International Journal of Research
Available at https://edupediapublications.org/journals

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 03 Issue 14
October 2016

Estimation of House Selling Price by Multiple Regression Analysis Using SAS Software

Dheyaa Mohammed Naeem

M.Sc Applied Statistic's, Department of Statistics,
University College of Science, Osmania University, India
Email Id: dheyaamohammed24@gmail.com

Abstract:

Regression analysis is one of the most widely used statistical techniques. Today, regression analysis is applied in the social sciences, medical research, economics, agriculture, biology, meteorology, marketing, retail, insurance and many other areas of academic and applied science. It is not only suited to suggesting decisions as to whether or not a relationship between two variables exists. It goes beyond this decision making and provides a different type of precise statement. Regression analysis specifies a functional form for the relationship between the variables under study that allows one to estimate the degree of change in the dependent variable that goes hand in hand with changes in the independent variable. At the same time, regression analysis allows one to make statements about how certain one can be about the predicted change in Y that is associated with the observed change in X.

The main objective of the present study is to investigate factors that contribute significantly to estimate the selling price of a house in a locality. The dependent variable is Average house selling price, the multiple regression analysis applied for exploring the factors affecting the house selling price. The power of SAS in analysing data patterns and developing such models is also demonstrated, appropriateand relevant portions of SAS code are included where possible.

1. Introduction to Regression:

In statistical modelling, regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modelling and analysing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors'). More specifically, regression analysis helps one to understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed. Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variables – that is, the average value of the dependent variable when the independent variables are fixed. Less commonly, the focus is on a quintile, or other location parameter of the conditional distribution of the dependent variable given the independent variables. In all cases, the estimation target is a function of the independent variables called the regression function. In regression analysis, it is also of interest to characterize the variation of the dependent variable around the regression function which can be described by a probability distribution.

Regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables. However this can lead to illusions or false relationships, so caution is advisable [8] for example, correlation does not imply causation.

Many techniques for carrying out regression analysis have been developed. Familiar methods such as linear regression and ordinary least squares regression are parametric, in that the regression function is defined in terms of a finite number of unknown parameters that are estimated from the data. Nonparametric regression refers to techniques that allow the regression function to lie in a specified set of functions, which may be infinite-dimensional.

The performance of regression analysis methods in practice depends on the form of the data generating process, and how it relates to the regression approach being used. Since the true form of the data-generating process is generally not known, regression analysis often depends to some extent on making assumptions about this process. These assumptions are sometimes testable if a sufficient quantity of data is available. Regression models for prediction are often useful even when the assumptions are moderately violated, although they may not perform optimally. However, in many applications, especially with small effects or questions of causality based on observational data, regression methods can give misleading results [9].

In a narrower sense, regression may refer specifically to the estimation of continuous response variables, as opposed to the discrete response variables used in classification [11]. The case of a continuous output variable may be more specifically referred to as metric regression to distinguish it from related problems.

Today, regression analysis is applied in the social sciences, medical research, economics, agriculture, biology, meteorology, marketing, retail, banking, insurance and many other areas of academic and applied science. Reasons for the outstanding role that regression analysis plays include that its concepts are easily understood, and it is implemented in virtually every all-purpose statistical computing package, and can therefore be readily applied to the data at hand. Moreover, regression analysis lies at the heart of a wide range of more recently developed statistical techniques such as the class of generalized linear models [1]. Hence a sound understanding of regression analysis is fundamental to developing one's understanding of modern applied statistics.

2. Simple Linear Regression:

Simple linear regression examines the linear relationship between two continuous variables: one response (y) and one predictor (x). When the two variables are related, it is possible to predict a response value from a predictor value with better than chance accuracy.

Regression provides the line that "best" fits the data. This line can then be used to:

1) Examine how the response variable changes as the predictor variable changes.
2) Predict the value of a response variable (y) for any predictor variable (x).

3. Multiple Linear Regression:

Multiple linear regression examines the linear relationships between one continuous response and two or more predictors.

If the number of predictors is large, then before fitting a regression model with all the predictors, you should use stepwise or best subsets model-selection techniques to screen out predictors not associated with the responses.

4. Ordinary Least Squares Regression:

In ordinary least squares (OLS) regression, the estimated equation is calculated by determining the equation that minimizes the sum of the squared distances between the sample's data points and the values predicted by the equation.

4.1Assumptions for OLS Regression:

OLS regression provides the most precise, unbiased estimates only when the following assumptions are met:

1) Residuals have a mean of zero. Inclusion of a constant in the model will force the mean to equal zero.
2) All predictors are uncorrelated with the residuals.
3) Residuals are not correlated with each other (serial correlation).
4) Residuals have a constant variance.
5) No predictor variable is perfectly correlated (r=1) with a different predictor variable. It is best to avoid imperfectly high correlations (multicollinearity) as well.
6) Residuals are normally distributed.

Because OLS regression will provide the best estimates only when all the assumptions are met, it is very important to test them. Common approaches include examining residual plots, using lack of fit tests, and viewing the correlation between predictors using the Variance Inflation Factor (VIF).

4.2 Multicollinearity Problem:

Multicollinearity is states of very high inter correlations or inter associations among the independent variables. It is therefore a type of disturbance in the data, and if present in the data the statistical inferences made about the data may not be reliable.

If the individual outcome of a statistic is not significant but the overall outcome of the statistic is significant. In this instance, the researcher might get a mix of significant and insignificant results that show the presence of multicollinearity. Suppose the researcher, after dividing the sample into two parts, finds that the coefficients of the sample differ drastically. This indicates the presence of multicollinearity. This means that the coefficients are unstable due to the presence of multicollinearity. Suppose the researcher observes drastic change in the model by simply adding or dropping some variable. This also indicates that multicollinearity is present in the data.

Multicollinearity can also be detected with the help of tolerance and its reciprocal, called variance inflation factor (VIF). If the value of tolerance is less than 0.2 or 0.1 and, simultaneously, the value of VIF 10 and above, then the multicollinearity is problematic.

Reasons for Multicollinearity:

- It is caused by an inaccurate use of dummy variables.
- It is caused by the inclusion of a variable which is computed from other variables in the data set.
- Multicollinearity can also result from the repetition of the same kind of variable.
- Generally occurs when the variables are highly correlated to each other.

Problems with Multicollinearity:

- The partial regression coefficient due to multicollinearity may not be estimated precisely. The standard errors are likely to be high.
- Multicollinearity results in a change in the signs as well as in the magnitudes of the partial regression coefficients from one sample to another sample.
- Multicollinearity makes it tedious to assess the relative importance of the independent variables in explaining the variation caused by the dependent variable.

5.Diagnostics of Multiple Regression model:

Std Error: if the regression were performed repeatedly on different datasets (that contained the same variables), this would represent the standard deviation of the estimated coefficients.

t-Ratio: the coefficient divided by the standard error, which tells us how large the coefficient is relative to how much it varies in repeated sampling. If the coefficient varies a lot in repeated sampling, then its t-statistic will be smaller, and if it varies little in repeated sampling, then its t-statistic will be larger.

Prob>|t|: The p-value is the result of the test of the following null hypothesis: in repeated sampling, the mean of the estimated coefficient is zero. E.g., if p = 0.001, the probability of observing an estimate of β that is at least as extreme as the observed estimate is 0.001, if the true value of β is zero.

In general, a p-value less than some threshold ⍰, like 0.05 or 0.01, will mean that the coefficient is “statistically significant”.

Confidence interval: the 95% confidence interval is the set of values that lie within 1.96 standard deviations of the estimatedβ.

R square: This statistic represents the proportion of variation in the dependent variable that is explained by the model (the remainder represents the proportion that is present in the error). It is also the square of the correlation coefficient.

7.Model Selection Methods:

Selection, on the other hand, allows for the construction of an optimal regression equation along with investigation into specific predictor variables. The aim of selection is to reduce the set of predictor variables to those that are necessary and account for nearly as much of the variance as is accounted for by the total set. In essence, selection helps to determine the level of importance of each predictor variable. It also assists in assessing the effects once the other predictor variables are statistically eliminated. The circumstances of the study, along with the nature of the research questions guide the selection of predictor variables.

Essentially, the multiple regression selection process enables the researcher to obtain a reduced set of variables from a larger set of predictors, eliminating unnecessary predictors, simplifying data, and enhancing predictive accuracy. Two criterions are used to achieve the best set of predictors; these include meaningfulness to the situation and statistical significance. By entering variables into the equation in a given order, confounding variables can be investigated and variables that are highly correlated can be combined into blocks.

