PREDICTION OF SALES PRICE OF HOUSES USING THE HEDONIC PRICING MODEL USING STATA

INSTRUCTOR: PROF. MICHAEL PARZEN, TF: ERIK OTAROLA-CASTILLO

ABSTRACT

The following report seeks to explore potential explanations for variation in the sales prices of houses in Springfield by analyzing the statistical significance of attribute data. The examined relationship is in the form of a regression equation with house sales price as the dependent variable and characteristics of the house as explanatory variables. The effects of the characteristics variables were compared individually and collectively at the 5% significant level ensuring that we met the assumptions of multiple linear regression analysis. This, in turn, was used in a Hedonic Pricing Model to predict future sales prices of houses in Springfield. This model like any other model has shortcomings, but on a positive side and with a further investigation, provides insight on how best to estimate value and demand of properties. A basic knowledge of algebra and the understanding of statistical concepts are required to grasp the analysis in this paper.

INTRODUCTION

Recent financial sector woes and real estate meltdown have left many current and prospective real estate owners, and managers wonder how to correctly value properties and real estate assets. Government agencies in the US and other communities in the world charge various taxes to help support services rendered, one of which is property taxes. These fees are estimated based on the value of the property. The way the value is determined must be cost-effective and relatively accurate below which there is a loss in tax revenue, above which will cause unnecessary burdens on taxpayers. Companies and organizations spread across different geographic locations also need to know the relative worth of same properties in various geographic areas to enable budgeting for future endeavors. These are just a few examples why estimation of the value of properties is essential. The key here is value and demand. And, are measured in monetary terms - how much can you afford to pay or what price should a property be sold at, considering all factors. It is for that reason that we have different models to ascertain for such. One of such models is the Hedonic Pricing Model.

The Hedonic Pricing Model is a method of pricing based on the principle that the price of a marketed good is affected by particular external environment or perceptual factors that can raise or lower the base price of that good. This method is commonly applied to the housing market, where the cost of a house can be affected by factors such as scenic views, house appearance, and neighborhood demand amongst other factors. The hedonic pricing is used to estimate the extent that price and market can be affected by such factors, i.e., how much people are willing to pay for that property when considering these factors. This model involves a technique called the regression analysis, which measures the connection between two or more phenomena. These characteristics, however, provide no perfect prediction of the sale price of properties, nor can we infer causation from the relationship. Reason for that is that all characteristics affecting the cost of ownership may not be accounted for systematically. Another reason could be the subjective nature of valuation, meaning features that are perceived valuable for one may not be for someone else. As a result, the market value of a property outweighs the actual or intrinsic value. Understanding the inherent value of a real estate asset or property and the characteristics that contribute to its potential transaction price (market value) is imperative for proper valuation and can only be calculated by fastidious underwriting.

The goal of this project report is to examine real estate transaction in Springfield investigating the correlation between sale prices of houses (a response variable) and external factors and characteristics (the explanatory variables) using regression analysis technique in the hedonic pricing model and to predict future sale prices. This model will be a regression equation with house sales price as the dependent variable, and characteristics of the house as regressors.

MATERIAL AND METHODS

The variable names of the fictitious dataset used in this project are from the records of a real estate office in Springfield, home to the famous television family - The Simpsons, and is as shown below:

Variable	Description		Variable	Description
salesprice	**Selling price in Dollars**		**bedrooms**	**Number of bedrooms**
lot_sq_ft	**Size of lot in square feet**		**fireplaces**	**Number of fireplaces**
condition	**House condition on 1-10 scale (10 best)**		**garage_cars**	**Number of car garages**
bsmt_sf	**Square footage of basement**		**deck_sf**	**Square footage of deck**
cac	**1 if home has central air conditioning**		**month_sold**	**Month of year sold (1=January)**
live_area_sf	**Square footage of living space**		**age**	**Age of house in years**
baths	**Number of bathrooms**		**dist_fire**	**Distance from closest fire station**
smoker	**1 if previous owner was a smoker**			

Stata statistical software build version MP – Parallel Edition was the software used for the analysis. Good reasons to use Stata are that it has a broad suite of features in one package, fast and easy to use, complete documentation, cross-platform and widely used and supported.

