MODEL FOR FORECASTING PRICE

OF HOUSES IN CITY OF

STILLWATER, OK.

By

YOGESH SINGH
Bachelor of Science in Statistics
M.S. University of Baroda
Gujarat, India
2002

Master of Science in Statistics
South Gujarat University
Gujarat, India
2004

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of

MASTER OF SCIENCE
May, 2006

MODEL FOR FORECASTING PRICE

OF HOUSES IN CITY OF

STILLWATER, OK.

Thesis Approved:

Dr. William D. Warde

Thesis Advisor

Dr. P. Larry Claypool

Dr. Carla Goad

A. Gordon Emslie

Dean of the Graduate College

ACKNOWLEDGEMENT

I am grateful to my advisor, Dr. William D. Warde for the assistance and guidance rendered to me in writing this report. Although he had a busy schedule, Dr. Warde was always ready to help and answer my questions. I would also like to thank Dr. P. L. Claypool and Dr. Carla Goad for serving on my committee.

I also would like to thank Raleigh Jobes for providing me the data set of selling prices of houses used in the analysis.

I thank other faculty members who helped me throughout my study at Oklahoma State University. I also thank Fran Mihura, Orlando Carrisalez and my fellow graduate students for their friendship and assistance. Furthermore, I appreciate the Department of Statistics for providing TA support during these two years of study.

TABLE OF CONTENTS

Chapter Page

I. INTRODUCTION....1

II. DEFINITION AND ASSUMPTIONS....5
III. STATISTICAL ANALYSIS....7

Step 1 Preliminary Data Analysis....7
Step 2 Model Adequacy for a predictor variable using Added Variable plot....12
Step 3 Variable Selection procedure Using All Possible Regression....20
Step 4 Addition of Quadratic and Interaction terms....23
Step 5 Checking the assumption of Linear association....25
Step 6 Checking the assumption of residuals with equal variances....32
Step 7 Checking the assumption of no Autocorrelation....46
Step 8 Checking the assumption of no Multicollinearity....46
Step 9 Assumption of normally distributed residuals....50
Step 10 Detection of Influential Observations....51
Step 11 Examination of Influential Observations....53
STEP 12 Validation of the Model....56

IV CONCLUSIONS....61

BIBLIOGRAPHY....65

APPENDIC — SAS PROGRAM....66

LIST OF TABLES

Table Page
1: Indicator matrix.....5

2: Correlation Coefficients for the variables.....11

3: Output of All Possible Regression on the Independent variables.....22

4: Output of All Possible Regression on the Independent variables and interaction and squared terms.....24

5: Output of the first Box-Cox transformations on the variable Selling Price.....27

6: Output of the second Box-Cox transformations on the variable Selling Price.....29

7: Output of the final Box-Cox transformations on the variable Selling Price.....31

8: Output of Breusch Pagan Test.....45

9: Condition Indices for the variables.....49

10: R Square obtained from the Least Square Fit.....58

11: Predicted error sum of squares for the validation data set.....58

12: Total sum of squares for the validation data set.....59

13: Output for the Average Percent Discrepancy.....59

14: Parameter Estimates for the variables included in the final model.....60

LIST OF FIGURES

Figure Page

1: Scatter Plot of Variables.....9

2: Normal quantile plot of the variable Area in Square Feet.....10

3: Normal quantile plot of the variable Selling Price.....10

4: Added Variable Plot of Selling price vs. Area in Square Feet.....14

5: Added Variable Plot of Selling price vs. Days on the Market.....15

6: Added Variable Plot of Selling price vs. Month.....16

7: Added Variable Plot of Selling price vs. X1.....17

8: Added Variable Plot of Selling price vs. X2.....18

9: Added Variable Plot of Selling price vs. X3.....19

10: Residual Plot for the variable Selling Price (untransformed).....26

11: Residual Plot for Predicted Values of the variable Selling Price (transformed).....34

12: Residual Plot for the variable Selling Price (transformed).....34

13: Residual Plot for variable Area in Square Feet.....35

14: Residual Plot for variable Days on the Market.....35

15: Residual Plot for variable Month.....36

16: Residual Plot for variable X1.....36

17: Residual Plot for variable X2.....37

18: Residual Plot for variable X3.....37

19: Residual Plot for variable SqFtX1.....38

20: Residual Plot for variable SqFtX2.....38

21: Residual Plot for variable SqFtX3.....39

22: Residual Plot for variable SqFtmonth.....39

23: Residual Plot for variable SqFtDOM.....40

24: Residual Plot for variable X1DOM.....40

25: Residual Plot for variable X2DOM.....41

26: Residual Plot for variable X3DOM.....41

27: Residual Plot for variable SqFt2.....42

28: Residual Plot for variable DOM2.....42

29: Residual Plot for variable month2.....43

30: Normal quantile plot of the residuals for the final model.....50

31: Scatter Plot of actual values of the variable Selling Price
vs. the predicted values for the validation data set.....57

LIST OF SYMBOLS FOR INTERACTION AND QUADRATIC TERMS CONSIDERED IN THE MODEL

1: SqFt – Area in Square Feet

2: DOM – Number of Days on the Market

3: SqFtmonth – Variable representing interaction between variables Area in Square Feet and Month

4: SqFtDOM – Variable representing interaction between variables Area in Square Feet and Days on Market

5: SqFtX1 – Variable representing interaction between variables Area in Square Feet and X1

6: SqFtX2 – Variable representing interaction between variables Area in Square Feet and X2

7: SqFtX3 – Variable representing interaction between variables Area in Square Feet and X3

8: monthDOM – Variable representing interaction between variables Month and Days on the Market

10: X1month - Variable representing interaction between variables X1 and Month

11: X2month - Variable representing interaction betweenvariables X2 and Month

11: X3month - Variable representing interaction between variables X3 and Month

12: X1DOM - Variable representing interaction between variables X1 and Days on the Market

13: X2DOM - Variable representing interaction between X2 and Days on the Market

14: X3DOM - Variable representing interaction betweenX3 and DOM

15: X1X2 - Variable representing interaction between variables X1 and X2

16: X1X3 - Variable representing interaction between variables X1 and X

17: X2X3 - Variable representing interaction between variables X2 and X3

18: SqFt2 - Variable representing quadratic effect of variable Area in Square Feet

19: month2 - Variable representing quadratic effect of variable Month

20: DOM2 - Variable representing quadratic effect of variable Days on the Market

CHAPTER I

INTRODUCTION

The housing market is one of the most important components of the US Economy and hence is the subject of a lot of studies. This report represents a small effort to understand the effect of the factors that affect the price of houses. The objective was to fit a multiple regression model which satisfies all the assumptions required for a multiple linear regression model, is simple to understand, easy to use and predicts the price of houses with a reasonable degree of accuracy. This report will explain the steps taken and the assumptions that were made in order to analyze the data.

The given real estate data has been collected from the City of Stillwater, Oklahoma from various local real estate agents and municipal corporation by Dr. Raleigh Jobes. The data has been collected over a period of 16 years from 1988 to 2004. From the time period 1988 to 2000, the data consists of only information about the number of houses sold in a particular month. For the time period 2000 to 2004, the data consists of 21,058 observations and is given in terms of following variable

> address of the house

> selling price of the house

- original price of the house
- number of bedrooms in the house
- area of the house in terms of square feet
- asking price of the house
- location of the house
- number of days on the market before a particular house was sold
- dollars per square feet
- The year in which the house was built

The data for the first time period i.e. from 1988 to 2000 can be treated as a time series problem in which the effect of successive years from 1988 to 2000 on the number of houses sold can be studied. The data from this time period is not part of the analysis. This report is based on the data collected from the 2000 to 2004 as data from this time period is more extensive in terms of number of variables and number of observations (21,058).

From the analysis some of the variables like 'asking price', 'original price' etc were excluded for the following reasons.

Information about the year in which the houses were built, was available for only 2 years. Hence this information could not be included in the analysis.

The reason for not considering the variable 'number of bedrooms' is that although it affects the selling price of a house, the actual effect depends on the size of the bedroom. For Instance, two houses having the same number of bedrooms may have different price if the size of the bedrooms are different.

The second variable that is not considered is the 'asking price'. The problem with this variable is based on expectation of the homeowner which may or may not be realistic. In this project, only those variables are considered whose effect on the selling price can be economically explained and additionally asking price may not be considered as a parameter by a real estate agent for assessing the value of a house.

The third variable that was not considered is 'dollars per square feet'. Technically, 'dollars per square foot' is a rate that serves as a base for assessing a property. But in this data set, the variable 'dollar per square foot' has been simply calculated by dividing the selling price with the variable 'area in square feet', hence this variable is not the variable that is used by the real estate agent for assessing the property. The variable 'dollars per square foot' considered by real estate agents is a measure of prevailing market rates in a particular location.

The variable 'original price' was not included in the analysis as a real estate agent may not consider it as a factor for assessing the price of property because a real estate agent assesses the property based on the underlying fundamentals like the area and condition of the house rather than what the owner paid for the property. Also there is an extremely high correlation between 'original price' and 'area in square feet' approximately 0.889. Including these two variables in the model will make the model unstable. Hence for all these reasons, the variable 'original price' is not included in the analysis.

In order to use the information related to the time period, a new variable called 'month' was created, which takes value 1 for January 2000, 2 for February 2000 and so on.

Three dummy variables X_1, X_2 and X_3 were created in order to use the information related to the location of houses. They are defined as

X_1= 1 if location of house is northeast otherwise 0

X_2= 1 if location is house is northwest otherwise 0

X_2= 1 if location is house is southeast otherwise 0

If all the three variables take value 0 then location of house is southwest.

	Northeast	Northwest	Southeast	Southwest
X1	1	0	0	0
X2	0	1	0	0
X3	0	0	1	0

Table 1 Indicator matrix

CHAPTER II

DEFINITION AND ASSUMPTIONS

Multiple linear regression is a technique to model the relationship between two or more independent variables and a dependent variable by fitting a linear equation to observed data. Every value of the independent variable **X** is associated with a value of the dependent variable **Y**.

A multiple linear regression model with p explanatory variables and n observations is given by the following equation.

$$Y = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots\ldots\ldots\ldots\ldots\ldots + \beta_p x_{i,p} + \varepsilon_i \qquad ,i=1,\ldots\ldots\ldots\ldots,n \qquad \mathbf{1}$$

where,

Y is the dependent variable

β's are the unknown parameter

X's are the independent variables

ε is a random error term

The basic assumptions required for a multiple linear regression are

> Relationship between the dependent and the independent variables is linear in parameters.

> The error terms are distributed with equal variance.

> The error terms are uncorrelated.

> The expectation of the error terms is zero.

> There is no serious multicollinearity among the independent variables.

> There is no outlier distortion i.e. the equation is not unduly affected by outlying observations.

> The error terms are normally distributed

For this analysis it is also assumed that various economic factors like interest rates, mortgage rates are stable. Since the data represents housing activity for only four years, any change in the overall economy will have a profound impact on the model, making it unfit for new data

CHAPTER III

STATISTICAL ANALYSIS

In order to analyze the data given from 2000 to 2004, the variables included in the analysis are Selling Price, Area in Square feet, Days on the Market, Month, X_1, X_2 and X_3. The total number of observations in the data set is 21,058.

STEP 1 - PRELIMINARY DATA ANALYSIS

Scatter plots were used for identifying possible relationships between the different variables. The scatter plot shown in figure 1 has slightly linear systematic pattern for the scatter plot of variable 'selling price and 'area in square feet'. A similar trend is observed in the scatter plot for variables 'days on the market' and 'area in square Feet'. For the rest of the variables there is no systematic pattern which means that the variables are possibly not linearly related. This information is also confirmed by the correlation coefficients given in table 2, the maximum of which is 0.228 between the variables 'area in square feet' and 'days on the market'. These coefficients indicate that there is no serious pairwise linear association between the independent variables under consideration.

From the normal quantile plots of ‘selling price’ and ’area in square feet’ shown in figures 2 and 3, it is clear that the variables are not normally distributed. This indicates that the final model may have the problem of non-normality.