7.1 Stepwise Selection:

This form of regression is used when we deal with multiple independent variables. In this technique, the selection of independent variables is done with the help of an automatic process, which involves no human intervention.

This feat is achieved by observing statistical values like R-square, t-stats and AIC metric to discern significant variables. Stepwise regression basically fits the regression model by adding/dropping covariates one at a time based on a specified criterion. Some of the most commonly used Stepwise regression methods are listed below:

1) Standard stepwise regression does two things. It adds and removes predictors as needed for each step.
2) Forward selection starts with most significant predictor in the model and adds variable for each step.
3) Backward elimination starts with all predictors in the model and removes the least significant variable for each step.

The aim of this modelling technique is to maximize the prediction power with minimum number of predictor variables. It is one of the methods to handle higher dimensionality of data set.

Stepwise regression procedures are used in data mining, but are controversial. Several points of criticism have been made.

The tests themselves are biased, since they are based on the same data [12]. Wilkinson and Dallal[14] computed percentage points of the multiple correlation coefficients by simulation and showed that a final regression obtained by forward selection, said by the F-procedure to be significant at 0.1%, was in fact only significant at 5%.

When estimating the degrees of freedom, the number of the candidate independent variables from the best fit selected is smaller than the total number of final model variables, causing the fit to appear better than it is when adjusting the r2 value for the number of degrees of freedom. It is important to consider how many degrees of freedom have been used in the entire model, not just count the number of independent variables in the resulting fit.

Models that are created may be over-simplifications of the real models of the data.

7.2 Forward Selection:

Forward Selection chooses a subset of the predictor variables for the final model. We can do forward stepwise in context of linear regression whether n is less than p or n is greater than p. It is a very attractive approach, because it's both tractable and it gives a good sequence of models.

- Start with a null model. The null model has no predictors, just one intercept (The mean over Y).
- Fit p simple linear regression models, each with one of the variables in and the intercept. So basically, you just search through all the single-variable models the best one (the one that results in the lowest residual sum of squares). You pick and fix this one in the model.
- Now search through the remaining p minus 1 variable and find out which variable should be added to the current model to best improve the residual sum of squares.
- Continue until some stopping rule is satisfied, for example when all remaining variables have a p-value above some threshold.

7.3 Backward Selection:

Unlike forward stepwise selection, it begins with the full least squares model containing all p predictors, and then iteratively removes the least useful predictor, one-at-a-time.

In order to be able to perform backward selection, we need to be in a situation where we have more observations than variables; because we can do least squares regression when n is greater than p. If p is greater than n, we cannot fit a least squares model. It's not even defined.

- Start with all variables in the model.
- Remove the variable with the largest p-value, that is, the variable that is the least statistically significant.
- The new (p - 1)-variable model is t, and the variable with the largest p-value is removed.
- Continue until a stopping rule is reached. For instance, we may stop when all remaining variables have a significant p-value defined by some significance threshold.

10. Merits and Demerits:

10.1 Merits:

- The estimates of the unknown parameters obtained from linear least squares regression are the optimal.
- Estimates from a broad class of possible parameter estimates under the usual assumptions are used for process modelling.
- It uses data very efficiently. Good results can be obtained with relatively small data sets.
- The theory associated with linear regression is well-understood and allows for construction of different types of easily interpretable statistical intervals for predictions, calibrations, and optimizations.

10.2 Demerits:

- Outputs of regression can lie outside of the range [0,1].
- It has limitations in the shapes that linear models can assume over long ranges
- The extrapolation properties will be possibly poor
- It is very sensitive to outliers
- It often gives optimal estimates of the unknown parameters.

11. Introduction to Data and Variables:

Our data represents house selling prices in a locality. In this data we have 550 observations with 12 variables. Performed a study is to investigate factors that contribute significantly to estimate the selling price of a house. The dependent variable is Average house selling price, the multiple regression analysis applied for exploring the factors affecting the house selling price. The entire analysis was done using SAS 9.2 software[19].

Variable Description:

Variable	Description
LSP_D	Local selling prices, in hundreds of dollars
NBR	Number of bathrooms
AS_TSF	Area of the site in thousands of square feet

SLS_TSF	Size of the living space in thousands of square feet
NG	Number of garages
NR	Number of rooms
NB	Number of bedrooms
Age Years	Age in years
Constuction_Type	Construction type
Architecture Type	Architecture type
NFP	Number of fire places
Selling Price	Selling price

12 .Data Analysis:

Below is SAS code to read the data into SAS environment.

```
DataHouse Price;
InputSno   LSP_D   NBR   AS_TSF   SLS_TSF   NG   NR   NB
Age_YearsConstuction_TypeArchitecture_Type NFP Selling Price;
Cards;

run;
```

Calculation of Correlation Matrix and Summary Statistics for each variable

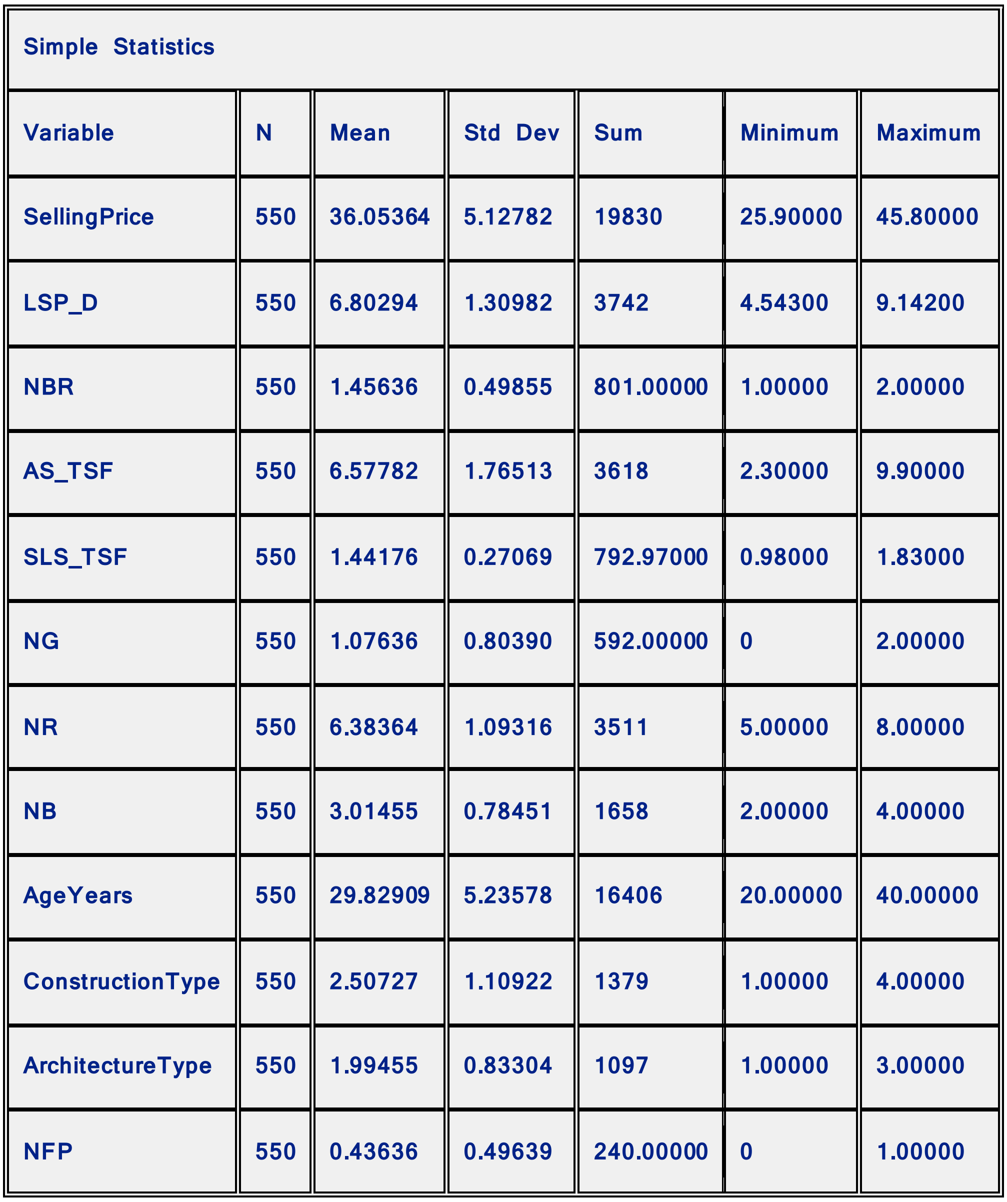
```
/* Computing Summary Statistics and Correlation Matrix */
odshtml;
odsgraphicson;
odsoutputSimpleStats=sspearsoncorr=co;
proccorrdata=HousePricenoprob;
varSellingPrice LSP_D NBR AS_TSF   SLS_TSF   NG   NR   NB
Age_YearsConstuction_TypeArchitecture_Type NFP;
run;
odshtmlclose;
odsgraphicsoff;

datacorMat;
set co;
format _numeric_ 5.2;
run;

dataSimpleStat;
```

```
setss;
format _numeric_ 5.2;
run;
```

