Estimate Price of Real Estate Property Using Linear Regression

By:
Hanna Song
EMSE 6765
December 1, 2022

Introduction: The price of Real Estate property can be determined by a number of attributes. From the house price dataset given, this report will discuss the following:

1. Develop a model for linear regression use of Log(Y) as the dependent variable as opposed to Y
2. Find the estimated linear regression of Log(Y) on an appropriate set of explanatory variables X1 to X10 using the 80 properties and interpret the result.
3. Perform Diagnostic analysis as per the techniques described on the regression analysis of the final selected model chosen
4. Forecast the median and average of price Y of a real estate property for the following values of the explanatory variables and provides a 95% prediction interval for Y and an approximate 95% confidence interval for E[Y]

Table 1 provides the 20 sample subset of the original house price data with the following variables:

Dependent variable is:

- Price of a Real Estate Property (Y)
- The Log of Real Estate Property Price is represented as Log (Y)

Independent variable are:

- Bedrooms (X1)
- Bathrooms (X2)
- Sqft Living (X3)
- Sqft Lot (X4)
- Floors (X5)
- Numbers of Times Viewed (X6)
- Quality Grade (X7)
- Sqft Above (X8)
- Sqft Basement (X9)
- Built or Renovated (X10)

Table 1: Original Performance 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	$ 5.64	3	2.5	1910	66211	2	0	7	1910	0	1997
2	$ 213,000.00	$ 5.33	2	1	1000	10200	1	0	6	1000	0	1961
3	$ 563,500.00	$ 5.75	4	1.75	2085	174240	1	0	7	1610	475	1964
4	$ 1,550,000.00	$ 6.19	5	4.25	6070	171626	2	0	12	6070	0	1999
5	$ 1,600,000.00	$ 6.20	6	5	6050	230652	2	3	11	6050	0	2001
6	$ 350,000.00	$ 5.54	3	2.25	1580	47916	1	0	7	1580	0	1979
7	$ 540,000.00	$ 5.73	3	2.25	2000	217800	2	0	8	2000	0	1996
8	$ 535,000.00	$ 5.73	3	1	1330	40259	1	0	7	1330	0	1977
9	$ 600,000.00	$ 5.78	2	2.5	2410	102366	1	0	7	1940	470	1989
10	$ 275,000.00	$ 5.44	3	1	1370	17859	1	0	7	1150	220	1930
11	$ 589,000.00	$ 5.77	3	2.5	2660	206480	1	0	8	2660	0	1989
12	$ 439,900.00	$ 5.64	2	2	1410	12282	1.5	0	8	1410	0	1988
13	$ 260,000.00	$ 5.41	3	1	1190	11120	1	0	6	1190	0	1947
14	$ 445,000.00	$ 5.65	4	1.75	2430	13211	1.5	0	7	2430	0	1978
15	$ 685,000.00	$ 5.84	3	2.75	3150	219978	2	0	9	3000	150	1990
16	$ 315,000.00	$ 5.50	3	1.5	1750	12500	1	0	7	1150	600	1954
17	$ 378,000.00	$ 5.58	3	1.5	1050	57499	1	0	7	1050	0	1975
18	$ 430,000.00	$ 5.63	3	2.5	2030	7770	2	0	8	2030	0	2003
19	$ 529,000.00	$ 5.72	3	2.25	1940	217800	2	0	9	1940	0	1990
20	$ 291,000.00	$ 5.46	3	1	1280	10500	1.5	0	5	1280	0	1941

With the analysis of initial house price dataset, the following observations were made concerning the Dependent Variable (Y) and its relationship with the Independent Variables (X1-10):

- There is large variability in the original price (Y) data. In the histogram of the Price (Y), we can see that it is skewed toward the left. This would be problematic in conducting the regression analysis.
- The Probability Plot of Price (Y) also has a large standard deviation which will also cause a problem conducting the regression analysis.

Figure 1: Histogram of Price of a Real Estate Property (Y)

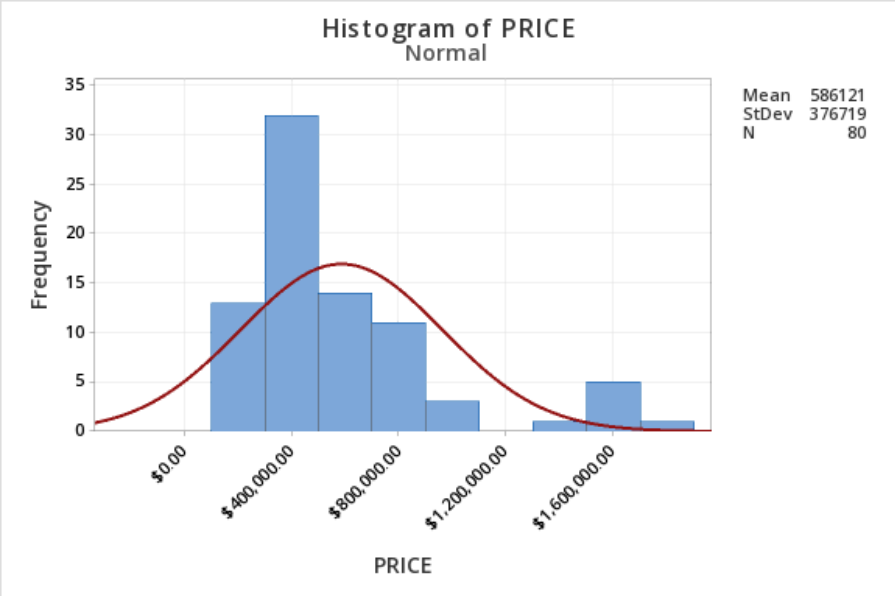
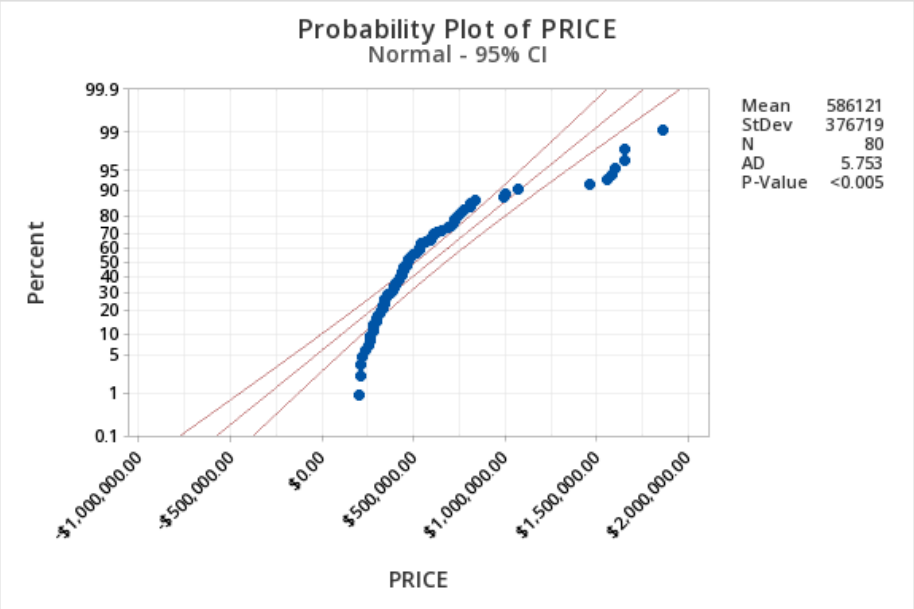


