

St. Cloud State University theRepository at St. Cloud State

Culminating Projects in Mechanical and
Manufacturing Engineering

Department of Mechanical and Manufacturing
Engineering

11-2015

Regression Analysis To Forecast the Demand of New Single Family Houses in USA

Lama Nayal

Follow this and additional works at: https://repository.stcloudstate.edu/mme_etds

Recommended Citation

Nayal, Lama, "Regression Analysis To Forecast the Demand of New Single Family Houses in USA" (2015). *Culminating Projects in Mechanical and Manufacturing Engineering*. 15.
https://repository.stcloudstate.edu/mme_etds/15

This Starred Paper is brought to you for free and open access by the Department of Mechanical and Manufacturing Engineering at theRepository at St. Cloud State. It has been accepted for inclusion in Culminating Projects in Mechanical and Manufacturing Engineering by an authorized administrator of theRepository at St. Cloud State. For more information, please contact rswexelbaum@stcloudstate.edu.

Regression Analysis To Forecast the Demand of New Single Family Houses in USA

Lama Nayal

A Starred Paper

Submitted to the Graduate Faculty Of

St. Cloud State University

In Partial Fulfillment of the Requirements

For the Degree

Master of Engineering Management

St. Cloud, Minnesota

November, 2015

Committee:

Ben Baliga, Chairperson

Hiral Shah

Balasubramanian Kasi

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor Dr. Ben Baliga for the continuous support of my Master study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis.

Besides my advisor, I would like to thank the rest of my thesis committee: Dr. Hiral Shah, and Dr. Balasubramaian, for their insightful comments and encouragement.

In addition, I would like to thank Randal D. Kolb, The statistic specialist in St. Cloud statistic station for his guidance and support through this study.

Last but not the least, I would like to thank my family: my parents, my brother, my sister, my husband and my kids for supporting me spiritually throughout writing this thesis and my life in general.

REGRESSION ANALYSIS

Abstract

Forecasting the market demand is a very critical step in planning all kinds of business including construction business. This study was conducted to develop a robust regression model that enables construction companies predicting the demand of new single family houses in the USA. The study identified each of inflation rate, mortgage rate, GDP, Personal consumption, unemployment rate, and population as independent variables that may affect the market demand of new single family houses. The data were collected over 21 years, evaluated, and sorted according the nature of the relationship between each independent variable factor and the market demand of new single family houses. The data reflected double conversion in relationship between GDP, Personal consumption, and population and the market demand due to the financial crises and the beginning of the recovery after it. The Dummy variables technique was used to identify the periods of before the financial crisis, during the financial crises, and after it. The dummy variables have been added to the model to handle the fluctuation in these data sets. The study concluded that the unemployment rate variable and the personal consumption variable are the most important factors that affect the market demand of new single family houses in the USA. A regression model was developed to be used to predict the market.

Table of Contents

List of Tables vi

List of Figures vii

Introduction..... 1

 Introduction..... 1

 Problem Statement..... 2

 Nature and Significance of the Problem 3

 Objective of the Project 4

 Project Questions/Hypotheses..... 4

 Limitations of the Project..... 4

 Definition of Terms..... 5

 Summary 5

Literature Review..... 6

 Introduction..... 6

 Literature Review Related to the Problem 6

 Literature Related to The Regression Analysis Methodology 10

 Summary 25

Methodology 26

 Introduction..... 26

 Design of the Study..... 26

 Data Collection 27

 Data Analysis 31

 Budget..... 35

 Timeline 36

 Summary 36

Data Presentation and Analysis 37

 Introduction..... 37

 Data Presentation and Analysis 37

 Summary 59

Results, Conclusion, and Recommendations 60

REGRESSION ANALYSIS

Introduction.....	60
Results.....	60
Conclusion	62
Recommendations.....	62
References.....	63

REGRESSION ANALYSIS

List of Tables

Table	Page
1. Pearson Values and the nature of relationship between two valuables.....	17
2. P-Value to accept or reject the null hypothesis	32
3. The Values of the Dummy Variables	44
4. Test Statistic for Model 1 & 6	53
5. Coefficients of Model 6	54
6. T- Values of Model 6	55
7. P- Values of Model 6	55

REGRESSION ANALYSIS

List of Figures

Figure	Page
1. American express categorie.....	10
2. Critical region right tail.....	25
3. Critical region left tail.....	25
4. Two tails hypothesis testing.....	25
5. Gantt chart of the time line of the project.....	36
6. Scatter plot of the Home sales VS Inflation.....	38
7. Scatter plot of the Home sales VS Mortgage Rate.....	39
8. Scatter plot of the Home sales VS Unemployment Rate.....	40
9. Scatter plot of the Home sales VS Population	41
10. Scatter plot of the Home sales VS Personal Consumption	45
11. Scatter plot of the Home sales VS GDP.....	47
12. Residual Plots for Home Sales USA	57
13. Probability Plot of Residuals	58
14. Home Sales Predicted VS Actual	59

Chapter I

Introduction

Introduction

Forecasting is the science of predicting the future. It invades all aspects of daily decisions of the personal life and business life. Every day forecasts are made and evaluated for future variables to be used by humans and organizations to predict things that might affect their decisions.

In business, Forecasting is a starting point of planning the business. Forecasting is needed in different aspects of business, such as cash flow, number of personnel, raw material prices, sales, etc. And it is used as a technique to estimate a certain value of an uncontrollable variables to be used in planning the business. Forecasting could be classified into two categories:

Qualitative forecasting: This type of forecasting aims to bring together in a logical, unbiased, and systematic way all information that are related to the factor of interest. This method is based on educated opinions of appropriate persons. Delphi method is a well-known method of qualitative forecasting.

Quantitative forecasting: This method relies on historical data to make predictions with the help of mathematical models. And it is further can be classified into two types, the statistical method and the deterministic method. The statistical method focuses on patterns and patterns changes. Regression models and exponential smoothing are classified as statistical method. On the other hand, the deterministic method is a mathematical model in which outcomes are determined through known relationships between the forecasted factor and the influencing

factors, without random variation. This method includes anticipation surveys and input output models.

Forecasting also could be categorized based on the time horizon of forecasting.

Short-range forecast: The time span of the short-range forecast is up to one year. It is used for job scheduling, planning purchasing, etc.

Medium-range forecast: The time span of the medium-range forecast is up to three years. The Medium- range forecast is useful in sales planning, production planning, budgeting, etc.

Long-range forecast: the time range of the long-range forecast is more than three years. The long-range forecast is useful for planning new product, capital expenditures, facility expansion, etc.

This study focuses on an important activity in construction business planning. It is about forecasting the market demand of new single family houses in the USA. The study attempts to understand what the customers will demand in future and how the market will behave.

Problem Statement

The construction industry is an important sector of the economy; a careful planning is required to run a successful construction business. Forecasting future values of many elements of the business is a very critical step prior to planning the business. The demand of the market is one of the elements to be forecasted prior to the planning process. This task is a complex task as the housing demand is affected by many social and economic factors and the market varies due to the variation of these affecting factors.

Nature and Significance of the Problem

As stated earlier, construction industry is one of the most important sectors of the economy. However, the failure risk in this sector of business is pretty high. One of the failure causes in construction industry is that construction activity rises and falls faster than the economy. Usually economic fluctuation affects property market demand, the prices in the market, and the cost of the projects. Forecasting the market demand enables companies to set their strategic plans, chose their future projects, define their needs of materials, calculate the expected cost and profit according to the forecasted market demand.

In addition, people are an important factor in the construction business; the business needs the right people in the right place at the right time. As the demand of the construction industry varies, sometimes the demand raises making more project opportunities available, that may create a problem of lack of workforce, especially that construction business needs technically qualified workers. In other cases, the demand is low, that creates a problem of finding work for the technically qualified experienced workers in the companies to keep them busy, which may causes staffing and financing challenges for the construction companies. Forecasting the market demand helps the company in defining their needs of workforce and its cost in the future.

In addition to the economic factors, the housing demand is also affected by other social factors as well, such as population, as it is believed that the more the population the more demand on housing market.

In summary, Identifying the factors that affect the demand, finding a mathematical relationship between the market demand and these factors, and developing a model to be used as

a tool to predict the market, will definitely help construction company's managers and top managers to make better decisions, and will guide them in setting the strategic plans for their companies, and will help them in better projects and resources planning.

Objective of the Project

The aim of this study is to build a prediction model to forecast the market demand of new single family houses in the USA using a quantitative method of forecasting.

To achieve this aim, the following objectives were defined to be achieved:

- 1- Identify the parameters that influence the demand of new single family houses in the USA.
- 2- Determine an appropriate tool of forecasting.
- 3- Develop the forecasting model.
- 4- Compare results of forecasting and exact demand of new single family houses in term of accuracy.

Project Questions/Hypotheses

- 1- What factors do affect the demand of building new single family houses in USA?
- 2- How strong is the relationship between these factors and the market?
- 3- How the variation of these factors will affect the demand of building new single family houses in USA?
- 4- How can we predict the demand of building new houses according to the prediction of the values of these factors?

Limitations of the Project

This study concentrated on medium term of forecasting to predict the market demand of new single family houses in the USA based on time series data of independent factors that have

been proven to affect the demand. The data of this project are secondary data collected by government agencies for the social factors and most of the economic factors. The model that has been developed in this study was able to explain 82.9 % of the variation in the demand. However 17.1 % of the variation have not been explained through the developed model.