Summary Statistics:



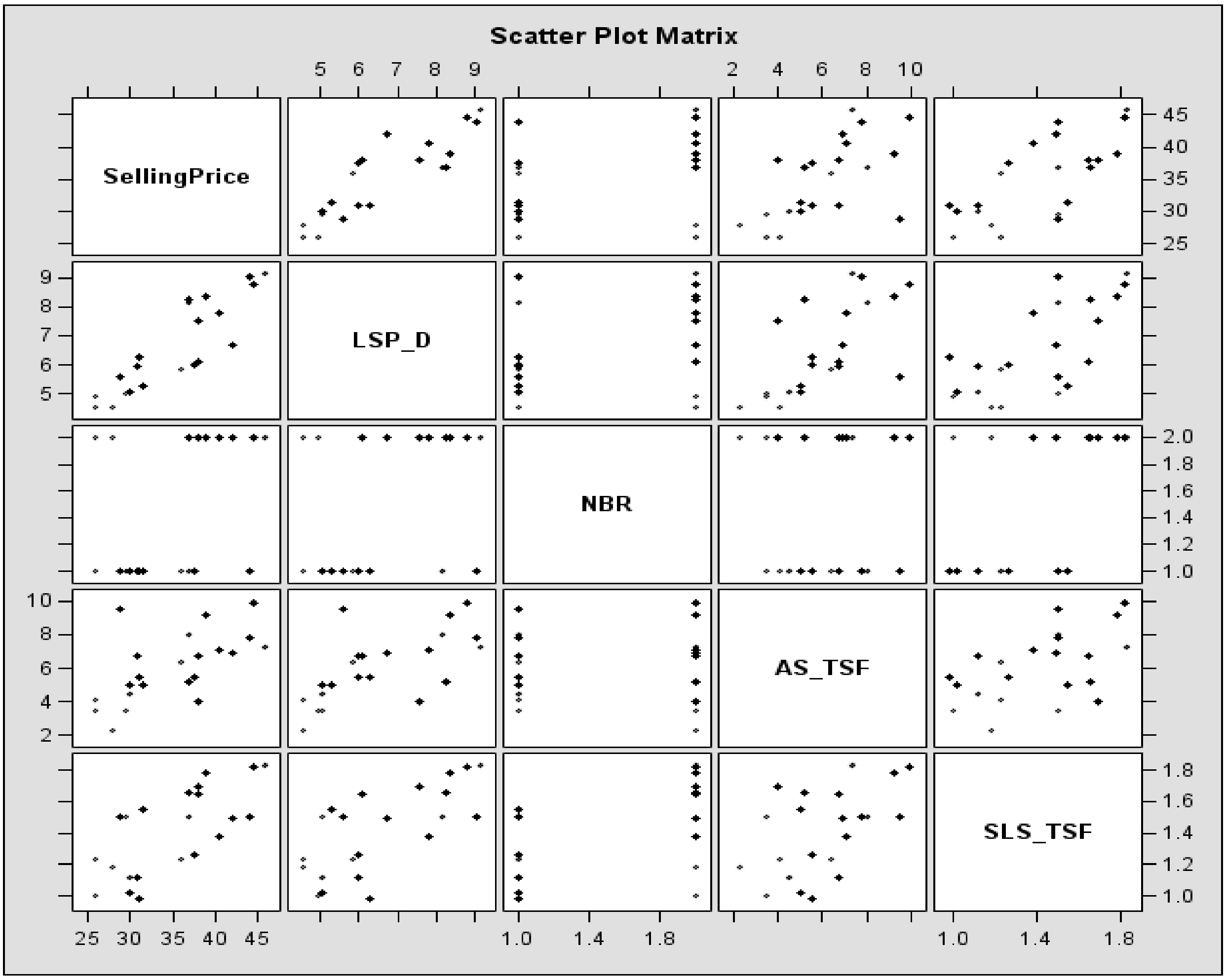
Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
SellingPrice	550	36.05364	5.12782	19830	25.90000	45.80000
LSP_D	550	6.80294	1.30982	3742	4.54300	9.14200
NBR	550	1.45636	0.49855	801.00000	1.00000	2.00000
AS_TSF	550	6.57782	1.76513	3618	2.30000	9.90000
SLS_TSF	550	1.44176	0.27069	792.97000	0.98000	1.83000
NG	550	1.07636	0.80390	592.00000	0	2.00000
NR	550	6.38364	1.09316	3511	5.00000	8.00000
NB	550	3.01455	0.78451	1658	2.00000	4.00000
AgeYears	550	29.82909	5.23578	16406	20.00000	40.00000
ConstructionType	550	2.50727	1.10922	1379	1.00000	4.00000
ArchitectureType	550	1.99455	0.83304	1097	1.00000	3.00000
NFP	550	0.43636	0.49639	240.00000	0	1.00000

Correlation Matrix:

Pearson Correlation Coefficients, N = 550

Variable	SellingPrice	LSP_D	NBR	AS_TSF	SLS_TSF	NG	NR	NB	Age_Years	Constuction Type	Architecture Type	NFP
SellingPrice	1	0.82	0.66	0.4	0.59	0.42	0.59	0.44	-0.28	0.07	0.03	0.25
LSP_D	0.82	1	0.59	0.46	0.55	0.37	0.66	0.51	-0.3	-0.06	0.09	0.06
NBR	0.66	0.59	1	0.23	0.66	0.26	0.54	0.49	-0.24	0	-0.08	0.16
AS_TSF	0.4	0.46	0.23	1	0.41	0.12	0.47	0.37	0.02	-0.2	-0.02	0.24
SLS_TSF	0.59	0.55	0.66	0.41	1	0.21	0.71	0.64	-0.2	-0.09	0.04	0.16
NG	0.42	0.37	0.26	0.12	0.21	1	0.41	0.35	-0.01	0.01	0.02	0.13
NR	0.59	0.66	0.54	0.47	0.71	0.41	1	0.63	-0.13	-0.01	0.02	0.26
NB	0.44	0.51	0.49	0.37	0.64	0.35	0.63	1	-0.02	-0.01	0.06	0.12
Age_Years	-0.28	-0.3	-0.24	0.02	-0.2	-0.01	-0.13	-0.02	1	0.22	-0.14	0.11
Constuction_Type	0.07	-0.06	0	-0.2	-0.09	0.01	-0.01	-0.01	0.22	1	-0.02	0.09
Architecture_Type	0.03	0.09	-0.08	-0.02	0.04	0.02	0.02	0.06	-0.14	-0.02	1	-0.12
NFP	0.25	0.06	0.16	0.24	0.16	0.13	0.26	0.12	0.11	0.09	-0.12	1

Scatter plot:



Below is SAS code to get frequency for each independent variable.

```
procfreqdata=HousePrice;
```