Figure 2: Probability Plot of Price of a Real Estate Property (Y)



We would like to have more normalized distribution for the dependent variable (Price (Y)). We decide to use rather Log(Y), and we do see some improvement in the distribution, especially increase in normality.

- The standard deviation for Log Price is much smaller than initial, but the P-value has increased.
- For the purposes of this report, we will use Log of Price as the dependent variable for the regression model. Choosing Log(Y) allows us to get a clear interpretation of the regression model.

Figure 3: Histogram of Log Price of a Real Estate Property (Log(Y))

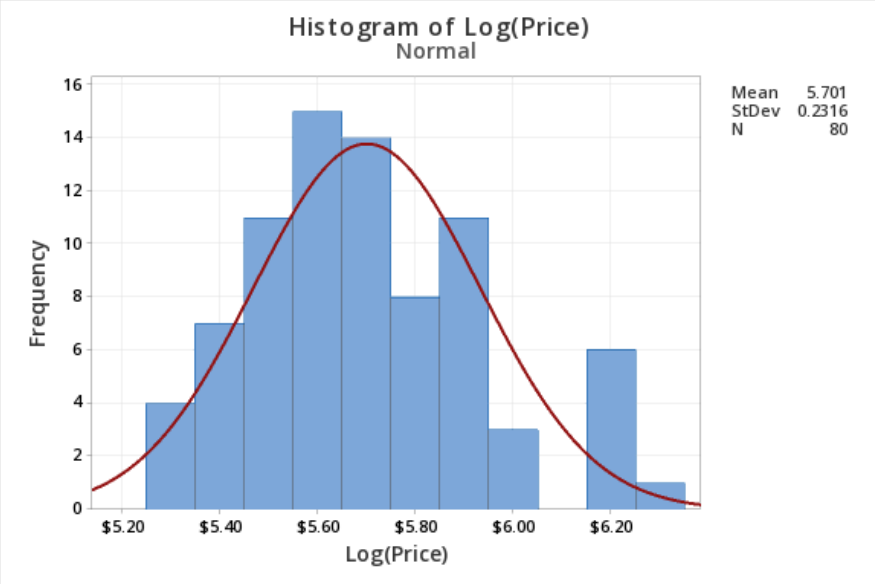
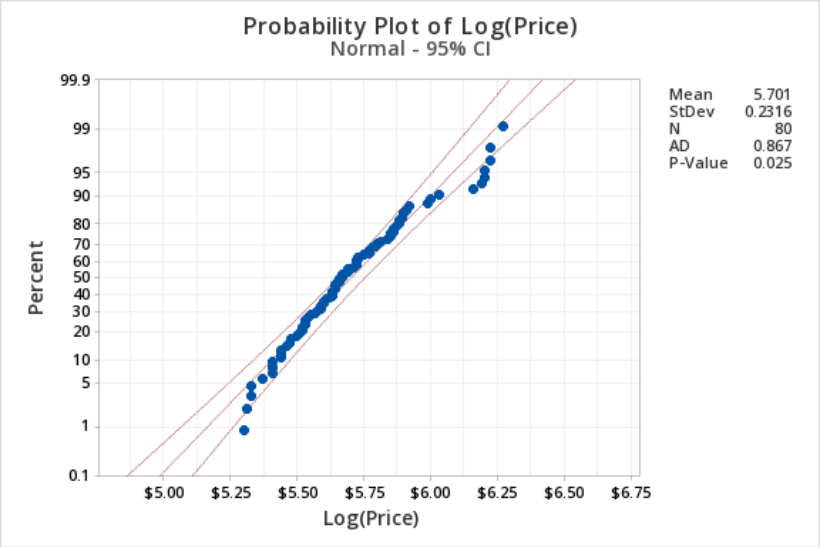


Figure 4: Probability Plot of Log Price of a Real Estate Property (Log(Y))



Correlation Analysis: Correlation Analysis is to determine the best predictors for the regression model, we completed a correlation analysis of the dependent variable Log(Y) and the independent variables (X1-X10). The figure below shows the correlation strengths between the dependent and independent variables.

As some independent variables show multicollinearity, I decide to use Threshold 0.6 rather than 0.8 for eliminating the independent variables with multicollinearity.

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

	Log(Price)	*bedrooms*	*bathrooms*	*sqft_living*	*sqft_lot*	*floors*	*ers of times vi*	*Quality Grade*	*sqft_above*	*qft_basement*	*ilt or Renovat*
Log(Price)	1										
bedrooms	0.52316644	1									
bathrooms	0.83892579	0.55905792	1								
sqft_living	0.89328459	0.60011722	0.88938022	1							
sqft_lot	0.61623017	0.11162056	0.42964328	0.47649725	1						
floors	0.66211037	0.41450755	0.62973124	0.62515949	0.47252586	1					
Numbers of ti	0.44552684	0.19675123	0.4129354	0.40151193	0.63496663	0.41916046	1				
Quality Grade	0.88757695	0.46197763	0.83022202	0.84656633	0.53892027	0.68420163	0.41466641	1			
sqft_above	0.88539791	0.58934551	0.8582438	0.96058309	0.45028536	0.66902532	0.37664804	0.84426044	1		
sqft_basemer	0.04356732	0.04871225	0.12609066	0.15743165	0.1014092	-0.1449437	0.0953152	0.02292904	-0.1233002	1	
Built or Renov	0.67187335	0.31314711	0.76970554	0.64453784	0.32128198	0.55196028	0.22154056	0.70102297	0.64754704	0.00053091	1

The highlighted variables represent significant correlation. Based on these findings, we should keep the following highlighted variables for initial regression models:

- Bathrooms X2
- Quality Grade X7
- Sqft Above X8
- Built or Renovated X10

Initial Regression Analysis:

Based on the decision we made from the correlation, we made three different initial models with finding the best fit of regression analysis.