In other for the data to be suitable for modeling and to produce valid results from the multiple linear regression analysis and predictions, the following assumptions are made:

1. The dependent variable should be measured at the continuous level
2. There should be two or more independent variables measured at the continuous level or a mix of a continuous and categorical level.
3. The data must show multicollinearity
4. There should be no significant outliers, high leverage points or highly influential points, which represent observations in the data set that are in some way unusual
5. There needs to be a linear relationship between the dependent variable and each of the independent variables
6. There also needs to be a linear relationship between the dependent variable and independent variables collectively.
7. The data as a whole must show homoscedasticity
8. The residuals should be approximately normally distributed
9. There should be independence of observations and residuals

The multiple linear regression model equation is shown below with Y being the dependent variable representing sale price of houses, and the X variables on the right are the independent variables. The $\beta_i(s)$ are the regression coefficients depicting the effects of each of independent variables on the response variable with all other things being equal and ε signifies noise in the system. β_o is the intercept, signifying the initial value of the dependent variable when the independent variable is zero.

$$\boldsymbol{Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_n X_n \pm \varepsilon = \hat{Y} \pm \varepsilon}$$

$$\textit{Where (yhat) } \boldsymbol{\hat{Y} = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_n X_n}$$

The dataset was investigated to ensure the assumptions associated with the above equation were met to validate prediction. Looking at the variables types provided, sales price show as the required continuous dependent variable and our independent variables have a mixture of both continuous and categorical variables meeting assumption 1 and 2. The variables types present are:

- Continuous: sales price, size of the lot in square feet, square footage of basement, square footage of living space, square footage of deck, age and distance from the closest fire station.
- Categorical: smoker, months sold, number bedrooms, number of bathrooms, central air condition, house condition, number of fireplaces and number of garages.

Further examination of the categorical variable shows that there are three types present: Nominal, Dichotomous and Ordinal. For this report, all categorical variables were converted into Dichotomous dummy variables of the form of 0 and 1. This conversion was done in Stata with the command: tabulate (variable name), generate (variable name)

Second, the data was investigated for multicollinearity between all the independent variables. The result of that is found in Appendix A. This is done first before any other data examination because if independent variables are highly correlated, then regression result

will be negatively impacted by high standard errors of the effect of not needing a variable when it is in fact required. Rule of thumb used is that if two independent variables have correlation > 0.80, then we pick the one that has a higher relationship with the dependent variable and drop the other. The Stata output with the dummy variables shows no significant correlation between the independent variables. So, assumption 3 passed!

Next, we examine the linear relationship assumption between the sales price variable and all independent predictive variables. Since there are so many explanatory variables and doing this for each of them will take a considerable amount of time, and we know the predictive variable is not highly correlated with each other, the chances of high standard errors are small. So it makes sense to remove insignificant predictive variables at the 5% significant level. We do this using the automatic backward stepwise regression in Stata as against the required manual backward stepwise regression for hypothesis testing of re-running the regression each time an insignificant predictive variable is removed. This also is to save time and reduce analysis on paper. This was done after every examination and transformation of the dataset. Results of that are found in Appendix B.

With the trimmed down dataset, it is now a good idea to examine the data closely to see if required variable were omitted, check for linear relationship between the dependent and independent variables and for constant variance of the distribution about the mean (heteroskedasticity) to satisfy assumptions 5, 6 and 7. This is carried out using the ovtest and hettest command in Stata. The ovtest and hettest outputs are shown in Appendix C, D and E respectively. Our dataset failed both tests at 5% significant level because their p-value was less than 0.05. So, we reject the null hypothesis and accept the alternative concluding that we need transformation of our variables. The easiest and time saving way of fixing that is to check scatter plots of the dependent vs each independent variable and use the rule of thumb for transformation of graph plots to transform the variables, then re-running both commands each time to check if our null hypothesis meets the 5% significant level. Another way is to check if each variable is skewed by looking at the histogram plots. Going first with the histogram plot, three variables were identified as being positively skewed. These were sales price, age, and lot_sq_ft. The rule of thumb for right skewed variable is logarithmic transformation.

Taking the logs of these variables and re-computing the histograms and tests for linearity and heteroskedasticity shows us passing the ovtest and hettest at the 5% significant level, meeting assumptions 5 and 7. As it was discussed in class, taking the logarithmic of the dependent variable should be a last resort because then you have issues comparing models due to change of units. The linear relationship between the log of the sales price and independent variables collectively was now investigated by computing the fitted line to our data (yhat) in Stata with the command: predict yhat, xb. A scatter plot of log of sales price and yhat showed a linear relationship satisfying assumption 6. Output for that is shown in result part of this report.