Definition of Terms

Multiple Linear regression analysis: An overall methodology that aim to predict the value of a variable (dependent variable or dependent factor) based on the value of two or more other variables (independent variables or independent factors).

The dependent factor/ dependent variable/ Market demand: It is the output value of the regression model; it is the value that the model aims to predict. The dependent factor of the regression model of this study is the new single family houses in the USA.

The independent factor(s)/ variable(s): They are the input values of the regression model. They are the factors that have been believed / proven to affect the factor of interest of the study.

Summary

In this chapter, the concept of forecasting and its types have been introduced, the problem statement has been defined, the nature and the significance of the problem have been explained, the objective of the project has been clarified, the questions of the project and the project's limitation have been addressed, and the definition of the terms has been specified.

The next chapter presents a literature review about the problem and about the methodology.

Chapter II

Literature Reveiw

Introduction

This chapter presents the literature review that has been conducted at the beginning of this study. The literature review is divided into two parts, one is related to the problem, the other is related to the methodology was used in this study.

The main topics that have been covered in the literature review related to the problem are:

- The importance of construction industry in the economy.
- The importance of forecasting in planning the construction business.
- The historical awareness about forecasting housing market.
- The different methods have been used in forecasting housing demand.
- Related studies and their findings.
- The factors to be evaluated as possible dependent factors that affect housing demand.

The literature review related to the methodology covers:

- Types of regression methods.
- Regression elements and key concepts.
- Assumptions of linear regression analysis.
- Residual analysis.
- Hypothesis testing.

Literature Review Related to the Problem

The construction industry is one of the important sectors of any economy (Jiang, & Liu, 2015). And in the USA, the construction industry makes a significant contribution to economic

output (Jiang, & Liu, 2015). It also represents one of the biggest industries in the United States (Hutchings, & Andrus, 2006). The construction industry is made of nearly 700,000 large and small companies (Hutchings, & Andrus, 2006).

As the construction industry has such importance, its believed that careful plans should be prepared and implemented to run a successful business (Prince, 2014). As planning without prior knowledge or reasonable expectations about the future leads to a risky business, forecasting many variables of the business is believed to be a very critical step in planning the business (Zainun, & Eftekhari, 2010). The tools and techniques used in forecasting the elements of the business have to trusted tools and techniques to help the decision makers to guess or arrive at a correct conclusion (Zainun, & Eftekhari, 2010).

Understanding construction market is required for a successful construction business (Prince, 2014). Decision makers of the construction companies are interested in understanding the dynamics of housing market due to its significant impact on the whole company (Boyed, 2014). Industry professionals and property academic have a long history of analyzing property markets, since the twentieth century, pioneers stressing the importance of real estate supply and demand (Boyed, 2014). The importance of market analysis become even more important during the evolution of the property discipline in the mid-20th Century (Boyed, 2014). The focus on analyzing the housing market and forecasting housing demand is still relevant today, lots of publications on property market analysis have continued over the years (Boyed, 2014).

In the construction economics sense, statistical forecasting for construction market can be classified into two types, the univariate method and the causal models (Jiang, & Liu, 2015). The univariate model, is technique that forecasts future values based on the past values of the time series (Jiang, & Liu, 2015). This technique of forecasting has been widely used for predicting

construction demand. Merkies and Poot (1990) used the univariate model to forecast construction activities in the Netherlands and New Zealand, they used the exponential smoothing technique in that study (Jiang, & Liu, 2015). On the other hand, causal modelling techniques identifies the related variables affecting the predicting variable and can develop statistical models to differentiate the relationship between these variables (Jiang, & Liu, 2015). The multi-regression analysis is one of the most commonly causal models used in forecasting (Jiang, & Liu, 2015). In the UK, the linear multi-regression model was adopted to predict demand for the many sectors of conduction buildings, such as residential, commercial and industrial construction markets (Jiang, & Liu, 2015).

In the USA, a descriptive study has been conducted to forecast the housing demand in the USA to estimate the housing in the USA in the 1990s by evaluating both demographic and socioeconomic changes in the demand housing(Eppli, Childs, 2001). Mark and Monty stated in that study's findings that owner-occupied demand in the 1990s is expected to be at average of 811,000 units per year, which has been addressed as a decline of 174,000 units per year from the prior decades' demand of 985,000 owner- occupied units (Eppli, Childs, 2001).

Another study has been conducted in 2007 to project the demand of new houses. That study examined the challenges of projecting the long-run demand for new residential construction and presented a range of estimates for the likely demand for new housing over the period from 2005 through 2014 (Belsky, & McCue, 2007).The study's finding stated that the estimate of the total demand for new housing from 2005 through 2014 is 19.5 million units ((Belsky, & McCue, 2007). That projection did not account for oversupply in the housing stock as of the beginning of the period, or the high level of construction that already occurred in 2005 and 2006 (Belsky, & McCue, 2007).

This current study is another attempt to forecast the demand of new single family houses in the USA by adopting the multi-regression analysis methodology, considering many social and economic factors. The following factors were identified to be as possible affecting factors on the housing market.

The population. The growth of a population has been identified as a key determinant of the demand for residential construction as it raises the basic need for new housing units (Jiang, & Liu, 2015).

The unemployment rate. The unemployment rate is believed to be one of the factors that affect housing demand. When unemployment rate rises, more people will be unable to afford a house (Jiang, & Liu, 2015). Therefore an increase in the unemployment rate represents a lowering of the purchasing power of the population, and it may discourage investment in the construction market (Jiang, & Liu, 2015).

The interest of mortgage rate. As the change in interest rates affects both the construction companies and the customers, the change in the interest rate affects the lending costs of the clients and their monthly mortgage payment (Jiang, & Liu, 2015). A lower interest rate will encourage investments in the construction market that will result in a raise in demand for construction (Jiang, & Liu, 2015). In contrast, any increase in interest rates will result in a raise of the cost of bank lending for construction projects and will also negatively affect the purchasing power. (Jiang, & Liu, 2015).

GDP. The overall health of the economy affects the construction industry (Pettinger, 2013). The overall health of economy is generally measured by economic indicators such as GDP (Pettinger, 2013). When the economic growth raises people will be able to spend more on houses; that will increase demand of housing market. (Pettinger, 2013).

Inflation. The increase in inflation rate raises the cost of new construction (Keefer, 2015). When inflation is high, the cost of materials rises, and the labor costs also rises as well (Keefer, 2015). Therefore, is believed that higher inflation rate may affects to some extent the construction industry and housing demand.

Personal consumption expenditures. It is an economic indicator that explains how consumers are spending on goods and services in the U.S (Amadeo,2015). It shows how much of the income earned by households is being spent on current consumption. The figure shows that almost one-third of the personal expenditure is to be spent on housing (“Where Americans spend their money”, 2015). This cost includes the explicit payments of rent by residential tenants and implicit space rent of owner-occupiers (Mayerhauser, & Reinsdorf, 2007). An assumption has been made that there is a relationship between the personal consumption and the market demand.

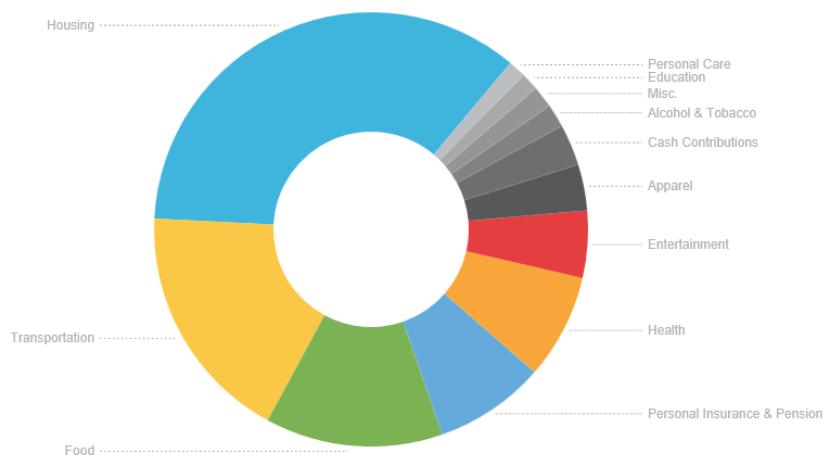


Figure 1 American expense categories (33)

Literature Related to The Regression Analysis Methodology

The literature review related to the methodology is divided into five sections, the first one covers the types of regression, the second one covers the key concepts of the multiple regression

analysis, the third one covers the assumptions of the multiple regression analysis, the fourth one covers the residual analysis, and the last one covers the hypothesis testing.

Types of regression analysis. The regression analysis is a statistical tool that aims to explore the strength of the relationship between one dependent variable which is the response variable and one or many other changing variables known as the independent variables or the explanatory variables (Suri, 2006). It is a powerful technique that uses the values from historical data of one or more variables to develop a model that helps in predicting the value of the dependent variable (Suri, 2006).

The regression analysis has different types according to the nature of the relationship between the dependent and the independent variable (Frost J.2015). The main two types of regression analysis according to the relationship between the dependent and the independent variables are (Linear Regression Analysis and Non-Linear Regression analysis) (Frost J.2015).

The Non-linear regression is out of the scope of this project and its literature review. The linear regression analysis has two main types according to the number of predictors (“Types of regression analysis”, 2015).