First model included independent variables of Bathrooms, Sqft living, Quality Grade, and Sqft Above with Regression Equation of Log(Y) = 4.9723 + 0.0119 bathrooms + 0.000056 sqft_living + 0.0654 Quality Grade + 0.000032 sqft_above

However, it shows high VIF and high individual p-value which concerns multicollinearity between independent variables so we decided to remove this model.

The second model included independent variables of Bathrooms, Sqft lot, floors, and built or renovated which are the lowest four independent variables from Threshold of 0.6 for eliminating multicollinearity of variables of correlation with 0.8. With regression equation of Log(Y) = 4.35 + 0.1368 bathrooms + 0.000001 sqft_lot + 0.0555 floors + 0.000465 Built or Renovated

Third model which included Bathrooms, Quality grade, sqft Above, and Built or Renovated, we have removed the Sqft_living which has a highest correlation with Log(Y) and between independent variables for a better fit of the regression model.

Comparing the second model to the third model, the third model seems to have a good fit of VIF in between 1-5, lower p-Value, higher R squared adjusted, less Standard deviation in the Residual probability Plot than the second model.

As doing so, we decide to move on to the third model for adjusting the regression model from the initial model.

The observations resulting from the initial regression model below are as follows:

- The Variance Inflation factors (VIF) from minitab for each independent variable are all in range 1 to 5, so there is little collinearity among variables and the estimates for coefficients are considered stable.
- R-Squared = 85.41% and Adjusted R-Squared = 84.63% which is moderately high, however, we would like to have a higher R-Squared Adjusted value for getting "goodness of fit" for the regression model.
- The Durbin-Watson statistic = 1.88511 which in between 1 to 2, indicates that very little of autocorrelation among the observations.
- Critical F value of regression seems high enough with 109.78 but we would like to have higher F value for significant F value.
- P-value is 0.000 for the regression, which is good where we prefer a lower P-value.
- However, Individual P-value of X2 and X10 are high, but X7 and X8 have 0.000 of P-value which we are satisfied with low p-value.
- The residual analysis appears to support the assumption of normality where
- Our P-value for Residual Probability plot is over 0.250 which indicates goodness of fit for normality test.
- There are some influential observations (outlier) identified within the probability plot for the residuals. This observation may require to be removed or reviewed further.
- There is no apparent heteroscedasticity in the plot of residual versus fitted values for Log(Y) which indicates there is constant variance in residuals.

The following figures support the observations above:

Figure 6: Initial Regression Analysis for Log(Y) and X2,X7,X8,X10

WORKSHEET 2

Model3

Regression Equation

Log(Price) = 4.77 + 0.0266 bathrooms + 0.0668 Quality Grade + 0.000074 sqft_above
+ 0.000100 Built or Renovated

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	4.77	1.22	3.92	0.000	
bathrooms	0.0266	0.0240	1.11	0.271	5.72
Quality Grade	0.0668	0.0140	4.78	0.000	4.29
sqft_above	0.000074	0.000018	4.18	0.000	4.91
Built or Renovated	0.000100	0.000636	0.16	0.875	2.58

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0907978	85.41%	84.63%	83.17%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	3.62024	0.905059	109.78	0.000
bathrooms	1	0.01015	0.010150	1.23	0.271
Quality Grade	1	0.18864	0.188637	22.88	0.000
sqft_above	1	0.14395	0.143950	17.46	0.000
Built or Renovated	1	0.00021	0.000206	0.02	0.875
Error	75	0.61832	0.008244		
Lack-of-Fit	74	0.60987	0.008241	0.98	0.685
Pure Error	1	0.00845	0.008450		
Total	79	4.23855			

Fits and Diagnostics for Unusual Observations

Obs	Log(Price)	Fit	Resid	Std Resid	
32	5.6200	5.7039	-0.0839	-1.06	X
41	5.6700	5.4736	0.1964	2.28	R
54	5.7900	5.6088	0.1812	2.04	R
67	6.2200	6.0377	0.1823	2.07	R
70	6.2700	6.1691	0.1009	1.24	X