At this point, it is now safe to check for significant outliers, high leverage points or highly influential points. There are so many ways to do this. The easiest way to identify outliers is using the scatter plot of the standardized residue vs. the yhat and the Cook's Distance plot in Stata. Remember, we repeat all tests after adjusting for outliers and after every transformation, including re-running the regression, ovtest, hettest, sktest and a check for linearity.

The sktest checks if our residual is normally distributed satisfying assumption 8 for the regression analysis. That is the ε in the multiple linear regression model equation above. This was examined by computing the residual in Stata with the command: predict res, r; and then computing the standard residual using the command: generate sres=res/[Se]. The residual Se is gotten from our regression output. Investigating for normality of the residual was now carried out by running the command: sktest sres in Stata. Results of that are found in Appendix F. At this point, all of hettest, ovtest, and sktest passed the hypothesis test at the 5% level with a p-value less than 0.05. This means our data meets the assumption 5, 6, 7, 8.

To examine the dataset for outliers, we invoke another rule of thumb we used in class. That was any point outside ± 2 on the scatter plot of the standardized residue (sres) vs. the fitted value (yhat) is regarded an outlier, leverage points or influential point. Those points were removed with the command in Stata: drop if sres>2 and the drop if sres<-2. After dropping those points and re-checking the assumptions, everything else passed the null hypothesis test except the sktest of normality of residuals. This case means some of the outliers caused the skewness of the distribution of the noise in the system, leverage points, and influential points removed. The most straightforward fix for that was to chisel away on the Cook's Distance plot in Stata, with the command: predict D,cooksd and then the plot with the command: graph twoway spike D (each of the variables). This allowed a way to see individual points in each variable and figure out which of them is causing the skewness. The dataset was cross-examined by re-running all the assumption tests of hettest, ovtest, sktest and the regression after the removal of each of the unusual spikes in data. All said and done, the required assumptions are now met and we now ready to compute the multiple linear regression analysis.

RESULTS

The final regression model is as shown below and interpreted hypothetically as thus:

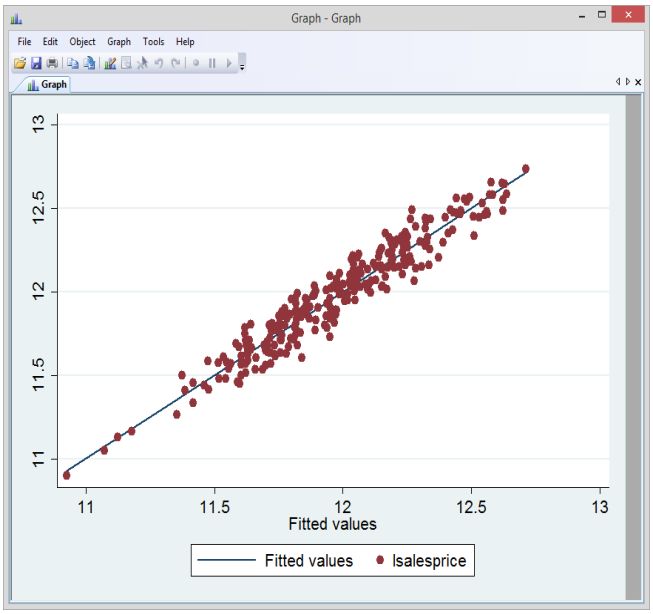
Source	SS	df	MS	Number of obs =	269
				F(16, 252) =	207.82
Model	26.6912543	16	1.6682034	Prob > F =	0.0000
Residual	2.02288452	252	.00802732	R-squared =	0.9296
				Adj R-squared =	0.9251
Total	28.7141389	268	.107142309	Root MSE =	.0896

lsalesprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
llot_sq_ft	.0773027	.0123627	6.25	0.000	.0529553	.10165
fireplaces3	.0766799	.0300799	2.55	0.011	.0174398	.13592
garage_cars4	.2190215	.0456914	4.79	0.000	.1290359	.3090071
condition5	.1412497	.0312327	4.52	0.000	.0797392	.2027601
condition6	.1434782	.0317411	4.52	0.000	.0809665	.2059899
condition7	.2471042	.0332944	7.42	0.000	.1815336	.3126749
condition8	.2643382	.0375496	7.04	0.000	.1903871	.3382893
condition9	.4379862	.0506181	8.65	0.000	.3382979	.5376746
bsmt_sf	.0001814	.0000169	10.70	0.000	.000148	.0002147
cac	.1781717	.0331399	5.38	0.000	.1129053	.2434381
live_area_sf	.0003582	.0000233	15.39	0.000	.0003123	.000404
garage_cars2	.0860496	.0405527	2.12	0.035	.0061842	.165915
garage_cars3	.1767929	.0397253	4.45	0.000	.098557	.2550288
fireplaces2	.0642765	.0125861	5.11	0.000	.0394893	.0890638
lage	-.1113828	.007701	-14.46	0.000	-.1265493	-.0962163
bedroooms4	-.0776635	.0263886	-2.94	0.004	-.1296338	-.0256931
_cons	10.4448	.120864	86.42	0.000	10.20677	10.68283

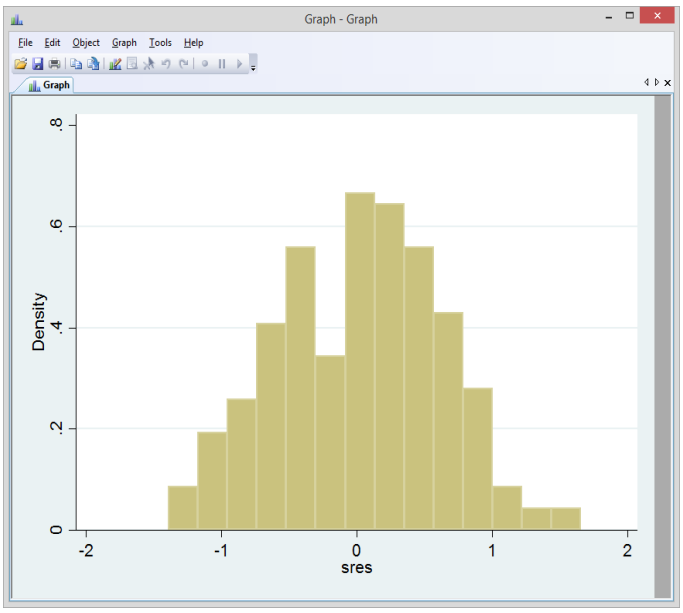
The Overall F test Prob>F of 0.0000 indicates that we accept the alternative hypothesis that we need at least an independent predictive variable in the model as against null hypothesis that all of the independent variables are not required. The coefficient of determination R^2 shows that about 93% of the variation of logarithmic of the sales price is explained collectively by all the dependent variables, with about 7% unaccounted for in the model. R^2 Adjusted, which shows a collective importance of each of the included variables in the model is very close to R^2, signifying we have less noise in the system. The residual shown (Root MSE) is far less than the residual of comparable models during our investigation. The values of R^2 and R^2 Adjusted in this model are also higher than models examined during our analysis. All predictive variables shown above are significant at the 5% significant level with different significance and degrees of effect on the overall model. This model contains dummy Indicator variables with dichotomous on and off impact on the log of the sale price such that there is a different effect of its coefficients depending on when they are off or on. The default for the log of the sale's price is the intercept (log of $10.44) which is common to all indicator variables, and changes based on the state of the interactive variable.

The bounded final multiple linear regression model at the 95% confidence level is:

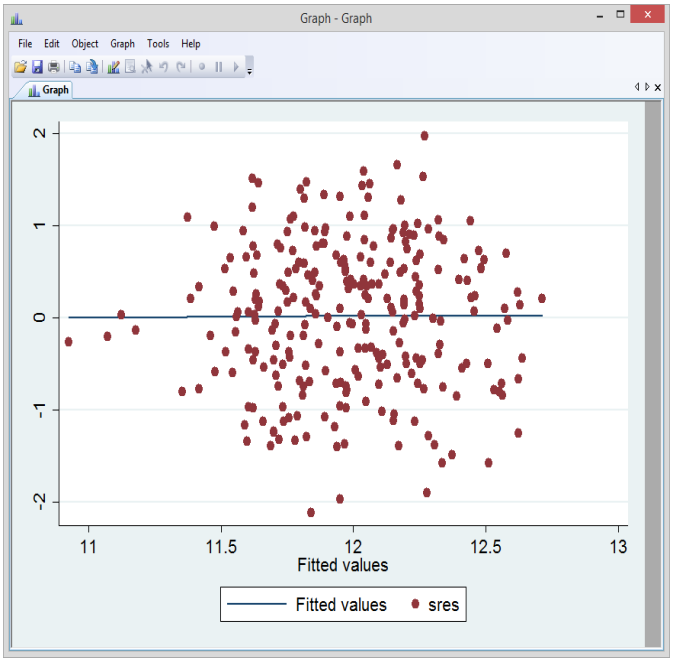
Log of Sales Price(Dollars) = 10.44 + 0.0773(Log of Lot size in Square Ft) + 0.0767(3 Fireplaces) + 0.219(4 Car Garages) + 0.141(House with condition 5) + 0.143((House with condition 6) + 0.247(House with condition 7) 0.264(House with condition 8) + 0.438(House with condition 9) + 0.000181(Basement in Square Ft.) 0.178(Air Conditioner in the House) + 0.000358(Living Area in Square Ft.) + 0.0860(2 Car Garages) + 0.177(3 Car Garages) + 0.0643(2 Fireplaces) – 0.111(Log of Age of House in Years) – 0.0777(4 Bedrooms in the House) ± 1.96(0.0896)



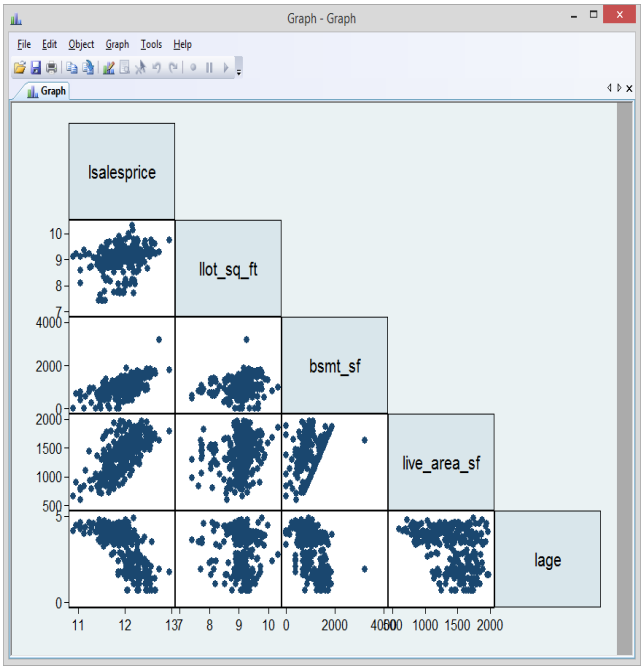
Scatter plot of Log of Sales Price vs Fitted Value showing a linear relationship



Histogram of the Standard Residual showing a normal distribution



Scatter plot of Standard Residual vs Fitted Value showing no linear relationship



Scatter plots after transformation of continuous variables

CONCLUSION

The final multiple linear regression analysis provides information about the relationships between the log of the sales price and all constituent predictive variable. Each of the included characteristics has a different effect on the log of the sales price. The model shows people in Springfield care more about the size of basements and living areas and the age of the house than any other factors. The bigger the size of the living area and the basement in square feet, the higher the log of the sales price in dollars and vice versa. The reverse is the case for the age of the house. The older the house in years, the lesser the log of the sales price in dollars you would get for it. Second, the more the amenities associated with the house - number of fireplaces, garages, and presence of air-conditioned unit, the more premiums you get for the house. The condition of the house is insignificant to the sales price up to condition 5, where an increase in that value provides a much better premium. These associated effects somewhat make econometric sense.

Other interesting findings of the model are the relevance of the month house was sold in and the number of bedrooms. Although significant, may have some lurking factors not considered in this model. The model shows the premium you get for a bedroom remain the same till you get a four-bedroom house where the log sales price in dollars drops. A possible explanation for that could be that people think they would incur more utility expenses or more maintenance with more than three bedrooms. The month a house was sold in has varying effects depending on how you tweak your dataset.

