	1910	66211	2	0	7	1910	0	1997
2	$ 213,000.00	5.32838	2	1	1000	10200	1	0	6	1000	0	1961
3	$ 563,500.00	5.75089	4	1.75	2085	174240	1	0	7	1610	475	1964
4	$ 1,550,000.00	6.19033	5	4.25	6070	171626	2	0	12	6070	0	1999
5	$ 1,600,000.00	6.20412	6	5	6050	230652	2	3	11	6050	0	2001
6	$ 350,000.00	5.54407	3	2.25	1580	47916	1	0	7	1580	0	1979
7	$ 540,000.00	5.73239	3	2.25	2000	217800	2	0	8	2000	0	1996
8	$ 535,000.00	5.72835	3	1	1330	40259	1	0	7	1330	0	1977
9	$ 600,000.00	5.77815	2	2.5	2410	102366	1	0	7	1940	470	1989
10	$ 275,000.00	5.43933	3	1	1370	17859	1	0	7	1150	220	1930

Methodology

Price/Y vs. Log (Price)/Log(Y)

After introducing the explanatory variables and dependent variables, the next step is to conduct exploratory data analysis on the two dependent variables *Price/Y* and *Log (Price)/Log(Y)*. When performing the histogram of the dependent variable *Price/Y* as the chart shown below:

- It is clear to see that the histogram is skewed to the left with a very long tail. This may potentially be problematic since it may highly increase the chance of errors in making prediction in the later stage. And it would be desirable to have a histogram of the dependent variable that is bell-shaped.
- The chart also demonstrates that the standard deviation is very large, and the distribution does not fit along with the fitting line very much.

When conducting the probability plot of *Price/Y* below on figure 2:

- The plot indicates that there is a very large standard deviation.
- By visually seeing from graph, most of the points do not follow the straight line.
- Lots of points are falling outside of the confidential boundaries.
- The p-value is relatively small which is smaller than 0.005.

These observations and findings are all displayed in the histogram and probability plot that are generated by Minitab below:

Figure 1: Histogram of Price (Y)

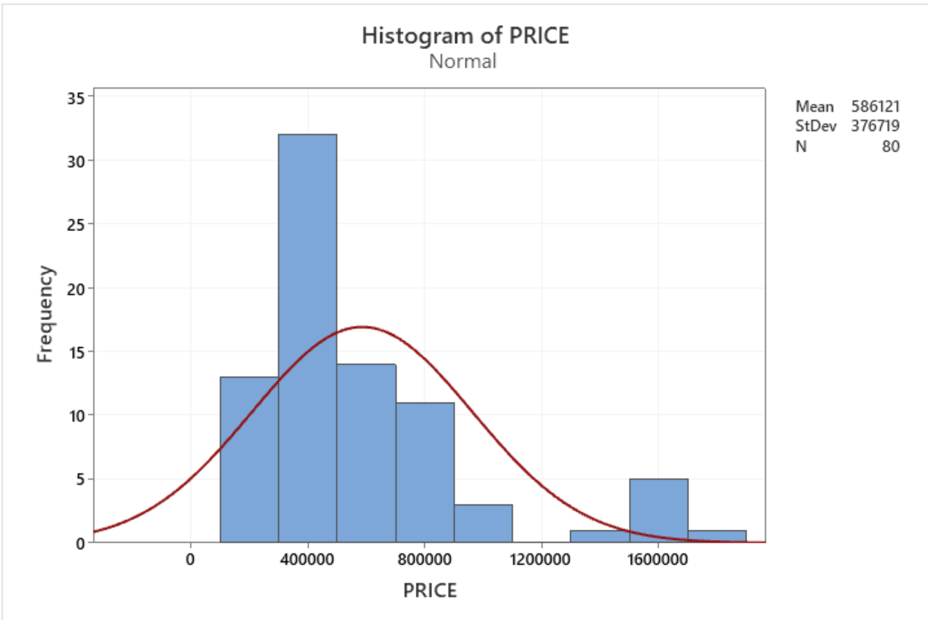
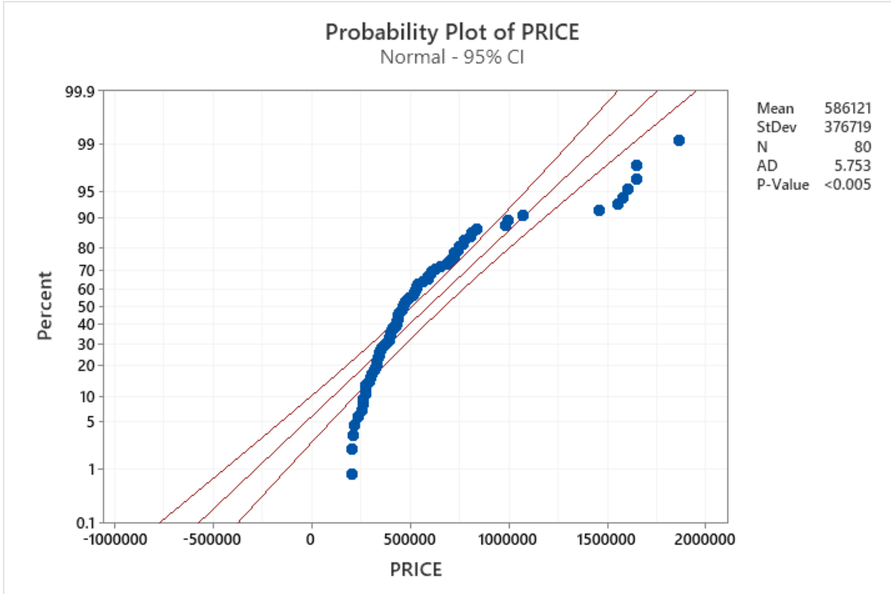