Simple regression analysis. This type of regression assesses the relationship between one dependent variable and only one independent variable (Frost J.2015). It simply has an equation of a line. (Suri, 2006).

$$Y = a + b X + E$$

Where **Y** is the value of the dependent variable to be predicted or explained (Suri, 2006).

Alpha (a) is a constant that equals the value of Y when the value of X=0 (Suri, 2006).

Beta β or is the coefficient of X. It is the slope of the regression line that can explain how much Y changes for each one-unit change in X (Suri, 2006).

X is the value of the Independent variable (X), which is the factor that predicting or explaining the value of Y (Suri, 2006).

E is the error in predicting the value of Y (Suri, 2006).

Multiple regression analysis. This type of regression assesses the relationship between one dependent variable and many independent factors (Frost J.2015). The form of this regression model for given (k) independent variables is. (Suri, 2006).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

β_0 is the intercept or the constant (Suri, 2006).

$\beta_1, \beta_2, \beta_3, \dots, \beta_k$ are the regression coefficients (Suri, 2006).

X_1, X_2, \dots, X_k are the independent variables to be used to predict the dependent variable (Suri, 2006).

Y is the response or the dependent variable which the factor of interest. (Suri, 2006).

Regression elements and key conspets. The Regression analysis has many elements and key concepts that need to be understood to be able to interpet the output of the regression. Following are the main elements and kley concepts.

The graphical interpretation. The graphical interpretation aims to visualize the regression model. In a simple regression the graph shows the line or curve that best it's the scatter of points (X_1, Y_1) (X_2, Y_2) , (X_n, Y_n) obtained on (n) individuals (Suri, 2006). The regression equation is the path described by the means of values of the distribution of Y when X varies(Suri, 2006).

By increasing the independent factors the graphical interpretation becomes a hyper surface in a dimensional space (Suri, 2006). The more the factors, the more complexity the

model becomes (Suri, 2006). Researches are recommended to keep the number of independent variables to be minimum with the respect of the model accuracy (Suri, 2006).

The regression coefficient β : The coefficient represents the slope of the regression line (Suri, 2006). The larger the coefficient, the steeper the slope, the more the dependent variable changes for each unit change of the independent variable (Suri, 2006). In a Multiple regression analysis, each coefficient become a partial coefficient (Suri, 2006).

The least square method: It is a method that minimizes the sum of squares of the distance between the predicted value by the fitted model and between the observed responses (Suri, 2006). Better fit will have smaller deviation of the predicted values from the observed values (12). Mathematically, the equation that represents the sum of square is:

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \dots - \hat{\beta}_k X_{ki})^2$$

(Suri, 2006)

The least some of square is also called the residual sum of squares or error sum of squares. It is referred as SSE\ESS (Suri, 2006)

Measure of goodness to fit. It describes how well the model fits a set of observations and summarizes the discrepancy between observed values and the predicted values by the mode (Suri, 2006).

In other words, for any observation (i), the difference between Y_i and \bar{Y} could be decomposed into the difference between Y_i and \hat{Y} , and the difference between \hat{Y} and \bar{Y} (Suri, 2006). That means that the difference between the observed value and the mean equals to the difference between the observed value and the fitted value, plus the difference between the fitted value and the mean (Suri, 2006). Mathematically, the equation is:

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

$$\text{TSS} = \text{ESS} + \text{RSS}$$

(Suri, 2006)

Y_i , is the observed 'i th' value of Y (Suri, 2006).

\bar{Y} , is the average value (Suri, 2006).

\hat{Y}_i , is the 'i th' fitted value of Y (Suri, 2006).

TSS, is the total sum of squares (Suri, 2006).

ESS, is the estimate sum of squares, and it represents variation in the Y values that is unexplained by the model (Suri, 2006).

RSS, is the residual sum of squares, and it represents the variation in the Y values that is explained by the model (Suri, 2006).

Multiple correlation coefficients. It measures how well the dependent variable is predicted by the data set of the independent variables using the regression model (Frost, 2013). The coefficient of multiple correlation is always between zero and one; as this value raises, the model gets better in predicting the dependent variable from the independent variables (Frost, 2013).

Mathematically, the equation that describes the R^2

$$R^2_{Y|X_1, X_2, \dots, X_k} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2 - \sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

(Suri, 2006).

Adjusted R^2 . R-squared measures the variation in the dependent variable explained by the independent variable for the linear regression model (Chen, 2013). The R-Squared value goes up just by adding more and more independent variables, even if they don't have a correlation with the dependent variable (Chen, 2013). The Adjusted R-squared provides an adjustment to the R-squared statistic, because the adjusted R-Squared does not go up unless the added variable has a correlation to dependent variable (Chen, 2013). In contrast, the adjusted R-squared goes down when the added variable does not have a strong correlation with the dependent variable (Chen, 2013).

Mathematically, the equation for the adjusted R- squared is.

$$\text{Adjusted } R^2 = 1 - \left((1 - R^2) \left(\frac{n - 1}{n - k - 1} \right) \right) \quad (\text{Suri, 2006})$$

Assumptions of linear regression. Regression. Linear regression analysis has four main assumptions.

Linearity. Some researchers stated that this assumption is the most important assumption in any linear regression model as it directly relates to the bias of the results of the whole analysis (Balance, 2012). Linearity requires the dependent variable to be a linear function of the independent variables. (Balance, 2012).

If linearity is violated all the estimates of the regression and its statistical output may be biased (Balance, 2012). Violation of this assumption results in serious error in the predicted values. (Balance, 2012). On the other hand, when a linearity relationship exists between the dependent and the independent variables, the linear multiple regressions can accurately estimate the dependent variable (Balance, 2012).

Mathematically, an assumption is made that the mean value of Y for each a given combination of $X_1, X_2, X_3, \dots, X_k$ is a linear function of $X_1, X_2, X_3, \dots, X_k$ (Suri, 2006)

$$\mu_{y|x_1, x_2, x_3} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_k X_k \text{ (Suri, 2006).}$$

Or

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_k X_k + E \text{ (Suri, 2006).}$$

Where E is the error that reflects the difference between an individual observed response Y and the true average response $\mu_{y|x_1, x_2, x_3}$. (Suri, 2006).

To evaluate the linearity a scatter plot with a trend line could be used (“Pearson product-moment correlation”, n.d)). Also the correlation coefficient, such as the Pearson Product Moment Correlation Coefficient can be used to test the linearity between the variables (“Pearson product-moment correlation”, n.d)). To quantify the strength of the relationship, Pearson value /

correlation coefficient (r) must be calculated (“Pearson product-moment correlation”,n.d). This numerical value ranges from +1.0 to -1.0 (“Pearson product-moment correlation”,n.d).

When $\text{Pearson} > 0$ indicates positive linear relationship (“Pearson product-moment correlation”,n.d).

$\text{Pearson} < 0$ indicates negative linear relationship (“Pearson product-moment correlation”,n.d).

$\text{Pearson} = 0$ indicates no linear relationship (“Pearson product-moment correlation”,n.d).

Table (1) presents the nature and strength between the independent and dependent factor associated with the different ranges of Pearson Values (“Finding the Pearson Correlation”, 2014).

Table (1) Pearson Value and the nature of relationship between two variables (20)

Pearson Value / r	The nature of the relationship
$r = +.70$ or higher	Very strong positive relationship
+.40 to +.69	Strong positive relationship
+.20 to +.39	Moderate positive relationship
-.19 to + .19	No or weak relation ship
-.20 to -.39	Moderate negative relationship
-.40 to -.69	Strong negative relationship
-.70 or lower	Very strong negative relationship

Independence of errors. Independence of errors refers to the assumption that errors are independent of one another. The violation of this assumption results in inaccurate standard scores and significance tests with an increased risk of Type I error (Balance, 2012). In other words, the

violations of this assumption will underestimate standard errors, and may label variables as statistically significant when they are not (Balance, 2012). So, the violation of this assumption threatens the interpretations of the analysis (Balance, 2012).

Equal variance or homoscedasticity. This assumption refers to an equal variance of errors across all levels of the independent variables (Balance, 2012). In other words, it assumes that errors are spread out consistently between the variables (Balance, 2012). This assumption is maintained when the variance around the regression line is the same for all values of the predictor variable (Balance, 2012).

When heteroscedasticity is marked it weakens the overall analysis, and may result in Type I error. (Balance, 2012). Homoscedasticity can be checked by examine the plot of the standardized residuals by the regression standardized predicted value (Balance, 2012). The Statistical software Minitab generates scatterplots of residuals with independent variables that enabled checking this assumption (Balance, 2012). In an Ideal situation, residuals should be randomly scattered around zero the horizontal line that has Zero value (Balance, 2012).

Mathematically, the variance of Y is to be same for any fixed combination of X1, X2, X3.....Xk, That is

$$\sigma^2_{y|x_1, x_2, \dots, x_k} = V_a (y|x_1, x_2, \dots, x_k) \text{ (Suri, 2006).}$$

Normality. Multiple regressions assume that the variables have normal distributions which means that the errors are normally distributed, and when the values of the residuals are plotted, it will approximately be in a normal curve (Balance, 2012).