R Large residual
X Unusual X

Durbin-Watson Statistic

Durbin-Watson Statistic = 1.88511

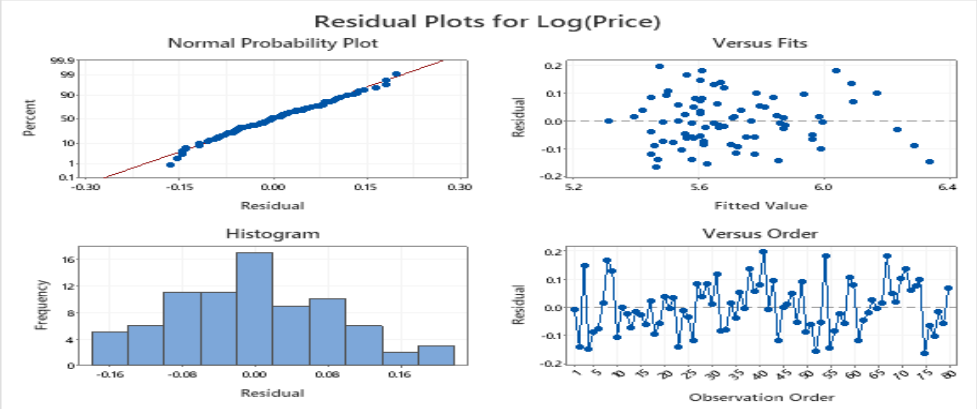
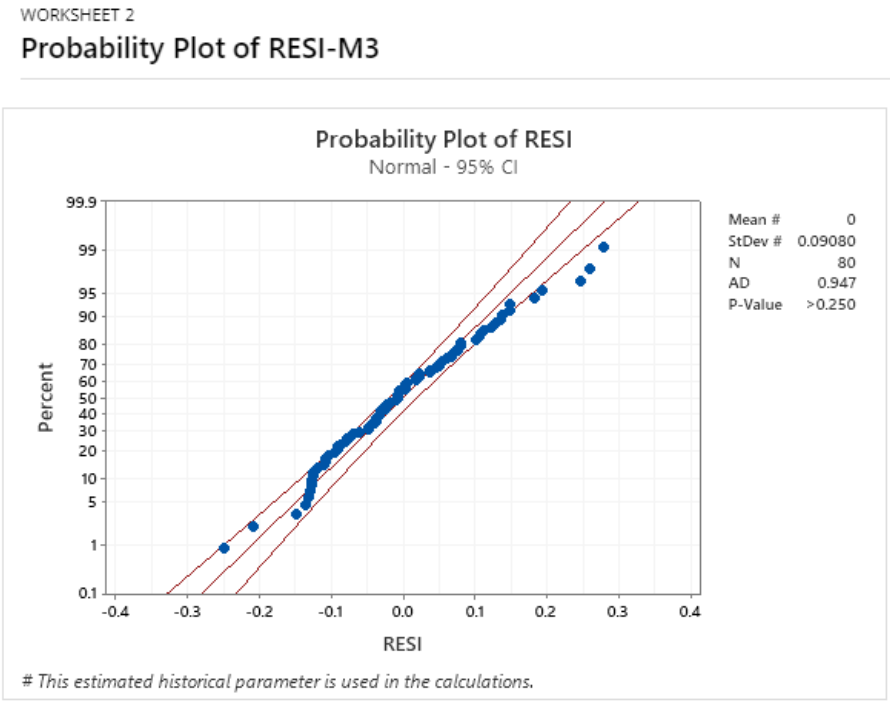


Figure 7: Probability Plot of Model 3 RESI



Diagnostic Analysis:

Initial Regression model indicates the regression equation of Log(Y) = 4.77 + 0.0266 bathrooms + 0.0668 Quality Grade + 0.000074 sqft_above + 0.000100 Built or Renovated

The initial regression model with residuals seems to need improvements from removing outliers, decreasing standard deviation of Residuals. However, the analysis of the residuals versus fitted values shows the majority of the values fall within expected threshold. As doing so, we decide to improve the model by looking at interaction between the independent variables.

These figures below shows scatter plots of Log(Y) and X10 with grouped other independent variables:

Figure 8: Scatterplot of Log(Y) vs X10

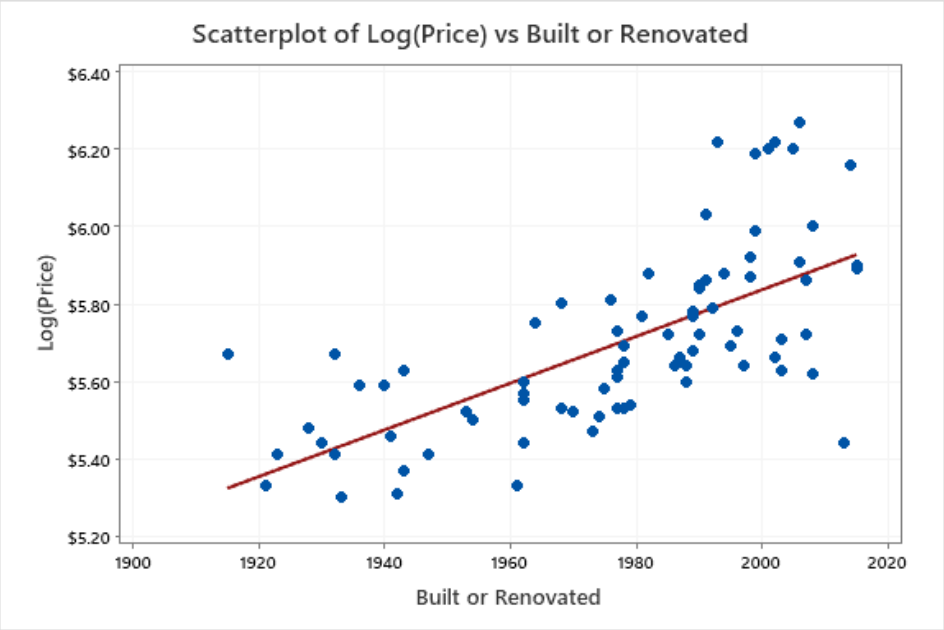
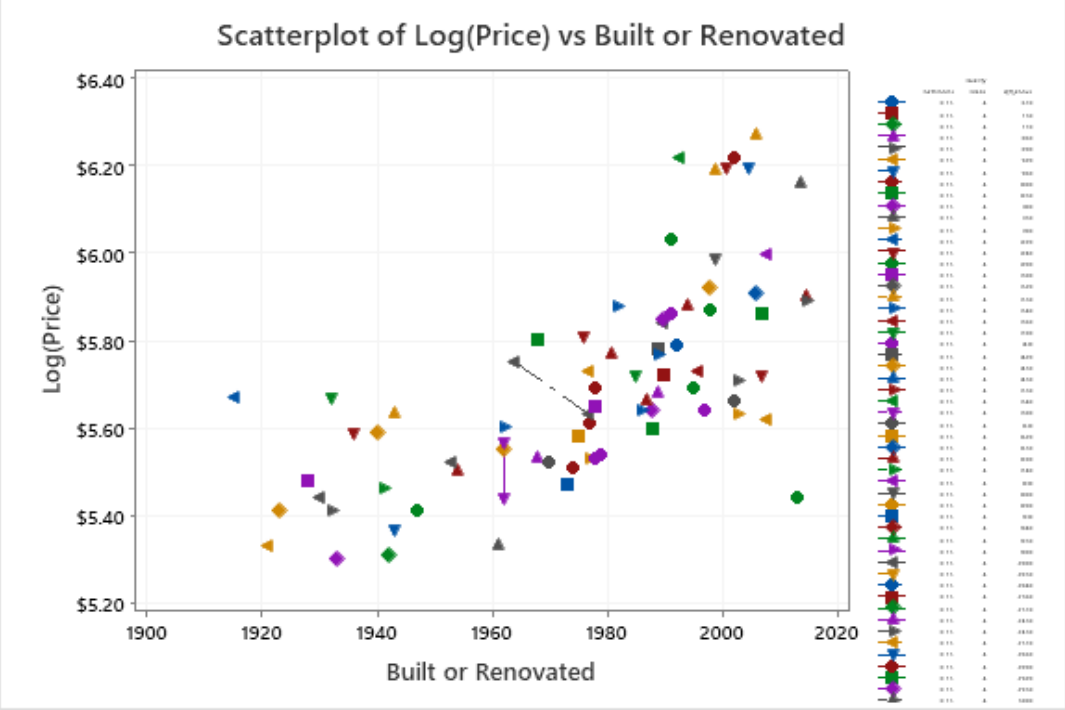