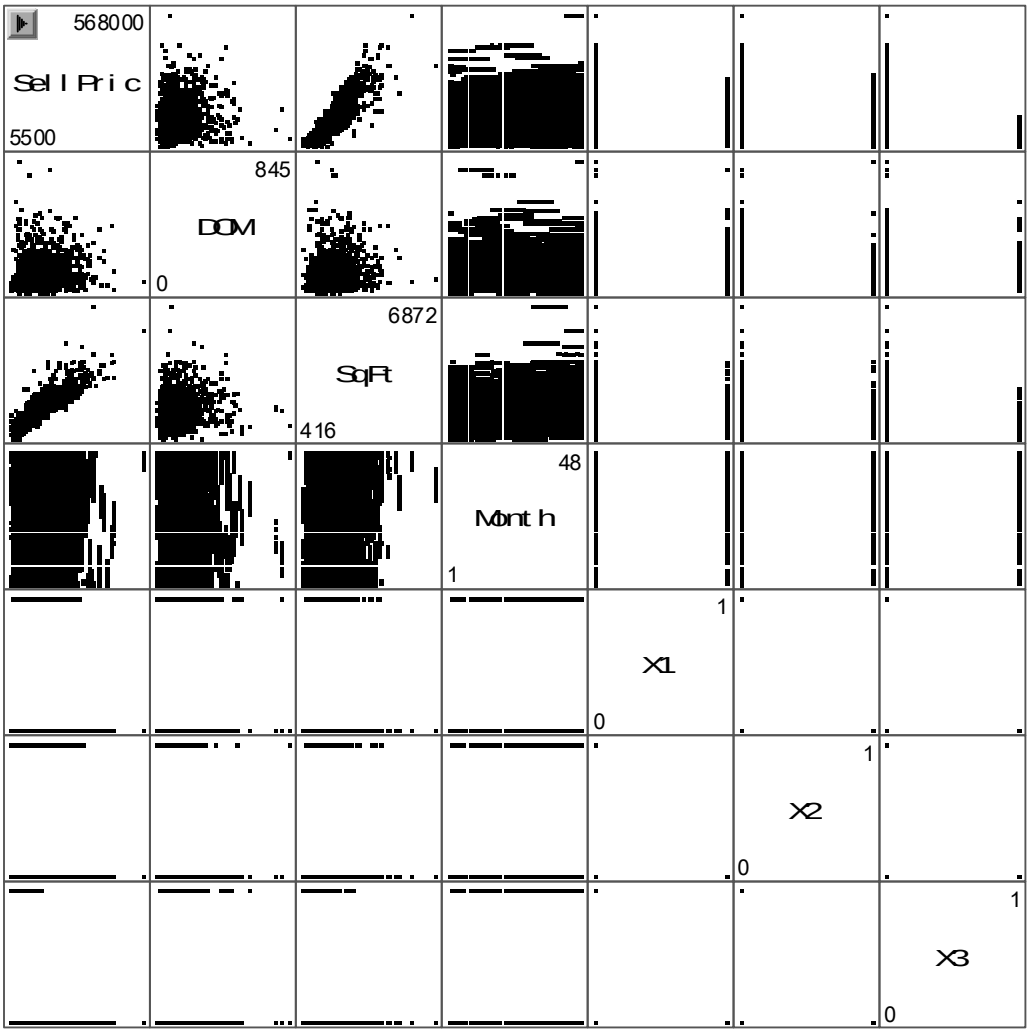

Figure 1 Scatter Plot of variables considered in the analysis

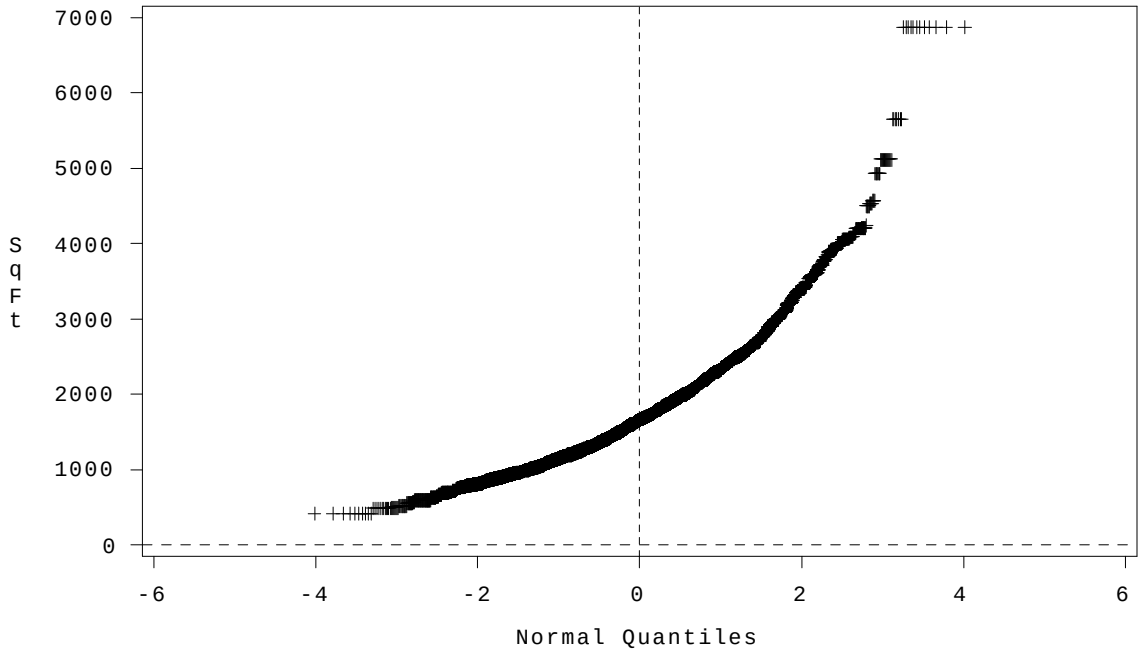

Figure 2 Normal qq plot of the Variable Area in Square Feet

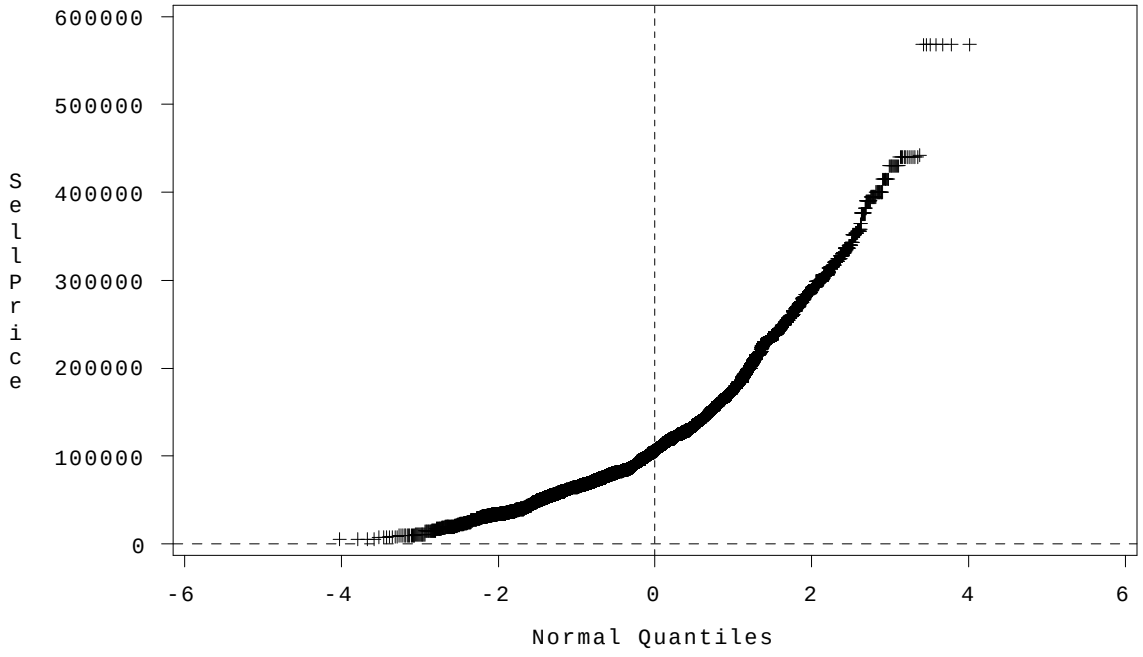

Figure 3 Normal qq plot of the Variable Selling Price

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations								
	SellPrice	**OriginalPrice**	**SqFt**	**DOM**	**Month**	**X1**	**X2**	**X3**
SellPrice SellPrice	1.000 21058	0.991 <.0001 21058	0.877 <.0001 21058	0.221 <.0001 21047	0.083 <.0001 21058	-0.245 <.0001 21058	-0.070 <.0001 21058	-0.242 <.0001 21058
OriginalPrice OriginalPrice	0.991 <.0001 21058	1.000 21058	0.889 <.0001 21058	0.253 <.0001 21047	0.073 <.0001 21058	-0.248 <.0001 21058	-0.066 <.0001 21058	-0.235 <.0001 21058
SqFt SqFt	0.877 <.0001 21058	0.889 <.0001 21058	1.000 21058	0.228 <.0001 21047	0.011 0.0936 21058	-0.251 <.0001 21058	-0.058 <.0001 21058	-0.208 <.0001 21058
DOM DOM	0.221 <.0001 21047	0.253 <.0001 21047	0.228 <.0001 21047	1.000 21047	-0.033 <.0001 21047	-0.061 <.0001 21047	-0.084 <.0001 21047	0.034 <.0001 21047
Month Month	0.083 <.0001 21058	0.0734 <.0001 21058	0.011 0.0936 21058	-0.033 <.0001 21047	1.000 21058	0.079 <.0001 21058	0.011 0.0828 21058	0.018 0.0064 21058
X1	-0.2450 <.0001 21058	-0.248 <.0001 21058	-0.251 <.0001 21058	-0.061 <.0001 21047	0.079 <.0001 21058	1.000 21058	-0.269 <.0001 21058	-0.177 <.0001 21058
X2	-0.070 <.0001 21058	-0.066 <.0001 21058	-0.058 <.0001 21058	-0.084 <.0001 21047	0.011 0.0828 21058	-0.269 <.0001 21058	1.000 21058	-0.086 <.0001 21058
X3	-0.242 <.0001 21058	-0.235 <.0001 21058	-0.208 <.0001 21058	0.034 <.0001 21047	0.018 0.0064 21058	-0.177 <.0001 21058	-0.086 <.0001 21058	1.000 21058

Table 2 Correlation Coefficients between the Variables and observed significance levels.

STEP 2 MODEL ADEQUACY FOR A PREDICTOR VARIABLE USING ADDED VARIABLE PLOT

As discussed by Chatterjee, Hadi and Price (1999), a limitation of residual plots is that they don't show the nature of marginal effect of an independent variable, given the other variables in the model. Added variable plots can be useful in identifying such marginal effects of an independent variable.

Added variable plots are constructed by using residuals obtained by regressing the dependent variable, Y and independent variable, X_k against the other independent variables (X_1,....,X_{k-1}) in the model. These residuals represent the part of each variable i.e. Y and X_k that is not linearly associated with other independent variables in the model. The plot of these residuals against each other

> Shows the marginal importance of the particular independent variable in reducing the residual variability.

> Provides information about the nature of the marginal regression relation for the independent variable under consideration for possible inclusion in the model.

The added variable plot for the variable 'area in square feet' shown in figure 4 clearly shows that linear term of the variable should be included in the model given the variables 'days on the market', 'area in square feet', X1, X2 and X3. For all the other added variable plots shown in figures 5 to 9, the residuals are distributed in a pattern which is somewhat circular and horizontal. This may mean that given the all the other independent variables, the independent variable under consideration does not provide additional information in explaining the variation in the dependent

variable ‘selling price’. These independent variables are not excluded the variables because the distribution of the residuals in the plots for these independent variables is not perfectly horizontal, which indicates that independent variable under consideration, has no effect given all the other independent variables.

Sometimes added variable plots do not show the proper form of the marginal effect of an independent variable if the functional relations for some or all of the independent variables already in the model are not properly specified.