Figure 2: Probability Plot of Price (Y)



While applying the logarithmic transformation of the dependent variable *Price/Y*, it simply transforms a highly skewed variable into a more normalized dataset which is the *Log (Price)/Log(Y)*. As we see from the figure 3 and figure 4 below, we can detect some of the improvement on the distribution and the normality plot:

- The histogram shows a much lower standard deviation comparing to the one in figure 1.
- It is clear to see that the distribution is relatively symmetric, and which it follows closely with the fitting line on the chart.
- The probability plot shows that most of the points on the graph follow the straight line comparing the points on figure 1.
- And the points have majority of them stay within the confidential boundaries.
- The p-value for the *Log (Price)/Log(Y)* also increases. It is 0.025 which is larger comparing to *Price/Y* of 0.005. It is preferable to choose the larger value of the p-value since larger p-value tends to indicate greater normality of the distribution.
- As summarizing from the observations above, we will choose the *Log (Price)/Log(Y)* as the dependent variable of the regression model. By choosing it as the dependent variable, we can have a less deviation from normality in *Log (Price)/Log(Y)* plot than in the *Price/Y*.

These observations and findings are all displayed in the histogram and probability plot that are generated by Minitab below:

Figure 3: Histogram of Log Price (Log(Y))

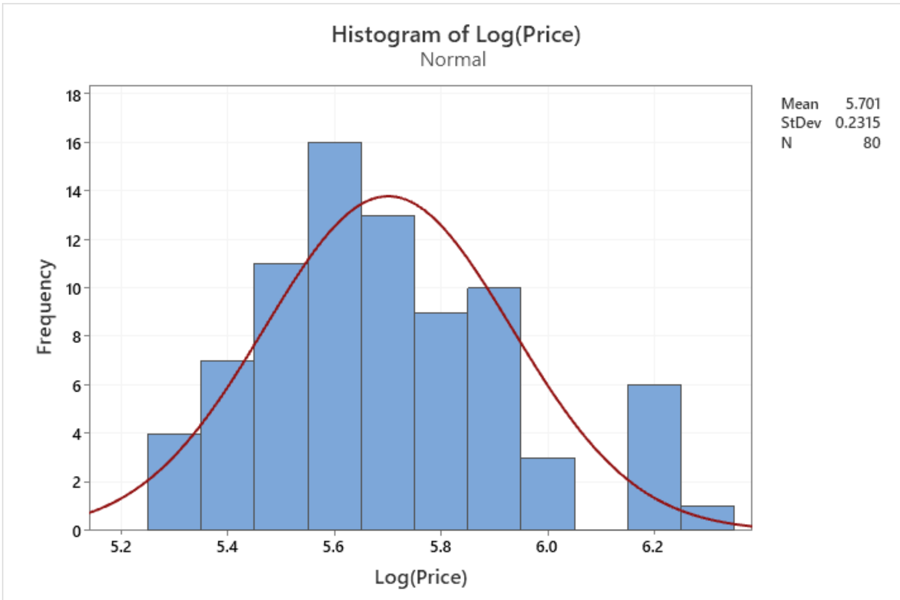
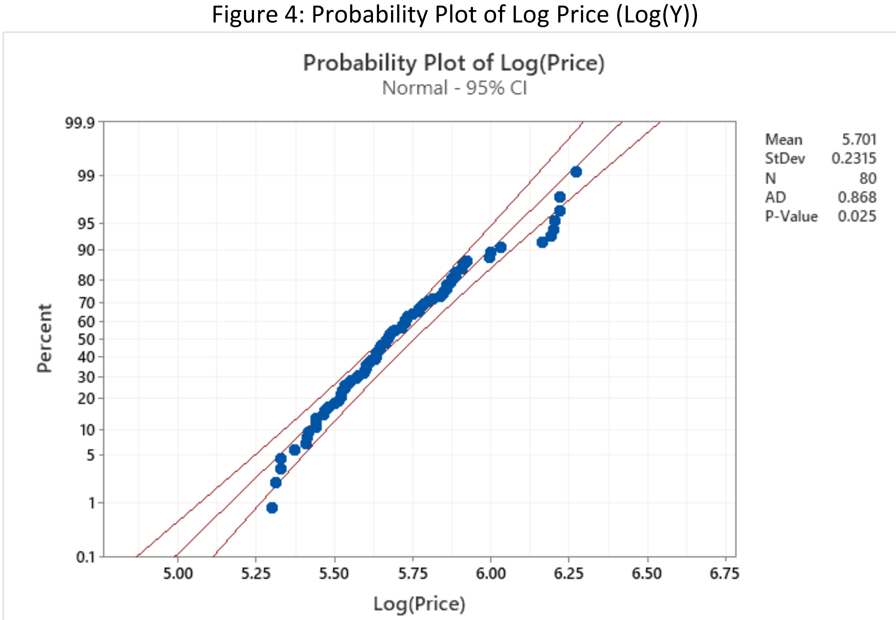


Figure 4: Probability Plot of Log Price (Log(Y))



Correlation Metrics Analysis

Before we conduct the linear regression model analysis, it is important to obtain a good understanding for the data through studying the correlations between the dependent variables and the explanatory variables. Correlation is a good measure for knowing the linear dependence. The figure below presents the correlation strengths between the dependent variable *Log (Price)/Log(Y)*, and explanatory variables (X1-10).