Figure 9: Scatterplot of Log(Y) vs X10 with grouped Independent Variables



The figure 9 : scatter plot with grouped independent variables indicates there is a positive correlation between X10 and other variables. We will try three different added variables regarding X10 by using the Minitab Fit Regression Model.

Based on interaction effect between Built or Renovated and Sqft Above, we decided to add another independent variable: Xnew1 : Built or Renovated * Sqft Above.

Following table displays the new independent variable along with the other remaining independent variables that comprise the regression model. We will now refer to the new adjusted model based on the additional independent variable.

Table 2: Adjusted Model Variables (Sample of first 20)

PRICE	Log(Price)	bedrooms	bathrooms	sqft_living	sqft_lot	floors	Numbers of times viewed	Quality Grade	sqft_above	sqft_basement	Built or Renovated	Built or Renovated*sqft_above
$ 440,000.00	$ 5.64	3	2.5	1910	66211	2	0	7	1910	0	1997	3814270
$ 213,000.00	$ 5.33	2	1	1000	10200	1	0	6	1000	0	1961	1961000
$ 563,500.00	$ 5.75	4	1.75	2085	174240	1	0	7	1610	475	1964	3162040
$ 1,550,000.00	$ 6.19	5	4.25	6070	171626	2	0	12	6070	0	1999	12133930
$ 1,600,000.00	$ 6.20	6	5	6050	230652	2	3	11	6050	0	2001	12106050
$ 350,000.00	$ 5.54	3	2.25	1580	47916	1	0	7	1580	0	1979	3126820
$ 540,000.00	$ 5.73	3	2.25	2000	217800	2	0	8	2000	0	1996	3992000
$ 535,000.00	$ 5.73	3	1	1330	40259	1	0	7	1330	0	1977	2629410
$ 600,000.00	$ 5.78	2	2.5	2410	102366	1	0	7	1940	470	1989	3858660
$ 275,000.00	$ 5.44	3	1	1370	17859	1	0	7	1150	220	1930	2219500
$ 589,000.00	$ 5.77	3	2.5	2660	206480	1	0	8	2660	0	1989	5290740
$ 439,900.00	$ 5.64	2	2	1410	12282	1.5	0	8	1410	0	1988	2803080
$ 260,000.00	$ 5.41	3	1	1190	11120	1	0	6	1190	0	1947	2316930
$ 445,000.00	$ 5.65	4	1.75	2430	13211	1.5	0	7	2430	0	1978	4806540
$ 685,000.00	$ 5.84	3	2.75	3150	219978	2	0	9	3000	150	1990	5970000
$ 315,000.00	$ 5.50	3	1.5	1750	12500	1	0	7	1150	600	1954	2247100
$ 378,000.00	$ 5.58	3	1.5	1050	57499	1	0	7	1050	0	1975	2073750
$ 430,000.00	$ 5.63	3	2.5	2030	7770	2	0	8	2030	0	2003	4066090
$ 529,000.00	$ 5.72	3	2.25	1940	217800	2	0	9	1940	0	1990	3860600
$ 291,000.00	$ 5.46	3	1	1280	10500	1.5	0	5	1280	0	1941	2484480

Adjusted Regression Analysis:

Now we have a new independent variable to analyze the regression model to determine how added independent variable results in an improvement of the regression model fit.

The figures below show the results and probability plot of residual from the adjusted Model3 regression analysis:

Figure 10: Adjusted Model 3 with new variable Xnew1 (X8*X10)

WORKSHEET 2

Adjusted Model 3

Regression Equation

Log(Price) = 1.72 + 0.0389 bathrooms + 0.0603 Quality Grade + 0.00266 sqft_above
+ 0.001634 Built or Renovated - 0.000001 sqft_above*Built or Renovated

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.72	1.82	0.95	0.346	
bathrooms	0.0389	0.0240	1.62	0.109	6.04
Quality Grade	0.0603	0.0139	4.33	0.000	4.48
sqft_above	0.00266	0.00117	2.28	0.026	22540.74
Built or Renovated	0.001634	0.000930	1.76	0.083	5.80
sqft_above*Built or Renovated	-0.000001	0.000001	-2.21	0.030	22923.57

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0885254	86.32%	85.39%	83.86%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	5	3.65864	0.731727	93.37	0.000
bathrooms	1	0.02058	0.020580	2.63	0.109
Quality Grade	1	0.14698	0.146984	18.76	0.000
sqft_above	1	0.04062	0.040617	5.18	0.026
Built or Renovated	1	0.02421	0.024211	3.09	0.083
sqft_above*Built or Renovated	1	0.03840	0.038399	4.90	0.030
Error	74	0.57992	0.007837		
Lack-of-Fit	73	0.57147	0.007828	0.93	0.698
Pure Error	1	0.00845	0.008450		
Total	79	4.23855			

Fits and Diagnostics for Unusual Observations

Obs	Log(Price)	Fit	Resid	Std Resid	
8	5.7300	5.5544	0.1756	2.05	R
32	5.6200	5.6960	-0.0760	-0.99	X
41	5.6700	5.4570	0.2130	2.55	R
52	5.4700	5.6462	-0.1762	-2.02	R
54	5.7900	5.6095	0.1805	2.08	R