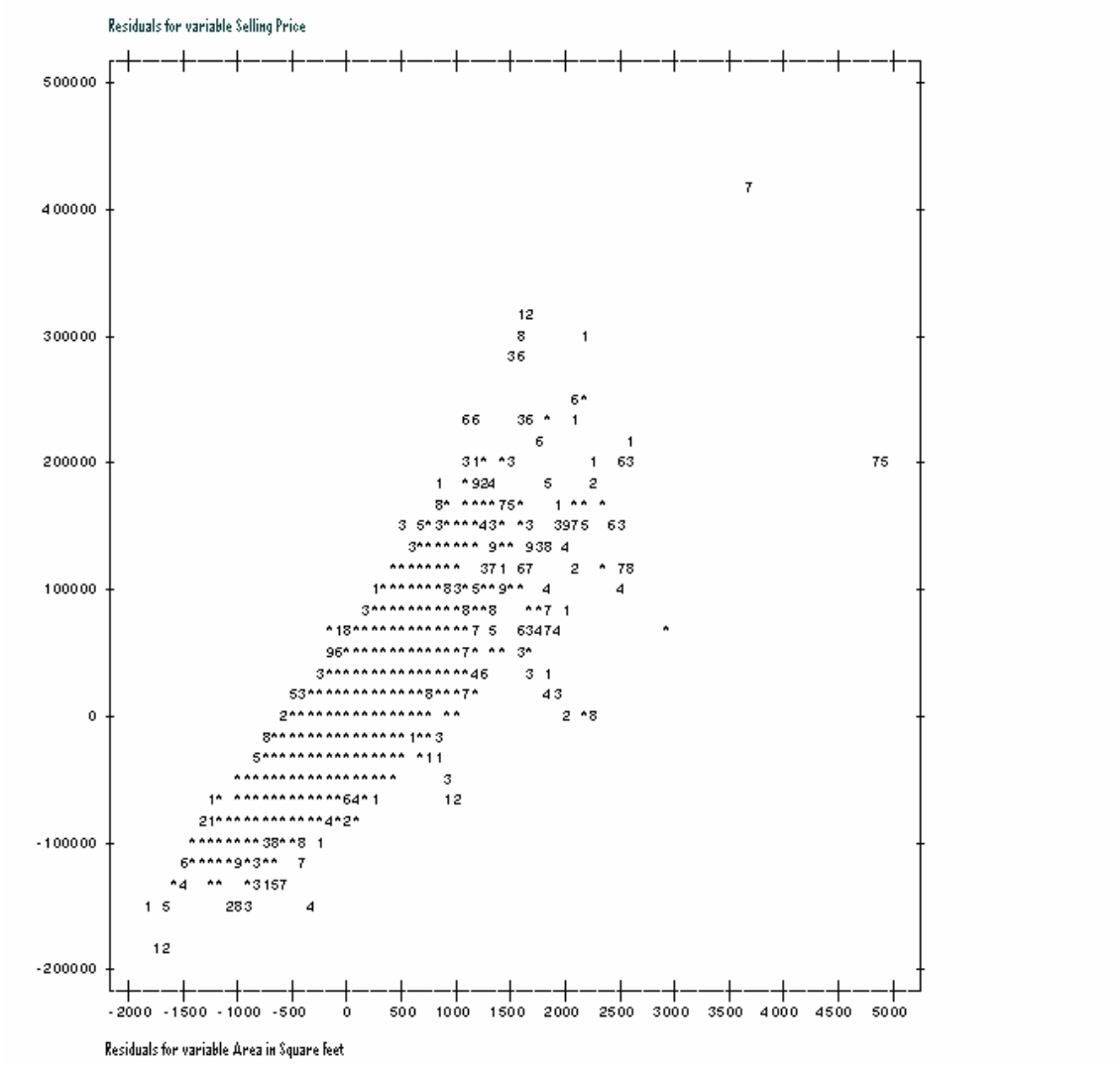

Figure 4 Added Variable Plot of Selling price vs. Area in Square Feet

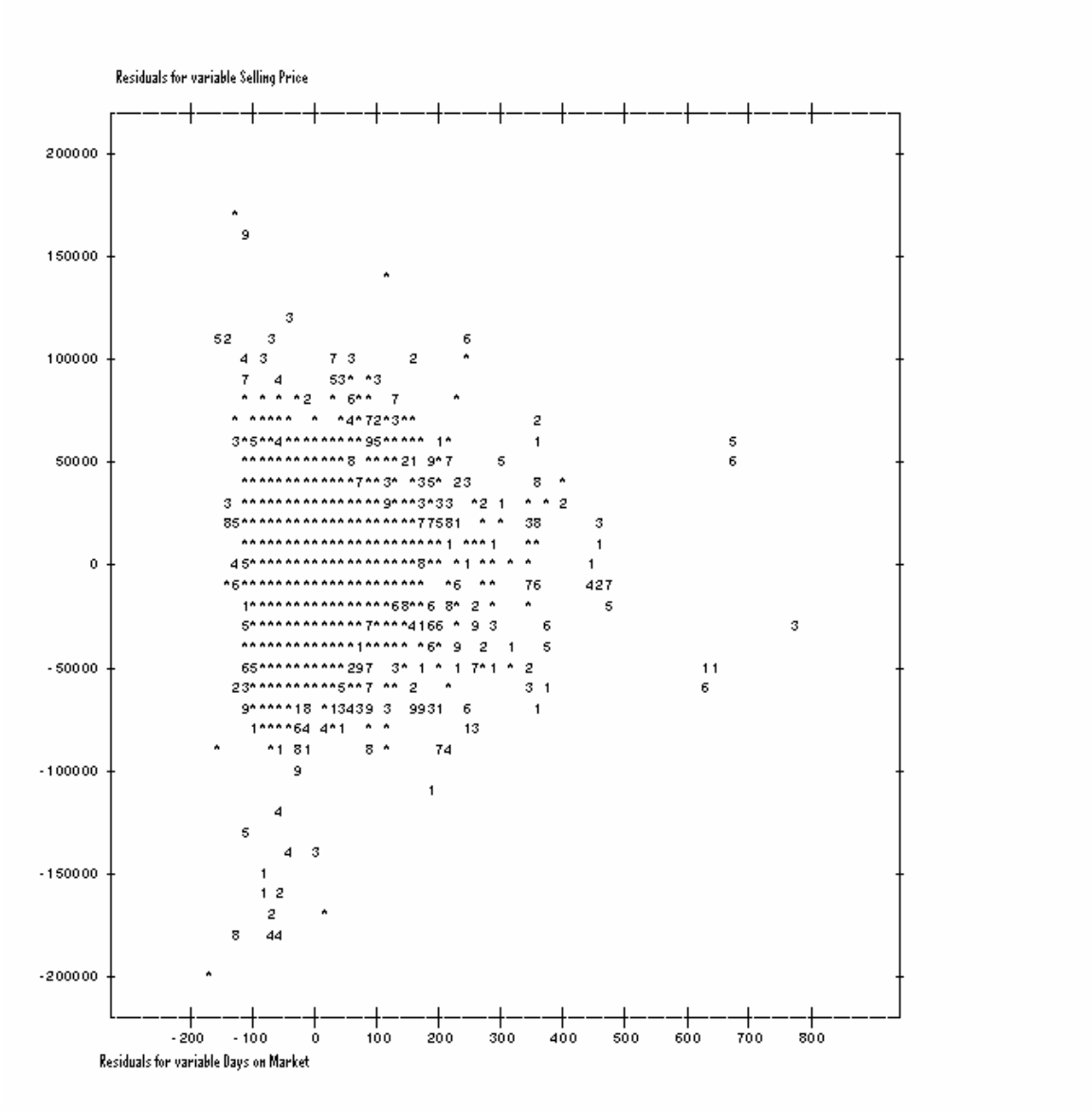

Figure 5 Added Variable plot of Selling Price vs, No. of Days on the Market

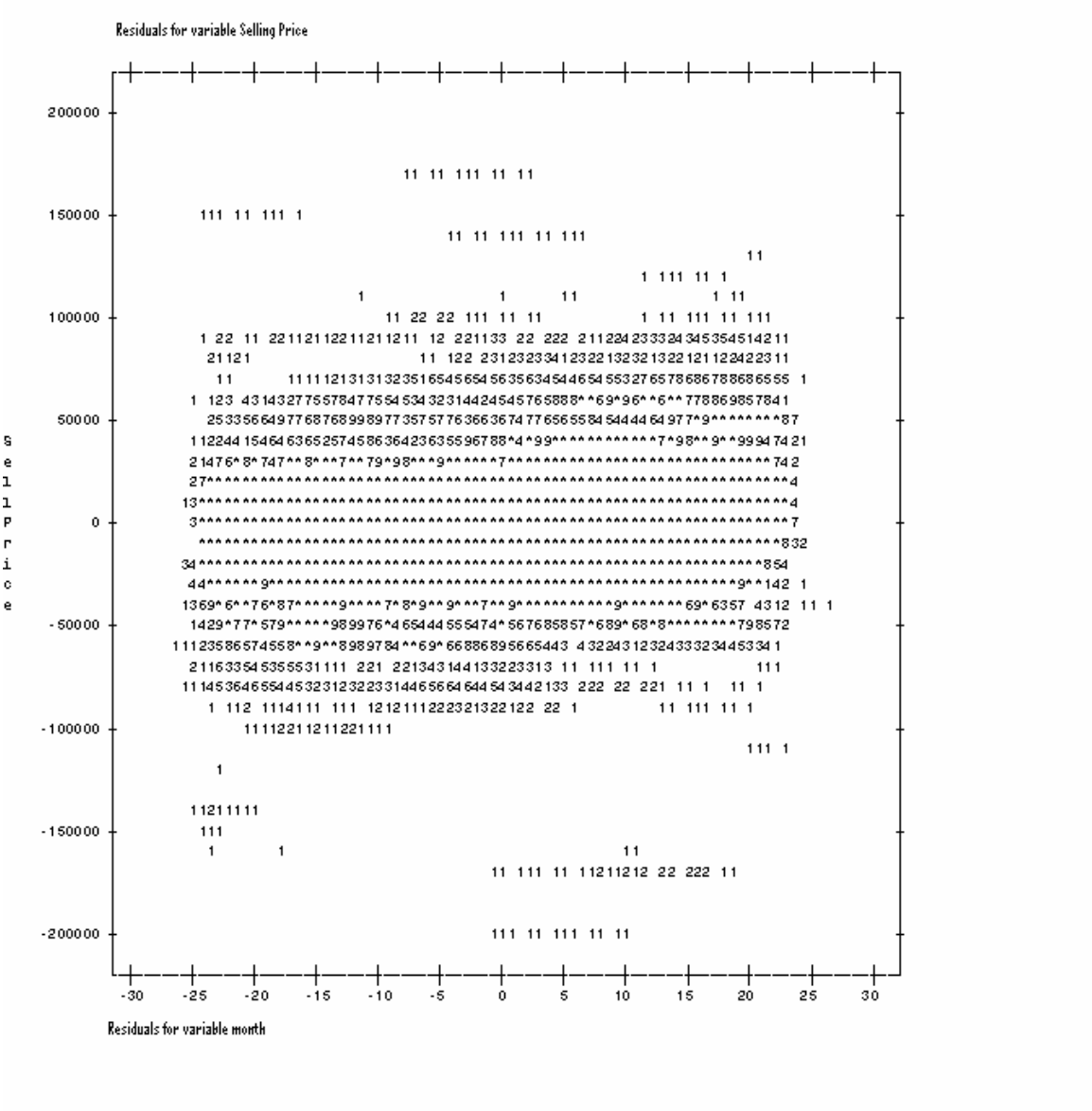

Figure 6 Added variable plot of Selling Price vs. Month

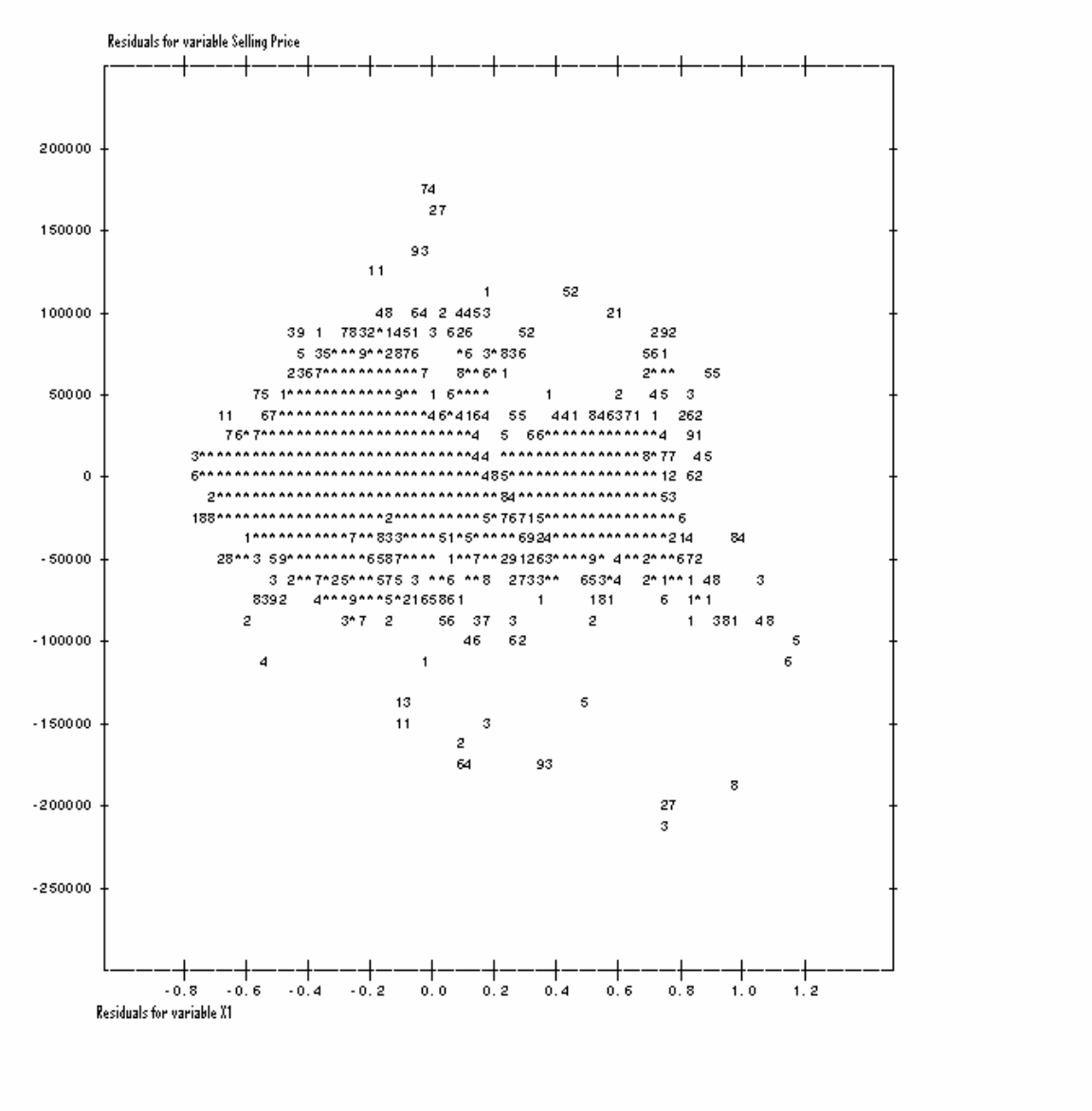

Figure 7 Added variable plot of Selling Price vs. X1

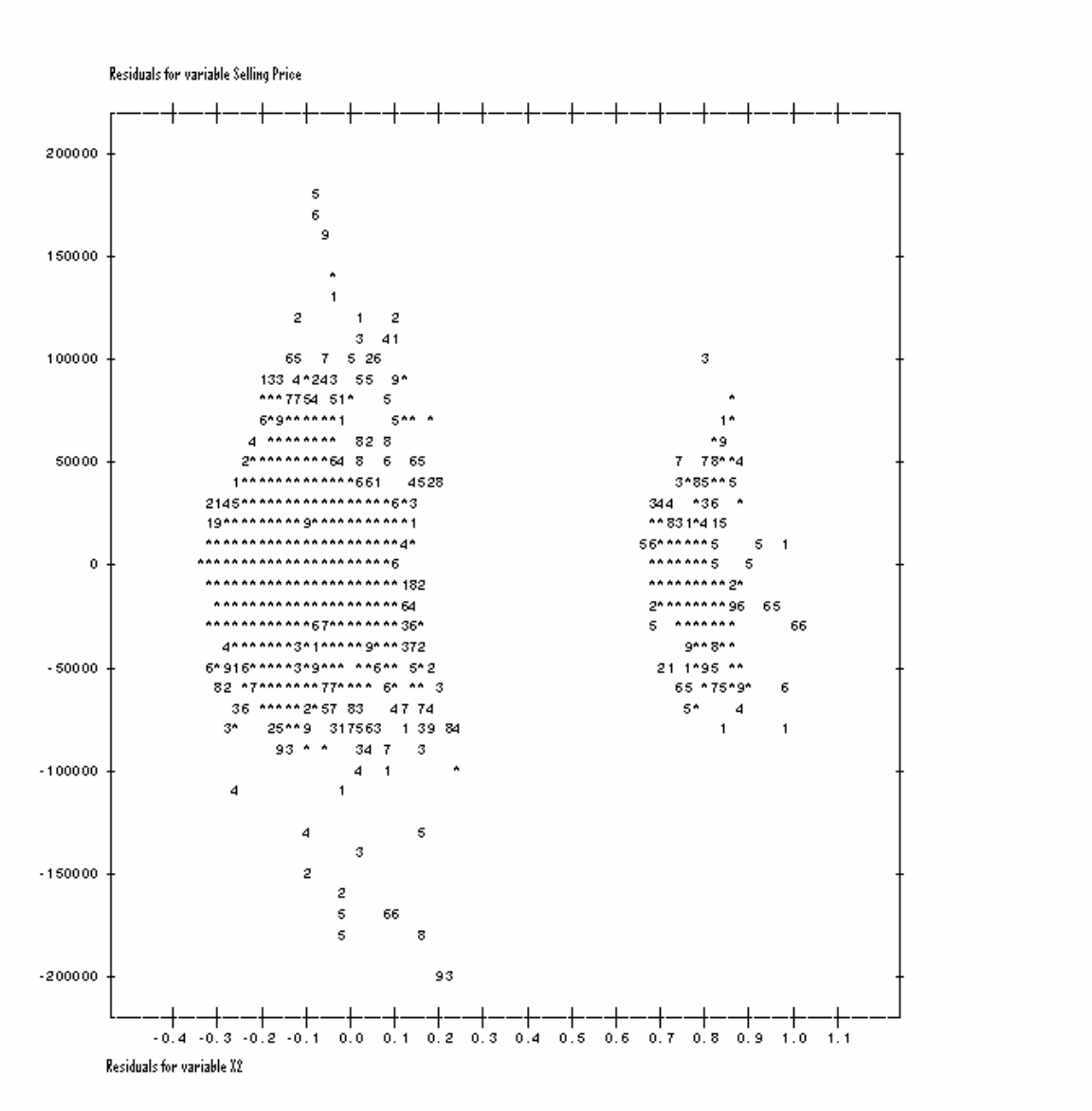

Figure 8 Added variable plot of Selling Price vs. X2

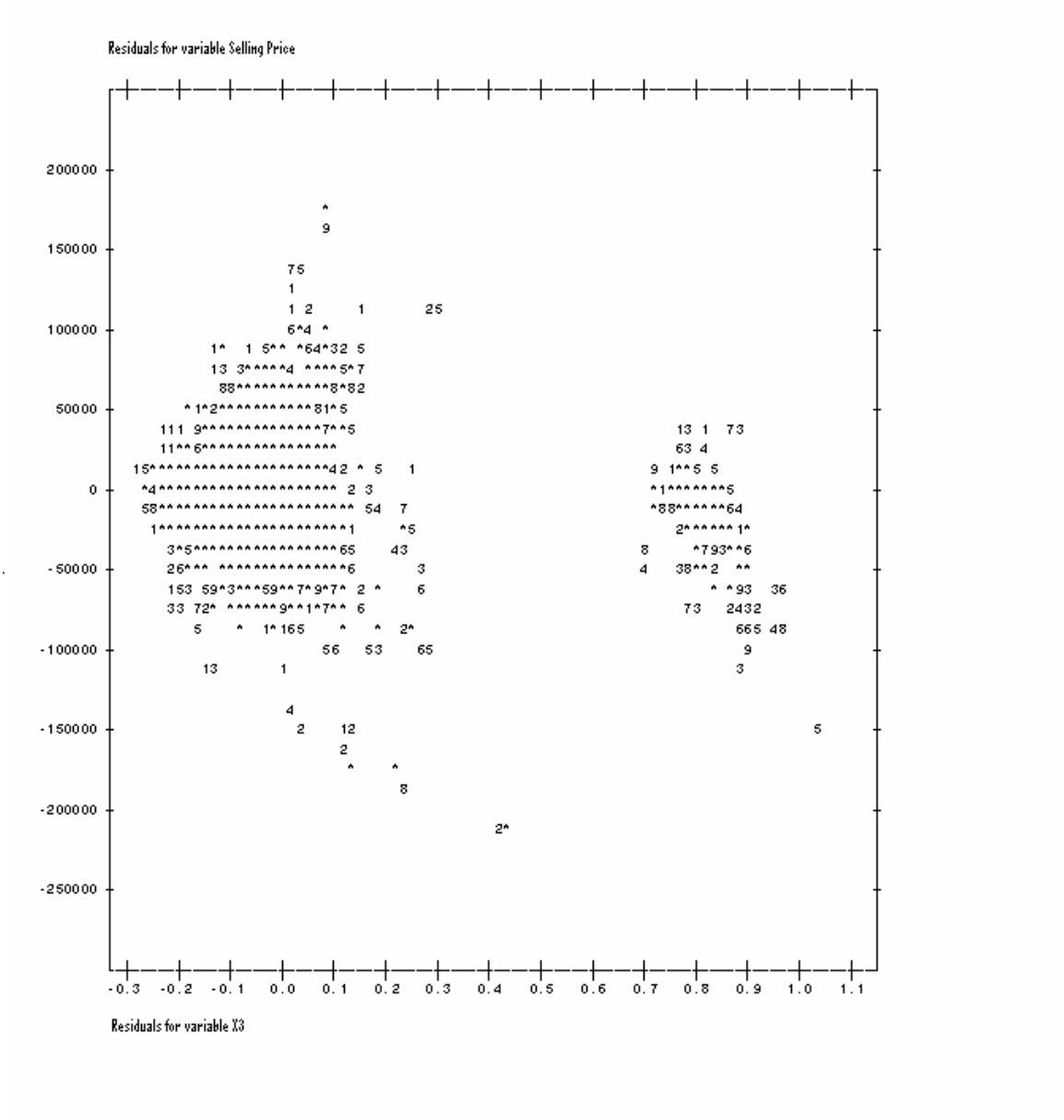

Figure 9 Added variable plot of Selling Price vs. X3

STEP 3 – VARIABLE SELECTION PROCEDURE USING ALL POSSIBLE REGRESSION

Since the added variables did not provide any conclusive information for all the independent variables other than the variable 'area in square feet', all possible regression model selection procedure discussed by Montgomery, Peck and Vining (2001) is used to check if some of these variables are extraneous. Of all the variable selection procedures all possible regression method is the most efficient method and hence the other variable selection procedures are not considered in this analysis.

There are many selection criteria for finding the best regression model using the all possible regression method like R-Square, Adjusted R-Square, Mallow's $C_{p,}$ Akaike's Information Criterion (AIC) and Schwarz's Bayesian Criterion (BIC).

In this analysis, Mallow's C_p, AIC and BIC criteria are used.

Mallow's C_p criterion: Mallow's C_p is a statistic which is a function of the error sum of squares for the full model and that for the reduced model.

The formula for C_p is given by

$$C_p = \frac{SSE_p}{s^2} - (n - 2p) \qquad \mathbf{2}$$

where SSE_p is the error sum of squares for the reduced model with p terms including the intercept term, s^2 is the estimate of MSE for the full model and n is the number of observations.

For full model i.e. the model containing all the independent variables Cp is equal to p. For an adequate model, C_p is approximately equal to p and otherwise is greater than p, reflecting bias in the parameter estimates in the regression equation. Thus, it is desirable to select a model in which the value of Cp is close to the number of independent variables p, including the constant term, in the model. Thus minimizing Mallow's Cp over all possible regressions can give a best subset model.

For AIC and BIC, models with small values of these criteria are selected. These criteria are given by

$$AIC_p = n\ln SSE_p - n\ln(n) + 2p \quad \textbf{3}$$

$$BIC_p = n\ln SSE_p - n\ln(n) + p\ln(n) \quad \textbf{4}$$

Using all these criteria, it is clear from the output given in table 3 model containing all the variables should be selected as it satisfies all the requirements of the three criterions i.e. C_p value closest to p and smallest AIC and BIC. All the other models have very large C_p, AIC and BIC values compared to the model with all the parameters.

Number in Model	C(p)	R-Square	AIC	BIC	Variables in Model
6	7.0000	0.7838	433744.231	433746.236	SqFt DOM Month X1 X2 X3
5	87.1018	0.7830	433824.201	433826.158	SqFt Month X1 X2 X3
5	189.4733	0.7819	433925.962	433927.861	SqFt DOM Month X1 X3
4	285.6497	0.7809	434021.093	434022.963	SqFt Month X1 X3
5	368.9928	0.7801	434103.231	434105.031	SqFt DOM Month X2 X3
4	425.7905	0.7795	434158.969	434160.775	SqFt DOM Month X3
4	449.4358	0.7792	434182.144	434183.938	SqFt Month X2 X3
3	515.8226	0.7785	434247.026	434248.836	SqFt Month X3
5	653.9649	0.7772	434381.602	434383.245	SqFt DOM Month X1 X2
5	661.0434	0.7771	434388.469	434390.109	SqFt DOM X1 X2 X3
4	702.9912	0.7766	434429.059	434430.739	SqFt Month X1 X2
4	724.6015	0.7764	434449.970	434451.640	SqFt X1 X2 X3
4	734.5891	0.7763	434459.628	434461.293	SqFt DOM Month X1
3	794.3803	0.7757	434517.278	434518.987	SqFt Month X1
4	801.8686	0.7756	434524.568	434526.203	SqFt DOM Month X2
4	812.7260	0.7755	434535.029	434536.659	SqFt DOM X1 X3
3	832.3954	0.7753	434553.892	434555.588	SqFt DOM Month

Table 3 Output of All Possible Selection procedure on the Independent Variables

STEP 4 ADDITION OF QUADRATIC AND INTERACTION TERMS

Since the relationship between the selling price and all the other independent variables can be quite complex, additional polynomial and interaction terms were included in the model. For evaluation, only squared terms and first order interaction terms were included. Some of these additional terms like square of indicator variables for location made no sense and they are excluded from the analysis. Even then there were 17 additional terms under consideration.

At this step there were 23 predictors in the full model, all possible regression method is used for selecting an optimum model with Adjusted R-Square, Mallow's $C_{p,}$ Akaike's Information Criteria (AIC) and Schwarz's Bayesian Criteria (BIC) as the selection criteria.

Using the all possible regression method with same criterions as mentioned above, following terms are included into the model. SqFt, DOM, Month, X1, X2, X3, SqFtX1, SqFtX2, SqFtX3, SqFtmonth, SqFtDOM, X1DOM, X2month, X3DOM, SqFt2, DOM2 and month2

From the above selected variables, X2DOM is included as the other two interaction terms of this kind are in this model. The variable X2month is excluded because the other two interaction terms of this kind are not in the model. This means that the model is no longer the optimum model found using the all possible regression method but this step is taken to make the model more meaningful by including all the similar terms in the mode Therefore the variables selected for the model are SqFt, DOM, Month, X1, X2, X3, SqFtX1, SqFtX2, SqFtX3, SqFtmonth, SqFtDOM, X1DOM, X2DOM, X3DOM, SqFt2, DOM2 and month2

Number in Mo Del	C(p)	R-Square	AIC	BIC	Variables in Model
17	19.0550	0.8054	431559.726	431561.755	SqFt DOM Month X1 X2 X3 SqFtX1 SqFtX2 SqFtX3 SqFtmonth SqFtDOM X1DOM X2month X3DOM SqFt2 DOM2 month2
19	19.5103	0.8054	431560.178	431562.217	SqFt DOM Month X1 X2 X3 SqFtX1 SqFtX2 SqFtX3 SqFtmonth SqFtDOM X1month X1DOM X2month X3month X3DOM SqFt2 DOM2 month2
18	19.6033	0.8054	431560.273	431562.306	SqFt DOM Month X1 X2 X3 SqFtX1 SqFtX2 SqFtX3 SqFtmonth SqFtDOM X1month X1DOM X2month X3DOM SqFt2 DOM2 month2
18	19.9598	0.8054	431560.630	431562.662	SqFt DOM Month X1 X2 X3 SqFtX1 SqFtX2 SqFtX3 SqFtmonth SqFtDOM X1DOM X2month X3month X3DOM SqFt2 DOM2 month2

Table 4 Output of All possible Selection procedure on the Independent Variables and interaction and squared terms

STEP 5 CHECKING THE ASSUMPTION OF LINEAR ASSOCIATION

In order to check the assumption of linear association, residual plots are used. This method is discussed by Montgomery, Peck and Vining (2001). The residual plot for predicted values of variable selling price against the residuals is shown in figure 10. The residual plot shows that residuals are not uniformly distributed.

Since it can be difficult to verify the true nature of relationship among the independent and dependent variables, Box-Cox transformations discussed by Ryan (1996) were used to find appropriate transformation from the family of power transformations on the dependent variable Y. The method of Box-Cox transformations is discussed by Kutner, Nachstshiem and Neter (2001). The family of power transformations is of the form

$$Y' = Y^{\lambda} \qquad \mathbf{5}$$

where, λ is a parameter to be determined from the data.

In SAS/STAT software the Box-Cox transformations are done using the PROC TRANSREG procedure as discussed in Freund and Littell (2000).

The results for PROC TRANSREG are given in table 5.

The results of the PROC TRANSREG procedure suggest that that the following transformation should be used for the variable 'selling price' based on the independent variables selected in step 4.

$$Y' = Y^{0.25}$$

This means that the dependent variable Selling Price needs to transformed in order to make the relationship between the dependent and independent variables linear.