Both cases could be subjective and need further investigation, which shows that no model is perfect. There are always rooms for improvement. The affects you get vary depending on variables you have and how you manipulate them. It may be impossible to include all possible factors affecting sales prices of a house in this model. However, bounding the model with twice the residue accounted for most in inaccuracies. Relationships established here are barely associations and does mean causation but with this model we have insight has to how various characteristics affect the variation in the sales price of a house.

APPENDIX A

Multicollinearity

```
. corr lot_sq_ft condition bsmt_sf cac live_area_sf baths bedroooms fireplaces garage_cars deck_sf month_sold age dist_fire smoker
(obs=300)
```

	lot_sq~t	condit~n	bsmt_sf	cac	live_a~f	baths	bedroo~s	firepl~s	garage~s	deck_sf	month_~d	age	dist_f~e	
lot_sq_ft	1.0000													
condition	-0.0352	1.0000												
bsmt_sf	0.2263	-0.1549	1.0000											
cac	0.0521	0.2579	0.2718	1.0000										
live_area_sf	0.2649	-0.0552	0.3567	0.2591	1.0000									
baths	0.0409	-0.1463	0.0163	0.1801	0.5969	1.0000								
bedroooms	0.2249	0.1078	-0.0020	0.1371	0.4613	0.2785	1.0000							
fireplaces	0.1948	-0.1214	0.2946	0.1433	0.4098	0.2382	0.1059	1.0000						
garage_cars	0.2051	-0.1463	0.4564	0.2648	0.4231	0.3485	0.0828	0.2797	1.0000					
deck_sf	0.1141	0.0093	0.2237	0.0870	0.2466	0.1955	0.0027	0.2775	0.1993	1.0000				
month_sold	-0.0062	0.0898	0.0359	0.0957	-0.0703	-0.1443	0.0303	0.0163	-0.0354	0.0242	1.0000			
age	-0.0676	0.3555	-0.4224	-0.2942	-0.3106	-0.5535	-0.0016	-0.2695	-0.5116	-0.3090	0.1077	1.0000		
dist_fire	-0.0643	-0.1581	-0.0227	-0.0820	-0.0590	0.0653	-0.1595	0.0702	-0.0155	0.0301	-0.0306	-0.1051	1.0000	
smoker	0.0324	0.0761	0.0039	-0.0465	0.0066	-0.1026	-0.0258	0.0369	0.0155	-0.0740	0.0536	0.1586	0.0091	

	smoker
smoker	1.0000

This was done to ensure no significant collinearity exist between dependent variables. Rule of thumb is if greater than 0.8, we select the variable with the stronger relationship with the response variable and drop the other. Repeating this test after transformation, and the creation of dummy variables shows the same effect.

APPENDIX B

The multiple linear regression was done automatically in Stata using the backward stepwise regression as against re-running the regression each time after manually removing insignificant predictive variables for hypothesis testing.

Source	SS	df	MS	Number of obs =	300
				F(16, 283) =	116.49
Model	8.4082e+11	16	5.2552e+10	Prob > F =	0.0000
Residual	1.2767e+11	283	451137431	R-squared =	0.8682
				Adj R-squared =	0.8607
Total	9.6850e+11	299	3.2391e+09	Root MSE =	21240

salesprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lot_sq_ft	1.500264	.3589061	4.18	0.000	.7937992	2.206728
fireplaces2	7427.887	2842.061	2.61	0.009	1833.624	13022.15
garage_cars4	57368.78	6041.062	9.50	0.000	45477.67	69259.9
month_sold4	9790.143	4697.714	2.08	0.038	543.2477	19037.04
condition5	16014.78	5776.933	2.77	0.006	4643.571	27385.99
condition6	13190.25	5952.626	2.22	0.027	1473.211	24907.3
condition7	30523.93	6184.31	4.94	0.000	18350.84	42697.01
condition8	40788.21	7170.22	5.69	0.000	26674.47	54901.94
condition9	78982.96	10364.97	7.62	0.000	58580.75	99385.18
bsmt_sf	31.77048	3.741299	8.49	0.000	24.40618	39.13479
garage_cars3	11470.97	3432.532	3.34	0.001	4714.44	18227.51
live_area_sf	60.40916	5.384614	11.22	0.000	49.81018	71.00813
bedroooms3	-7548.982	3087.117	-2.45	0.015	-13625.61	-1472.357
age	-633.9504	62.16939	-10.20	0.000	-756.3235	-511.5772
fireplaces3	17065.86	6146.294	2.78	0.006	4967.602	29164.11
bedroooms4	-28482.17	5985.255	-4.76	0.000	-40263.44	-16700.91
_cons	32981.43	9044.739	3.65	0.000	15177.93	50784.93