R Large residual
X Unusual X

Durbin-Watson Statistic

Durbin-Watson Statistic = 1.86965

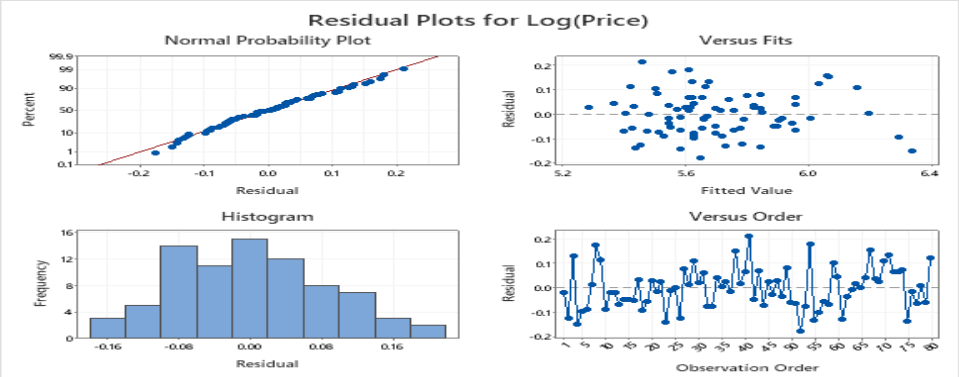
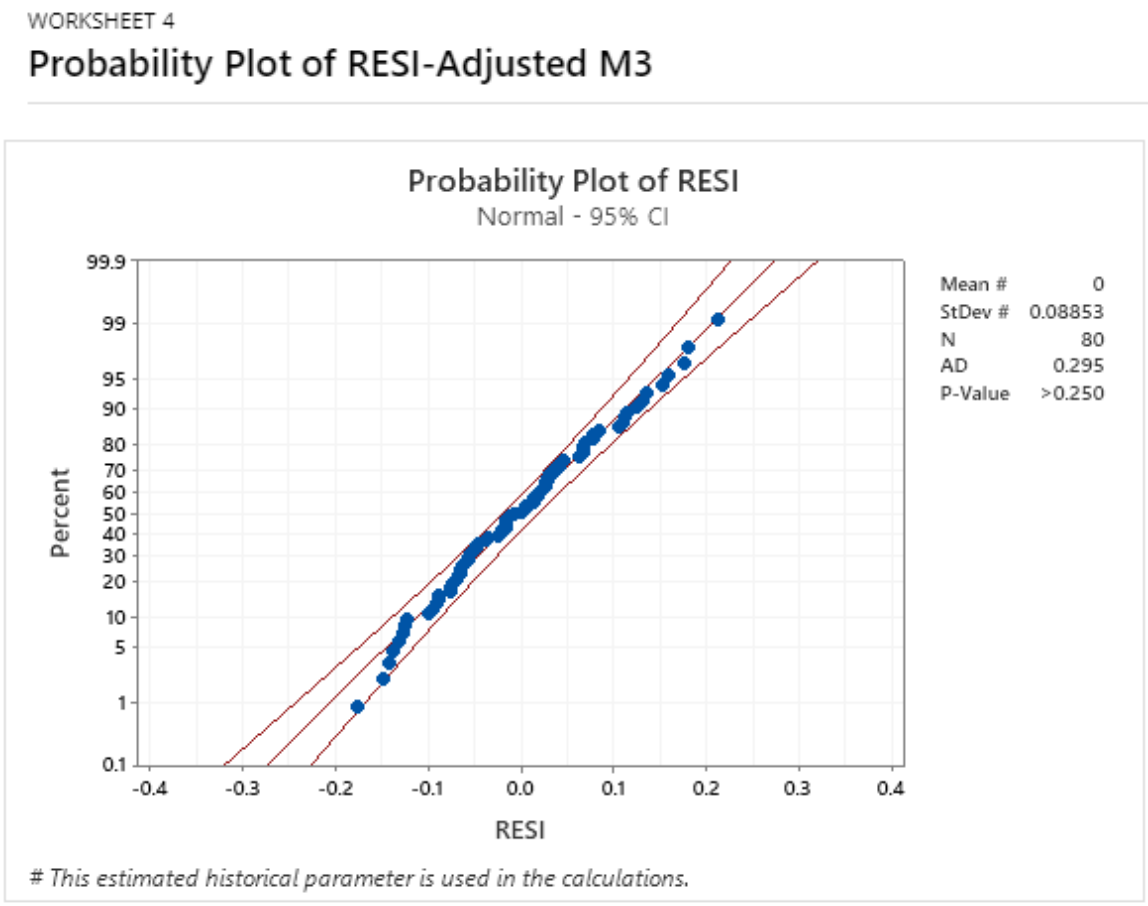


Figure 11: Probability Plot of Adjusted Model 3 with new variable Xnew1(X8*X10)



The observations from the adjusted regression model and probability plot of adjusted model above are as follows:

- The R-Squared value has slightly increased to 86.32%
- The Adjusted R-Squared value has increased to 85.39%, indicating that the new model seems little improvement.
- P-Value has increased 0.346 which gives the initial model a better fit than adjusted one.
- F-value decreased to 93.37 which gives the initial model a better fit than adjusted one.
- DW statistics = 1.86965, which is close to 2 indicating that very little of auto-correlation.
- Individual P-values for the independent variables, X2 and X10 still have high p-values but lower than initial model while new independent variable(X8*X10), X7, and X8 have low p-values.

- The residual analysis appears to support normality for residuals.
- Normal probability plot of Residuals shows quite decent normality which indicates goodness of fit.
- However, some deviations might be still removed or revised.
- We still see a high p-value of Probability plot of Residuals which indicates a better fit of model and there is no outlier identified for probability plot of Residuals
- There is no apparent heteroscedasticity in the plot of Residual versus fitted values for Log(Y)

Overall, the addition of new independent variable of Built or Renovated * Sqft Above results in a better model fit. However, we would like to decrease individual p-value and increase F-value for a better model fit.

Test for Adding a second new independent variable:
Now, we would like to add another independent variable if additional improvement to the regression model can happen. We can test by adding another independent variable to compare the regression model with the previous model. In particular, we would like to carefully look at the R-Squared Adjusted value and p-value to determine whether or not an additional variable can improve the model.