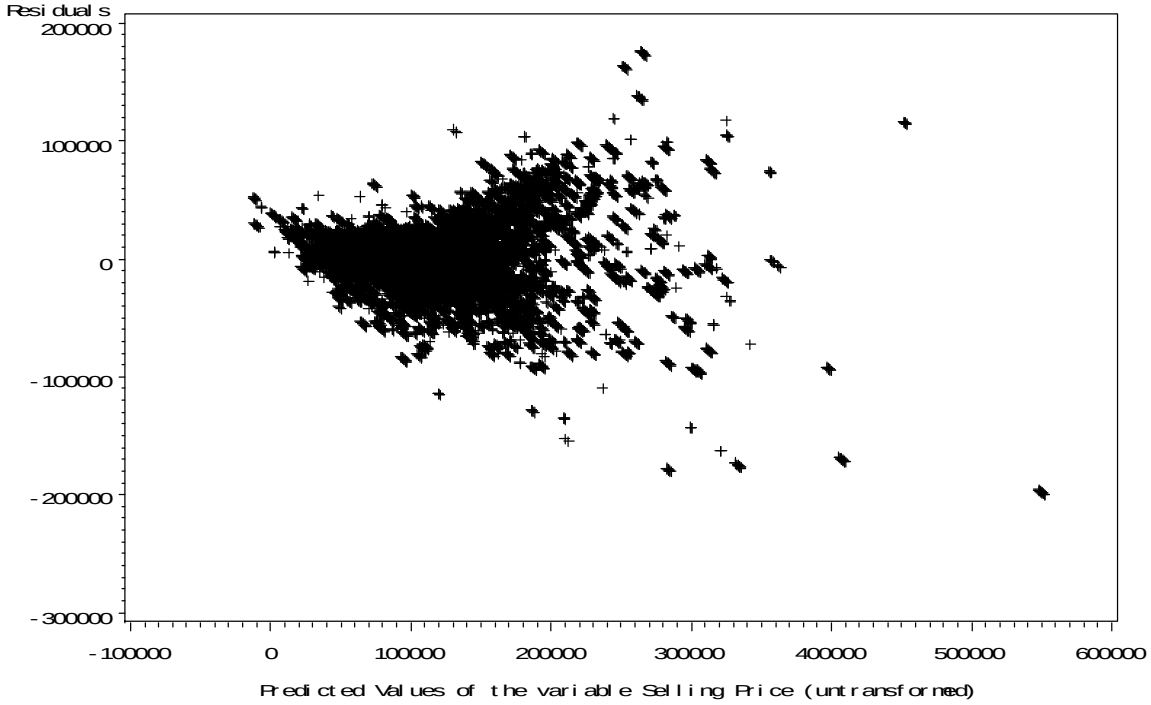

Figure 10 Residual Plot for the variable Selling Price (untransformed)

Transformation Information for BoxCox(SellPrice)				
Lambda		**R-Square**	**Log Like**	
-3.00		0.01	-319247	
-2.75		0.02	-306284	
-2.50		0.03	-293699	
-2.25		0.05	-281583	
-2.00		0.08	-270044	
-1.75		0.14	-259213	
-1.50		0.23	-249230	
-1.25		0.35	-240233	
-1.00		0.48	-232341	
-0.75		0.59	-225645	
-0.50		0.69	-220220	
-0.25		0.75	-216143	
0.00		0.79	-213486	
0.25		0.81	-212261	<
0.50		0.82	-212375	
0.75		0.82	-213634	
1.00		0.80	-215802	
1.25		0.79	-218672	
1.50		0.77	-222091	
1.75		0.74	-225960	
2.00		0.72	-230220	
2.25		0.69	-234836	
2.50		0.66	-239789	
2.75		0.62	-245062	
3.00		0.59	-250642	

< - Best Lambda
*** - Confidence Interval**
+ - Convenient Lambda

Table 5 Output of first Box-Cox transformations on the variable Selling Price

After transforming the variable selling price, once again the assumption of linear association is checked. Again using the PROC TRANSREG procedure, the model is evaluated to see if the Box-Cox transformation of 0.25 is valid. The results given in table 6 indicate that the transformed variable 'selling price' needs to be further transformed by using the following transformation.

$$Y' = Y^{1.50}$$

Transformation Information for BoxCox(SellPrice)				
Lambda		**R-Square**	**Log Like**	
-3.00		0.59	-13911.1	
-2.75		0.62	-12432.7	
-2.50		0.64	-11035.0	
-2.25		0.67	-9719.0	
-2.00		0.69	-8486.0	
-1.75		0.71	-7337.4	
-1.50		0.72	-6274.2	
-1.25		0.74	-5297.9	
-1.00		0.75	-4409.5	
-0.75		0.76	-3610.0	
-0.50		0.77	-2900.3	
-0.25		0.78	-2280.8	
0.00		0.79	-1751.7	
0.25		0.80	-1312.7	
0.50		0.80	-963.2	
0.75		0.81	-701.8	
1.00		0.81	-526.8	
1.25		0.82	-436.2	
1.50		0.82	-427.1	<
1.75		0.82	-496.6	
2.00		0.82	-641.4	
2.25		0.82	-857.9	
2.50		0.82	-1142.3	
2.75		0.82	-1490.9	
3.00		0.82	-1899.9	

< - Best Lambda
*** - Confidence Interval**
+ - Convenient Lambda

Table 6 Output of second Box-Cox transformations on the variable Selling Price

Using both the transformations i.e. 0.25 and 1.5 a new transformation of 0.25*1.5=0.375 is tested for the original variable ‘selling price’.

Again using the PROC TRANSREG procedure, the model is evaluated to see if the Box-Cox transformation is valid. The results given in table 7 clearly indicate that the suggested transformation for the original variable ‘selling price’ is valid.

The residual plot of predicted values of selling price shown in figure 11 although not perfectly horizontal, does not show any systematic pattern indicating that the transformation may be valid.

Transformation Information for BoxCox(SellPrice)				
Lambda		**R-Square**	**Log Like**	
-3.00		0.41	-63368.9	
-2.75		0.46	-60490.0	
-2.50		0.51	-57778.1	
-2.25		0.55	-55237.1	
-2.00		0.59	-52871.0	
-1.75		0.63	-50683.6	
-1.50		0.67	-48678.9	
-1.25		0.70	-46861.0	
-1.00		0.72	-45234.1	
-0.75		0.74	-43802.5	
-0.50		0.76	-42569.9	
-0.25		0.78	-41539.2	
0.00		0.79	-40711.6	
0.25		0.80	-40086.8	
0.50		0.81	-39661.7	
0.75		0.81	-39431.0	
1.00	+	0.82	-39387.0	<
1.25		0.82	-39519.7	
1.50		0.82	-39817.8	
1.75		0.82	-40268.7	
2.00		0.82	-40859.8	
2.25		0.81	-41578.5	
2.50		0.81	-42412.9	
2.75		0.80	-43352.1	
3.00		0.80	-44386.2	

< - Best Lambda
*** - Confidence Interval**
+ - Convenient Lambda

Table 7 Output of final Box-Cox transformations on the variable Selling Price

STEP 6 CHECKING THE ASSUMPTION OF RESIDUALS WITH EQUAL VARIANCES

In order to check the assumption that residuals are distributed with equal variances, the residual plots of residuals against independent variables are used. This method is discussed by Chatterjee and Hadi (1987). If the error variances are constant, then the residuals in the residual plot fall within a narrow band centered around zero with no systematic tendencies to be positive and negative.

The residual plot for the variables 'selling price' and 'area in square feet' shown in figures 12 and 13 although do not have perfectly horizontal band but there is no 'megaphone' pattern of residuals, which is an indication of non-constant variance.

The residual plot for the variable DOM shown in figure 14 does not indicate any systematic pattern except for larger variation at smaller values of DOM.

The residual plot of the variable month shown in figure 15 has bars of almost equal lengths with no apparent pattern.

The residual plots for the variables X_1, X_2 and X_3 shown in figures 16, 17 and 18 respectively have a pattern of bars of slightly unequal lengths for the levels of each variable. The bars of residuals have slightly longer length when each variable takes value zero. This indicates that residuals are more spread for the southwest location compared to all the three other locations. But this is natural as different locations will have different residual distribution pattern.

For all the other terms i.e. the squared and residual terms the distribution of residuals are more or less horizontally distributed in the residual plots shown in figures 19 to 29. Indicating the residuals may have constant variances with respect to these terms.

To summarize the observations made from these residuals plots it is clear that distributions of residuals are more or less constant.

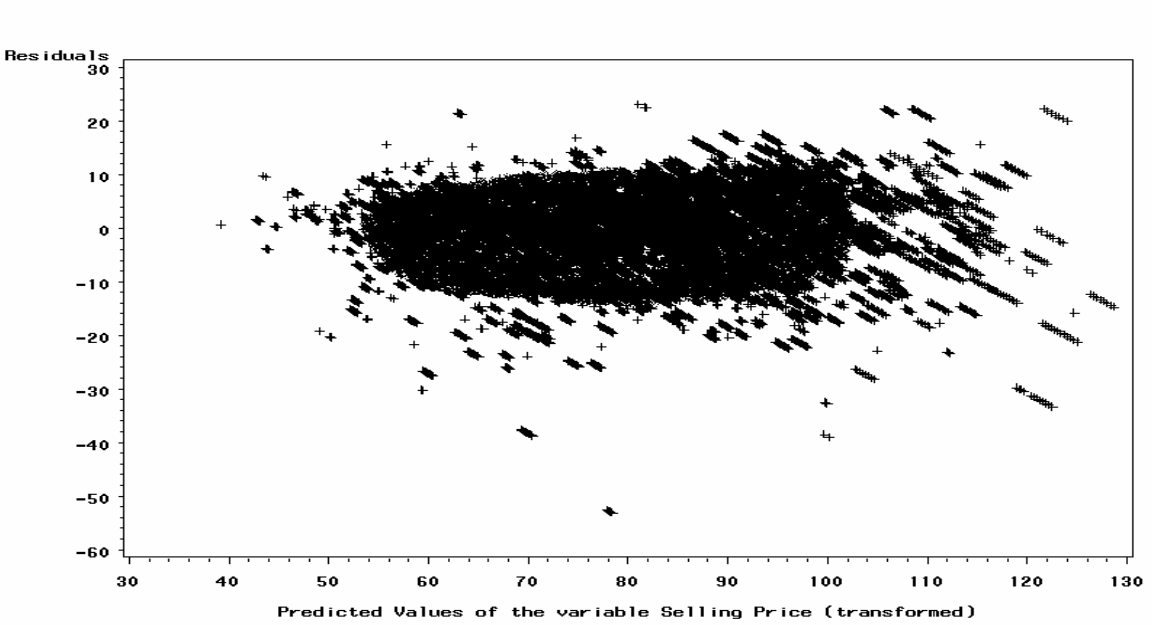

Figure 11 Residual Plot for Predicted Value of variable selling price (transformed)

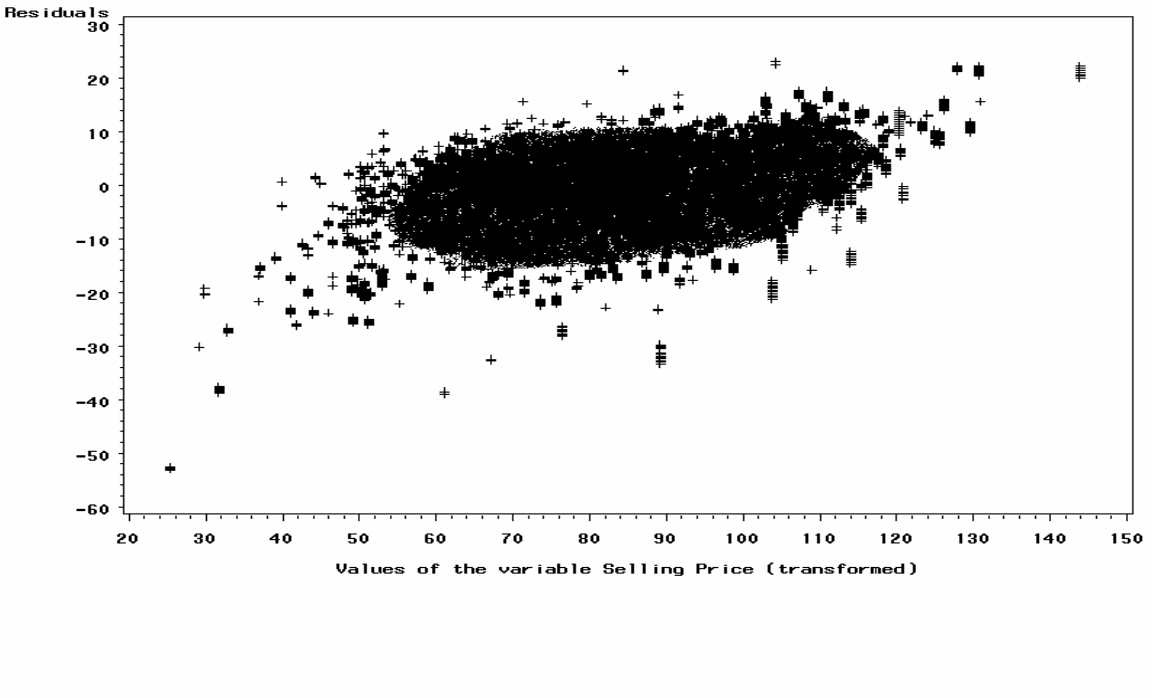

Figure 12 Residual plot for the variable Selling Price (transformed)