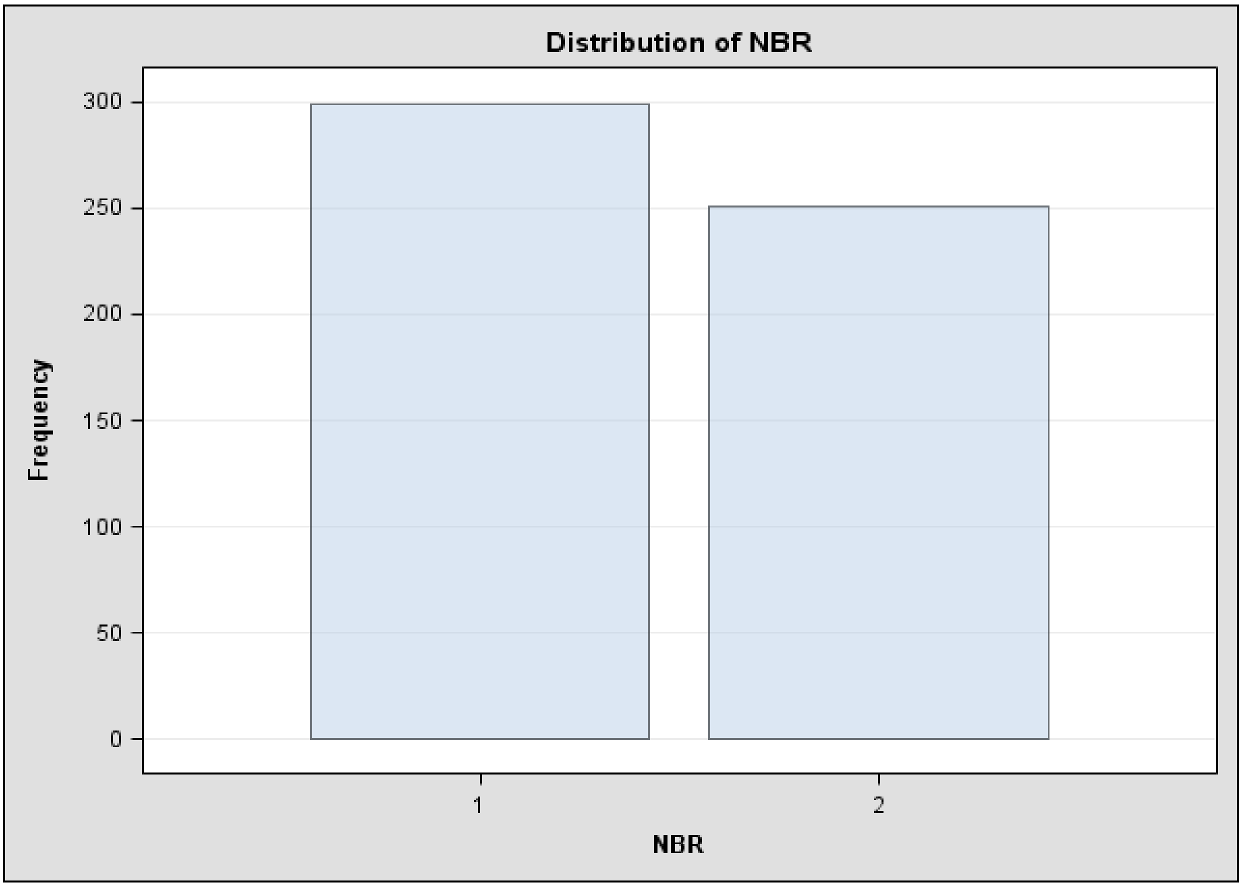
```
tables NBR NG NRNBConstuction_TypeArchitecture_TypeNFP;
run;
```

Variable: NBR (Number of bathrooms).

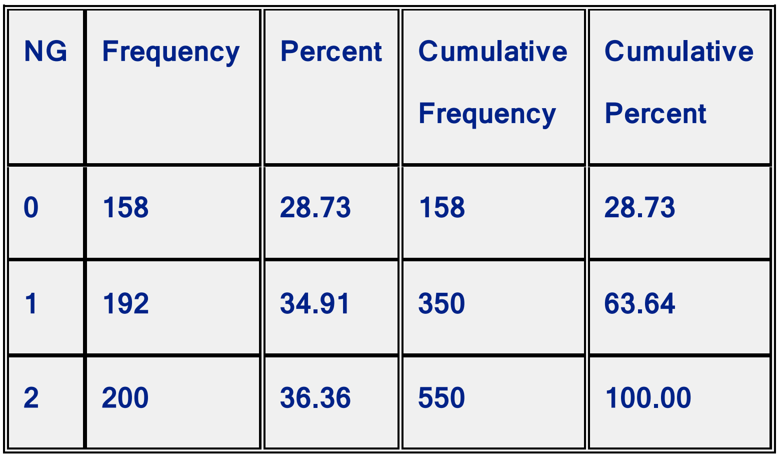
The SAS System

The FREQ Procedure

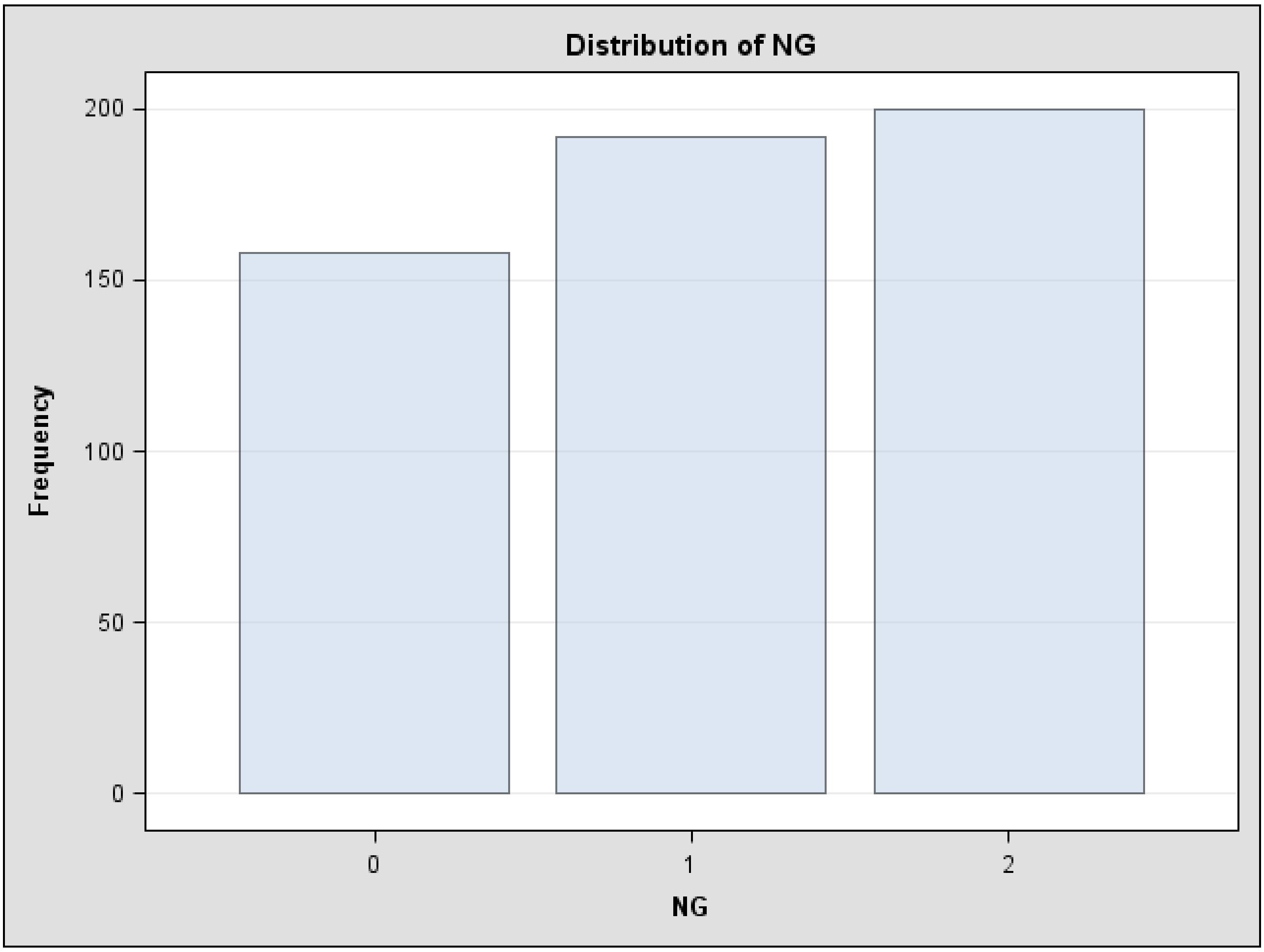
NBR	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	299	54.36	299	54.36
2	251	45.64	550	100.00



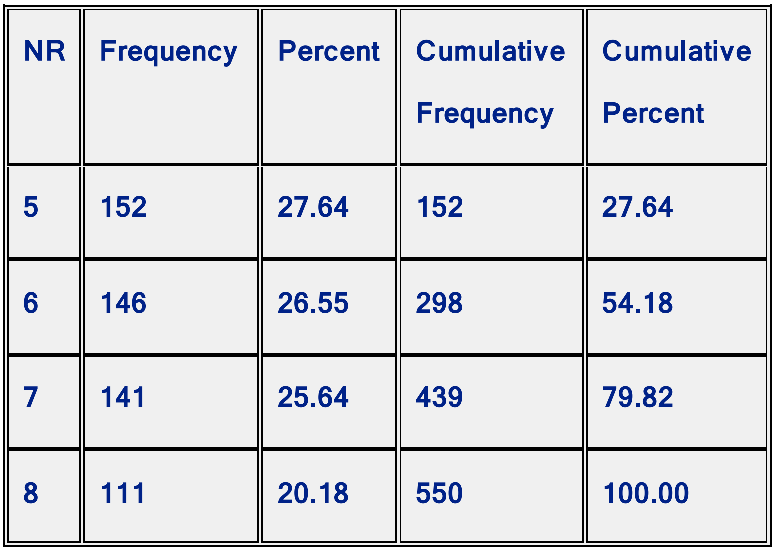
Variable: NG (Number of garages)