Collinearity. It assumes that the independent variables are uncorrelated (Balance, 2012). When collinearity is low, the researcher is able to interpret regression coefficients as the effects

of the independent variables on the dependent variables (Balance, 2012). It insures the causes and effects of variables reliably (Balance, 2012).

Multicollinearity occurs when two or more independent variables are highly correlated each other (Balance, 2012). The more variables correlates the less the researchers can separate the effects of variables(Balance, 2012).

However, the independent variable are allowed to be correlated to some degree (Balance, 2012). The multicollinearity is measured by the VIF index that measures how much a variable is contributing to the standard error in the regression (Balance, 2012). When high multicollinearity exist, the variance inflation factor of the variables involved will be very large (Balance, 2012). Many statistics stated that having the VIF value more than ten is alarming about multicollinearity (Balance, 2012). While others argued that (VIF) values greater than 10 are indicator for Multicollinearity while (VIF) values greater than 100 means certain presence of Multicollinearity (“Assumption of multiple linear regression”, 2015). Mathematically, The VIF is given by the equation:

$$VIF_j = \frac{1}{1-R^2} \text{ (Suri, 2006)}$$

To solve a multicollinearity issue, it is recommended to combine the correlated variables in the analysis, or to drop one of them from the model (Balance, 2012).

Residual analysis. Residuals from multiple regression models help the researcher to evaluate the adequacy of the model in respect to the data and the assumptions that the researcher made (Hoerl, 2008). Simply, a residual is the difference between the observed value of y (the dependent variable) and the value of y predicted by the model (Hoerl, 2008).

Residual = y observed - y predicted (Hoerl, 2008).

According to this concept, there is one residual for each observation (Hoerl, 2008). Minitab typically standardizes residuals and put them on a common scale (Hoerl, 2008). In the Ideal case, if the regression model fit the data perfectly, the residuals would all be zero (Hoerl, 2008). Therefore, the analysis of the residuals is an effective method to evaluate the fit of the model to the data, and to decide whether the model is useful (Hoerl, 2008). There are a variety of residual plots to look for patterns and trends (Hoerl, 2008). Here are four plots to consider:

Residuals vs. predicted values. The good regression model that fits to the data satisfied the typical assumption of independent normally distributed residuals, and the plot of the residuals versus predicted values shows no pattern or trend (Hoerl, 2008). When nonrandom pattern is observed, the form of the model should be changes (Hoerl, 2008).

Histogram of residuals. A histogram highlights the overall distribution of the residuals, examines the expected bell-shaped of the distribution, checks the outliers of the residuals that are formed when the model does not adequately fit one or more observations (Hoerl, 2008). Outliers suggest that observations are due to a special cause that could be a measurement error or may indicate that the model is inadequate (Hoerl, 2008).

Residuals vs. normal probability scale. The normal probability plot of the residuals is used to check the assumption that the residuals are normally distributed (Hoerl, 2008). This technique is better than using the histogram (Hoerl, 2008)

In the normal probability plot, the software calculates the normal probability scale, and plot the residual versus it (Hoerl, 2008). The model is considered that it stratified the assumption that the residuals follow a normal distribution if the residual plot follows approximately a straight line (Hoerl, 2008). Any real data will never follow an exact normal distribution and no

perfect line in this plot will be found, the presence of a general linear trend satisfies this assumption (Hoerl, 2008).

Residuals vs. observation sequence. This plot show no trends when the model is adequate and no special causes if you are using the individuals control chart (Hoerl, 2008). A trend would suggest there is a variable which is not currently in the model that have changed during the time spanned by these data, adding this variable will it would improve the predictability of the model if possible (Hoerl, 2008).

Hypothesis testing. Defining and evaluating hypotheses is a very critical part of statistical inference (Easton, & McColl, n.d). In order to perform this task, some theory should be setup put forward, either because that theory is believed to be true or because it is to be used as a basis for argument, but has not been proved (Easton, & McColl, n.d). Each problem leads simply into two competing hypotheses; the null hypothesis, denoted H_0 , against the alternative hypothesis, denoted H_1 (Easton, & McColl, n.d).

Null hypothesis. The null hypothesis, H_0 , represents a theory that the researcher puts forward, either because it is believed that it is true or because it will be used as a basis for argument, but has not been proved so far (Easton, & McColl, n.d). The null hypotheses had a special consideration because the null hypothesis relates to the statement being tested, while the alternative hypothesis relates to the statement to be accepted if the null was rejected (Suri, 2006).

After carrying out the test, the decision should be given in terms of the null hypothesis (Suri, 2006). The researcher either "Rejects H_0 in favor of H_1 " or "Does not reject H_0 "; that means that the researcher never decide to "Reject H_1 ", or "Accept H_1 "(Suri, 2006).

If the researcher concludes "Do not reject H0", this does not always mean that the null hypothesis is true, it just suggests that there is not sufficient evidence against H0 in favor of H1(12). In other words, rejecting the null hypothesis then, suggests that the alternative hypothesis may be true (Suri, 2006).

Alternative hypothesis. The alternative hypothesis, H1, is a statement of what a statistical hypothesis test is set up for (Suri, 2006).

Type 1 error. In a hypothesis test, a type I error occurs when the researcher rejects the null hypothesis while it is true in fact (Easton, & McColl, n.d). Type I error is often considered to be more serious than Type 2 error. therefore researchers should be careful in rejecting H0 (Easton, & McColl, n.d). The significance level is considered to guarantee 'low' probability of rejecting the null hypothesis wrongly; this probability is never have zero value (Easton, & McColl, n.d). The probability of a type I error could be computed as $P(\text{type I error}) = \text{significance level} = \alpha$ (Suri, 2006).

Type 2 Error. Sometimes, if the researcher did not reject the null hypothesis, it may still be false (a type II error) that may happen because the sample is not big enough to identify the falseness of the null hypothesis(Suri, 2006). The exact probability of a type II error is usually unknown. It is usually symbolized as β (Suri, 2006). For any Hypotheses testing for any set if data, type I and type II errors are inversely related; the higher the risk of one, the smaller the risk of the other (Suri, 2006).

$$P(\text{type II error}) = \beta$$

Test Statistic. “A test statistic is a standardized value that is calculated from sample data during a hypothesis test. You can use test statistics to determine whether to reject the null hypothesis” (“What is a test statistics”, n.d).

F- Value: It is the statistic value that is generated through (ANOVA) analysis (“What is a test statistics?” n.d). It evaluates the overall significance of the model by comparing it with a model with no predictors (intercept model), and it is usually used to compare between two models as well (Frost, 2015). The P value of the F-test of overall significance test should be examined, if the P- Value is less than the significance level of the study, the null-hypothesis is rejected and a conclusion could be made that the model of the study provides a better fit than the intercept-only model (Frost, 2015).

T-Value: It is part of the test statistics, this value is computed by the coefficient associated with the parameter by its standard error (Dallal, 2000). It helps in making the decision of rejecting or not to reject the hypothesis test (“Reading and using STATA output, n.d). The H₀ could be rejected when a T-Test value is greater than 2 that happen when the coefficient is almost twice as large as the standard error. Acceptable T- values should be associated with p-values less the 0.05 to make the conclusion that these factors are important in our model (“Reading and using STATA output, n.d).

Significance Level: The significance level of a statistical hypothesis test represents the fixed probability of wrongly rejecting the null hypothesis H₀, while it is true in reality (Suri, 2006). The researcher needs to make the significance level as small as possible in order to protect the null hypothesis by minimizing the probability of Type I error (Suri, 2006). As stated earlier, the significance level is usually denoted by α (Suri, 2006). Significance Level = P (type I error) = α and it is usually chosen to be 0.05 (Suri, 2006).

P-Value: “The probability value (p-value) of a statistical hypothesis test is the probability of getting a value of the test statistic as extreme as or more extreme than that observed by chance alone” (Suri, 2006). if the null hypothesis H_0 is true. It is the probability of wrongly rejecting the null hypothesis if it is in fact true” (Suri, 2006).

The critical P-value is usually equal to the significance level of the test at the researcher will reject the null hypothesis (Easton, & McColl, n.d). The p-value of the test should be compared with the significance level, if the P-value is smaller, the result is significant, and therefore, if the null hypothesis were to be rejected at the 5% significance level, this is reported as $p < 0.05$ (Easton, & McColl, n.d).

Smaller p-values suggest that the null hypothesis is unlikely to be true, and the smaller the P-Value the more confident the researcher can reject the null hypothesis (Easton, & McColl, n.d).

The rejection region represents the area of the values of the test statistic that leads to reject the null hypothesis (Zaiontz, 2015). With 95% significance level, P-value of 0.05 or smaller leads to rejecting the null hypotheses. A P-value that is greater than 0.05 leads to failure in rejecting the null hypotheses (Zaiontz, 2015).

The rejection areas are classified to right tail, left tail, and two tails sampling distribution (Zaiontz, 2015). The following figure explains the concept of each of those the rejection areas (Zaiontz, 2015).