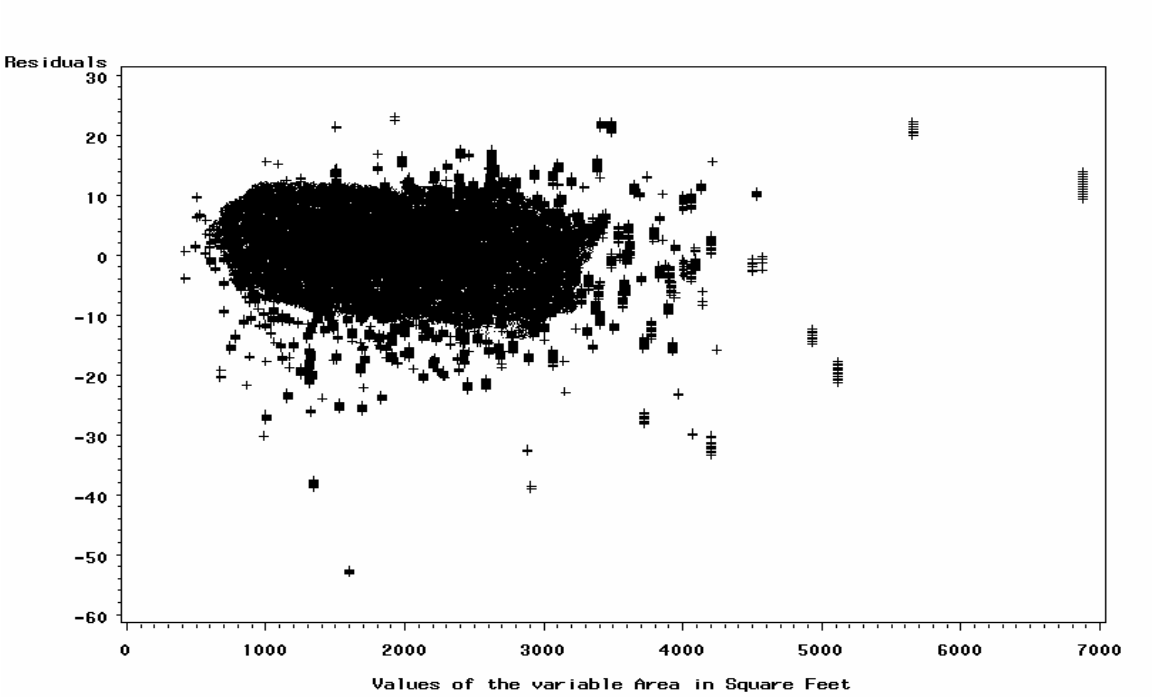

Figure 13 Residual Plot for variable Area in Square Feet

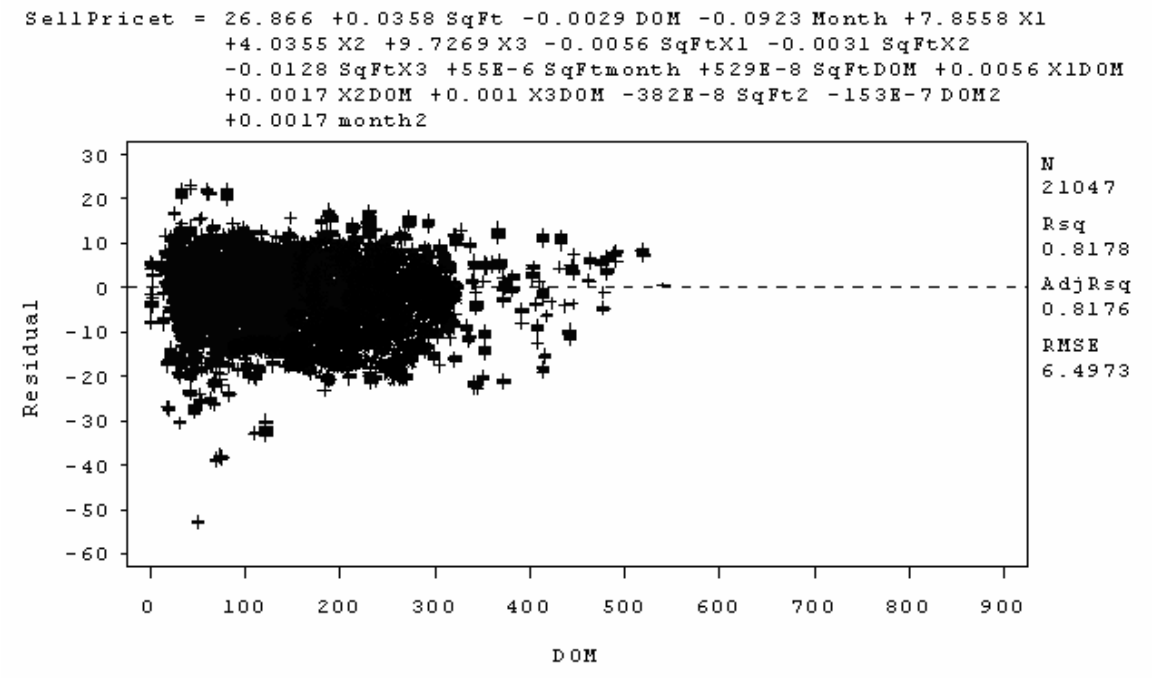

Figure 14 Residual Plot for variable Days on the Market

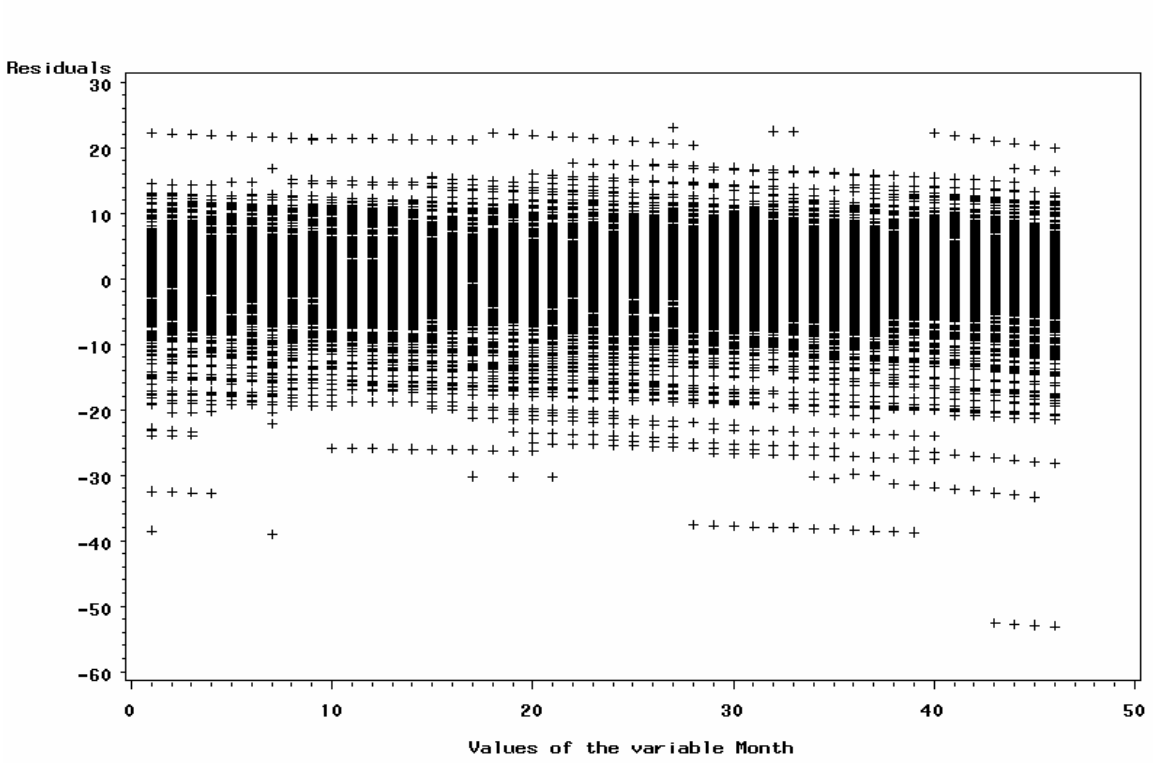

Figure 15 Residual Plot for variable Month

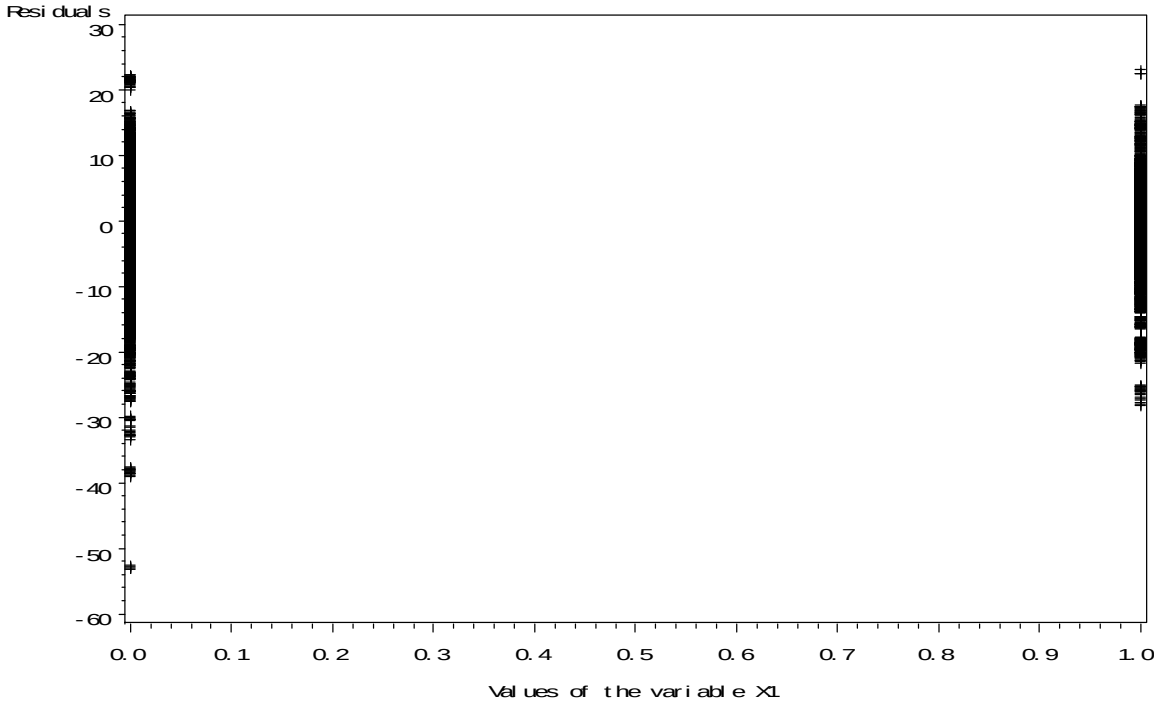

Figure 16 Residual Plot for variable X1

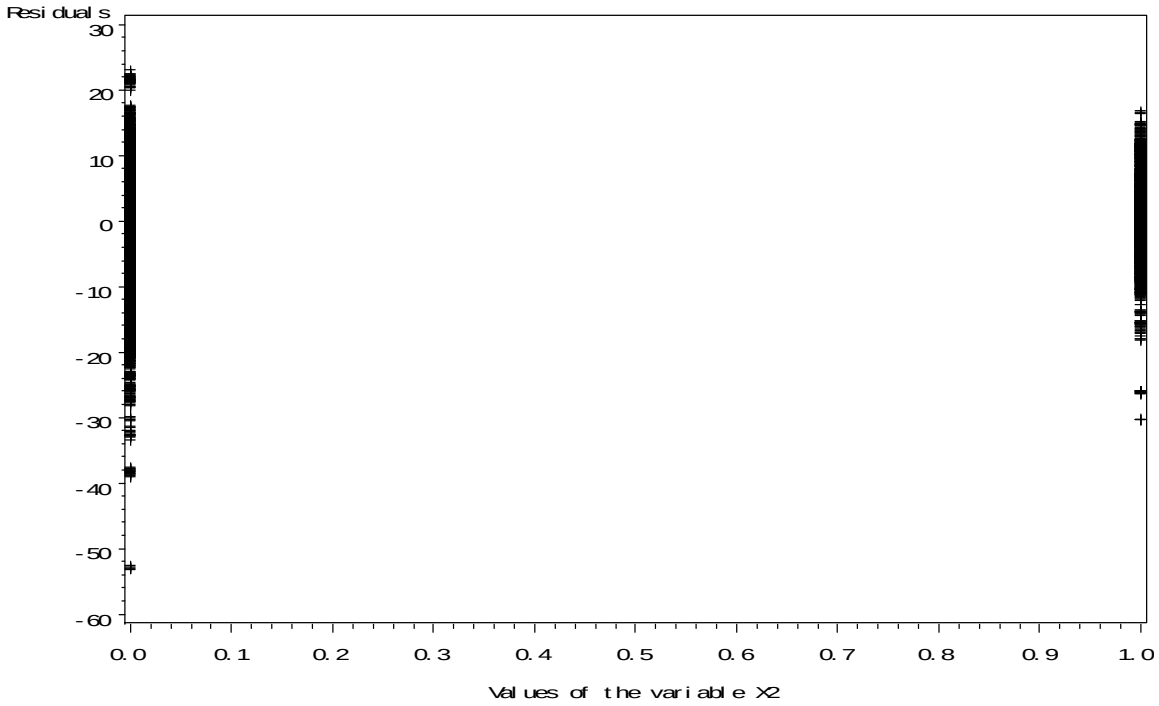

Figure 17 Residual Plot for variable X2

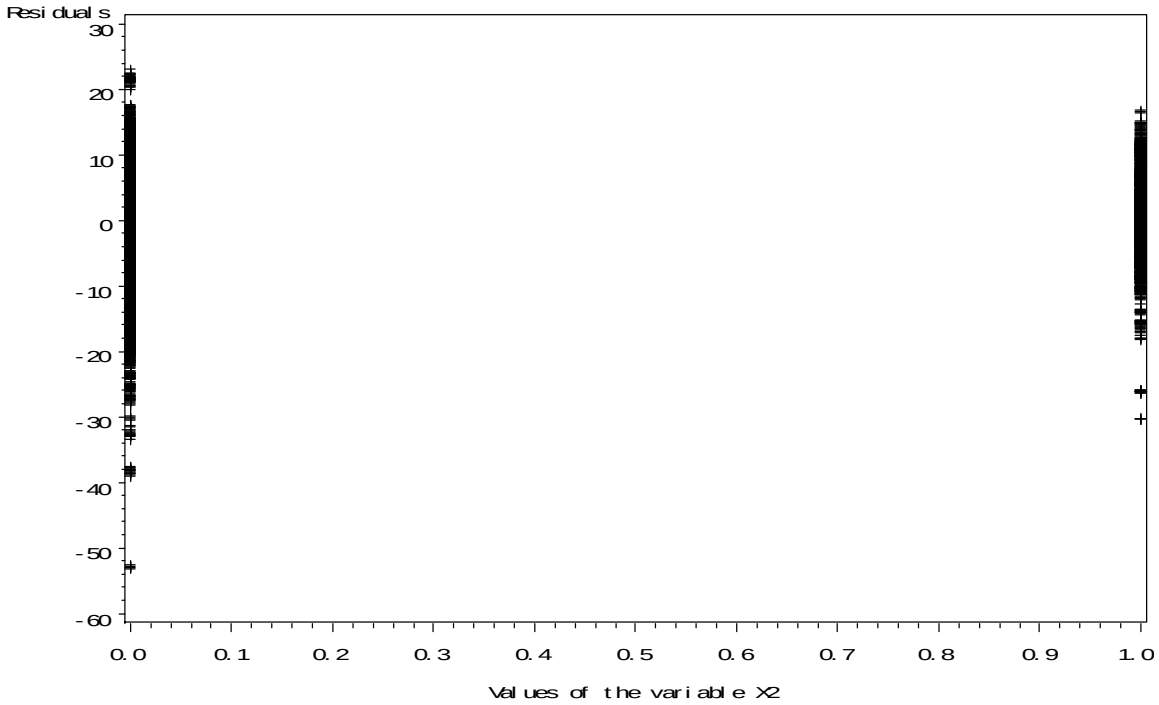

Figure 18 Residual Plot for variable X3

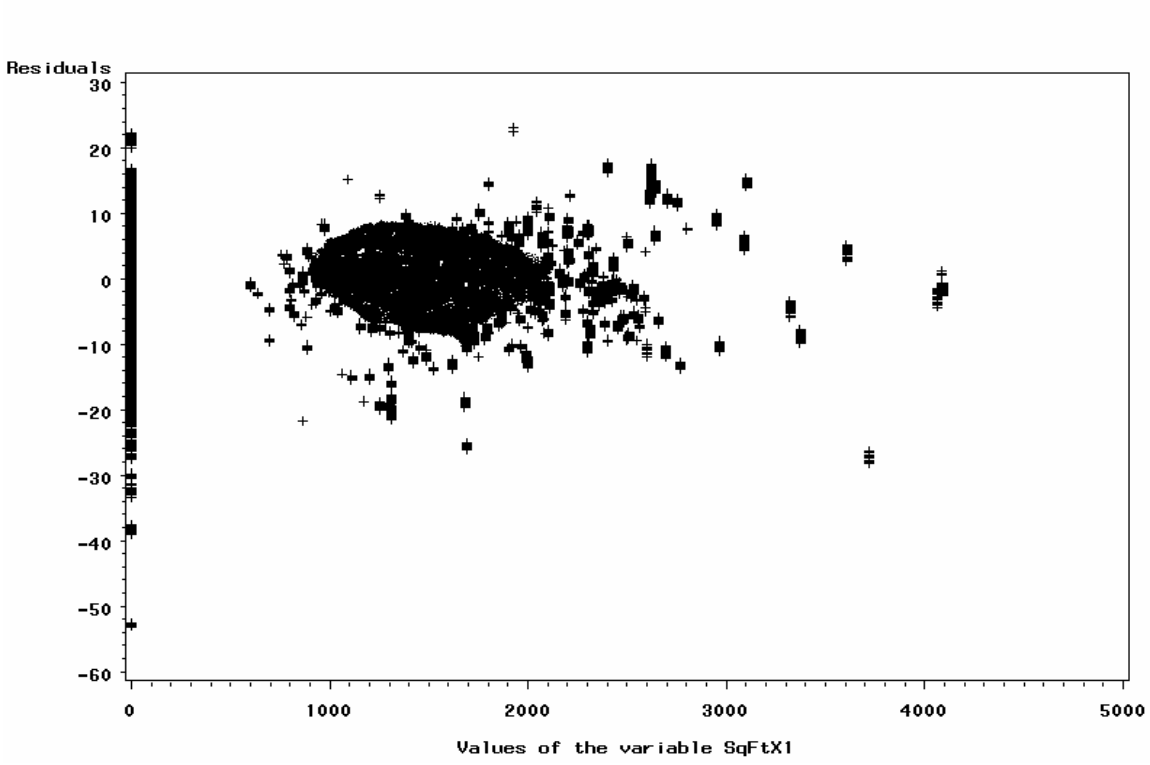

Figure 19 Residual Plot for variable SqFtX1

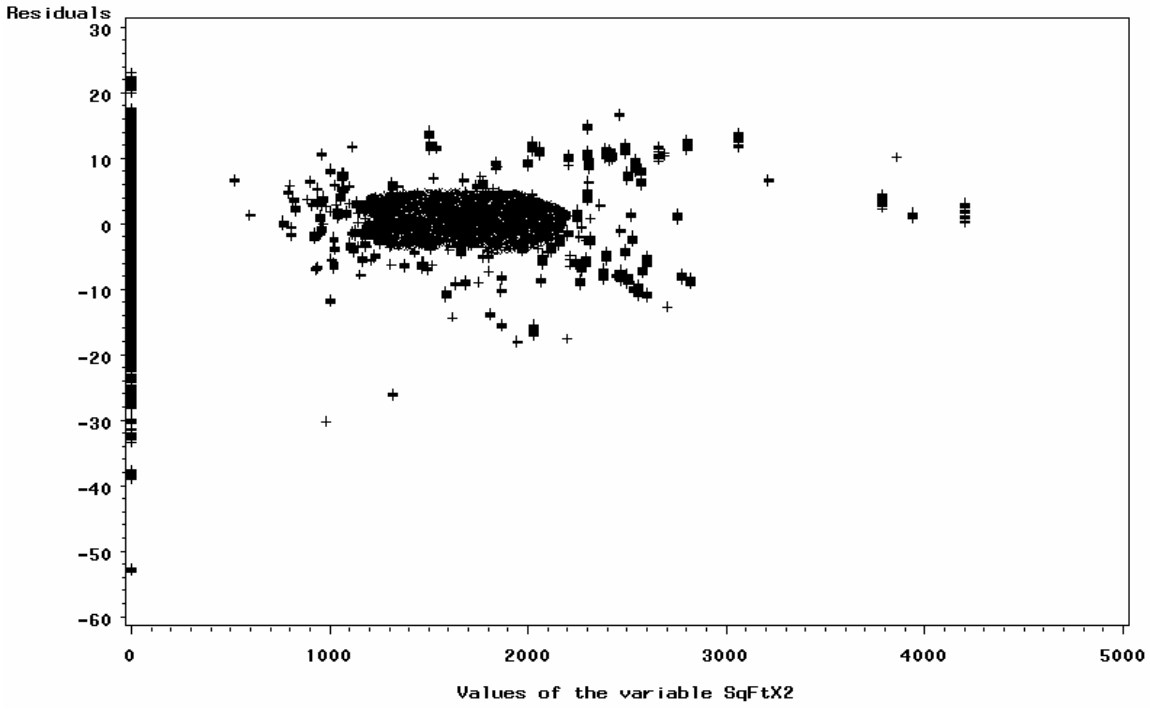

Figure 20 Residual Plot for variable SqFtX2

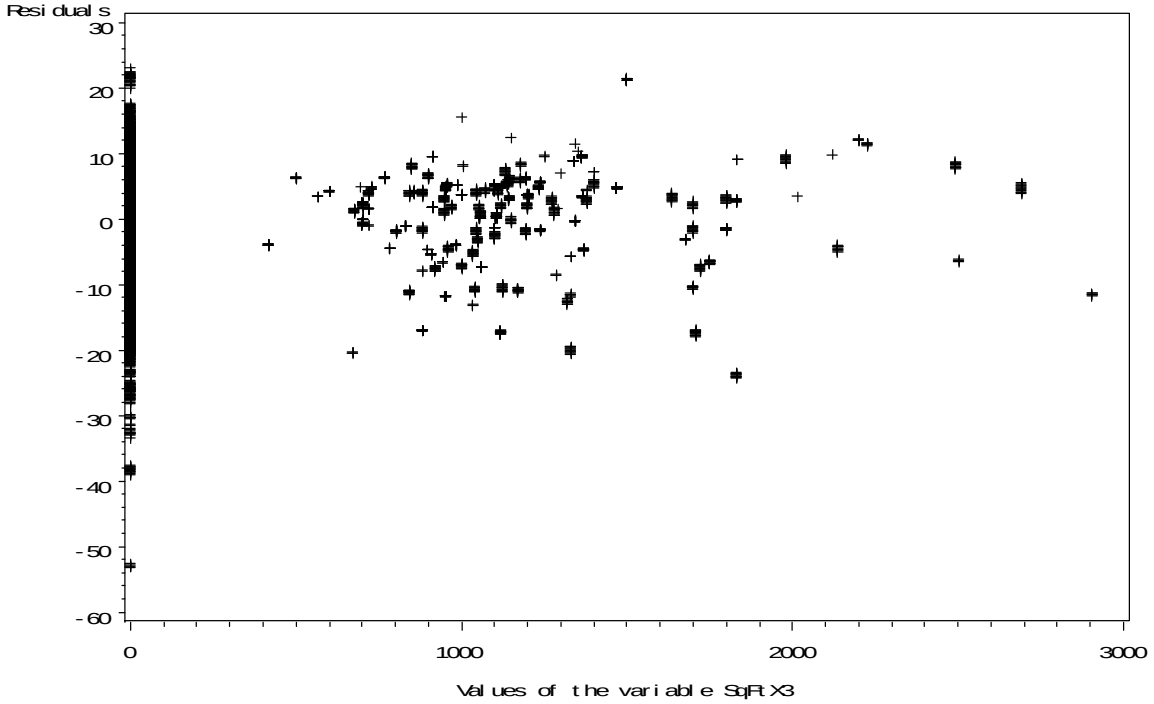

Figure 21 Residual Plot for variable SqFX3

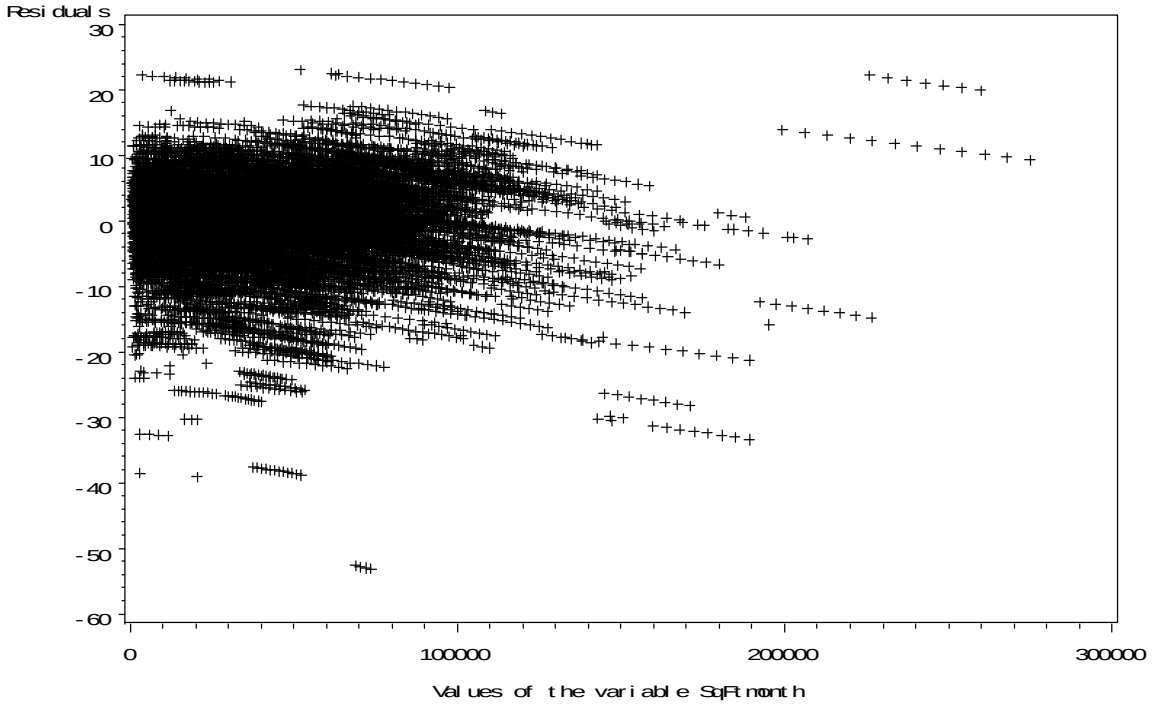

Figure 22 Residual Plot for variable SqFtmonth

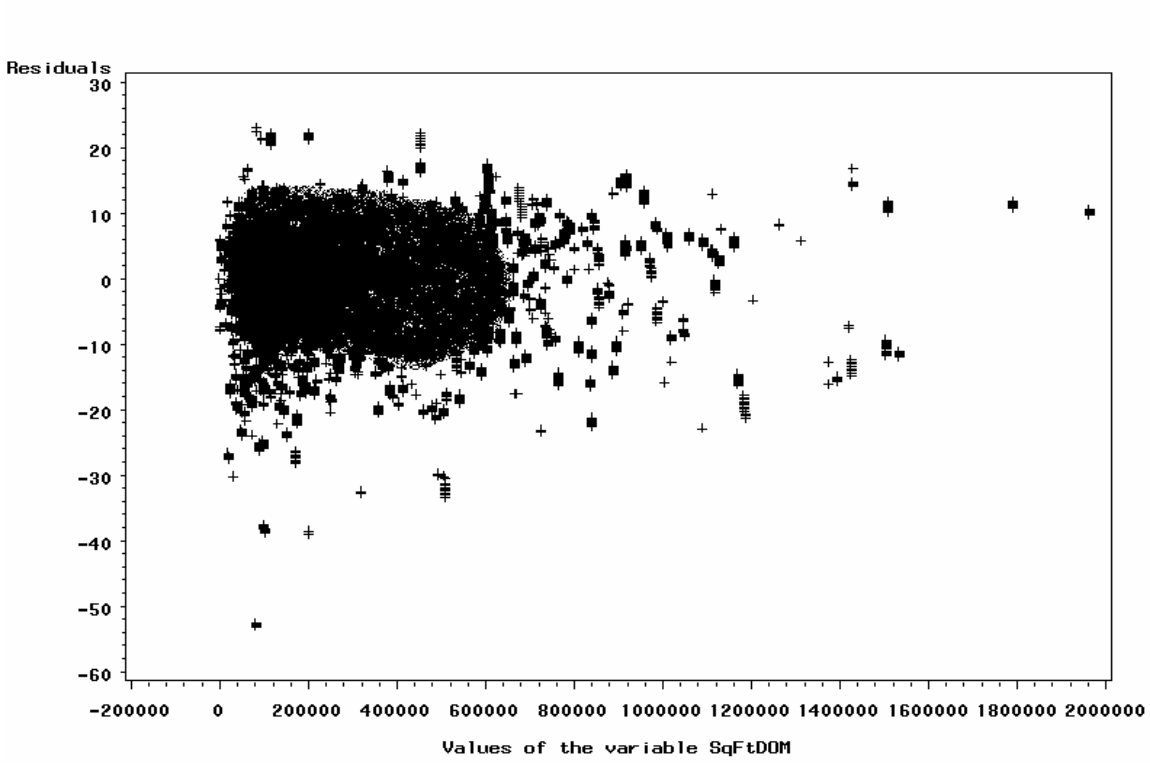

Figure 23 Residual Plot for variable SqFtDOM

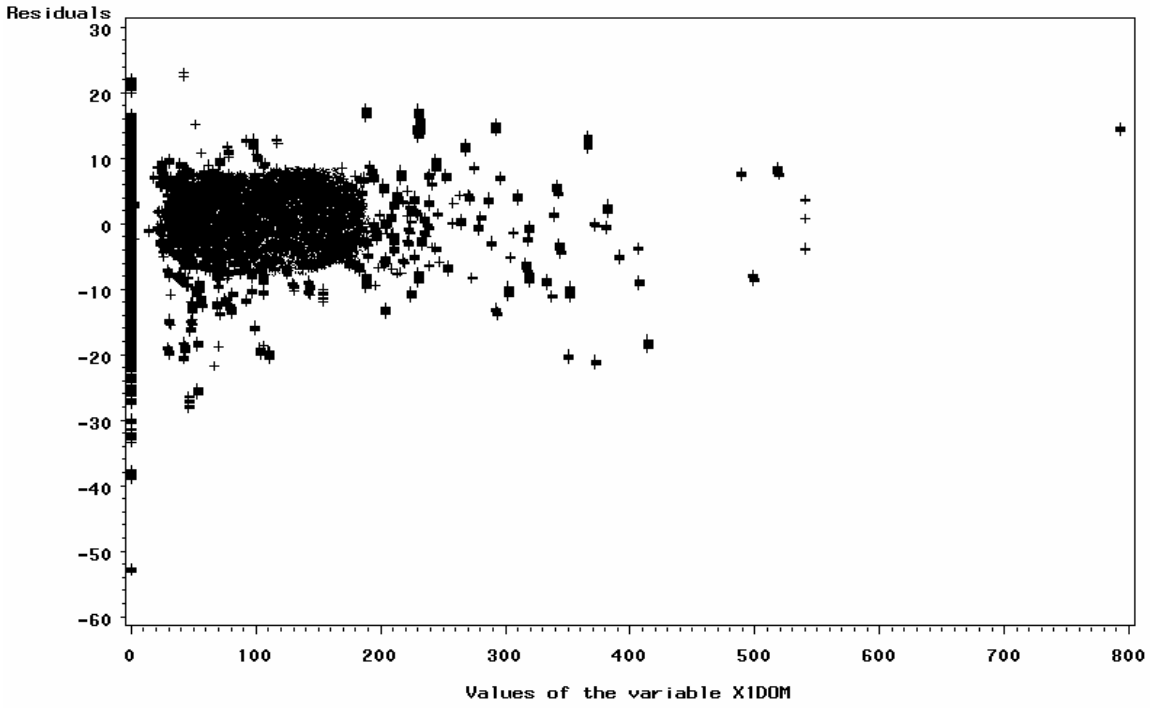

Figure 24 Residual Plot for vs. variable X1DOM

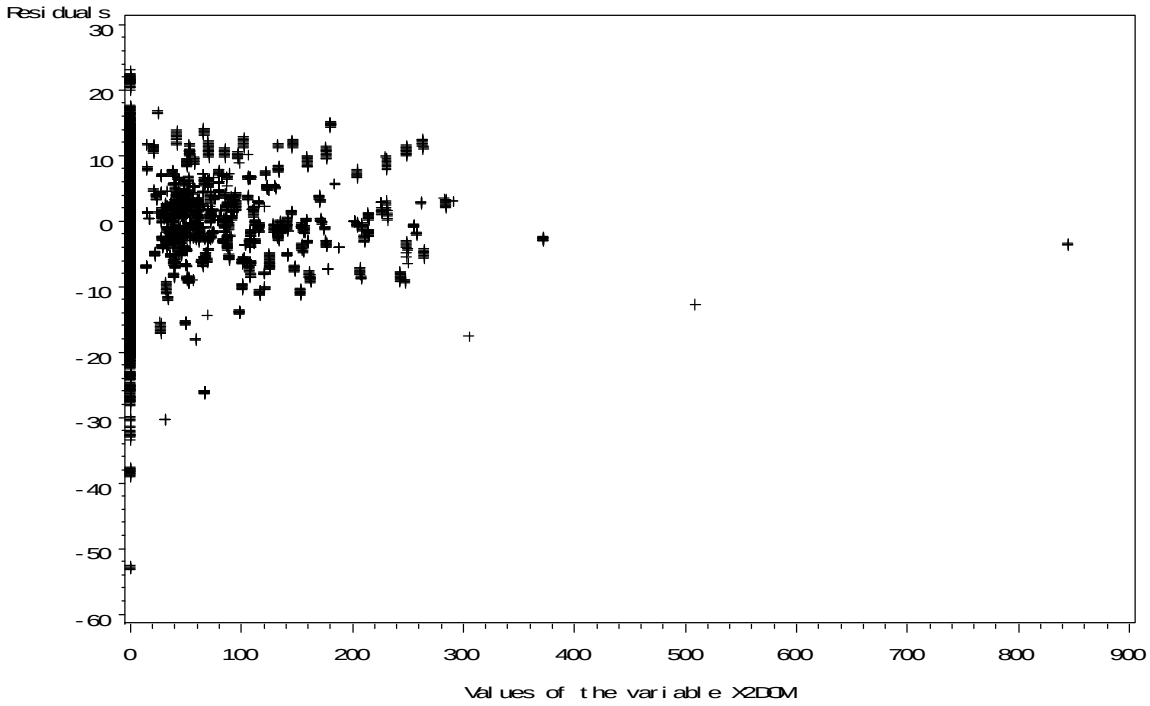

Figure 25 Residual Plot for variable X2DOM

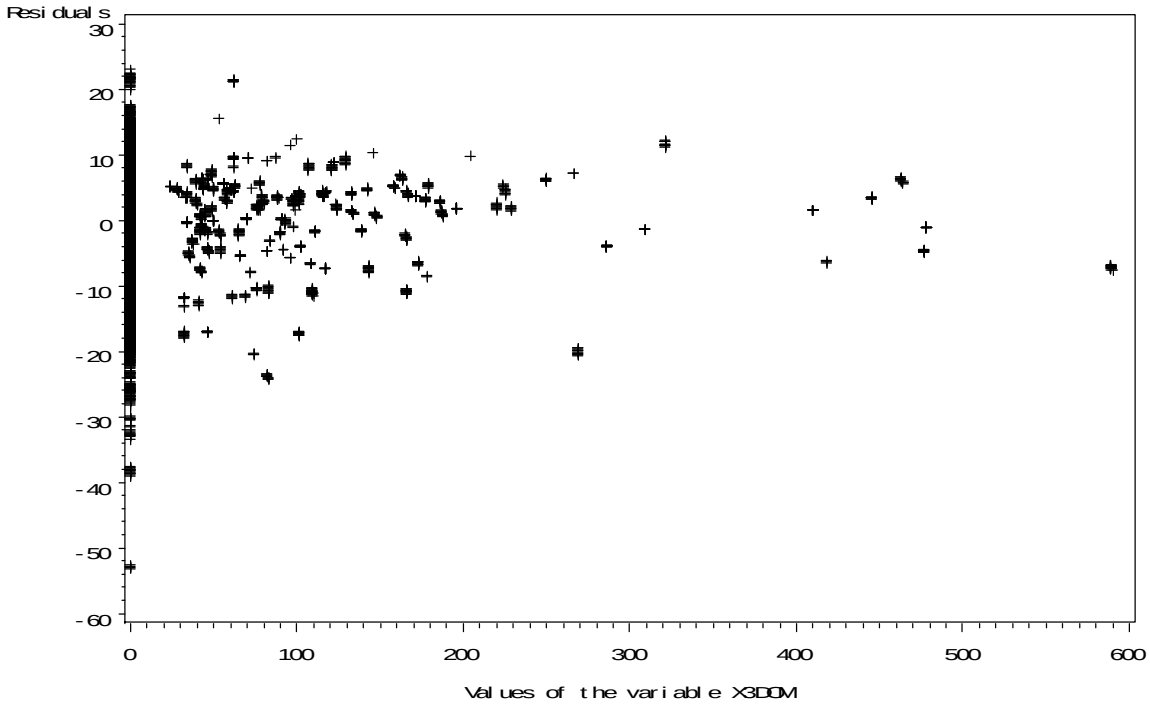

Figure 26 Residual Plot for variable X3DOM

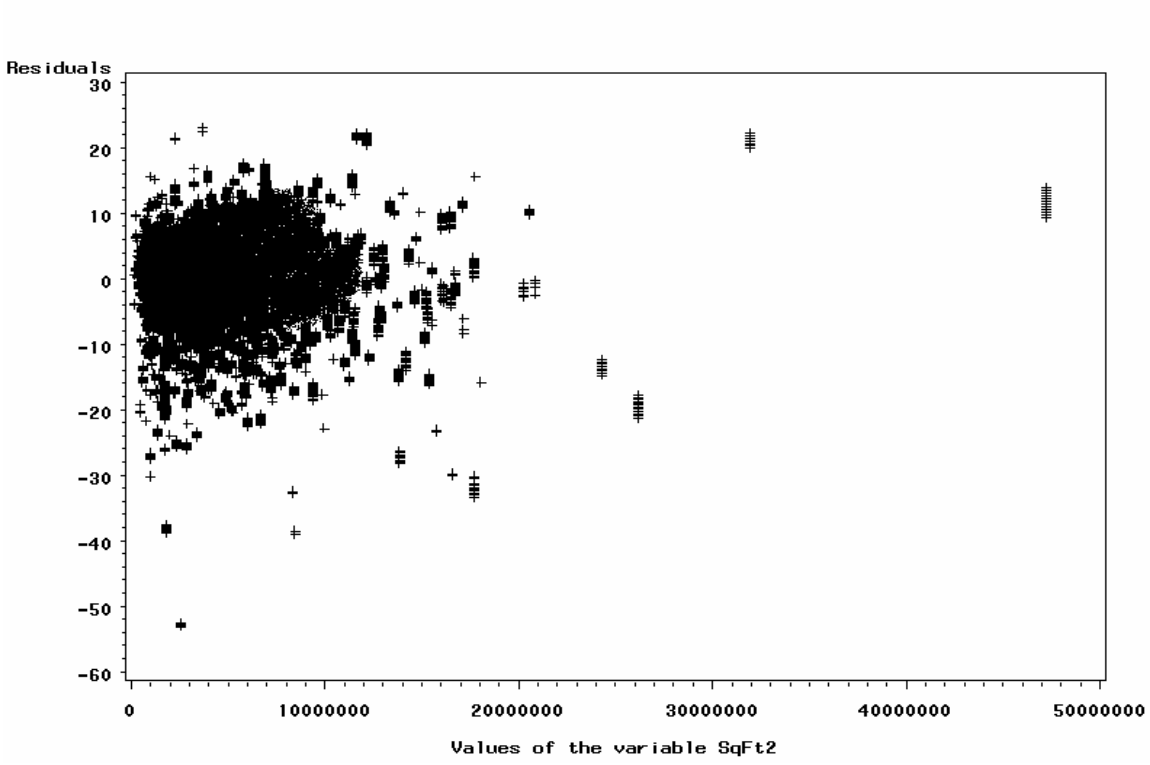

Figure 27 Residual Plot for variable SqFt2

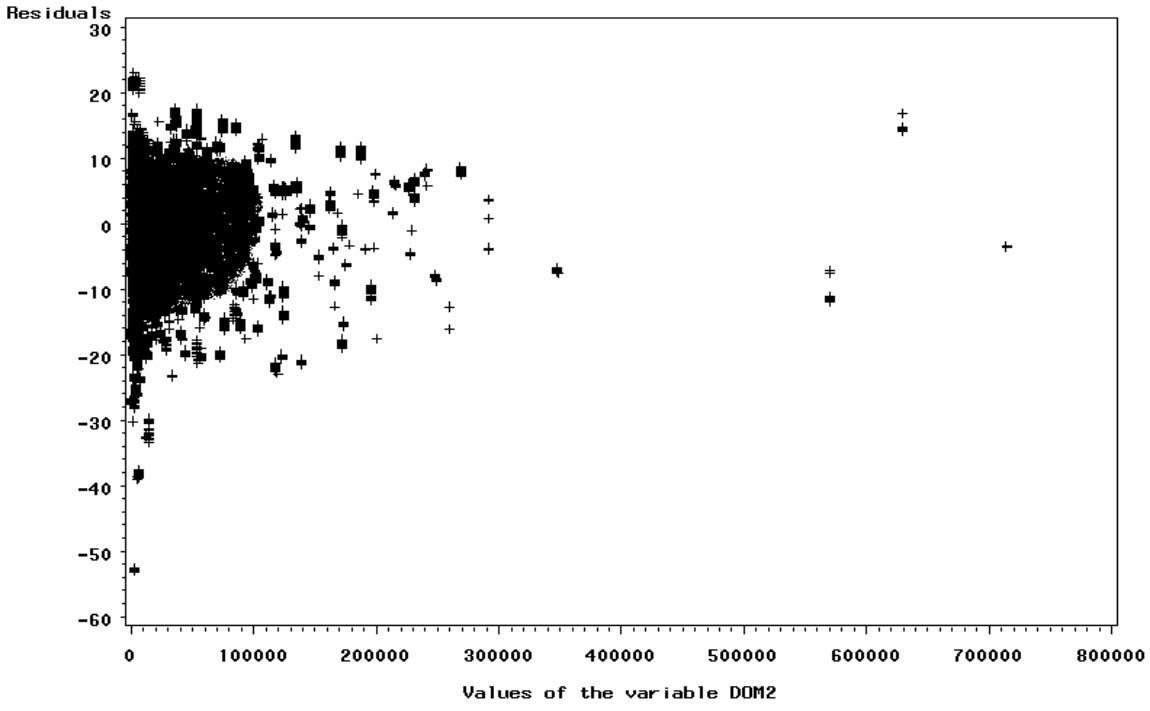

Figure 28 Residual Plot for variable DOM2

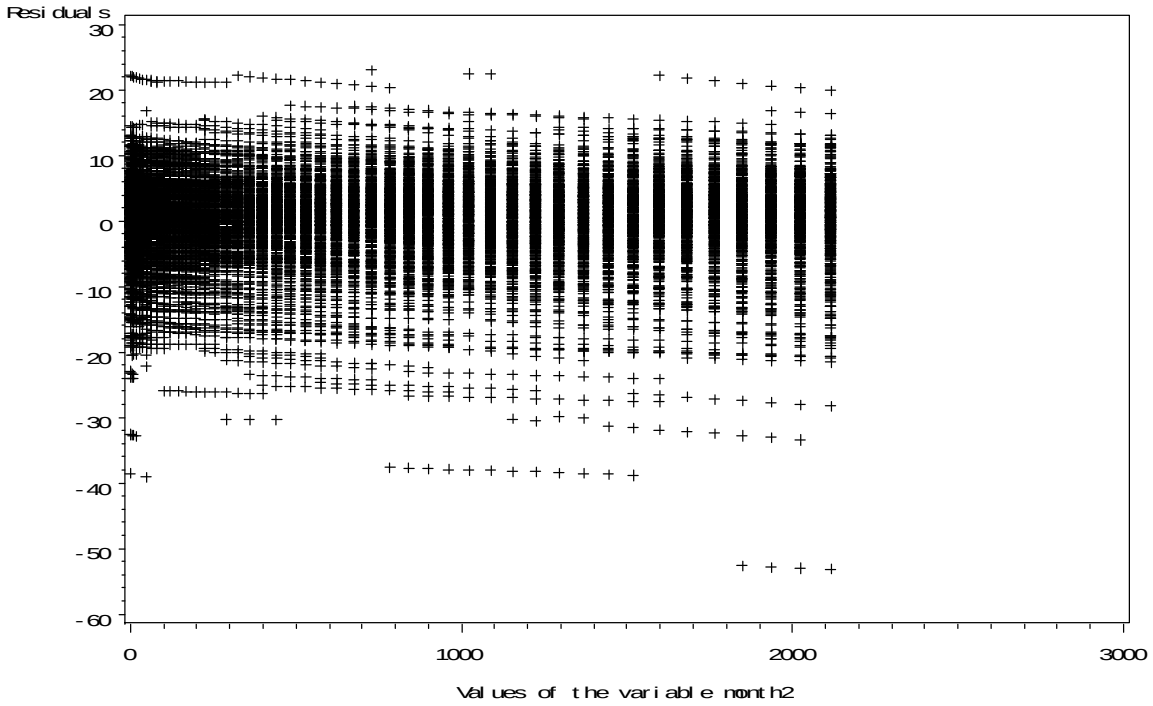

Figure 29 Residual Plot for variable month2

In order to verify observation of equal variances, the Breusch Pagan test is used. For the Breusch Pagan test the null hypothesis is that error terms are homoscedastic.

The Breusch Pagan test is a large sample test that assumes that the residuals are independently and normally distributed. The Test Statistic for the Breusch Pagan test is given by

$$\chi^2_{BP} = \left(\frac{SSR^*}{2}\right) \div \left(\frac{SSE}{n}\right)^2 \qquad \textbf{6}$$

where, SSR^* is the regression sum of squares obtained by regressing squared residuals, e^2 on the independent variables and SSE is the error sum of squares obtained by regressing dependent variable, selling price against the independent variables. The test statistic follows a chi-square distribution with 1 degree of freedom when n is large and the null hypothesis of constant variance holds true. Large value of the test statistic lead to a conclusion that error variance is not constant.

The assumption of independent errors cannot be verified because the data is only for 46 months, which is not enough to show that the errors are correlated. Even then, if we look at the plot of residuals against the variable month shown in figure 15, there is not any systematic pattern in distribution of residuals clearly indicating the independence of residuals and even though the normal probability plot clearly indicates that the residuals are not normally distributed, the Breusch-Pagan test can be still be applied because of the large sample size (approximately 21,000).

Since the p value is 1 for the Breusch Pagan test, we will conclude that error variances are constant.

Therefore, using the residual plots and the result of Breusch Pagan test, it is safe to conclude that the residuals are distributed with equal variances.

Result of Breusch-Pagan Test

Obs	SSR*	SSE	pvalue
1	10089160.38	887732.60	1

Table 8 Output of Breusch-Pagan Test

STEP 7 CHECKING THE ASSUMPTION OF NO AUTOCORRELATION

It is common knowledge that selling price is affected by time. The data given are only for 46 months which is too small a period to capture the effect of time. Even then, this assumption is checked using method of residual plots discussed by Kutner, Nachstshiem and Neter (2001).

A residual plot of residuals against time is used for the detection of correlated error terms. If there is positive correlation, residuals of identical sign occur in clusters. That is there are not enough changes of sign in the pattern of residuals. On the other hand if there is negative correlation, the residuals will alternate signs too rapidly.

The residual plot of residuals against the variable month shown in figure 15 contains bars of more or less equal length, without any systematic pattern. This shows that error terms are unaffected by months i.e. time and hence are not autocorrelated.

STEP 8 CHECKING THE ASSUMPTION OF NO MULTICOLLINEARITY

Multicollinearity is the problem of near dependencies among the independent variables. Due to multicollinearity, the estimated regression coefficients tend to have large sampling variability. Thus, the estimated regression coefficients tend to vary widely from one sample to other sample. Also, the common interpretation of a regression coefficient as measuring the change in the expected value of the dependent variable when the given independent variable is increased by one unit while all other independent are held constant is not fully applicable.

DETECTION OF MULTICOLLINEARITY

One of the most widely used measures for detecting multicollinearity is Variance Inflation Factors (VIF) which is discussed by Kutner, Nachstshiem and Neter (2001).

The Variance inflation factor is given by

$$VIF_i = \frac{1}{1 - R_i^2} \qquad \textbf{7}$$