Figure 5: Correlation between Log(Y) and X1-10

	Log(Price)	*bedrooms*	*athrooms*	*sqft_living*	*sqft_lot*	*floors*	*rs of times*	*uality Grac*	*qft_above*	*ft_baseme*	*or Renovated*
Log(Price)	1										
bedrooms	0.523166	1									
bathrooms	0.838926	0.559058	1								
sqft_living	0.893285	0.600117	0.88938	1							
sqft_lot	0.61623	0.111621	0.429643	0.476497	1						
floors	0.66211	0.414508	0.629731	0.625159	0.472526	1					
Numbers of times viewed	0.445527	0.196751	0.412935	0.401512	0.634967	0.41916	1				
Quality Grade	0.887577	0.461978	0.830222	0.846566	0.53892	0.684202	0.414666	1			
sqft_above	0.885398	0.589346	0.858244	0.960583	0.450285	0.669025	0.376648	0.84426	1		
sqft_basement	0.043567	0.048712	0.126091	0.157432	0.101409	-0.14494	0.095315	0.022929	-0.1233	1	
Built or Renovated	0.671873	0.313147	0.769706	0.644538	0.321282	0.55196	0.221541	0.701023	0.647547	0.000531	1

The threshold that was chosen to form the above correlation analysis is 0.65. The cells that are highlighted represent with the significant correlation that the value is not in between -0.65 and 0.65. The reason for choosing 0.65 as the threshold is because we do not want to have too few or too many explanatory variables for the initial model. The six explanatory variables are:

X2: Bathrooms
X3: Sqft_living
X5: Floors
X7: Quality Grade
X8: Sqft_above
X10: Built or Renovated

For the next step, the regression model analysis will be conducted in excel following by the initial findings and conclusion from the correlation analysis. The graphs below demonstrate the results from the initial regression analysis and here presents the regression equation:

$$\begin{aligned}\text{Log (Price)} \quad = \; & 4.36 + 0.0030 \text{ bathrooms} + 0.000064 \text{ sqft\_living} \\ & + 0.0196 \text{ floors} \\ & + 0.0615 \text{ Quality Grade} + 0.000026 \text{ sqft\_above} \\ & + 0.000320 \text{ Built or Renovated}\end{aligned}$$

Figure 6: Analysis of the Initial Regression Model for Log(Y) and X2, X3, X5, X7, X8, X10

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.928964498
R Square	0.862975038
Adjusted R Square	0.851712712
Standard Error	0.089130216
Observations	80

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	6	3.65234104	0.608723507	76.62494083	1.68781E-29
Residual	73	0.579926268	0.007944195		
Total	79	4.232267308			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	4.358444532	1.224418269	3.559604298	0.000657697	1.918182441	6.798706622
bathrooms	0.002975967	0.026407061	0.112695876	0.910581003	-0.04965323	0.055605163
sqft_living	6.41466E-05	3.28986E-05	1.949825103	0.05503788	-1.4203E-06	0.000129713
floors	0.019644216	0.029473569	0.666502796	0.507190816	-0.039096522	0.078384954
Quality Grade	0.061507933	0.014334188	4.290995162	5.37934E-05	0.032939937	0.09007593
sqft_above	2.55291E-05	3.00079E-05	0.850748059	0.397691273	-3.42764E-05	8.53347E-05
Built or Renovated	0.00031956	0.000640369	0.499024874	0.619262392	-0.000956694	0.001595814

After conducting the initial regression analysis, we have the observations and findings as follow:

- The regression equation shows the positive relationship between the dependent variables and explanatory variables since the coefficients for these four variables are all positive.
- As looking at the F-value and P-value from the ANOVA section, the p-value is very small which equals to 0.001. It indicates that the set of explanatory variables that were chosen is fairly a good set from the perspective that not all them being equal to zero. And it also indicates to reject the null hypothesis in favor of the alternative hypothesis that at least one coefficient is not equal to zero.
- The F-value is moderately high and we are hoping to get a higher F-value.
- The R-Squared is equal to 86.3% and which is fairly high. We would like to increase it at the later stage to improve the goodness of fitting the mode.
- For the individual explanatory variables from the Summary Output, both X2 variable (Bathrooms) and X10 variable (Built or Renovated) have a relatively high P-value (Bathrooms: 0.91, Built or Renovated: 0.62).
- When looking at the variance inflation factors (VIF) values, both X3 (sqft_living) and X4 (sqft_above) are very high which are bigger than 5. It indicates there is some collinearity between explanatory variables and there is a need to at least remove one of them.
- The Durbin-Watson Statistic that was obtained from Minitab is equal to 2.00581. It is in the range of 1.5 to 2.5 which indicates very little to no presence of auto-correlation between the variables and observations.

- The four in one plot talks about the normality of the residuals and the constant variables of the residuals. The normal probability plot shows that a normal distribution with a mean of zero and standard deviation of 0.08913 fits the residuals well. The points are mostly following along with the straight line but there is one point seems to fall outside of the confidential boundaries. The P-value is bigger than 0.250 which indicates a good measure of the normality test.
- For the Versus Fits scatter plot, there is no apparent heteroscedasticity can be found on the graph. It means there is evidence to support the residuals have a constant variance.
- The Versus order graph shows that the points are relatively chaotic. It indicates that the residuals are not infected individually distributed.