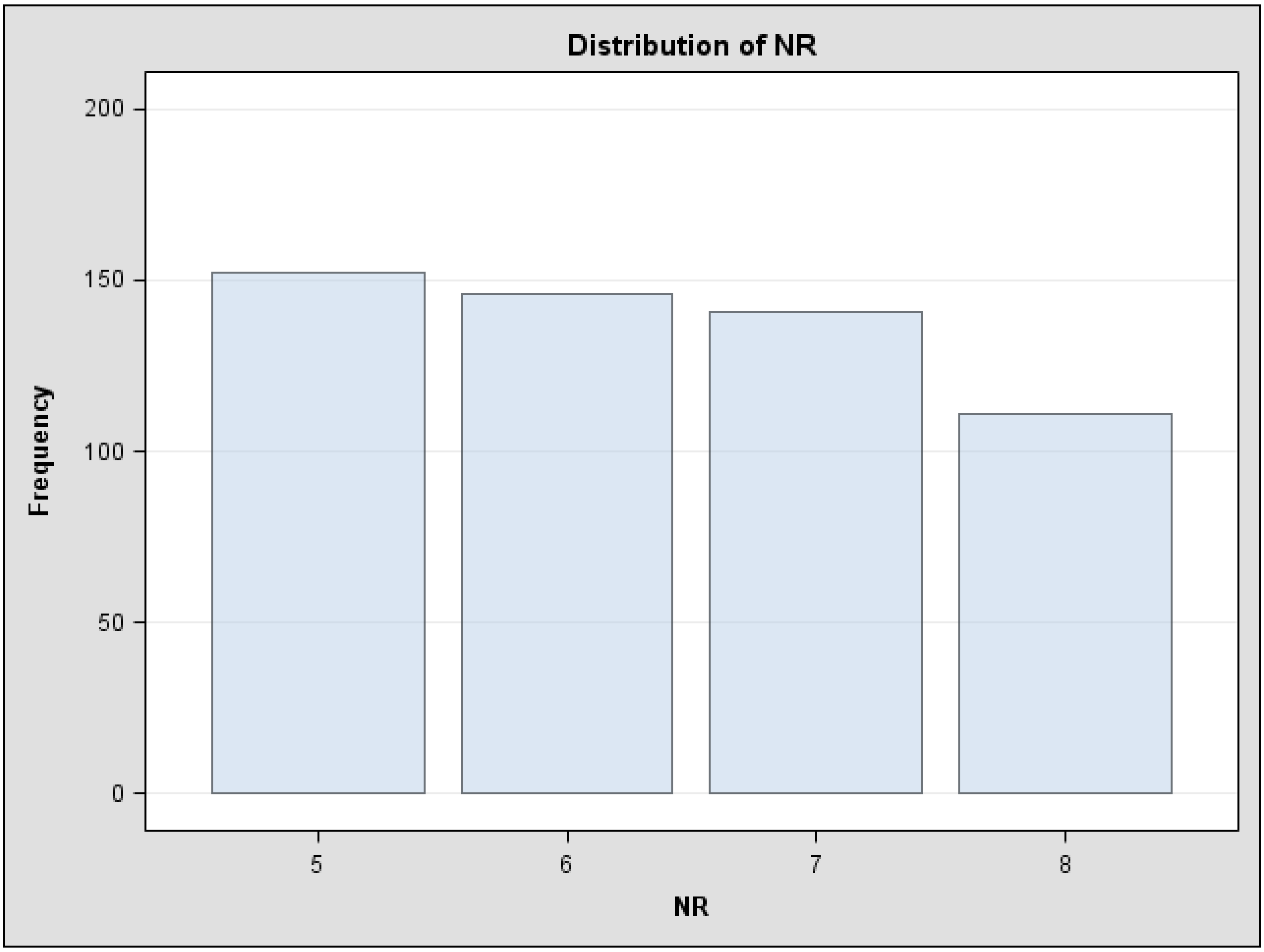
NG	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	158	28.73	158	28.73
1	192	34.91	350	63.64
2	200	36.36	550	100.00



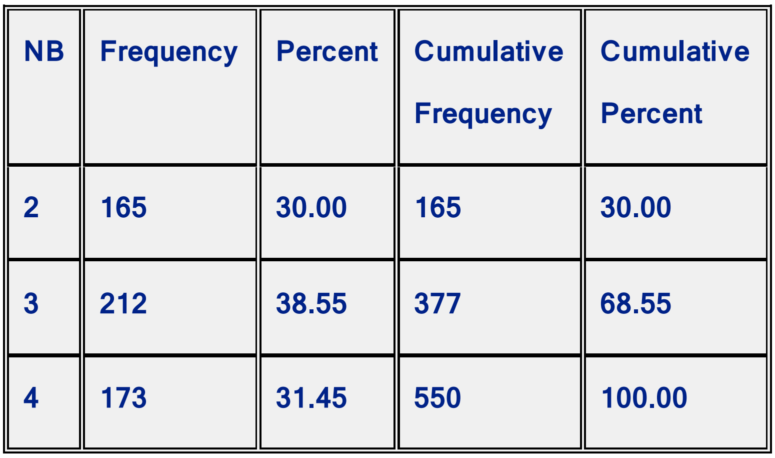
Variable: NR (Number of rooms)



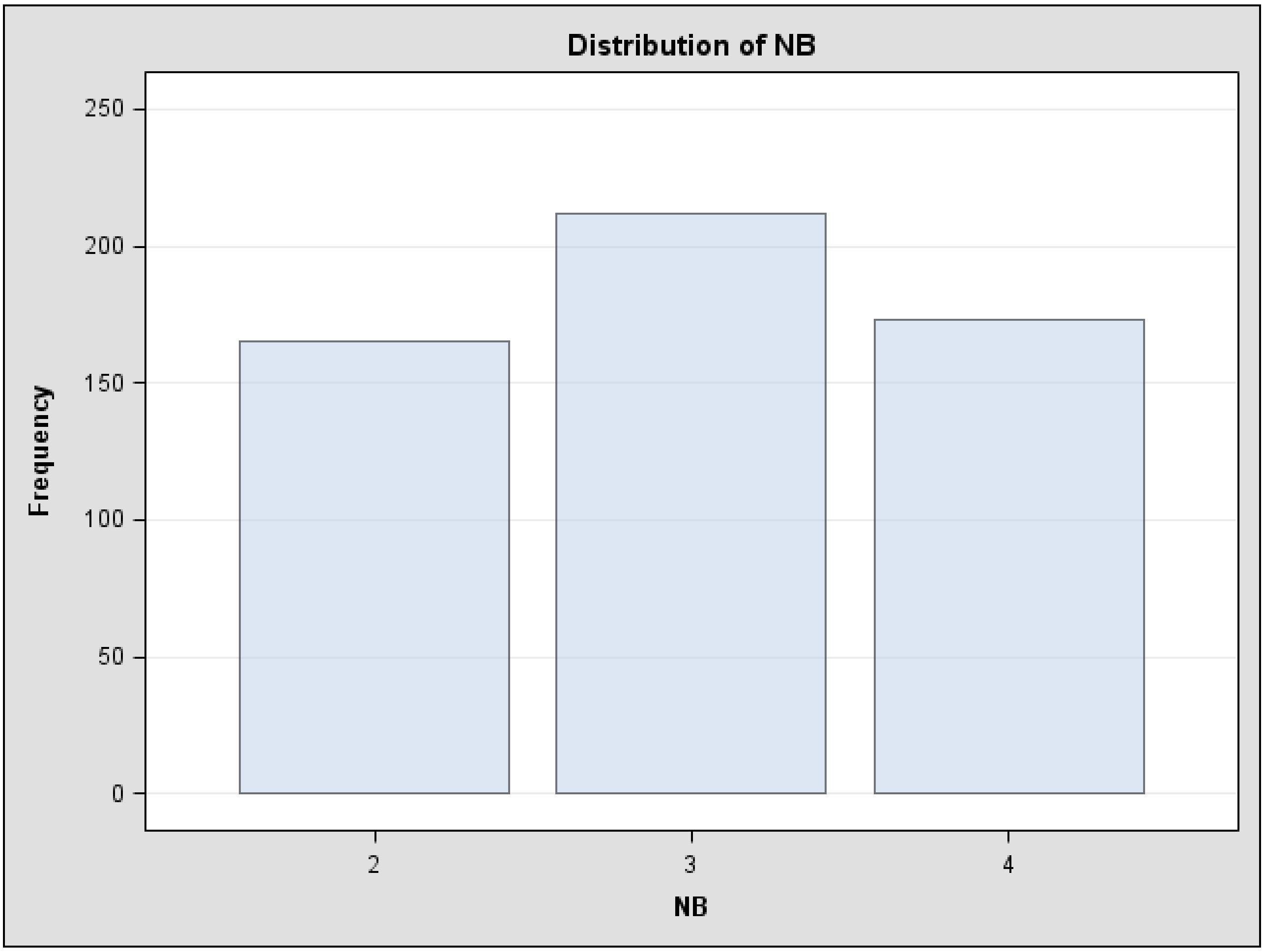
NR	Frequency	Percent	Cumulative Frequency	Cumulative Percent
5	152	27.64	152	27.64
6	146	26.55	298	54.18
7	141	25.64	439	79.82
8	111	20.18	550	100.00