Where, R_i^2 is the coefficient of determination when X_i regressed on the remaining independent variables. A high VIF indicates a R_i^2 near unity, and hence points to multicollinearity. But according to Belsley, Kuh and Welsch (1980), VIF is not a good measure for detecting Multicollinearity as it is unable to distinguish among several coexisting near dependencies and also there is lack of meaningful boundary for distinguishing between values of VIF that can be considered high and those that can be considered low. They have proposed a new measure called Condition Index. Condition Index is given by

$$K^{th}\text{ Condition Index} = \frac{\lambda_{max}}{\lambda_k} \quad k=0,1,2,\ldots\ldots p-1 \qquad \textbf{8}$$

Where, λ's are the characteristic roots of the X'X matrix.

The logic behind this measure is that if there are one or more near-linear dependencies then one or more of the characteristic roots will be small. Also, there are as many near dependencies among the columns of a data matrix X as there are high condition indexes. Using experiments, they determined that moderate to strong relations are associated with condition indexes of 30 to 100.

So for this analysis, Condition Index is used as the measure for detecting multicollinearity.

Using the Collin option available with PROC REG in SAS, the condition index is calculated and the result is given in table 9. From the results it is clear that multicollinearity is present but is characterized by small condition indices. The maximum condition index is 13.02941 which is well within the limits for small multicollinearity.

Since the multicollinearity is not high, there is no need for remedial measures. This analysis illustrates the most basic way of dealing with multicollinearity, variable selection. In the beginning, the variable 'original price' and 'asking price', which were highly collinear with the variable 'area in square feet' were dropped. This shows that Subject matter knowledge can be an effective tool to deal with problem of Multicollinearity.

Number	Eigenvalue	Condition Index
1	3.56255	1.00000
2	3.34429	1.03212
3	2.95394	1.09820
4	2.57864	1.17540
5	2.14322	1.28928
6	1.15900	1.75323
7	0.35467	3.16934
8	0.29177	3.49429
9	0.17290	4.53926
10	0.11457	5.57618
11	0.07213	7.02769
12	0.06560	7.36947
13	0.05891	7.77669
14	0.05465	8.07418
15	0.02812	11.25505
16	0.02404	12.17311
17	0.02099	**13.02941**

Table 9 Condition indices for the variables

STEP 9 CHECKING THE ASSUMPTION OF NORMALLY DISTRIBUTED RESIDUALS

From the normal probability plot shown in figure 30, it is clear that the residuals are not normally distributed. The normality assumption is even though very important, is not a big issue as all the test statistics are fairly robust against non-normality. Also, the sample size is very large which satisfies the requirement of large sample for the Central Limit theorem to hold true.

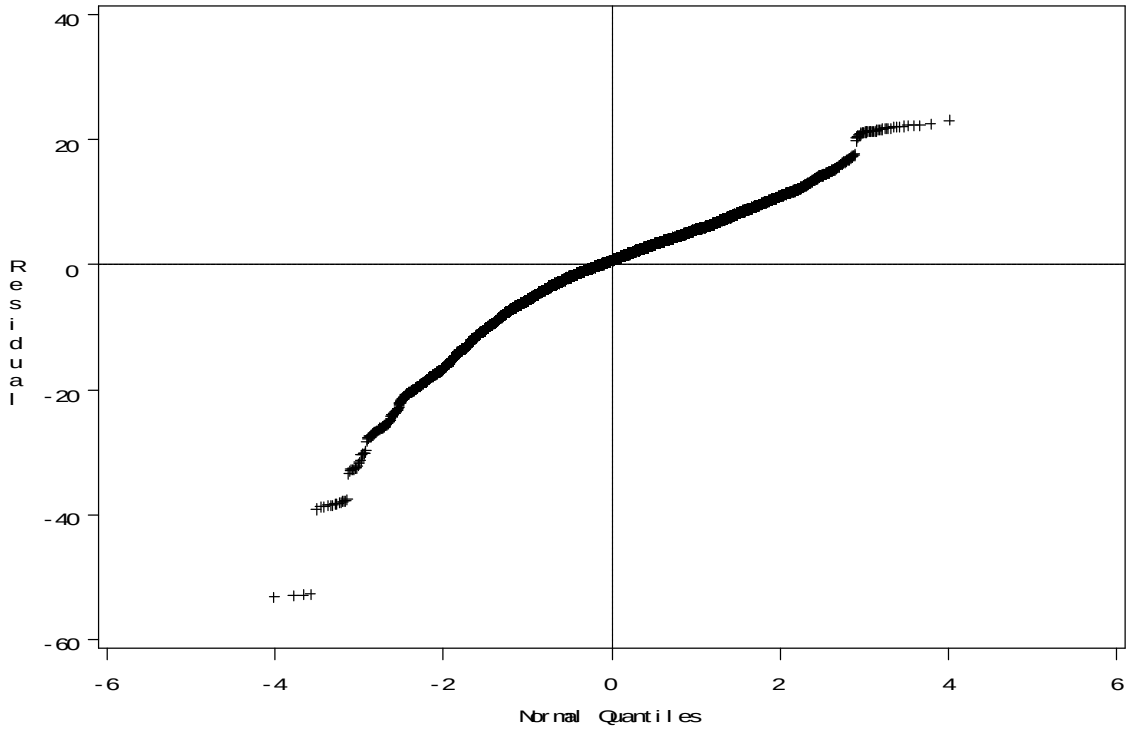

Figure 30 Normal qq plot for the residuals of the final model

STEP 10 DETECTION OF INFLUENTIAL OBSERVATIONS

Influential observations are those observations that have a disproportionate influence on the model coefficients. Thus, this is an undesirable situation as a regression model should be representative of all of the sample observations. Hence, it is important to find these points and assess their impact on the model. If these points genuinely have unusual values then it is important to know about them as they would affect the end use of the regression model.

There are many measures that are used for detecting influential observations, but the four most widely used are Cook's D, DFFITS, DFBETAS and Covariance Ratios discussed by Belsley, Kuh and Welsch (1980). For this analysis, DFFITS and Covariance Ratios are used.

DFFITS measures the change in the fitted value $\hat{Y}_i$ when the ith observation is deleted.

It is given by

$$(\text{DFFITS})_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_i h_{ii}}} \qquad \textbf{9}$$

where, $\hat{Y}_i$ is ith fitted value when all the observations are used,

$\hat{Y}_{i(i)}$ is the ith fitted value when the ith observation is omitted,

MSE_i mean error sum of square when ith observation is omitted, and

h_{ii} is the ith diagonal element of Hat matrix, H where H= $X(X'X)^{-1}X'$.

If for any observation for which $|DFFITS_i| > 2\sqrt{p/n}$ then that particular observation is classified as an influential observation.

COVARIANCE RATIO measures the change in estimated variance of $\hat{\beta}$ when the i[th] observation is removed. Thus, it measures the precision of estimation.

It is given by

$$COVRATIO_i = \frac{\left|\left(X_{(i)}^{'}X_{(i)}\right)^{-1} MSE_i\right|}{\left|\left(X^{'}X\right)^{-1} MSE\right|} \quad \text{,i=1,2……,n} \qquad \textbf{10}$$

If any observation i, $COVRATIO_i > 1+3\,p/n$ or if $COVRATIO_i < 1-3\,p/n$, then the observation should be considered influential.

For DFFITS the cutoff values are (-0.058489, 0.058489)

For Covariance Ratio the cutoff values are (0.99743419, 1.00256582)

Using the above mentioned Statistics the influential observations are identified.

STEP 11 EXAMINATION OF INFLUENTIAL OBSERVATIONS IDENTIFIED IN STEP 10

After careful examination, the following characteristics of the influential observations are observed. There are lots of houses having 4 and 5 bedrooms that are classified as influential observations. Some of the 4 bedroom houses have very small area and small selling price like the house with 1123 square feet area and $32,500 selling price. There are lots of 4 bedroom houses with area like some of the 3 bedroom houses. For instance, house with an area of 1368 square feet and selling price of $55,000. But rest of the 4 and 5 bedroom houses have very large area and high selling price.