The following figures support the observations listed above:

Durbin-Watson Statistic

Durbin-Watson Statistic = 2.00581

Figure 7: Regression Results Conducted from Minitab

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	4.36	1.22	3.56	0.001	
bathrooms	0.0030	0.0264	0.11	0.911	7.20
sqft_living	0.000064	0.000033	1.95	0.055	17.81
floors	0.0196	0.0295	0.67	0.507	2.08
Quality Grade	0.0615	0.0143	4.29	0.000	4.68
sqft_above	0.000026	0.000030	0.85	0.398	14.67
Built or Renovated	0.000320	0.000640	0.50	0.619	2.71

Figure 8: Four in One Plots for Residuals of Log (Price)

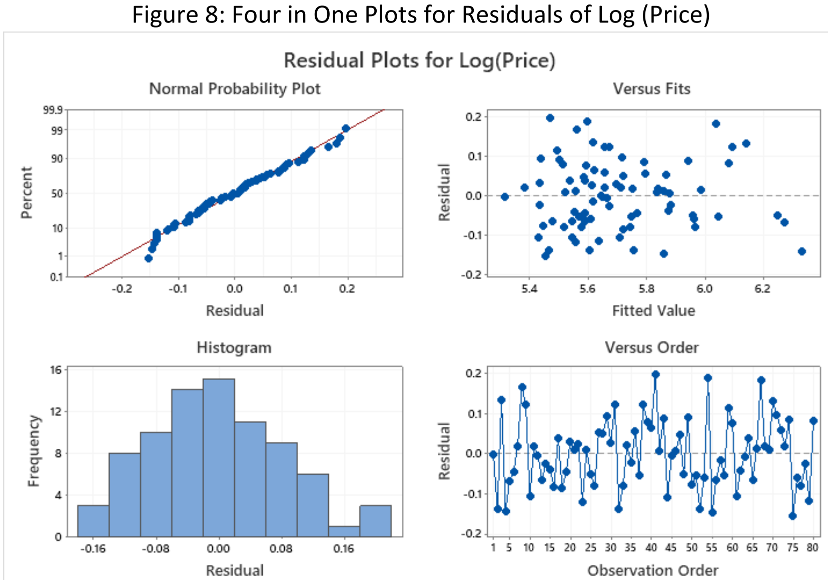
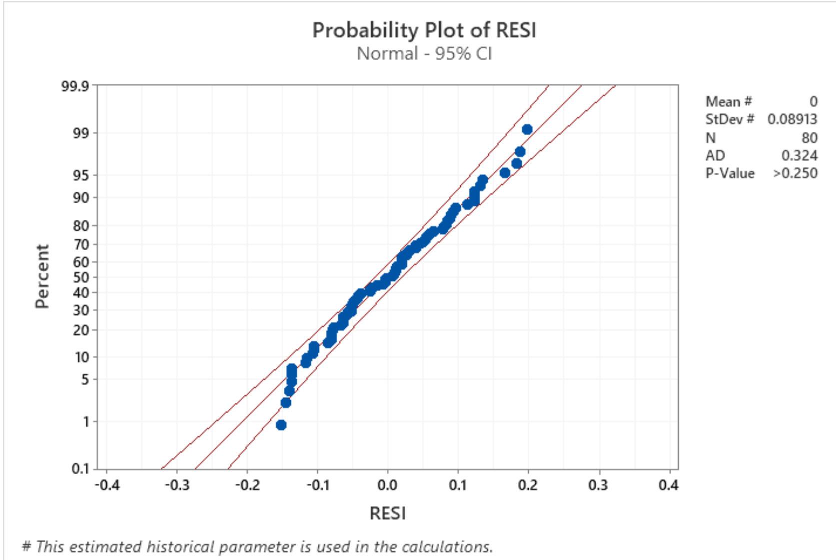


Figure 9: Probability Plot of Residuals



The analysis result of the initial regression model indicates there is a need to remove certain explanatory variable in order to improve the performance of the existing model. As our previously observation on comparing the value of VIF, the explanatory variable which has the highest is X3 (sqft_living) so we decide to remove it.

Remove X3 (sqft_living) Variable

Figure 10: Adjusted Regression Analysis for Log(Y) and X2, X5, X7, X8 and X10

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.925115571					
R Square	0.855838819					
Adjusted R Square	0.846098199					
Standard Error	0.090801882					
Observations	80					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	5	3.622138655	0.724427731	87.86286603	1.06942E-29	
Residual	74	0.610128653	0.008244982			
Total	79	4.232267308				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	4.823914863	1.223444503	3.942896349	0.000181096	2.386148478	7.261681248
bathrooms	0.026332478	0.02397523	1.098320155	0.275625498	-0.02143921	0.074104167
floors	0.009145793	0.029521073	0.309805583	0.757579561	-0.04967623	0.067967816
Quality Grade	0.066161268	0.014399211	4.594784155	1.74343E-05	0.037470213	0.094852323
sqft_above	7.27719E-05	1.80356E-05	4.034906594	0.000131658	3.68352E-05	0.000108709
Built or Renovated	7.19916E-05	0.000639427	0.112587619	0.910662364	-0.001202095	0.001346079

- As we can initially see from the results below, the VIF values among the explanatory variables have a significant drop and they are all either at below 5 or at around 5. It indicates the collinearity among explanatory variables drop.
- For the individual explanatory variables of X5 variable (floors) and X10 variable (Built or Renovated), they both have a relatively high P-value (floors: 0.76, Built or Renovated: 0.91). We should consider removing one of these variables in the next regression model.