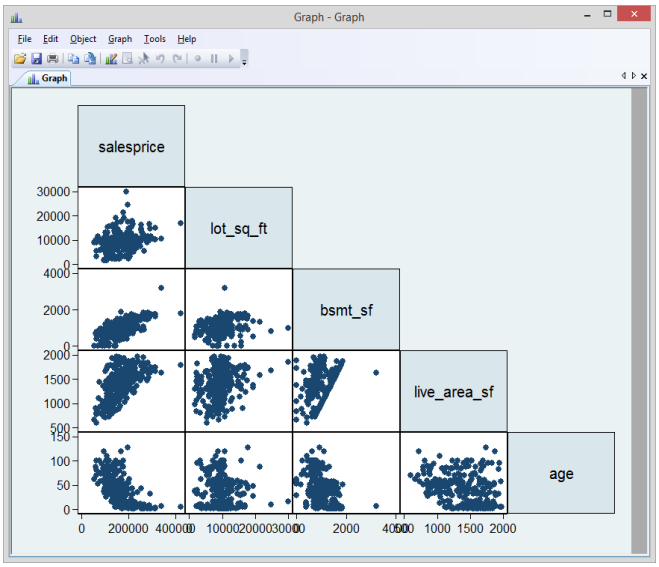
Initial multiple linear regression of the variables before transformation.

Source	SS	df	MS	Number of obs =	300
				F(18, 281) =	143.04
Model	32.0826925	18	1.7823718	Prob > F =	0.0000
Residual	3.50144082	281	.012460643	R-squared =	0.9016
				Adj R-squared =	0.8953
Total	35.5841333	299	.119010479	Root MSE =	.11163

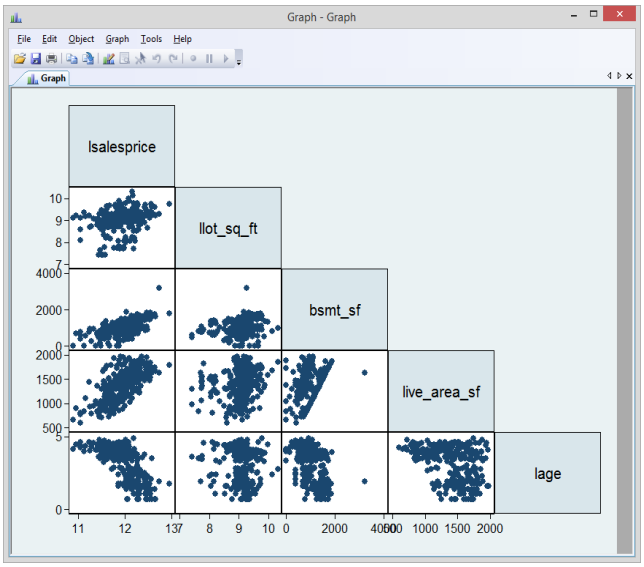
lsalesprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
llot_sq_ft	.0815873	.0144798	5.63	0.000	.0530846	.11009
garage_cars3	.1925754	.0360978	5.33	0.000	.121519	.2636318
month_sold5	-.0468483	.0223791	-2.09	0.037	-.0909003	-.0027963
garage_cars5	.3211357	.117777	2.73	0.007	.0892984	.5529729
condition5	.1473644	.0323819	4.55	0.000	.0836226	.2111063
condition6	.1632335	.0329867	4.95	0.000	.0983011	.228166
condition7	.2542366	.0352423	7.21	0.000	.1848643	.323609
condition8	.2908048	.0408185	7.12	0.000	.210456	.3711535
condition9	.4749565	.0557844	8.51	0.000	.3651481	.5847648
bsmt_sf	.0001784	.0000199	8.97	0.000	.0001393	.0002176
cac	.173739	.0354316	4.90	0.000	.103994	.243484
live_area_sf	.0003278	.0000271	12.09	0.000	.0002744	.0003811
garage_cars4	.2478743	.045765	5.42	0.000	.1577885	.3379602
fireplaces2	.0616352	.0148917	4.14	0.000	.0323217	.0909487
lage	-.115236	.0090941	-12.67	0.000	-.1331372	-.0973348
garage_cars2	.0971782	.0364041	2.67	0.008	.0255189	.1688375
fireplaces3	.1004884	.0322019	3.12	0.002	.0371009	.1638759
bedroooms4	-.0825178	.0289895	-2.85	0.005	-.1395819	-.0254536
_cons	10.44411	.1360559	76.76	0.000	10.1763	10.71193