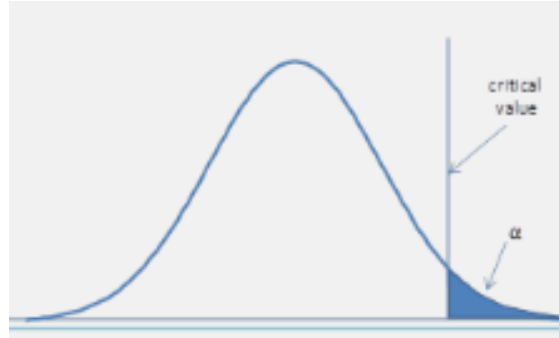


Figure 2 – Critical region right tail (28)

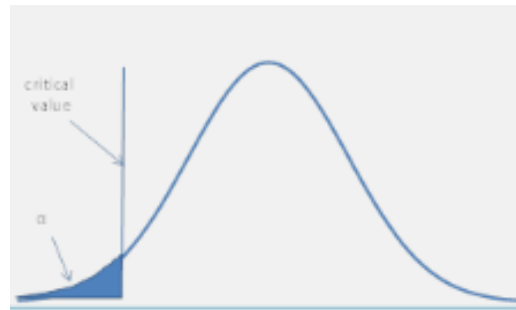


Figure 3 – Critical region left tail (28)

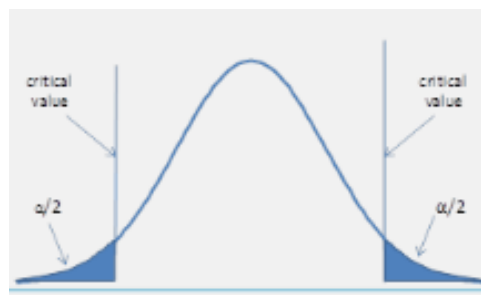


Figure 4 – Two tailed hypothesis testing (28)

Summary

This chapter represents the literature review conducted by the beginning of this study. It covered a back ground of the problem, a literature review related to the problem, and a literature review related to the methodology.

The next chapter covers the methodology of data collection and initial analysis.

Chapter III

Methodology

Introduction

This chapter covers the approach that has been used to conduct this study and the design of the study. It clearly identifies the steps of building the forecasting model including the data collection. In addition, it verifies the reliability of the sources of the data, and describes the tools and techniques that have been used to analyze the data. It also defines the hypotheses that have been tested and statistical values that have been used to analyze the data and interpret the output of the regression analysis.

Design of the Study

This study was designed to enable the decision makers to predict a numerical output of the future demand of new single family houses in the USA using a quantitative approach of analysis. The frame work of the study could be summarized in the following steps:

- 1- Identify several economic and social variables that were believed to affect the market demand of new single family houses to be used as independent factors in the regression model.
- 2- Collect the historical data of this these independent variables along with the historical data of the dependent factor of the model from reliable sources of data.
- 3- Evaluate the nature of the relationship between each of the independent variable and the dependent variable using the historical data of each of the independent factors and the dependent factor. Scatter plots were used as technique evaluate the relationship between each independent factor and the dependent factor as it visualize the relationship between the two factors. In addition, a correlation test was performed between each of the

independent factor and the dependent factor to generate Pearson Values, P-Values to assess the nature and the strength of the relationship between each of the independent factor and the dependent factor.

- 4- Build different regression models using the data of the independent variables that have been proven to have a linear relationship with the dependent variable to reach to a model that can affectively predict the demand and satisfies all the assumptions of a linear regression model.

This type of study could be classified as associative/causal forecast, the quantitative approaches usually are used for associative / casual forecasts as they aim to draw numerical conclusion from numerical data by using quantitative analysis that is steady and robust.

Data Collection

The first step in this study was to identify the possible factors that may affect the demand of the sales of the single family (the dependent variable). A Brainstorm session took place to define possible factors that may affect the dependent factor. A consultation of many experienced individuals in the area of houses marketing was considered to verify that the identified independent factors sounds reasonable to have a relationship with the independent variable. A literature review related to the problem was conducted to assure possible effects of the chosen independent variables on the dependent variable. The factors that have been identified as independent variables were:

- GDP
- Personal Consumption
- Inflation
- Unemployment Rate

- mortgage Rate
- Population USA

The historical data were collected for these factors along with the data of sales of the new single family houses in the USA through 21 years, from 1994 till 2014.

These data sets could not be primary data as they could not be collected by the researcher; they are secondary data that have been collected by other sources. However the researcher is responsible for collecting the data from reliable sources.

The data were collected from reliable data. Following is a brief explanation about each factor chosen to be evaluated in the study and a description of the source of the data to verify its reliability.

Population. The source of this data is the Bureau of Labor Statistics / US (“US. Bureau of the Census”, n.d). Bureau of the Census is responsible for conducting census of population through a monthly population survey, estimating and projecting the population numbers in the future. This data covers all ages of people in the USA, in thousands, on monthly basis. An average was used to find the population of the USA for each year.

The data source is well known source of information, it is a reliable source as this data is based on the survey by sampling method approved by the statisticians. There is no doubt that this data is trusted to be included in the model if it was proven to be correlated with the dependent variable.

GDP. Gross domestic product (GDP), it is simply the monetary value of all the finished goods and services produced within the USA country's borders in a specific time period. GDP is usually calculated on an annual basis. It is an economic index that measure of the size of an

economy by measuring the economic performance of a whole country. GDP was considered as a possible factor that may affect the dependent factor of our model.

The source of this data is the USA Department of Commerce (“US. Bureau of Economic Analysis” n.d). The mission of this department is to improve living standards for all Americans by creating an infrastructure that supports economic growth, competitiveness in technologies. This department is responsible for gathering economic and demographic data for business and government decision-making, and helps in setting industrial standards by using approved methods by the statisticians.

Definitely this source considered as a reliable source of data. This data is trusted to be used in the regression model if the correlation relationship between GDP and the dependent variable was proven.

Personal expenditure income. it is an economic index that measures the price changes in consumer goods and services. It is essentially measures the cost of goods and services consumed by individuals.

The source of this data is the US. Bureau of Economic Analysis (“US. Bureau of Economic Analysis” n.d). It is an agency that provides different economic statistics including the gross domestic product of the United States. It is a principal agency of the U.S. Federal Statistical System. It produces many economic statistics that influence decisions of government officials, business people, and even individuals.

For sure, this source considered as a reliable source of data. This data is trusted to be used in the regression model if the correlation relationship between personal expenditure income and the dependent variable was proven.

Unemployment rate. It is the percentage of the workforce that is unemployed and is looking for a paid job. Unemployment rate is a very important economic index because a rising in the unemployment rate is considered as a sign of weakening economy, while a falling rate in unemployment indicates a growing economy.

The source of this data for this factor is the Bureau of Labor Statistics (“Unemployment rate in the United States”, n.d). This source has been discussed earlier in this paper and considered as a reliable source of data. Therefore, this data could be trusted and included in the regression model if a correlation relationship was proven between the unemployment rate and the dependent variable of the regression model.

Fixed mortgage rate. It is the rates of interest mortgage that lender charges upon lending money, where the interest rate remains the same through the term of the loan. In other words, payment amounts and the duration of the loan are fixed. Each lender has different mortgage rate.

The source of this data is the Federal Reserve of bank of St. Louis (“15- years fixed rate mortgage”, n.d). Which is one of twelve regional Reserves Banks relates to Board of Governors in Washington, D.C. This data could be considered as a reliable source of data and could be used in the model if a correlation relationship with was proven with the dependent factor.

Inflation. Inflation in economics is defined as a sustained increase in the general price level of goods and services in an economy over a period of time. As the price level of goods or services rises, the unit of currency buys fewer goods and services. Therefore inflation is an index that reflects a reduction in the purchasing power per unit of money.

The source of this data is the Bureau of Labor Statistics (“ Annual inflation rate”, n.d.), which is a unit of the United States Department of Labor. This agency is responsible for

collecting, processing, analyzing, and disseminating essential statistical data to the American public, the U.S. Congress, State and local governments, business, other Federal agencies and labor representatives. It conducts researches into how much families need to earn to be able to enjoy a suitable standard of living.

This data could be identified as a very reliable data, and it is trusted to be used in the regression model if inflation was proven to be correlated to the dependent factor.

New Single Family home sales in USA. This is the factor of interest of this study as it is the dependent factor and the model. The source of this data is the US, Bureau of the Census (“US. Bureau of the Census”, n.d) It is a principal agency of the U.S. Federal Statistical System and it is responsible for producing data about the American people and economy.

This source of data is completely reliable. This data will be used in the model as it is dependent / predicted factor of the regression model.

Data Analysis

The main data analysis of this study is to analyze the regression model of the study; however, the regression model requires checking the assumptions about the data that will be included in the model in order to verify that the regression model will be robust.

The linearity assumption has been checked prior to the model building, while the other assumptions have been checked after building the model.

Correlation analysis. This step is extremely important in building any regression model as it will help in sorting the data. A very careful analysis should be conducted to assess the relationship between each independent variable and the dependent variable. Forecasters must be completely careful in adding and eliminating the independent variables, that’s because adding an

unnecessary variable in the model will make the model more complex without making it more accurate, while eliminating a useful factor will lead to an unreliable model from accuracy aspect.

Minitab was used to generate scatter plots to visualize the relationship that exists between each the independent variable and the dependent variable. Minitab was used to conduct a correlation test. Pearson values were generated in this step to determine if there is a linear relationship between each independent variable and the dependent variable, and to define the level of strength of that relationship.

The three possible cases of relationship between the dependent and independent factors are:

- Positive linear relationship (very strong, strong, moderate, weak, negligible)
- Negative linear relationship(very strong, strong, moderate, weak, negligible)
- No linear relationship

A hypotheses testing was required in this step, the two hypotheses were:

H₀: There is no correlation between the two factors.