Figure 11: Adjusted Regression Results Conducted from Minitab

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	4.82	1.22	3.94	0.000	
bathrooms	0.0263	0.0240	1.10	0.276	5.72
floors	0.0091	0.0295	0.31	0.758	2.01
Quality Grade	0.0662	0.0144	4.59	0.000	4.55
sqft_above	0.000073	0.000018	4.03	0.000	5.11
Built or Renovated	0.000072	0.000639	0.11	0.911	2.61

Overall, the removal of the X3 (sqft_living) variable results in improving the performance of the regression model. While the VIF value among the explanatory variables have significantly decreased, the individual P-value of explanatory variable X10 (Built or Renovated) remains very high (at around 0.91) and we decide to remove it.

Remove X10 (Built or Renovated) Variable

Figure 12: Adjusted Regression Analysis for Log(Y) and X2, X5, X7, and X8

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.925102224
R Square	0.855814125
Adjusted R Square	0.848124211
Standard Error	0.09020223
Observations	80

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	3.622034142	0.905508536	111.290477	9.53539E-31
Residual	75	0.610233166	0.008136442		
Total	79	4.232267308			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	4.961416802	0.072124875	68.78926001	1.6253E-69	4.817736683	5.105096922
bathrooms	0.027599718	0.021029128	1.312451826	0.193370079	-0.014292454	0.06949189
floors	0.009495662	0.029163188	0.325604402	0.745630356	-0.048600388	0.067591713
Quality Grade	0.066459679	0.014059705	4.726961209	1.04342E-05	0.038451311	0.094468047
sqft_above	7.24581E-05	1.77012E-05	4.093400092	0.000106124	3.71955E-05	0.000107721

- As we see from the individual VIF value among the explanatory variables, all of them remain low which are all smaller than 5. It indicates the collinearity among explanatory variables remain low and it is a good set of explanatory variables.
- The R-squared of the model at this point is 85.58%.
- As looking at the individual explanatory variables, X5 variable (floors) still remains high in P-value (floors: 0.746). We may consider to remove it in the next regression model.

Figure 13: Adjusted Regression Results Conducted from Minitab

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	4.9614	0.0721	68.79	0.000	
bathrooms	0.0276	0.0210	1.31	0.193	4.46
floors	0.0095	0.0292	0.33	0.746	1.99
Quality Grade	0.0665	0.0141	4.73	0.000	4.40
sqft_above	0.000072	0.000018	4.09	0.000	4.98

At this point, we are still detecting a relatively high P-value in explanatory variable, X5 (floors). So, we then decide to remove this variable. We also want to compare with the R-squared of the regression after removing the variable. The result of the regression analysis is followed:

Remove X5 (floors) Variable

Figure 14: Adjusted Regression Analysis for Log(Y) and X2, X7, and X8

Regression Statistics	
Multiple R	0.924992058
R Square	0.855610307
Adjusted R Square	0.849910714
Standard Error	0.089670139
Observations	80

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	3.621171531	1.207057177	150.1177865	7.51701E-32
Residual	76	0.611095777	0.008040734		
Total	79	4.232267308			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	4.962804785	0.071574083	69.3380142	1.82673E-70	4.820252644	5.105356925
bathrooms	0.027836915	0.020892533	1.332385822	0.186714782	-0.013774171	0.069448001
Quality Grade	0.067681278	0.013469956	5.024610077	3.25623E-06	0.040853535	0.09450902
sqft_above	7.35136E-05	1.72992E-05	4.249548123	6.0156E-05	3.90593E-05	0.000107968

- The R-squared does not change much after dropping the explanatory variable X5 (from 85.58% to 85.56%). It indicates that variable X5 does not contribute much to the model so it is a good move to drop it.
- The individual P-value of the remaining explanatory variables look fine. They are relatively small value and it indicates that they are significant variables to the model.
- The overall P-value decreases from 0.001 (initial model) to nearly 0. It indicates that the set of explanatory variables that were chosen is fairly a good set from the perspective that not all them being equal to zero. And it also indicates to reject the null hypothesis in favor of the alternative hypothesis that at least one coefficient is not equal to zero.
- The F-value is moderately high and it indicates a better fit.
- The Durbin-Watson Statistic is equal to 1.89534 and it indicates very little to no presence of auto-correlation between the variables and observations.
- The normal probability plot shows that a normal distribution with a mean of zero and standard deviation of 0.0896701 fits the residuals well. The points are mostly following along with the straight line but there is one point seems to fall outside of the confidential boundaries. The P-value is bigger than 0.250 which indicates a good measure of the normality test.
- There is no apparent heteroscedasticity can be found on the graph and the points on Versus order graph are relatively chaotic. It means there is evidence to support the residuals have a constant variance and they are not infected individually distributed.

Figure 15: Adjusted Regression Results Conducted from Minitab

Regression Equation

Log(Price) = 4.9628 + 0.0278 bathrooms + 0.0677 Quality Grade + 0.000074 sqft_above

Durbin-Watson Statistic

Durbin-Watson Statistic = 1.89534

Figure 16: Four in One Plots for Residuals of Log (Price)