Multiple linear regression after transformation of the variables

APPENDIX C



Scatter plots before transformation of variables



Scatter plots after transformation of continuous variables

APPENDIX D

Ho: No transformation of the predictive variable needed
Ha: Transformation of the predictive variable needed.

Decision rule states that we reject the Ho and accept Ha if the p-value is below 5% significant level

```
. ovtest

Ramsey RESET test using powers of the fitted values of salesprice
       Ho:  model has no omitted variables
                 F(3, 280) =     13.19
                  Prob > F =      0.0000
```

Before transforming the salesprice response variable, ovtest failed at 5% significant level

```
. ovtest

Ramsey RESET test using powers of the fitted values of lsalesprice
       Ho:  model has no omitted variables
                 F(3, 278) =      0.09
                  Prob > F =      0.9637
```

Transformation of salesprice, age and lot_sq_ft variables, ovtest shows we do not need any more transformation of the predictive variables

```
. ovtest

Ramsey RESET test using powers of the fitted values of lsalesprice
       Ho:  model has no omitted variables
                 F(3, 249) =      0.86
                  Prob > F =      0.4608
```

This is the final result for the ovtest after chiseling outliers out with Cook's D plot. Hence at 5% significance level, we fail to reject Ho concluding there is evidence that we do not need transformation of the predictive variables.

APPENDIX E

Ho: Homoscedasticity
Ha: Heteroskedasticity

Decision rule states that we reject the Ho and accept Ha if the p-value is below 5% significant level

```
. hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of salesprice

         chi2(1)      =    34.29
         Prob > chi2  =   0.0000
```

Before transforming the salesprice, age and lot_sq_ft variables, our dataset failed the homoscedasticity test at 5% significant level

```
. hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of lsalesprice

         chi2(1)      =     0.24
         Prob > chi2  =   0.6212
```

After the transformation of salesprice, age and lot_sq_ft variables, hettest shows we now have homoscedasticity at the 5% significant level

```
. hist sres
(bin=16, start=-2.1234627, width=.25617725)

. hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of lsalesprice

         chi2(1)      =     0.03
         Prob > chi2  =   0.8534
```

This is the final result for the hettest after chiseling outliers out with Cook's D plot. Hence at 5% significance level, we fail to reject Ho concluding that there is evidence that we have constant variance or homoscedasticity in the system

APPENDIX F

Ho: Normality of the residual distribution
Ha: Residual not normal

Decision rule states that we reject the Ho and accept Ha if the p-value is below 5% significant level

```
. sktest sres

                    Skewness/Kurtosis tests for Normality
                                                              ------- joint ------
    Variable |   Obs   Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+-------------------------------------------------------------------
        sres |   300      0.0306         0.0000        20.77         0.0000
```

Before transforming the salesprice, age and lot_sq_ft variables, sktest failed the normality test at 5% significant level

```
. generate sres=res/[0.11163]

. scatter sres yhat

. sktest sres

                    Skewness/Kurtosis tests for Normality
                                                              ------- joint ------
    Variable |   Obs   Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+-------------------------------------------------------------------
        sres |   300      0.2650         0.1876         3.00         0.2236
```

After transformation of salesprice, age and lot_sq_ft variables, sktest passed beautifully.

```
. sktest sres

                    Skewness/Kurtosis tests for Normality
                                                              ------- joint ------
    Variable |   Obs   Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+-------------------------------------------------------------------
        sres |   269      0.2675         0.0456         5.26         0.0721
```

After playing around with dataset trying to make R^2 Adj go up with a better fit and chisel off some outliers off with Cook's D my sktest dropped drastically. This is the final result of the sktest. Hence at 5% significance level, we fail to reject Ho concluding there is evidence that our residual is normally distributed.