Variable: NB (Number of bedrooms)

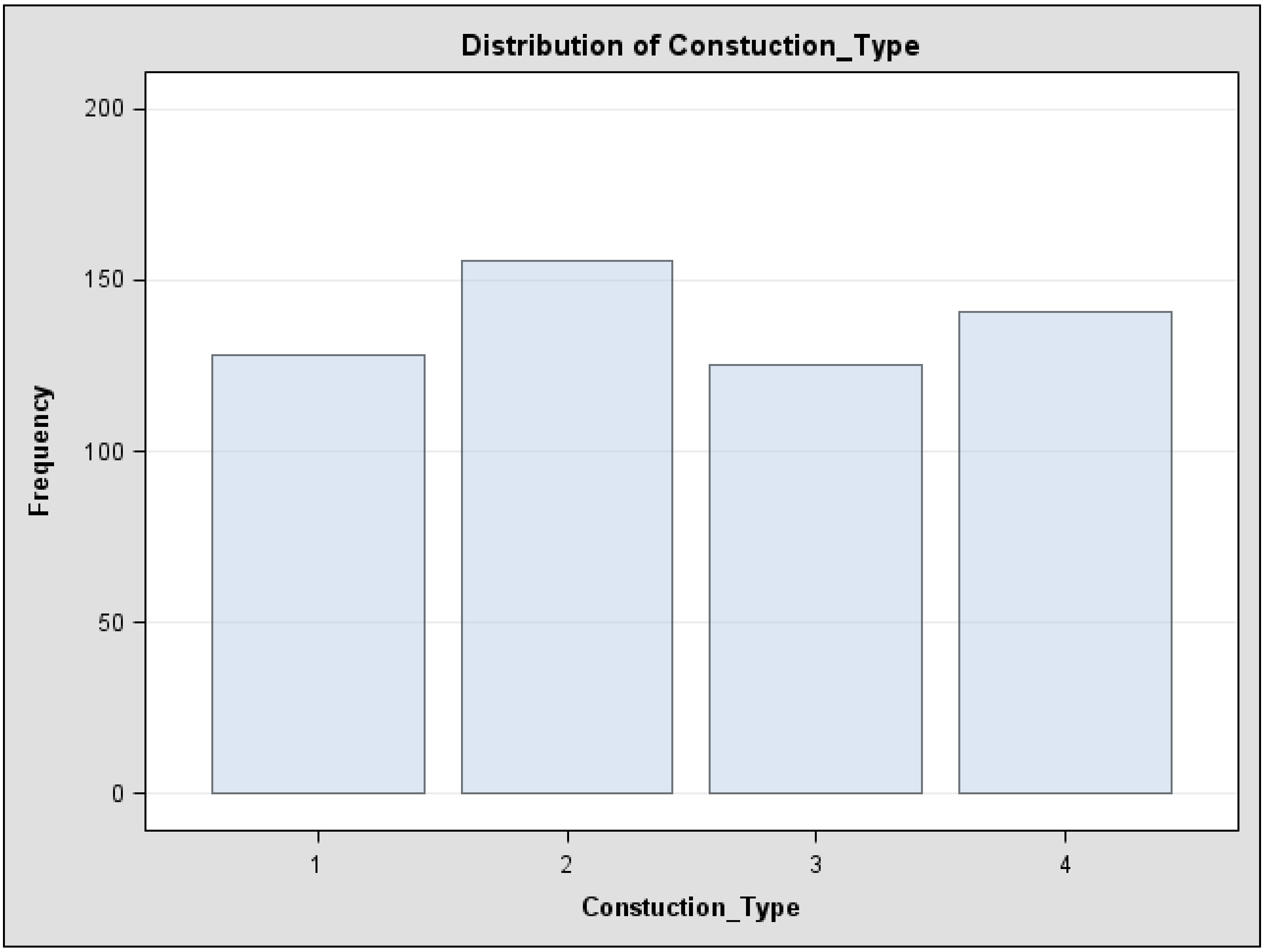


NB	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2	165	30.00	165	30.00
3	212	38.55	377	68.55
4	173	31.45	550	100.00



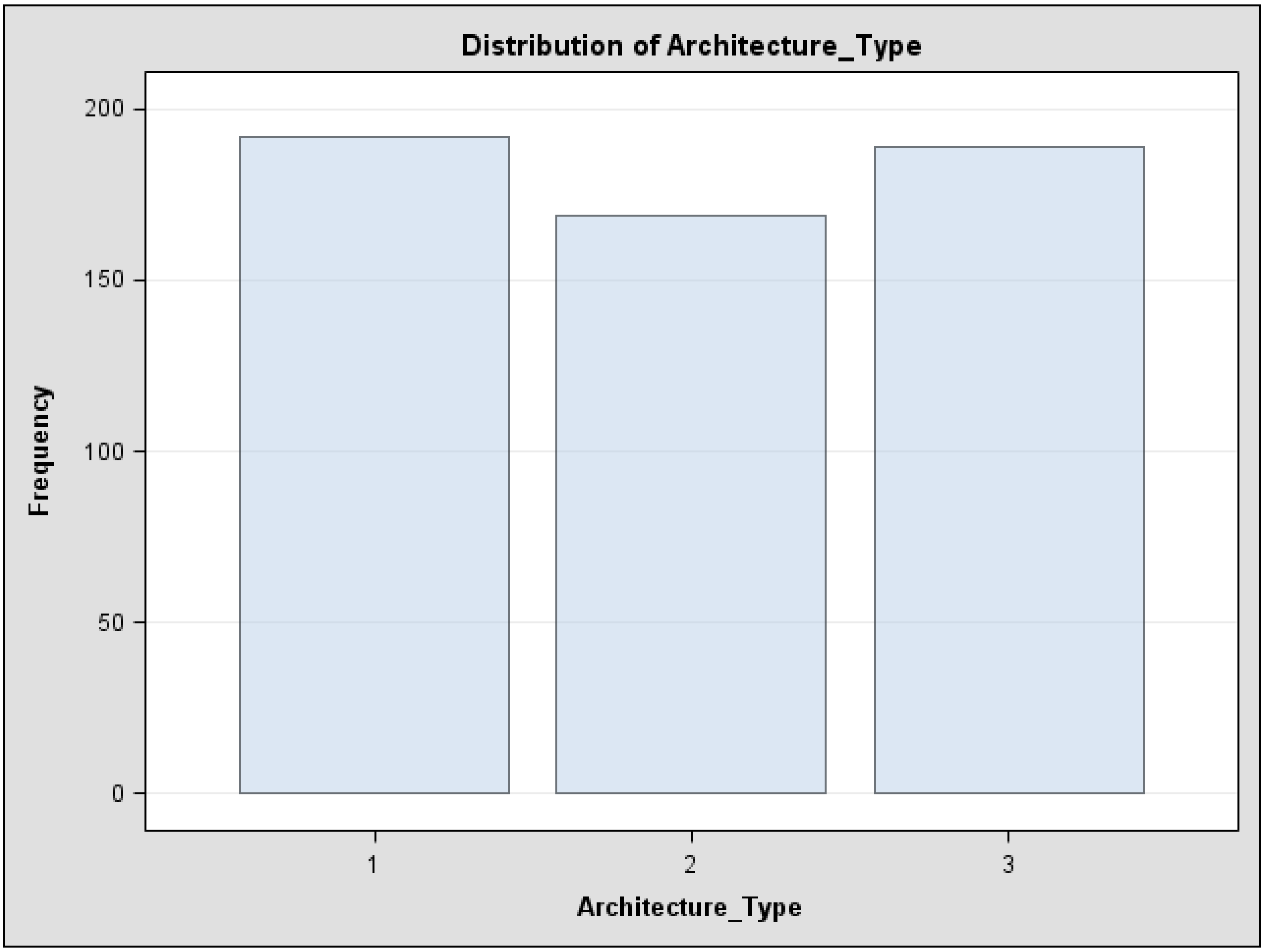
Variable: Construction Type

Construction Type	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	128	23.27	128	23.27
2	156	28.36	284	51.64
3	125	22.73	409	74.36
4	141	25.64	550	100.00