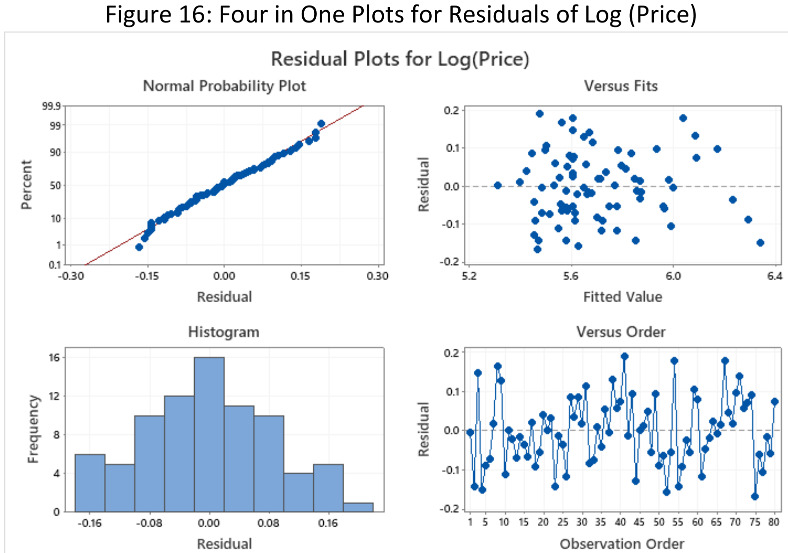
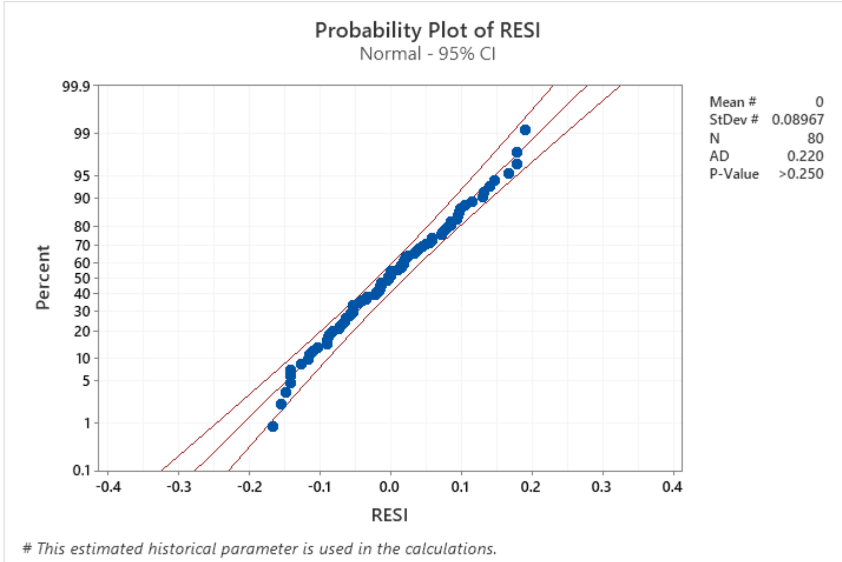
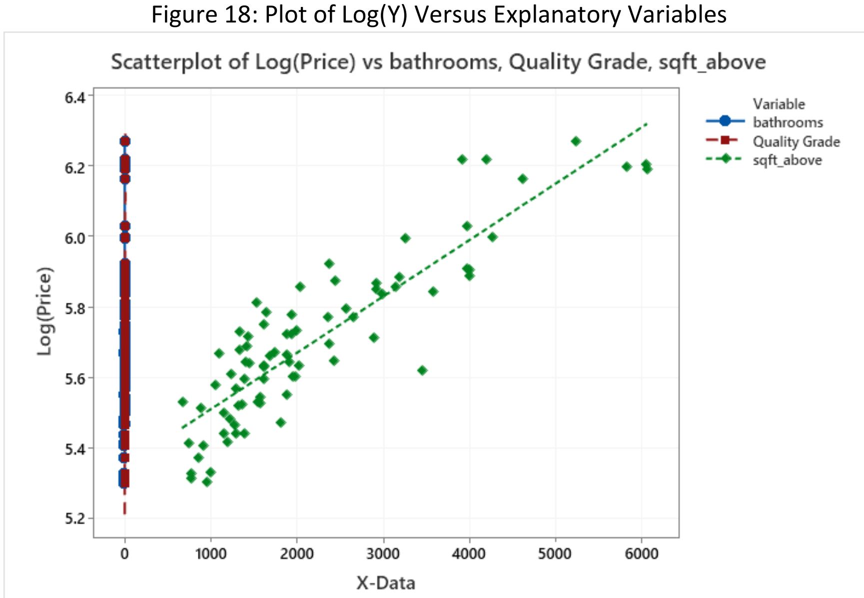


Figure 17: Probability Plot of Residuals



The above graphs and figures support the observations list previously. Now the regression model looks great with reasonable low individual P-values and VIF value after removing certain explanatory variables. Next, to further improve the performance of the existing model, we look at the interaction between explanatory variables when plotted against the explanatory variables. The scatter plot below demonstrates the interaction between X2 (bathroom) and X8 (sqft_above). We then add an addition explanatory variable to see how the performance of the existing regression model will be affected.

Figure 18: Plot of Log(Y) Versus Explanatory Variables



Add a New Explanatory Variable (Xnew): X2 (bathroom) * X8 (sqft_above)

The following is results of the regression model that consists of new added variable, Xnew:

Figure 19: Adjusted Regression Analysis for Log(Y) and X2, X7, X8, and Xnew

Regression Statistics						
Multiple R	0.926780937					
R Square	0.858922905					
Adjusted R Square	0.851398793					
Standard Error	0.089224509					
Observations	80					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	4	3.635191331	0.908797833	114.1560539	4.22555E-31	
Residual	75	0.597075977	0.007961013			
Total	79	4.232267308				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	4.929093566	0.07561335	65.18813905	8.5879E-68	4.778464048	5.079723083
bathrooms	0.050769433	0.027033273	1.878034977	0.06426361	-0.003083609	0.104622476
Quality Grade	0.063126216	0.013835559	4.562606734	1.93509E-05	0.035564369	0.090688064
sqft_above	0.000107509	3.0863E-05	3.483414625	0.000829834	4.60264E-05	0.000168991
Xnew	-9.56397E-06	7.20695E-06	-1.327048331	0.188518649	-2.3921E-05	4.79301E-06