There are lots of houses with 3 bedrooms that are classified as influential observations. These houses have unusual figures for different variables. For instance, there are houses with 3 bedrooms but very high Selling price but were on the market for around 60-70 days like the house having 3 bedrooms with selling price $415,000. This house was on the market for only 59 days. The figures for this are unusual as it is too expensive for a 3 bedroom house as there are 4 bedroom houses with lower prices. Also it was sold in just 2 months.

There are also 3 bedrooms houses with reasonable selling price that were on the market for more than 300-400 days like the house having a selling price of $54,000. This house was on market 337 days which is unusual as 3 bedroom houses with lower selling prices should not be on the market for such a long period.

There are 3 bedroom houses with extremely small areas like the house having an area of 914 square feet and selling price of $65,000. This house is highly unusual because for such small area

it has 3 bedrooms yet the selling price is more than the house mentioned above with selling price of $54,000.

There are 2 bedroom houses with very high selling price and large area and were on the market for a short time like the house having an area of 2584 square feet and $105,000 selling price but was on the market for just 67 days.

There are some 2 bedroom houses with selling price of around 8000 like the house having a selling price of $8500 with 672 square feet area which is too small and cheap for a 2 bedroom house. So this house will have large influence on the overall fit.

Using the covariance ratio, similar kinds of influential observations are observed. Most of the influential observations for both the measures have unusual figures. Either the size of the house is too small for number of bedrooms or has very high selling price compared to the number of bedrooms. In some cases, there are houses more expensive than houses with more bedrooms. Some of the houses are very expensive but were sold in just 2 or 3 months which is a short time period compared to houses that were of lesser price but were sold in 7 or 8 months. Thus any house having values for different variables, different from the conventionally expected practice are classified as influential observation. It is expected that more bedrooms mean more area, higher selling price and consequentially more days on the market. The only difference is the number of influential observations identified, for DFFITS it was 1308 and 1811 for Covariance Ratio. Using both DFFITS and Covariance Ratio, the total number of influential observations identified with both the technique were 692. Although the number of influential observations dropped, the nature of the influential observations is the same.

After finding the influential observations the next step would be to examine them for their validity. Even though this is an important step, this is not done in this analysis. Most of the observations that are classified as influential may not be influential because there are many factors other than the ones included in the model, influence the selling price. Factors like the age of the house, proximity to schools and public places etc play important role in determining the Selling price but information related to these factors is not available. Due to this, the influential observations are not dropped from the model.

One way to deal with such observations is to use Robust Regression. Robust Regression is a technique in which individual observations are weighted according to some weight function. This is done to limit the influence of influential observations. Although, Robust Regression is technically a good method, there are some problems associated with it. First problem is the relative efficiency of Robust Regression over the OLS Regression. The relative asymptotic efficiency of Robust Regression over OLS regression is less than 1. Also, in Robust Regression there is no clear cut rule for selecting the different estimators, weight functions and tuning constants used for weighing the observations. Since Robust Regression is still more or less an experimental procedure, it is not used in this analysis.

Due to inability to take a definite step to deal with the issue of influential observations, the model will be unduly affected by influential observations.

STEP 12 VALIDATION OF THE MODEL

The most effective method of validating a regression model with respect to its prediction performance is to collect new data and directly compare the model predictions against them. For this purpose, a new set of data is used. The new data set contains information about the selling price of houses over the period January 2005 to February 2006.

There are many ways of comparing the predictive performance of a model when new set of data is present. One of the ways is to plot the actual values of the dependent variable against the predicted values of the dependent variable. This method is discussed by Freund and Littell (2000). The plot of the actual values against the predicted values for selling price is shown in figure 30. In the plot there is a clear linearly increasing trend which confirms that the model performs well.