H₁: There is a correlation between the two factors.

This test was conducted at 95% significance level, the P-Value generated through this test was used to accept or reject the null hypothesis. Table (2) presents the P-values and the associate determination regarding H₀.

P- Value	Null Hypothesis
P- Value \leq 0.05	Reject the H ₀
P- Value \geq 0.05	Accept the H ₀

(Table 2) P-Value to accept or reject the Null Hypothesis

Pearson value generated from the test used to determine the direction and the strength of the relationship. Please refer to (Table 1).

Step-wise regression. Stepwise regression is an automated tool in Minitab that is usually used in the exploratory stages of regression model building to identify a useful subset of predictors. The process systematically adds the most significant variable or removes the least significant variable during each step.

Standard step wise was used to identify the most important factors in the model. The independent variables to be included in this step are only the factors that have been proven to have a strong linear relationship with the dependent factor.

The step wise regression usually provides an understanding about the degree of importance of each independent factor. However, the weakness of this test is that it does not consider the multicollinearity between the independent factors. The multicollinearity should be checked while conducting the general regression analysis.

Regression analysis. Using the recommended independent factors by the step wise regression test, a general regression test was conducted to evaluate the model and its components. The two Hypotheses that have been tested in the ANOVA analysis for each independent variable at this step are:

H0 the independent variable has no effect on the dependent variable.

H1 the independent variable has an effect on the dependent variable.

The significance level of the test was 95%, The P- Value of (0.05) was used to accept or reject the null Hypothesis. Please refer table (2)

Other test statistics were evaluated to understand the output of the regression test. These elements has been covered thoroughly in the literature review. Here is a quick reminder about their meaning and interpretation.

Regression coefficients. As these values represent the mean change in the response variable for one unit of change in the predictor variable while holding other predictors in the model constant, then, The greater the coefficient, the steeper the slope, the greater change in the response/dependent variable.

Coefficient of multiple determination / R-Squared. As this value measures how close the data are to the fitted regression line and presents the percentage of the response variable variation that is explained by a linear mode, then, the higher the R-squared, the better the model fits the data. However, The R- Squared is not the best value to draw a conclusion about the model as it increases by increasing the numbers of the independent variables even if they are not important. The adjusted R- squared is usually used to draw a conclusion about the model.

Adjusted R-squared. The adjusted R-squared is a modified version of R-squared for the number of predictors in a model. The greater this value the better the model.

T-Value. As this value represent the value of the estimated coefficient for each independent variables divided by its own standard, and it measures how many standard deviations the estimated coefficient is from zero, then the bigger the t- value the more significant the coefficient. This value is used in an association to P-Value to accept or reject the null hypothesis. An absolute t- value less than two leads to reject the null hypothesis.

F- Value. As this evaluates the overall significance of the model by comparing the model to the intercept model. This Value is usually used to compare between different models, the greater this value, the better the model.

VIF. These the independent values of the model. There is a debate about the values that presents multicollinearity, some researches consider VIF values more than (10) are alarming, while others consider values could be generated in the regression test, and are used to check the

collinearity assumption between (VIF) Values in range of 10 to 100 are just an indicator of Multicollinearity, but (VIF) values greater than (100) assure that the multicollinearity is definitely exists between the independent variables.

Residual analysis for checking assumptions. The assumptions of the linear regression were checked through the residual plot.

Equal variance or homoscedasticity assumption. Residual verses fit assumption could be generated individually for each independent factor in the model through Minitab. A random patten should be noticed in the plot to fulfill this assumption.

Normality assumption. Although the normality assumption has been checked through the regression, an individual normality test through Minitab has been conducted to verify that the residuals of the regression are normally distributed. The normality test generates a normality plot and provides information about the mean value, the standard deviation, the Anderson value, the P-value.

The two hypotheses to be tested at this step are:

H0 the data are normally distributed.

H1 the data are not normally distributed.

The P-value was used to reject or accept the null hypothesis. Please refer to table (2). The Anderson Value should be less than 0.75.

Budget

This project has been conducted individually by the researcher at no cost for data collecting or analysis.

Timeline

This project has been conducted through several months. It went through couple of important phases

- Defining the problem and preparing a proposal.
- Literature review and data collection.
- Data analysis and Calculation.
- Preparing the paper and presentation.

The Gantt chart down explains these phases.

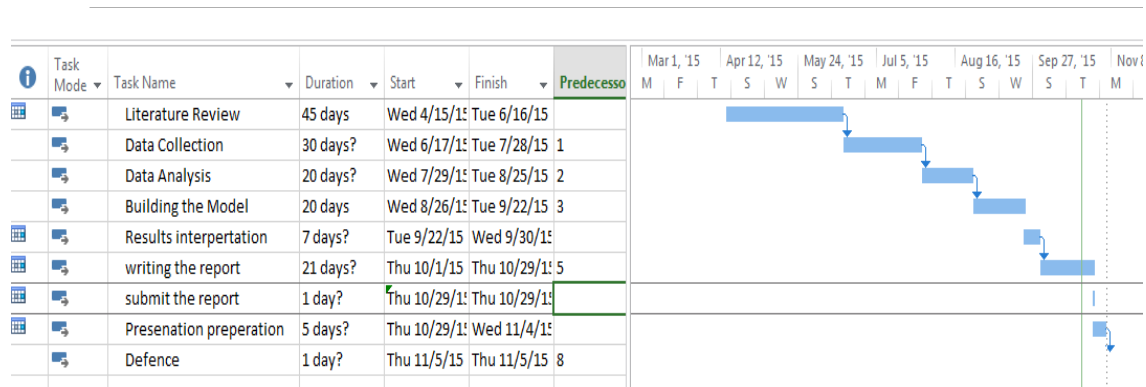


Figure 5 Gantt chart of the time line of the project

Summary

This chapter covered the design of the study and the approach of conducting the search, defined the framework of the study and presented the steps of building the forecasting model and collecting the data, verified the reliability of the sources of the data, and described the tools and techniques that have been used to perform the statistical tests and described the statistic values that have been used to analyze the data in this study.

In Addition, this chapter addressed that this project was conducted with no cost and presented the timeline of conducting this study.

The next chapter covers an in- depth data analysis.

Chapter IV

Data Presentation and Analysis

Introduction

After collecting the data from reliable sources, the next step was to analyze the collected data to determine the independent factors to be included in the model, and run the regression analysis to predict the demand of buying new single family homes in the United States. This chapter covers the in-depth analysis of the nature of the relationship between each independent factor and the dependent factor. The steps of building the regression model, and analyze the results.

Data Presentation and Analysis

The next step in building the model after choosing the independent factor was to check the linear regression assumptions for the data sets.

Linearity assumption. As the most important assumption to consider a factor as an independent variable in the regression model is to prove a correlation relationship between that factor and the dependent factor, a correlation analysis was conducted between each independent factor and the dependent factor to sort the data and find the data that could be considered in the model.

The hypotheses have been set up to check this assumption as the following for each independent variable.

H0: There is no correlation (linear relationship) between the independent factor the dependent factor.

H1: Hypotheses there is a correlation (linear relationship) between the independent variable and the dependent variable

Inflation Vs New single family home sales.

Data presentation. The following scatter plot, Pearson Value, P- value were generated through using correlation analysis between inflation and the sales of a new single family home in USA.

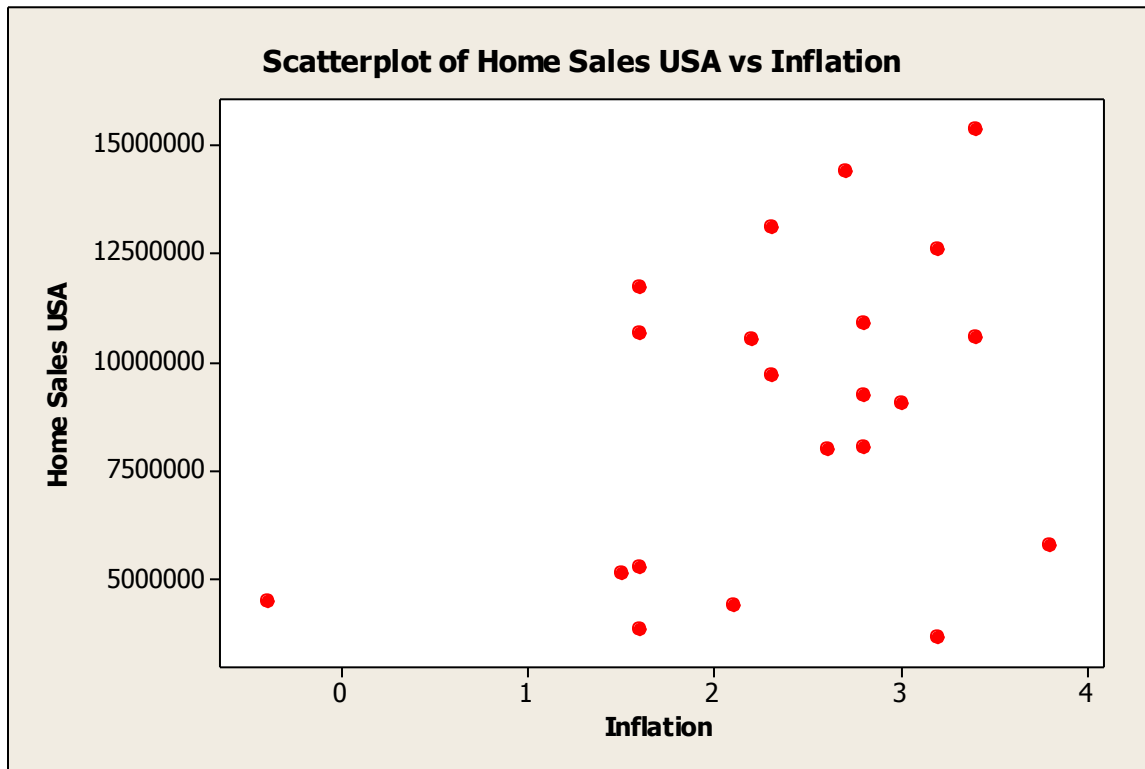


Figure 6 Scatter plot of the Home sales VS Inflation

Pearson correlation of Home Sales USA and Inflation = 0.361

P-Value = 0.108

Data analysis. The scatter plot indicates that there is no linear relationship between the inflation and the sales of new single family homes in USA.