The observations resulting from the adjusted regression model are as follows:

- The R-Squared value has increased from 85.56% to 85.89% indicates a better fit.
- The adjusted R-Squared value has increased from 84.99% to 85.13% indicates better fit.
- The P-value has still remained very low and F-value is moderately high. Each value looks fine to the model.
- The Durbin-Watson statistic is equal to 1.88456 which is still close to 2 indicating very little to no presence of auto-correlation.
- The individual P-value of explanatory variables are all small enough indicating we can reject the null hypothesis in favor of the alternative hypothesis that at least one coefficient is not equal to zero.
- The residuals analysis seems to support the normality test assumption for residuals.
- The normal probability plot shows that a normal distribution with a mean of zero and standard deviation of 0.08922 fits the residuals well. The points are mostly following along with the straight line but there is one point seems to fall outside of the confidential boundaries. The P-value is bigger than 0.250 which indicates a good measure of the normality test.
- There is no apparent heteroscedasticity can be found on the graph and the points on Versus order graph are relatively chaotic.

Figure 20: Four in One Plots for Residuals of Log (Price)

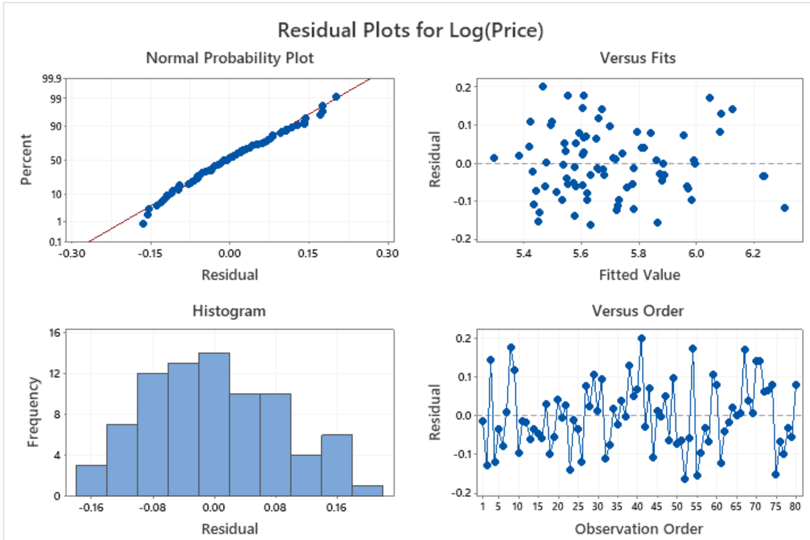
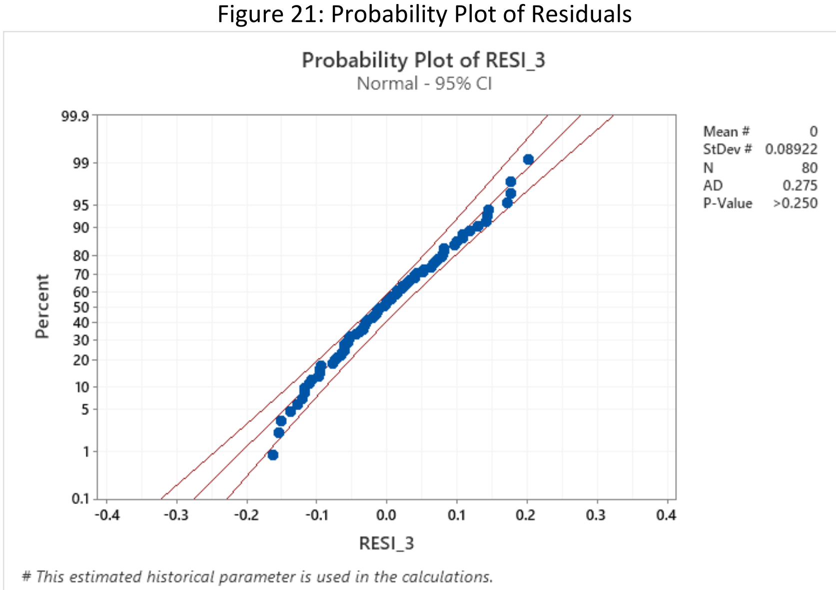


Figure 21: Probability Plot of Residuals



Add a Second New Explanatory Variable (Xnew2): X7 (Quality Grade) * X8 (sqft_above)

We are trying to find out if adding a second variable would result in additional improvement to the model. Again, by conducting a regression analysis and comparing the R-Squared results to the previous model can give us a clear understanding. In addition, we will also like to focus on comparing the individual P-value of the explanatory variables to determine whether or not an additional variable improves the existing model.

The regression analysis results are as follow:

Figure 22: Adjusted Regression Analysis for Log(Y) and X2, X7, X8, Xnew and Xnew2

Regression Statistics						
Multiple R	0.926796659					
R Square	0.858952047					
Adjusted R Square	0.84942178					
Standard Error	0.089816076					
Observations	80					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	5	3.635314668	0.727062934	90.12885358	4.79318E-30	
Residual	74	0.59695264	0.008066928			
Total	79	4.232267308				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	4.916356137	0.128081818	38.38449694	1.3268E-50	4.661147544	5.171564729
bathrooms	0.046745107	0.042423743	1.101861915	0.274092634	-0.037786047	0.131276261
Quality Grade	0.06580836	0.025777684	2.552919798	0.012742998	0.014445203	0.117171517
sqft_above	0.000112967	5.39821E-05	2.092679106	0.039806657	5.40553E-06	0.000220529
Xnew	-8.32254E-06	1.23867E-05	-0.671893714	0.503743693	-3.30036E-05	1.63585E-05
Xnew2	-9.92776E-07	8.02894E-06	-0.123649739	0.901927849	-1.69908E-05	1.50052E-05

As comparing on the results of the new model to the existing one, we can conclude that adding a second new variable does not improve the regression mode. The following observations substantiate the conclusion:

- Both the R-Squared and Adjusted R-Squared value remain exactly the same as the previous model. It indicates adding this variable does not improve the model.
- The P-value of the Xnew2 variable is extremely high, it indicates this variable is not significant and does not contribute much to the performance of the model.