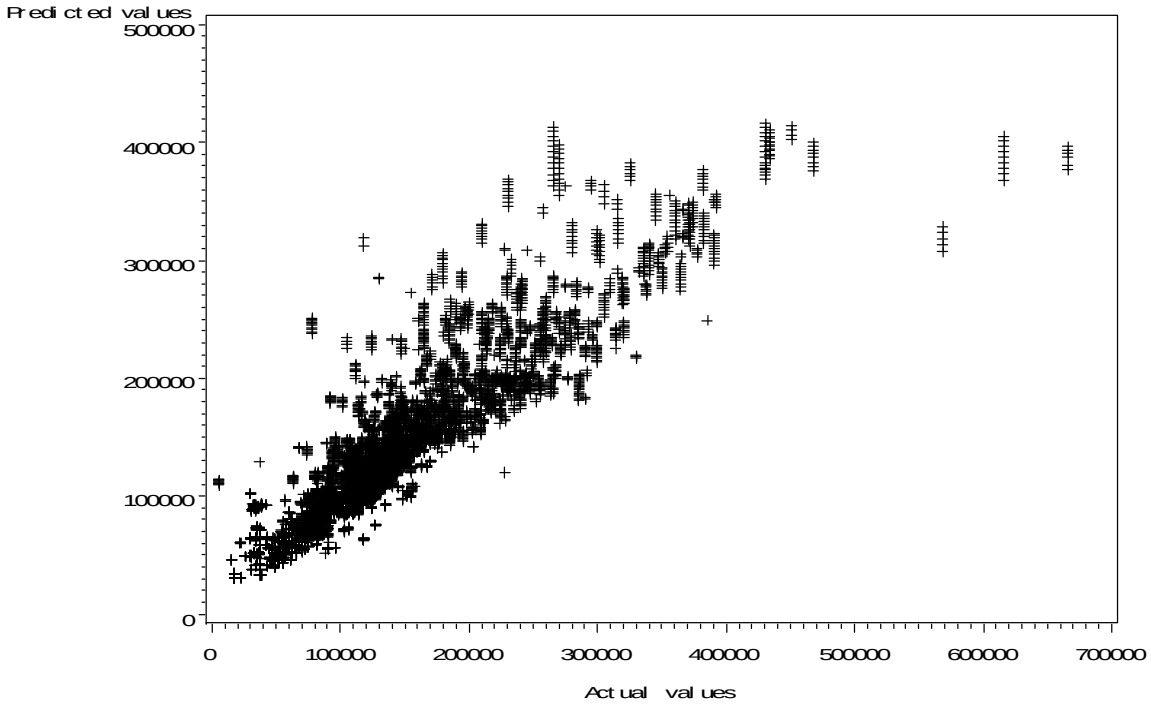

Figure 31 Scatter Plot of predicted values of Selling Price vs. the actual values for the validation data set

There is another method discussed by Montgomery, Peck and Vining (2001) in which the R^2 from the least squares fit is compared with the percentage of variability in the new data explained by the selected model.

$R^2_{prediction}$ is given by

$$R^2_{prediction} = 1 - \frac{PRESS}{SST} \quad \textbf{11}$$

where, PRESS is the predicted error sum of squares and SST is the total sum of squares for the validation dataset.

For the given set of new data, the $R^2_{prediction}$ is

$$1 - \frac{8.5571271E12}{4.7655917E13} = 1 - 0.1796 = 0.8204$$

Root MSE	6.49728	**R-Square**	0.8178
Dependent Mean	77.95607	**Adj R-Sq**	0.8176
Coeff Var	8.33454		

Table 10 R Square obtained from the least squares fit

PRESS
8.5571271E12

Table 11 Predicted error sum of squares for the validation data set

SST
4.7655917E13

Table 12 Total sum of squares for the validation dataset

Since the R Square obtained from the Least Squares fit is very close to the $R^2_{prediction}$, the model is a good fit i.e. the model performs well.

There is one more method suggested by Dr. William D. Warde in which Average Percent Discrepancy is used. Average Percent Discrepancy is the average of the percent absolute difference between the actual value and the predicted values of the dependent variable.

$$\text{Average Percent Discrepancy} = \frac{\sum_{i=1}^{n}\left(\frac{\left|y_i - \hat{y}_i\right|}{y_i}\right)}{n} \qquad \textbf{12}$$

The Average Percent Discrepancy for the validation model is given in table 13

Analysis Variable : APD				
N	**Mean**	**Std Dev**	**Minimum**	**Maximum**
7938	0.0631201	0.0941924	3.4293746E-7	.21558260

Table 13 Output for the Average Percentage Discrepancy

The Average Percent Discrepancy is 6.31% which means on an average there is 6.31% difference between the actual value and the predicted value of Selling Price given by the selected Model. This figure is reasonable given that information like condition of the houses, economic factors were not included in the model and also influential observations were not removed from the data. So, all the three validation methods indicate that the model is a good fit to the available data.

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	26.86566	0.46148	58.22	<.0001
SqFt	SqFt	1	0.03585	0.00031115	115.21	<.0001
DOM	DOM	1	-0.00294	0.00232	-1.27	0.0205
Month	Month	1	-0.09231	0.01628	-5.67	<.0001
X1		1	7.85579	0.35011	22.44	<.0001
X2		1	4.03547	0.47125	8.56	<.0001
X3		1	9.72686	0.67712	14.37	<.0001
SqFtX1		1	-0.00563	0.00020109	-27.99	<.0001
SqFtX2		1	-0.00311	0.00025530	-12.17	<.0001
SqFtX3		1	-0.01281	0.00049295	-25.98	<.0001
SqFtmonth		1	0.00005502	0.00000516	10.66	<.0001
SqFtDOM		1	0.00000529	8.815424E-7	6.00	<.0001
X1DOM		1	0.00558	0.00131	4.27	<.0001
X2DOM		1	0.00167	0.00214	0.78	0.3804
X3DOM		1	0.00104	0.00217	0.48	0.3280
SqFt2		1	-0.00000382	5.937875E-8	-64.35	<.0001
DOM2		1	-0.00001532	0.00000289	-5.31	<.0001
month2		1	0.00171	0.00026201	6.54	<.0001

Table 14 Parameter Estimates for the variables included in the final model

CHAPTER IV

CONCLUSION

One of the problems with the dataset used in this analysis is that it contains duplicate observations. In this dataset a house is listed until it is sold i.e. if a house is sold in four months then that particular house is listed four times in the dataset. Since duplicate observations were not removed from the dataset, the analysis will be biased towards the observations with large values for the variable 'days on the market'. This is because observations with large values for the variable 'days on the market' are on the market for large time period and consequently will be listed in the dataset more number of times compared to the observations with smaller values for the variable 'days on the market'.

In this report, the model is created using well accepted methods for analyzing the data. Although these methods are widely used there are some newer methods like Robust Regression which may be more appropriate but due to lack of literature and thus acceptance, these newer methods were not used.

The data for most part conformed to the requirements of multiple regression model theory. One of the reasons for this can be that the data was collected from a short duration. Due to this there is not much variation in the data. Usually, price of houses change in a cyclical fashion with each cycle lasting for almost a decade. The data collected is for only 4 years, which is not enough to

capture the effect of time. Also Stillwater being a small city may not be very sensitive to upswings and downswing of the economy compared to bigger cities. Hence, the price of houses in Stillwater may not exhibit the volatility associated with house prices of bigger cities.

Since the data is from a short period of time, the results drawn from the analysis should be used with care. The results are meant only to give an idea about the importance of factors involved in the study when the economic factors like interest rates are stable. Also as mentioned before, the model will be affected by the presence of influential observations. Although these factors would limit the usefulness of the model, the main purpose behind the development of model was to come up with a model that is simple to understand and easy to use. Also considering the fact that process of pricing houses is inherently a subjective process; this model will serve the purpose of determining a base price using which the actual price can be calculated.

The equation for predicting the selling price is

$$\hat{Y}_{initial} = 26.86566 + 0.03585*SqFt - 0.00294*DOM - 0.09231*Month + 7.85579*X1 + 4.03547*X2 + 9.72686*X3 - 0.00563*SqFtX1 - 0.00311*SqFtX2 - 0.01281*SqFtX3 + 0.00005502*SqFtmonth + 0.00000529*SqFtDOM + 0.00558*X1DOM + 0.00167*X2DOM + 0.00104*X3DOM -0.00000382*SqFt2 – 0.00000382*DOM2 + 0.00171*month2$$

After obtaining the initial predicted value, it needs to be transformed in terms of dollars by using the following transformation.

Predicted value of selling price = $\left(\hat{Y}_{initial}\right)^{2.667}$

Demonstration of the equation for predicting selling price

The equation is tested for a house with the following values for different independent variables.

selling price = $176090

area in square feet = 2000

number of days on the market = 244

month = 60

location = southwest

X1 = 0 X2 = 0 X3 = 0

$$\hat{Y}_{initial} = 26.86566 + 0.03585*SqFt - 0.00294*DOM - 0.09231*Month$$
$$+ 0.00005502*SqFtmonth + 0.00000529*SqFtDOM - 0.00000382*SqFt2$$
$$- 0.00000382*DOM2 + 0.00171*month2$$

$$= 26.86566 + 0.03585*2000 - 0.00294*244 - 0.09231*60$$
$$+ 0.00005502*2000*60 + 0.00000529*2000*244 - 0.00000382*2000^2$$
$$- 0.00000382*244^2 + 0.00171*60^2$$
$$= 92.14219248$$

Predicted value of the selling price, $\hat{Y} = \left(\hat{Y}_{initial}\right)^{2.667}$

$$= (92.14219248)^{2.667}$$
$$= \$173464.62882982$$
$$\approx \$173464$$

If it is assumed that for the house under consideration

DOM = 0 i.e. the house has just been put on sale. Then,

$\hat{Y}_{initial}$ = 26.86566 + 0.03585*SqFt - 0.09231*Month + 0.00005502*SqFtmonth

- 0.00000382*SqFt2 + 0.00171*month2

= 26.86566 + 0.03585*2000 - 0.09231*60 + 0.00005502*2000*60

- 0.00000382*2000^2 + 0.00171*60^2

= 90.50546

Predicted value of the selling price, $\hat{Y} = \left(\hat{Y}_{initial}\right)^{2.667}$

= $(90.14219248)^{2.667}$

= $165368.06

BIBLIOGRAPHY

Belsley, D. A., Kuh, E. and Welsch, R. E. Regression Diagnostics: Identifying Influential Data and the Sources of Collinearity John Wiley & Sons1980

Chatterjee, S. and Hadi, A. S. Sensitivity Analysis in Linear Regression John Wiley & Sons 1987

Chatterjee, S., Hadi, A. S. and Price, B. Regression Analysis by Example John Wiley & Sons 1999

Freund, R. J. and Littell, R. C. SAS System for Regression SAS Institute Inc. 2000

Kutner, M. H., Nachstshiem, C. J. and Neter, J. Applied Linear Regression Models McGraw Hill 2001

Montgomery, D. C., Peck, E. A. and Vining, G. Introduction to Linear Regression Analysis John Wiley & Sons 2001

Ryan, T. P. Modern Regression Methods John Wiley & Sons 1996

APPENDIX

SAS PROGRAM USED IN THE ANALYSIS

/*Program for creating the indicator variables X1, X2 and X3*/

```
dm ' log ; clear ; output ; clear ; ' ;

libname regress 'C:\Documents and Settings\Yogesh Singh\My Documents\My SAS Files\9.1';


data regress.transmonth;
set regress.Monthlydata;
if Area='NE' then X1=1; else X1=0;
if Area='NW' then X2=1; else X2=0;
if Area='SE' then X3=1; else X3=0;

proc print data=regress.transmonth;
```

/* program for preliminary data analysis*/

```
proc univariate data=regress.transmonth normal;
var SqFt SellPrice;
qqplot/href=0 vref=0;

proc corr data=regress.transmonth spearman;
var SellPrice SqFt DOM month X1 X2 X3;
```

/*Program for creating the added variable plots*/

```
data regress.reviseddata1;
set regress.transmonth;

proc reg data=regress.reviseddata1;
model SellPrice= DOM SqFt X1 X2 X3 month/partial;
```

/* Program for selecting the independent variables*/

```
proc reg data=regress.reviseddata;
model SellPrice=SqFt DOM month X1 X2 X3 / selection = cp aic bic;
```

/* program for selecting the Interaction and Quadratic terms*/

```
data regress.reviseddata1;
set regress.transmonth;
SqFtX1=SqFt*X1;
SqFtX2=SqFt*X2;
SqFtX3=SqFt*X3;
```

```
SqFtmonth=SqFt*month;
SqFtDOM=SqFt*DOM;
X1X2=X1*X2;
X1X3=X1*X3;
X1month=X1*month;
X1DOM=X1*DOM;
X2X3=X2*X3;
X2month=X2*month;
X2DOM=X2*DOM;
X3month=X3*month;
X3DOM=X3*DOM;
SqFt2=SqFt**2;
DOM2=DOM**2;
month2=month**2;

proc reg data=regress.reviseddata1;
model SellPrice=SqFt DOM month X1 X2 X3 SqFtX1 SqFtX2 SqFtX3 SqFtmonth
          SqFtDOM X1X2 X1X3 X1month X1DOM X2X3 X2month X2DOM
          X3month X3DOM SqFt2 DOM2 month2 /selection =cp aic bic;

/* program for creating Residual Plots for the model containing
  selected variables */

proc reg data=regress.reviseddata1;
model SellPrice=SqFt DOM Month X1 X2 X3 SqFtX1 SqFtX2 SqFtX3 SqFtmonth
          SqFtDOM X1DOM X2DOM X3DOM SqFt2 DOM2 month2;

plot r.*(p. SqFt DOM Month X1 X2 X3 SqFtX1 SqFtX2 SqFtX3 SqFtmonth
   SqFtDOM X1DOM X2DOM X3DOM SqFt2 DOM2 month2);

/* program for checking the correct form of the dependent variable
  selling price */

proc transreg data=regress.reviseddata1;
model boxcox(SellPrice)=identity(SqFt DOM Month X1 X2 X3 SqFtX1 SqFtX2
                SqFtX3 SqFtmonth SqFtDOM X1DOM X2DOM X3DOM
                SqFt2 DOM2 month2);

data regress.reviseddata2;
set regress.reviseddata1;
SellPrice=SellPrice**0.375;

proc transreg data=regress.reviseddata2;
model boxcox(SellPrice)=identity(SqFt DOM Month X1 X2 X3 SqFtX1 SqFtX2
                SqFtX3 SqFtmonth SqFtDOM X1DOM X2DOM X3DOM
                SqFt2 DOM2 month2);

/* program for conducting Breusch Pagan test */

proc reg data=regress.reviseddata2;
model SellPrice=SqFt X1 X2 X3 SqFtX1 SqFtX2 SqFtX3 SqFtmonth X1month
```

```
            X1DOM X2month X2DOM X3DOM X3month month2;

ods output ANOVA = temp1;
output out = temp r=e;

run;

ods listing;
data temp1;
 set temp1;
 if source='Error' then call symput ('sse' , ss);
run;
data temp2;
 set temp nobs=total;
 e2 = e**2;
run;

ods listing close;

proc reg data = temp2;
 model e2 = SqFt X1 X2 X3 SqFtX1 SqFtX2 SqFtX3 SqFtmonth X1month X1DOM
          X2month X2DOM X3DOM X3month month2;
 ods output anova = temp3;

run;

ods listing;

data temp3;
 set temp3;
 if source='Model' then call symput ('ssr', ss);

run;

data tempf;
 set temp3;
pvalue = probchi( (&ssr/2)/(&sse/21046)**2 , 1 );
 ssr = &ssr;
 sse = &sse;

run;

proc print data = tempf ;
var ssr sse pvalue;

/* program for creating residual plots for model containing transformed
   selling price and for obtaining collinearity diagnostics*/

proc reg data=regress.reviseddata2;
model SellPrice =SqFt DOM Month X1 X2 X3 SqFtX1 SqFtX2 SqFtX3 SqFtmonth
            SqFtDOM X1DOM X2DOM X3DOM SqFt2 DOM2 month2/collin;
```

```
plot r.*(p. SqFt DOM Month X1 X2 X3 SqFtX1 SqFtX2 SqFtX3 SqFtmonth
    SqFtDOM X1DOM X2DOM X3DOM SqFt2 DOM2 month2);

/* program for obtaining influential observations*/

proc reg data=regress.reviseddata2;
model SellPrice =SqFt DOM Month X1 X2 X3 SqFtX1 SqFtX2 SqFtX3 SqFtmonth
            SqFtDOM X1DOM X2DOM X3DOM SqFt2 DOM2 month2/influence;
output out=influentialdata COOKD=cookd DFFITS=dffit covratio=covratio;

proc print data=influentialdata;

/* program for obtaining predicted values of the variable Selling Price
  from the validation data set and for creating scatter plot of actual
  values of the variable selling price against predicted values*/



proc reg data = regress.reviseddata2 outest=shortest;
model SellPrice =SqFt DOM Month X1 X2 X3 SqFtX1 SqFtX2 SqFtX3 SqFtmonth
            SqFtDOM X1DOM X2DOM X3DOM SqFt2 DOM2 month2;

proc print data=shortest;

proc score data = transverify score = shortest out=predtransverify type=parms predict;
var SellPrice SqFt DOM Month X1 X2 X3 SqFtX1 SqFtX2 SqFtX3 SqFtmonth
   SqFtDOM X1DOM X2DOM X3DOM SqFt2 DOM2 month2;

proc plot data = predtransverify;
plot model1*SellPrice;

/* program for calculating the predicted R Square */

data predtransverify1;
set predtransverify;
SellPrice=SellPrice**2.67;
errorsquare=(SellPrice-model1)**2;

proc means data=predtransverify1;
var errorsquare SellPrice;

data predtransverify2;
set predtransverify1;
devSellPrice=(SellPrice-140965.12)**2;

proc means data=predtransverify2;
var devSellPrice;

/* program for obtaining Average Percent Discrepancy */

data predtransverify3;
```

```
set predtransverify;
diff=(abs(SellPrice-model1)/SellPrice);

proc means data = predtransverify3;
var diff;

run;
quit;
```

VITA

YOGESH SINGH

Candidate for the Degree of
Master of Science

Report: MODEL FOR FORECASTING PRICE OF HOUSES IN CITY OF STILLWATER, OK.

Major Field: Statistics

Biographical:

Education: Received Bachelor of Science degree in Statistics from M.S. University of Baroda, Gujarat, India in May 2002; received Master of Science in Statistics from South Gujarat University, Gujarat, India in May 2004. Completed the requirement for the Master of Science in Statistics at Oklahoma State University in May 2006.

Experience: Employed by Oklahoma State University as a Graduate Teaching Assistant in the Department of Statistics (Fall 2004 – Spring 2006)

ABSTRACT

Name: Yogesh Singh Date of Degree: May 2006

Institution: Oklahoma State University Location: Stillwater, Oklahoma

Pages in Study: 70 Candidate for the Degree of Master of Science

Major Field: Statistics

Scope and Method of Study: The purpose of this study was to construct a model for forecasting selling price of houses in city of Stillwater, Oklahoma which estimates the selling price with reasonable degree of accuracy and which is easy to understand and apply. The data for this study was collected from local real estate agents and Municipal Corporation in Stillwater. The data was collected from 1988 to 2004. The analysis was done using method of Multiple Regression. The validation of the model was done using data collected from January 2005 to February 2006.

Findings and Conclusions: The predicted values obtained from the forecasting model, on an average were 6% different from the actual values of the selling prices for the validation data set. Since the data related to factors like age and condition of the house and economic factors like the interest rate were not included, the 6% average difference between the actual values and the predicted values is reasonable. The data used in this study is from a small period of time and hence results obtained from this study should be used with care.

Advisor's Approval: Dr. William D. Warde