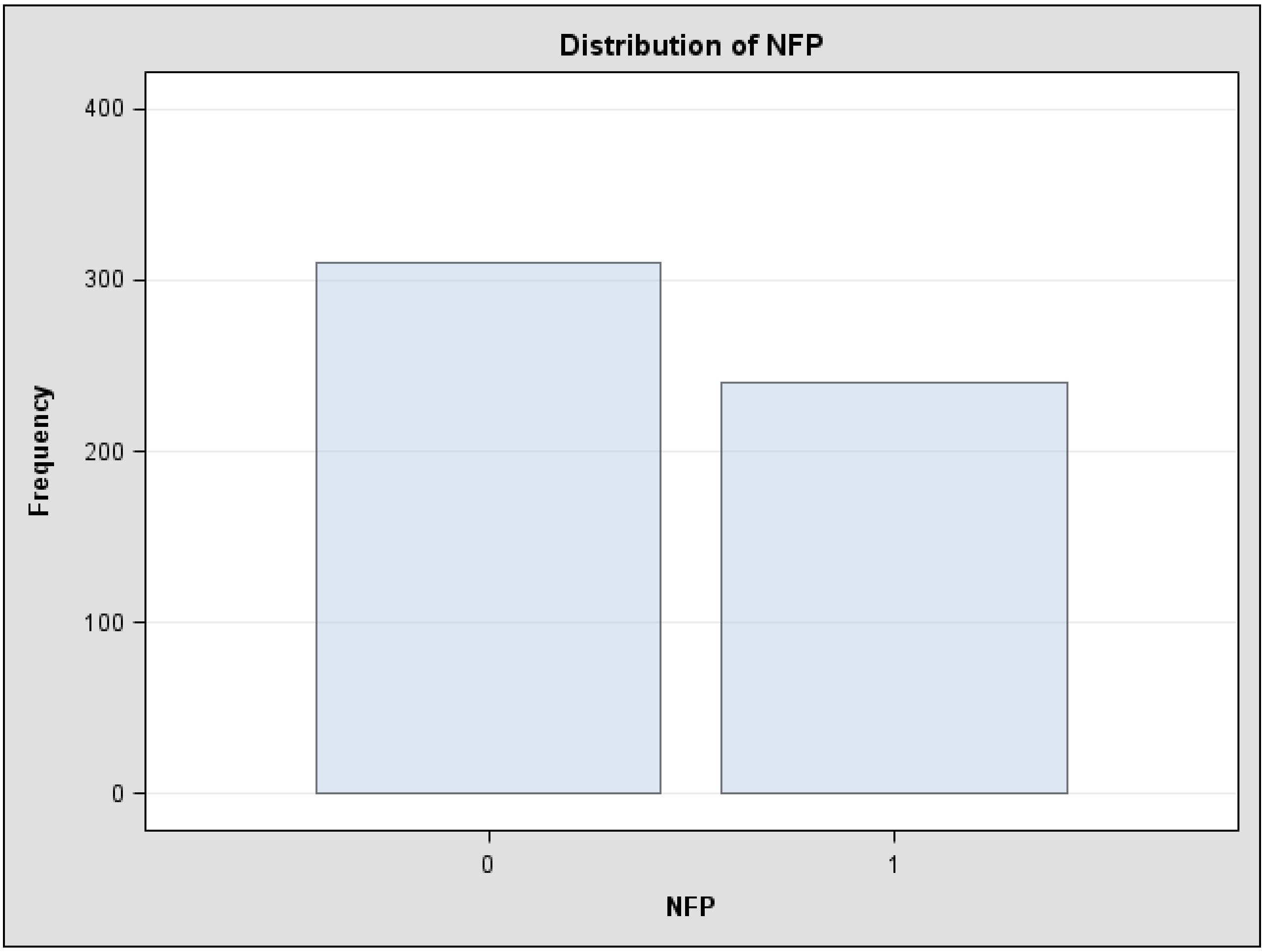
Variable: Architecture Type

ArchitectureType	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	192	34.91	192	34.91
2	169	30.73	361	65.64
3	189	34.36	550	100.00



Variable: NFP (Number of fire places)

NFP	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	310	56.36	310	56.36
1	240	43.64	550	100.00



Fitting regression model:

```
/* Regression Analysis with Original Data */


odsoutputFitStatistics = t0;
odshtml;
odsgraphicson;
procregdata = HousePrice;
modelSellingPrice   =   LSP_D   NBR   AS_TSF   SLS_TSF   NG   NR   NB
Age_YearsConstuction_TypeArchitecture_Type NFP;
plotresidual. * predicted.;
run;
quit;




/* Store the estimated r-square */


data_null_;
set t0;
if label2 =  ″R-Square″then
callsymput(’r2bar’, cvalue2);
run;
```

Regression Output:

The SAS System

The REG Procedure

Model: MODEL1

Dependent Variable: SellingPrice

Number of Observations Read	550
Number of Observations Used	550

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	P_r > F
Model	11	11387	1035.22100	182.71	<.0001
Error	538	3048.27676	5.66594		
Corrected Total	549	14436			

Root MSE	2.38032	R-Square	0.7888
Dependent Mean	36.05364	Adj R-Sq	0.7845
Coeff _Var	6.60217		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	P_r > |t|
Intercept	1	14.73812	1.09633	13.44	<.0001
LSP_D	1	2.52881	0.12600	20.07	<.0001
NBR	1	1.81473	0.30816	5.89	<.0001
AS_TSF	1	0.18635	0.07425	2.51	0.0124

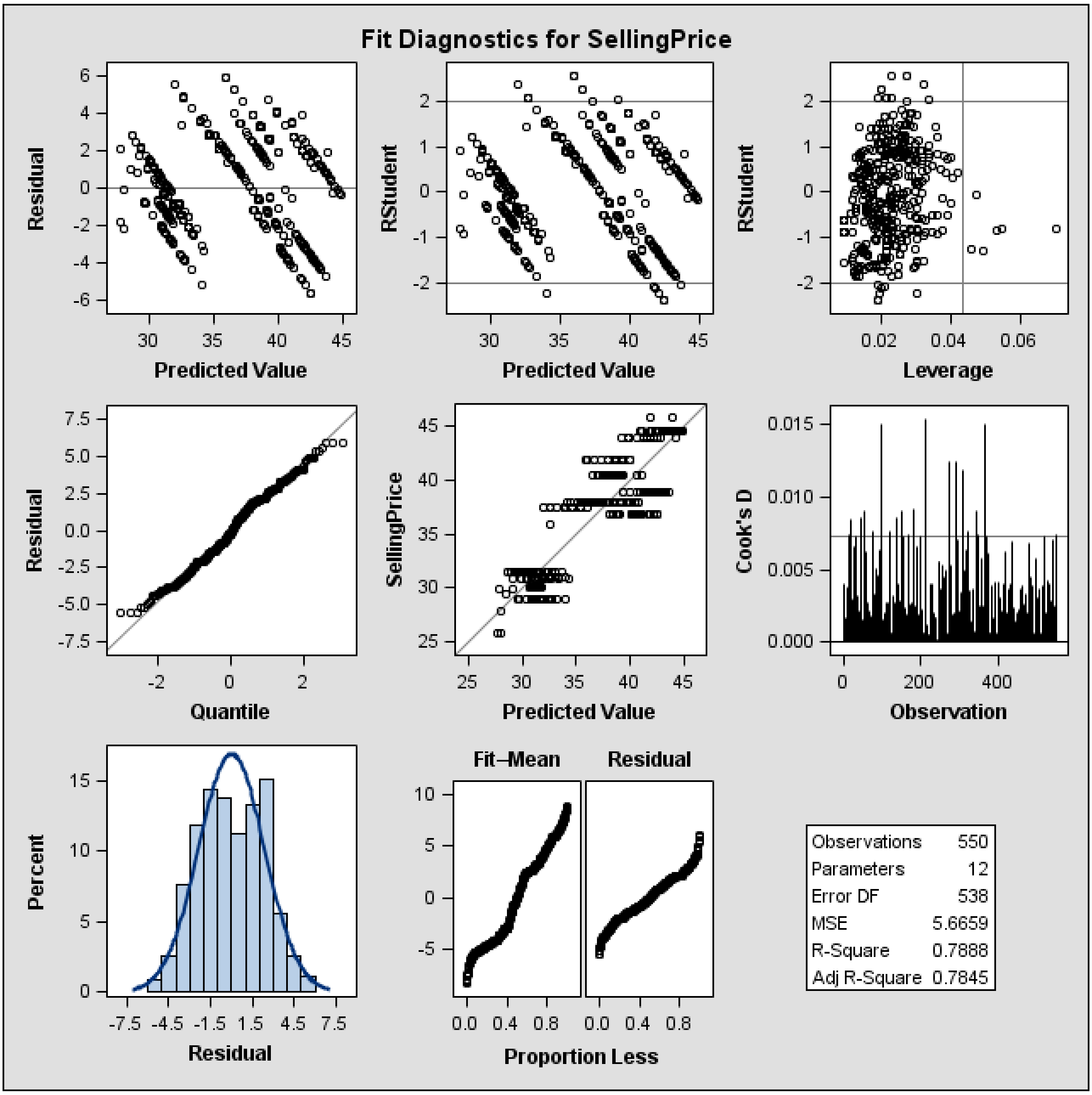
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	P_r > |t|
SLS_TSF	1	3.96816	0.66467	5.97	<.0001
NG	1	1.00855	0.14629	6.89	<.0001
NR	1	-0.74574	0.16537	-4.51	<.0001
NB	1	-0.68358	0.18654	-3.66	0.0003
AgeYears	1	-0.06789	0.02221	-3.06	0.0024
ConstructionType	1	0.64189	0.09831	6.53	<.0001
ArchitectureType	1	-0.01814	0.12699	-0.14	0.8865
NFP	1	1.62658	0.22348	7.28	<.0001