Therefore, we will not add the second new explanatory variable.

Diagnostic Analysis

Before the final regression model is selected, we want to conduct a diagnostic analysis on the model and to identify influential observations. By calculating the TRES and DFIT value from Minitab and pasting them on excel, we use the threshold to filter the values that are not in the range of the selected threshold. The threshold for TRES is 1.6657 and for DFIT is 0.5. We will then find these highlighted rows accordingly on the residual plot to identify if they are influential observations. The highlighted rows that are generated from excel is below:

Table 2: Highlighted Rows for the Influential Observations in TRES and DFIT

Propert	TRES	DFIT
4	-1.519767745	-0.83263865
8	2.072423235	0.478395859
32	-1.439509733	-0.832547065
38	1.546958842	0.528711773
41	2.371581547	0.491490845
52	-1.871329339	-0.262159406
54	2.025871657	0.25585561
55	-1.788000759	-0.372254508
67	2.017511182	0.505805355
70	1.941997453	1.270955387
75	-1.763890556	-0.386419516

There are a number of influential observations detected. For instance, row 52 is detected as the point that is located outside of the confidential boundaries on the residual plot. Row 54, 55, and 67 are all detected as influential observations since they do not follow strictly of the straight line on residual plot. We then check back at the data source and ensure that these data are all collected correctly. So, we are keeping all these data point eventually.

At this point, we can conclude that the best fit model is the one presented by the following equation:

Log (Price) = 4.9291 + 0.0508 bathrooms + 0.0631 Quality Grade
+ 0.000108 sqft_above
- 0.000010 bathrooms * sqft_above

The explanatory variables are:

X2: bathrooms
X7: Quality Grade
X8: sqft_above
Xnew: bathrooms*sqft_above

Forecasting Dependent Variable Values

On this section, we will predict the median and average of Price/Y of a real estate property for the following values of the explanatory variables and provide a 95% prediction interval for Price/Y and an approximate 95% confidence interval for E[Y].

The provided values for each of the explanatory variables are below:

Table 3: Numbers of Attributes to estamate on Price/Y

	Y	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
Forecast	??????	3	2	1500	50000	2	0	6	1000	500	2000

As conducting the prediction both from Minitab and excel, we have the results as follow:

Figure 23: Prediction Analysis for the Model Using Provided Inforamtion

PFITS	PSEFITS	CLIM	CLIM_1	PLIM	PLIM_1
5.497770371	0.021131402	5.45567446	5.539866283	5.315109168	5.680431574
LOG(PRICE) - hat	5.498	5.498			
MEDIAN[PRICE]	$314,608.44		Standard Error Residuals		0.089224509
E[PRICE]	$321,699.15		Var[Log(Price)]		0.008408
			Standard Deviation [Log(Price)]		0.091693
95% Confidence Interval			**95% Prediction Interval (or Credibility Interval)**		
LB E[LOG(PRICE)]	5.45567		LB LOG(PRICE)	5.315109	
UB E[LOG(PRICE)]	5.53987		UB LOG(PRICE)	5.680432	
Approximate 95% Confidence Interval			**95% Prediction Interval (or Credibility Interval)**		
LB E[PRICE]	$285,544.93		PRICE	$206,589.94	
UB E[PRICE]	$346,630.11		PRICE	$479,105.96	

The median of the price of the real estate property is $314,608.44 and the average price of the real estate is $321,699.15 using the information from Table 1. The 95% prediction interval for Price/Y has a lower bound of $206,589.94 and an upper bound of $479,105.96. For the approximate 95% Confidence Interval of E[Price], it has a lower bound of $285,544.93 and an upper bound of $346,630.11.

Conclusion

Based on the results we derived from prediction analysis, it is clear to notice that the range for the prediction interval of Price/Y is very large. This is because the price of real estate property is relatively expensive in reality and can fluctuate a lot as well. The price of a real estate property can be influence by many factors. Using the adjusted final model that is selected in the previous step, we can see that the numbers of bathrooms, the quality grade of the property, and the sqft_above are the factors that generate the prediction of the property price.

When looking at the final equation of the selected final model, it is important to notice that variable X7 (Quality Grade) has the highest coefficient which indicates that it has the largest influence on the final prediction of the property price, comparing with other explanatory variables in the model. It is quite reasonable since we know that properties with higher quality grade tends to have a higher price. For the other two explanatory variables, X2 (Bathroom) and X8 (sqft_above), they all have a positive relationship with the property price. They both indicate that the increase in both variables will results in the increase of the property price. However, the last explanatory variable, Xnew (Bathroom*sqft_above), has a negative relationship with the property price. It indicates that it is not necessarily good or desirable when the number of bathrooms and the sqft_above of a property are very large. This explanatory variable helps adjust the price of property when both the numbers of bathrooms and the sqft_above are too large. Therefore, the adjusted model for predicting the property price is recommended for homeowners or invertors.