Hence, we test the additional variable of X7*X10 (Quality Grade * Built or Renovated). Below figures are the result of adjusted Model3 with new variable:

Figure 12: Adjusted Model 3 with Knew1 and Xnew2 (X7*X10)

WORKSHEET 2

Adjusted M3-2 Add

Regression Equation

Log(Price) = 10.46 + 0.0382 bathrooms - 1.71 Quality Grade + 0.00450 sqft_above
- 0.00277 Built or Renovated + 0.000894 Quality Grade*Built or Renovated
- 0.000002 sqft_above*Built or Renovated

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	10.46	5.47	1.91	0.060	
bathrooms	0.0382	0.0237	1.61	0.112	6.04
Quality Grade	-1.71	1.05	-1.63	0.107	26054.28
sqft_above	0.00450	0.00159	2.84	0.006	42682.03
Built or Renovated	-0.00277	0.00276	-1.00	0.319	52.42
Quality Grade*Built or Renovated	0.000894	0.000529	1.69	0.095	28397.55
sqft_above*Built or Renovated	-0.000002	0.000001	-2.79	0.007	43734.39

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0874336	86.83%	85.75%	84.37%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	6	3.68050	0.613416	80.24	0.000
bathrooms	1	0.01981	0.019809	2.59	0.112
Quality Grade	1	0.02040	0.020395	2.67	0.107
sqft_above	1	0.06152	0.061517	8.05	0.006
Built or Renovated	1	0.00769	0.007687	1.01	0.319
Quality Grade*Built or Renovated	1	0.02186	0.021861	2.86	0.095
sqft_above*Built or Renovated	1	0.05947	0.059469	7.78	0.007
Error	73	0.55806	0.007645		
Lack-of-Fit	72	0.54961	0.007633	0.90	0.704
Pure Error	1	0.00845	0.008450		
Total	79	4.23855			

Fits and Diagnostics for Unusual Observations

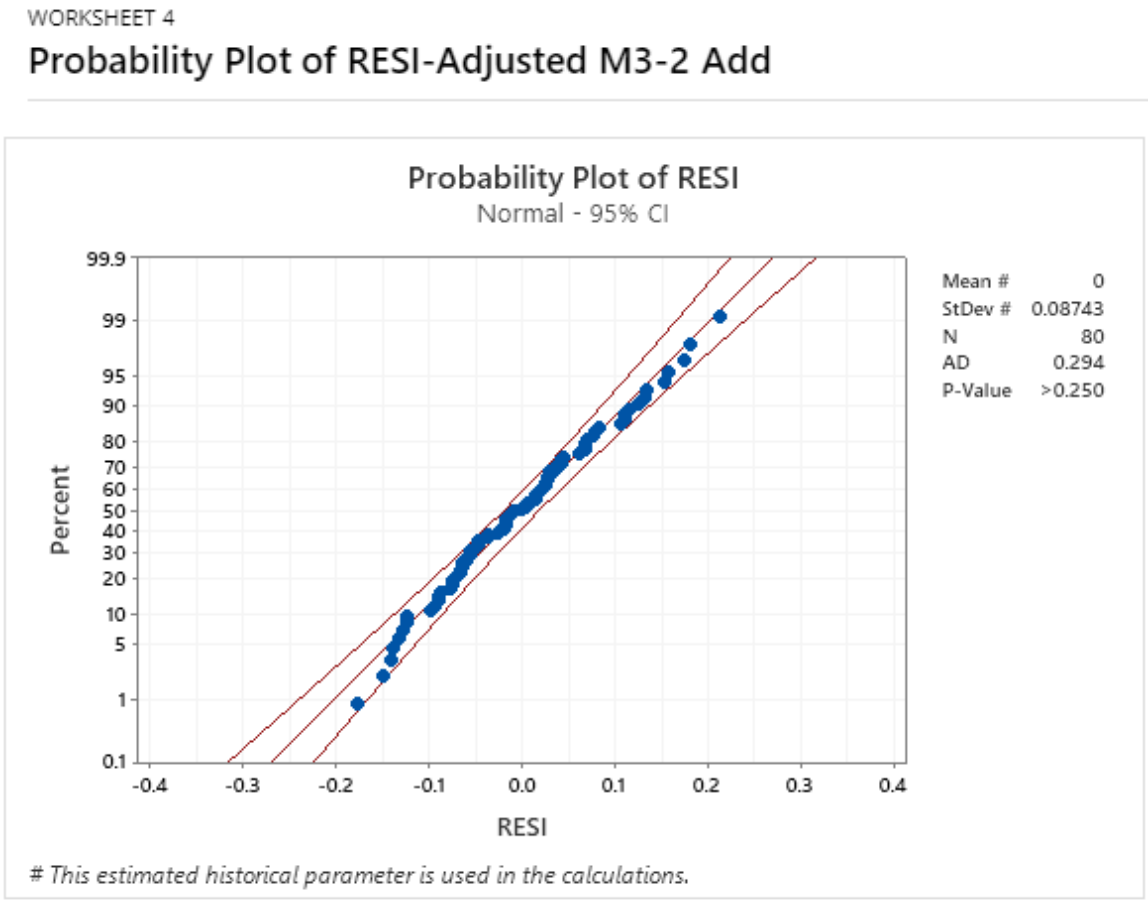
Obs	Log(Price)	Fit	Resid	Std Resid	
8	5.7300	5.5484	0.1816	2.15	R
32	5.6200	5.6409	-0.0209	-0.31	X
41	5.6700	5.4598	0.2102	2.55	R
45	5.3100	5.3471	-0.0371	-0.52	X
54	5.7900	5.6100	0.1800	2.10	R

R Large residual
X Unusual X

Durbin-Watson Statistic

Durbin-Watson Statistic = 1.85367

Figure 13: Probability Plot of Model 3 with Xnew1 and Xnew2 (X7*X10)



These figures show that adding a second new independent variable does improve the regression model.

- Both R-Squared and Adjusted R-Squared value slightly increased, indicating that adding new variables can slightly improve the model.
- Further, individual variable p-value decreased enough to be normality, but F-value decreased. However, p-value of regression decreased so we are using this model as a final model.