The SAS System

The REG Procedure

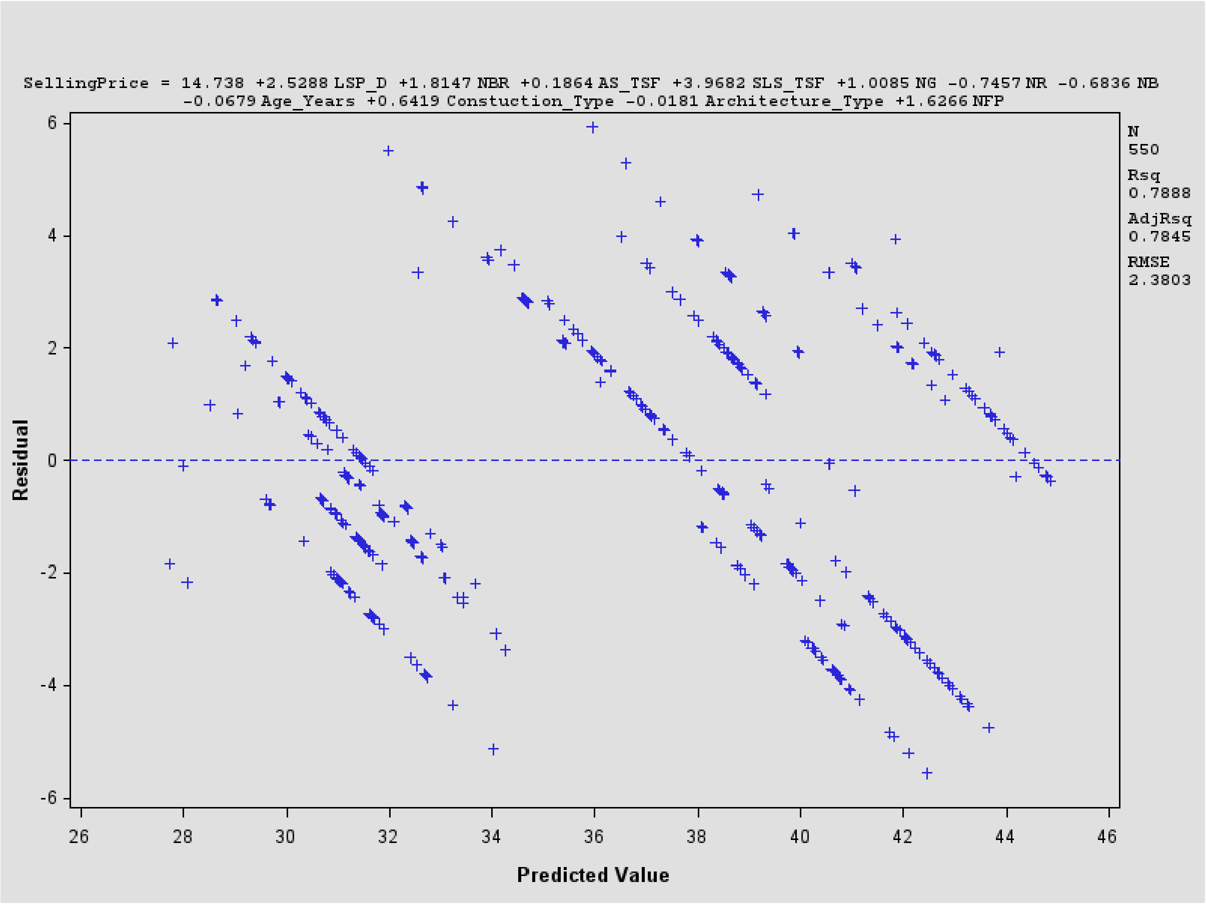
Model: MODEL1

Dependent Variable: SellingPrice



Observations	550
Parameters	12
Error DF	538
MSE	5.6659
R-Square	0.7888
Adj R-Square	0.7845

Plot for Predicted values and Residuals:



Conclusions:

The main objective of the present study is to investigate factors that contribute significantly to estimate the selling price of a house. The dependent variable is Average house selling price, the multiple regression analysis applied for exploring the factors affecting the house selling price.

Below is fitted regression model. From regression coefficients we can observe that, except variable Architecture_type, all other variables were turned as significant in order to predict the selling price of a house.

SellingPrice = 14.738 + 2.5288 LSP_D + 1.8147 NBR + 0.1864VAS_TSF + 3.9682 SLS_TSF + 1.0085 NG – 0.7457 NR – 0.6836 NB – 0.0679 Age_Years + 0.6419 Construction_Type – 0.0181 Architecture_Type + 1.6266 NFP

The coefficient of determination R^2 values is 0.78, so it means that approximately 78% of variation in SellingPrice can be explained by all included independent variables also this R^2

Value indicates that the model fitted to be good and the R^2 is significant.

References:

[1] "Generalized Linear Models" – McCullagh P and Nelder J A (1989).

[2] "An Introduction to Generalized Linear Models" – Dobson A J (1990).

[3] "Applied Statistics" – Mogull, Robert G (2004).

[4] "On the theory of Correlation" – Yule G, Udny (1987).

[5] "The Law of Ancestral Heredity" – Pearson Karl, Yule GU, Blanchard, Norman, Lee Alice (1903).

[6] "The Goodness of Fit of Regression Models" – Fisher R A (1922).

[7] "Fisher and Regression" – Aldrich, John (1954).

[8] "Illusions in Regression Analysis" – Armstrong J Scott (2012).

[9] "Statistical Models - Theory and Practice" – David A, Freedman (2005).

[10] "Criticism and Influence Analysis in Regression" – Dennis Cook, Sanford Weisberg (1982).

[11] "Pattern Recognition and Machine Learning" – Christopher M Bishop (2006).

[12] "Inflation of R2 in Best Subset Regression" – Rencher AC and Pun FC (1980).

[13] "Regression Prediction and Shrinkage" – Copas JB (1983).

[14] "Test of Significance in Forward Selection Regression with an F to enter stopping rule" – Wilkinson L and Dallal GE (1981).

[15] "The Impact of Model Selection on Inference in Linear Regression" – Hurvich CM and Tsai C (1990).

[16] "Prediction Error and its Estimation for Selected Models" – Roecker, Ellen B (1991).

[17] "Linear Regression" – Paul Glasserman (2001).

[18] "Linear Regression and the Minimum Sum of Relative Errors" – Narula SC and Wellington JF.