The Pearson value indicates a moderate positive relationship between the inflation and the dependent variable.

The P-Value is greater than 0.05. That's indicates that we failed to reject the null hypotheses that states that there is no correlation between the two factors. So we accept the conclusion that Inflation factor has no linear relationship with the dependent variable. Therefore, the inflation factor cannot be considered as an independent factor in the regression model to predict the demand of the sales of the new single family homes in the USA.

Correlation analysis: mortgage rate Vs new single family homes sales.

Data presentation. The following scatter plots, Pearson Value, P- value were generated through using correlation analysis between the mortgage rate and the sales of a new single family home in USA.

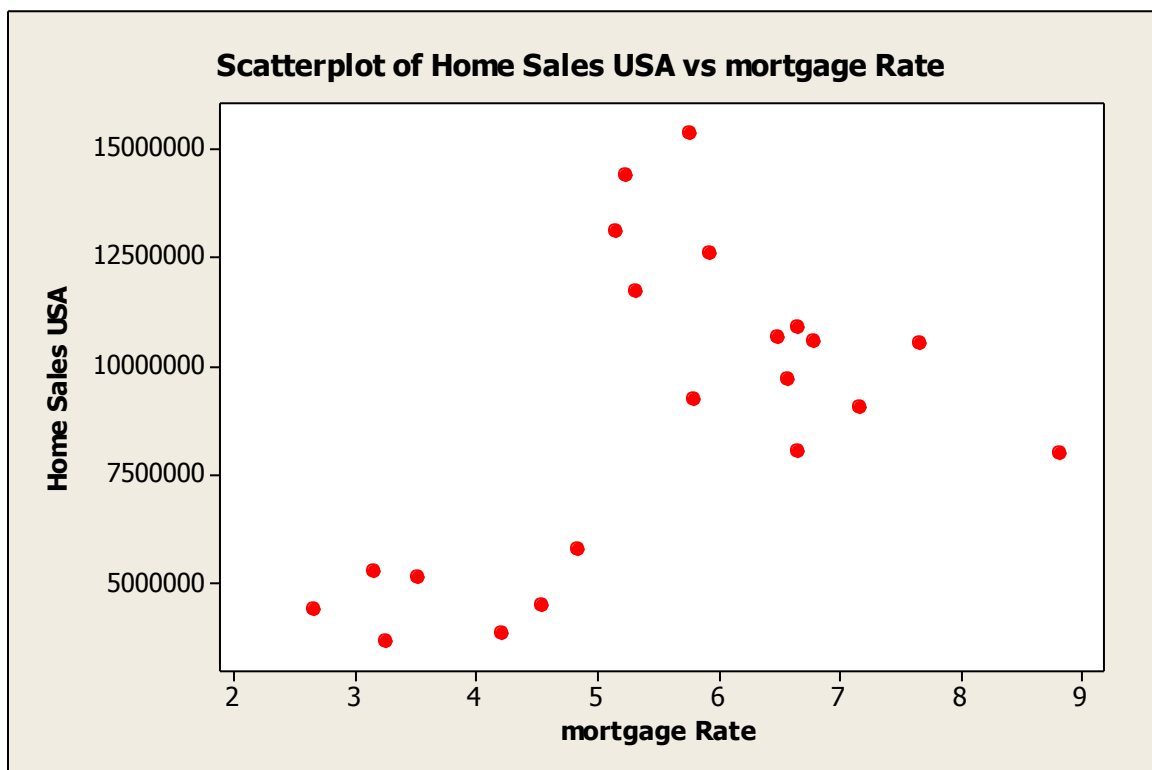


Figure 7 Scatter plot of the Home sales VS Mortgage Rate

Pearson correlation of Home Sales USA and mortgage Rate = 0.527

P-Value = 0.014

Data analysis. The scatter plot indicates possible linear relationship between the inflation and the sales of new single family homes in USA.

The Pearson value indicates a strong positive relationship between the inflation and the dependent variable.

The P-Value is less than 0.05. That's indicates that to null hypotheses (There is no correlation between the two factors) is rejected. The mortgage rate is considered as an independent variable to be included in the model.

Unemployment rate Vs new single family homes sales.

Data presentation. The following scatter plots, Pearson Value, P- value were generated through using correlation analysis between the unemployment rate and the sales of a new single family home in USA.

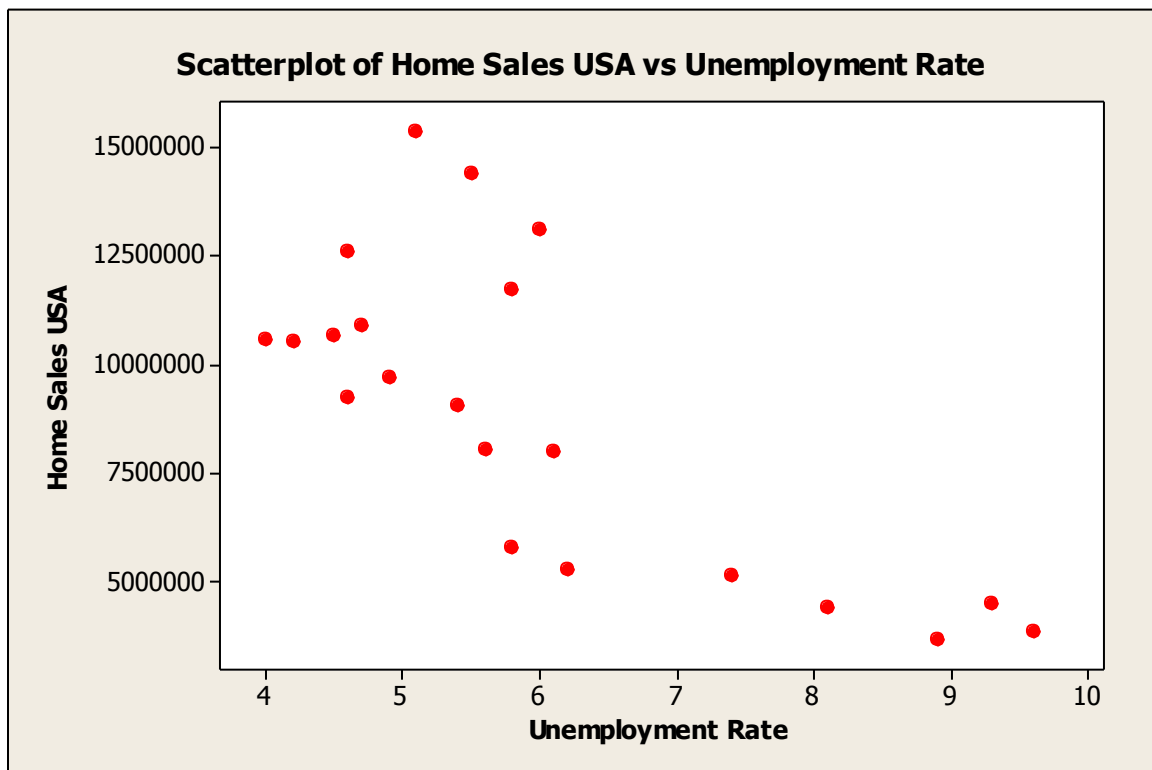


Figure 8 Scatter plot of the Home sales VS Unemployment Rate

Pearson correlation of Home Sales USA and Unemployment Rate = -0.748

P-Value = 0.000

Data analysis. The scatter plot indicates that there is a negative linear relationship between the unemployment rate and the sales of new single family homes in USA.

The Pearson value indicates a very strong negative relationship between the inflation and the dependent variable. The P-Value is 0, this value indicates that the null hypotheses that this factor has no effect on the dependent factor should be rejected with a very high level of confidence.

According to this analysis, the unemployment rate factor should be considered as an independent factor in the regression model to predict the demand of the sales of the new single family homes in the USA.

Correlation analysis: population Vs new single family homes sales.

Data presentation. The following scatter plots, Pearson Value, P- value were generated through using correlation analysis between the population and the sales of a new single family home in USA.