Adjusted Model 3 Test Drop Variable:

On the other hand, we would like to consider if we remove the independent variable of X2 (bathrooms) which has a high correlation from independent variables used in the model. Would subtraction of an existing variable improve the model? The following figures are result of the independent variable X2 removed:

Figure 14: Adjusted Model 3 with removing X2 (bathrooms)

WORKSHEET 2

Adjusted M3-2 Drop

Regression Equation

Log(Price) = 1.35 + 0.0664 Quality Grade + 0.00223 sqft_above + 0.001833 Built or Renovated
- 0.000001 sqft_above*Built or Renovated

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.35	1.82	0.74	0.463	
Quality Grade	0.0664	0.0136	4.89	0.000	4.17
sqft_above	0.00223	0.00115	1.94	0.056	21405.45
Built or Renovated	0.001833	0.000931	1.97	0.053	5.70
sqft_above*Built or Renovated	-0.000001	0.000001	-1.87	0.066	21689.87

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0894799	85.83%	85.08%	83.66%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	3.63806	0.909514	113.59	0.000
Quality Grade	1	0.19133	0.191325	23.90	0.000
sqft_above	1	0.03020	0.030197	3.77	0.056
Built or Renovated	1	0.03101	0.031007	3.87	0.053
sqft_above*Built or Renovated	1	0.02797	0.027969	3.49	0.066
Error	75	0.60050	0.008007		
Lack-of-Fit	74	0.59205	0.008001	0.95	0.693
Pure Error	1	0.00845	0.008450		
Total	79	4.23855			

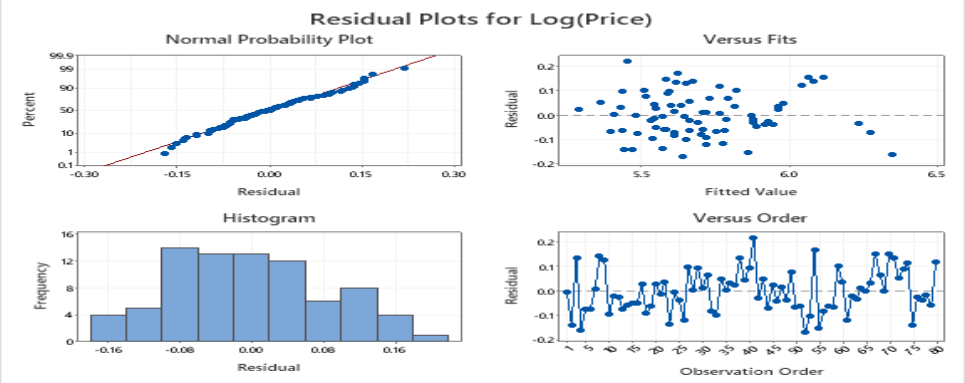
Fits and Diagnostics for Unusual Observations

Obs	Log(Price)	Fit	Resid	Std Resid	
32	5.6200	5.7008	-0.0808	-1.04	X
41	5.6700	5.4514	0.2186	2.59	R

R Large residual
X Unusual X

Durbin-Watson Statistic

Durbin-Watson Statistic = 1.81878



We can conclude the following observations from the results from the regression model with X2 removed:

- R2 value decreased, and the adjusted R2 value also decreased which indicates that we would not use this new model as a final model.
- Further supporting this decision, p-value of regression highly increased which gives the previous model a better fit than this model.

Best Regressions Model Fit:
We can now test for improved fit in the adjusted model by the initial model.

As a result, we can say with confidence the best fit model is the following regression equation:

Log(Price) = 10.46 + 0.0382 bathrooms - 1.71 Quality Grade + 0.00450 sqft_above - 0.00277 Built or Renovated + 0.000894 Quality Grade*Built or Renovated - 0.000002 sqft_above*Built or Renovated

With Dependent Variables Log(Y)
And independent variables:
- Bathrooms (X2)
- Quality Grade (X7)
- Sqft Above (X8)
- Built or Renovated (X10)
- Sqft Above (X8)*Built or Renovated (X10)
- Quality Grade (X7)*Built or Renovated (X10)

Forecasting Dependent Variable Values:
The figure below shows Prediction of initial Model 3 and Adjusted Model:

Using following given values of independent variables:
- Bathrooms (X2): 2
- Quality Grade (X7): 6
- Sqft Above (X8): 750
- Built or Renovated (X10): 2000

Figure 15. Prediction for Log(Y) of Adjusted Model3 with Additional independent variable

WORKSHEET 4

Prediction for Log(Price)-Adjusted Model3-2 Add

Regression Equation

Log(Price) = 10.46 + 0.0382 bathrooms - 1.71 Quality Grade + 0.00450 sqft_above
- 0.00277 Built or Renovated + 0.000894 Quality Grade*Built or Renovated
- 0.000002 sqft_above*Built or Renovated

Settings

Variable	Setting
bathrooms	2
Quality Grade	6
sqft_above	750
Built or Renovated	2000

Prediction

Fit	SE Fit	95% CI	95% PI
5.48972	0.0286446	(5.43263, 5.54680)	(5.30635, 5.67308)

Figure 16. Prediction for Log(Y) of initial Model3

WORKSHEET 4

Prediction for Log(Price)-Model3

Regression Equation

Log(Price) = 4.77 + 0.0266 bathrooms + 0.0668 Quality Grade + 0.000074 sqft_above
+ 0.000100 Built or Renovated

Settings

Variable	Setting
bathrooms	2
Quality Grade	6
sqft_above	750
Built or Renovated	2000

Prediction

Fit	SE Fit	95% CI	95% PI
5.48288	0.0296343	(5.42385, 5.54191)	(5.29261, 5.67315)

	x0	bhat
int	1	10.46
x1	2	0.0382
x2	6	-1.71
x3	750	0.0045
x4	2000	-0.00277
x5	12000	0.000894
x6	1500000	-0.000002

With this X0 and bhat from the regression equation:
And using variance of Log(price) = 0.007644634
We can conclude the median Price of 690875.831, and E[Price] of 705019.632.

Converting 95% Confidence interval for Adjusted Model is (5.43263, 5.54680) and 95% prediction Interval is (5.30635, 5.67308) to normal values:
(270788.365,352208.6356) and (202465.0197, 471064.0916)

However, our given values result seems not to be contained into 95% confidence interval and 95% prediction interval which means we have to reject the null hypothesis. This indicates our model 3 seems to fail to have 95% chance of estimated price of real estate. This may result from multicollinearity even if we decide to remove the highest correlation independent variable X3. This result is indicative that using correlation matrix to prediction would have significant impact on multicollinearity from other high correlation independent variables such as X2, X7 and X8.