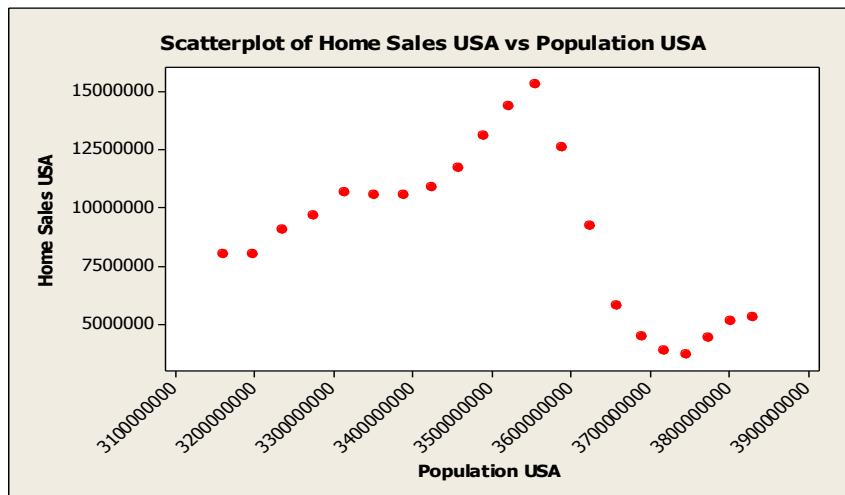


Figure 9 Scatter plot of the Home sales VS Population

Pearson correlation of Home Sales USA and Population USA = -0.480

P-Value = 0.028

Data analysis. The Pearson value indicates that there is a moderate negative relationship between Population and the sales of a new single family home in USA. The P- Value indicates that this relationship is not significant statistically.

According to this analysis the population should not be included in the regression model. However, the scatter plot indicates that there was an important positive relationship between the populations and the sales of new single family houses for a period of time. Then this relationship was converted into a negative relationship between the two factors, again, the relationship between the two factors appeared again.

An investigation was conducted to understand this fluctuation in the relationship between the two factors. It has been noticed that the conversion of the relationship began at observation (Frost, 2013), it is the year of 2006, and it is the year before the financial crisis occur in the United States on 2007. This financial crisis affected severely the all the economic factors, especially the construction industry and property market. By 2010 the effect of the financial crisis reached to its end, and beginning 2011 the normal positive relationship between the two factors showed up again. That's indicates that under normal conditions there is a very important relationship between the population and the dependent variable.

An in-depth analysis was conducted to prove this assumption, by dividing the data of population Vs the sales of new single family houses into three periods.

Period one: From 1994 to 2005.

Period two: from 2006 till 2010.

Period three: from 2011 till 2014.

The following Pearson's Values and P-Values were generated through conducting a correlation analysis between the two factors through Minitab, considering the three periods.

Correlations: Home sales 1, Population 1

Pearson correlation of Home sales 1 and Population 1 = 0.925

P-Value = 0.000

Correlations: Home Sales 2, Population 2

Pearson correlation of Home Sales 2 and Population 2 = -0.541

P-Value = 0.166

Correlations: Home Sales 3, GDP3

Pearson correlation of Home Sales 3 and GDP3 = 0.947

P-Value = 0.053

The Pearson Values for period one and three indicates a very strong positive relationship between the population and the dependent factor. The P- Values for both periods one and three indicates that this relationship is highly significant, and the H_0 should be conveniently rejected. This analysis indicated that even though there is a conversion in the relationship between the two factors in period two, this factor cannot be excluded as a dependent factor in the model as population was believed to be an important factor the affect out dependent factor. However, a suitable technique should be used to allow the use of this fluctuated set of data in the regression model. The technique of dummy factors was used to overcome the fluctuation in the relationship between the dependent and independent variables in this set of data.

Three new dummy factors were considered to be added to the model to balance the fluctuation in the data. The following table presents the values of the dummy periods added to the model.

Table 3. The Values of the Dummy Variables

Year	Period 1	Period2	Period3
994	1	0	0
1995	1	0	0
1996	1	0	0
1997	1	0	0
1998	1	0	0
1999	1	0	0
2000	1	0	0
2001	1	0	0
2002	1	0	0
2003	1	0	0
2004	1	0	0
2005	1	0	0
2006	0	1	0
2007	0	1	0
2008	0	1	0
2009	0	1	0
2010	0	1	0
2011	0	0	1
2012	0	0	1
2013	0	0	1
2014	0	0	1

Personal consumption Vs New single family homes sales

Data presentation. The following scatter plots, Pearson Value, P- value were generated through using correlation analysis between the personal consumption and the sales of a new single family home in USA.

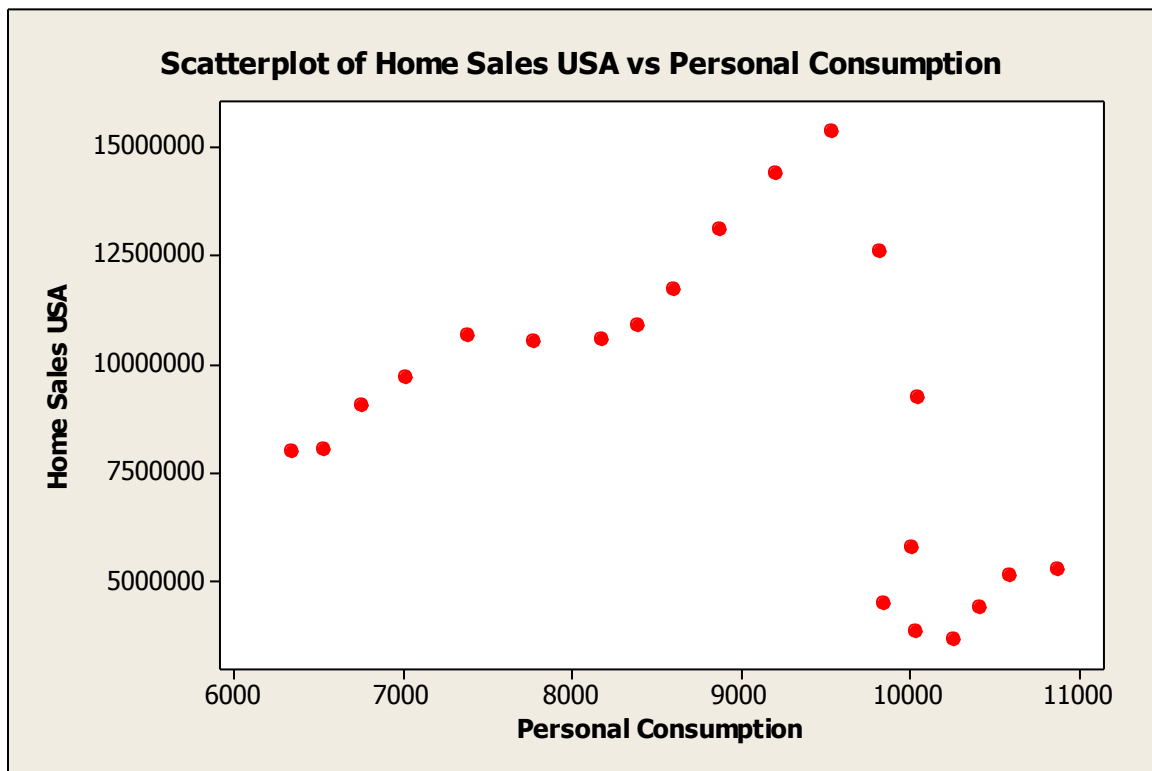


Figure 10 Scatter plot of the Home sales VS Personal Consumption

Pearson correlation of Home Sales USA and Personal Consumption = -0.353

P-Value = 0.1

Data analysis. The Pearson value indicates that there is a moderate relationship between the personal consumption and the sales of a new single family home in USA. The P- Value indicates that this relationship is not significant statistically.

According to this analysis the personal consumption should not be included in the regression model. However, the scatter plot indicates that there was an important positive relationship between the personal consumption and the sales of new single family houses for a

period of time. Then this relationship was converted into a negative relationship between the two factors, again, the relationship between the two factors appeared again.

The conversion of the relationship between the two factors is again due the financial crisis that took place on 2007. Its effect started on 2006 and last till 2010. So this factor should not be excluded as an independent variable. In order to verify the linearity between this factor and the dependent factor, another correlation analysis was conducted after dividing the set of data to three periods (the same three periods have been identified through population correlation analysis). The following Pearson's values and P-Values were generated through this analysis.

Correlations: Home sales 1, Personal Consumption 1

Pearson correlation = 0.920

P-Value = 0.000

Correlations: Home Sales 2, Personal consumption 2

Pearson correlation = -0.119

P-Value = 0.779

Correlations: Home Sales 3, Personal Consumption 3

Pearson correlation = 0.917

P-Value = 0.083

The Pearson Values for both Periods one and three indicates a very strong positive relationship between the personal consumption and the dependent factor. The P- Value of period one indicates that this relationship is highly significant, and the null hypotheses could be rejected with a very high level of confidence. This analysis indicated that even though there is a conversion in the relationship between the two factors in period two, this factor cannot be

excluded as a dependent factor in the model. However, a suitable technique should be used to allow the use of this set of fluctuated data.

The technique of dummy factors that have been used to overcome the fluctuation in the population set of data was used to handle the fluctuation in the personal consumption set of data as well.

GDP Vs New single family homes sales.

Data presentation. The following scatter plots, Pearson Value, and P- value were generated through using correlation analysis between the GDP and the sales of a new single family home in USA.

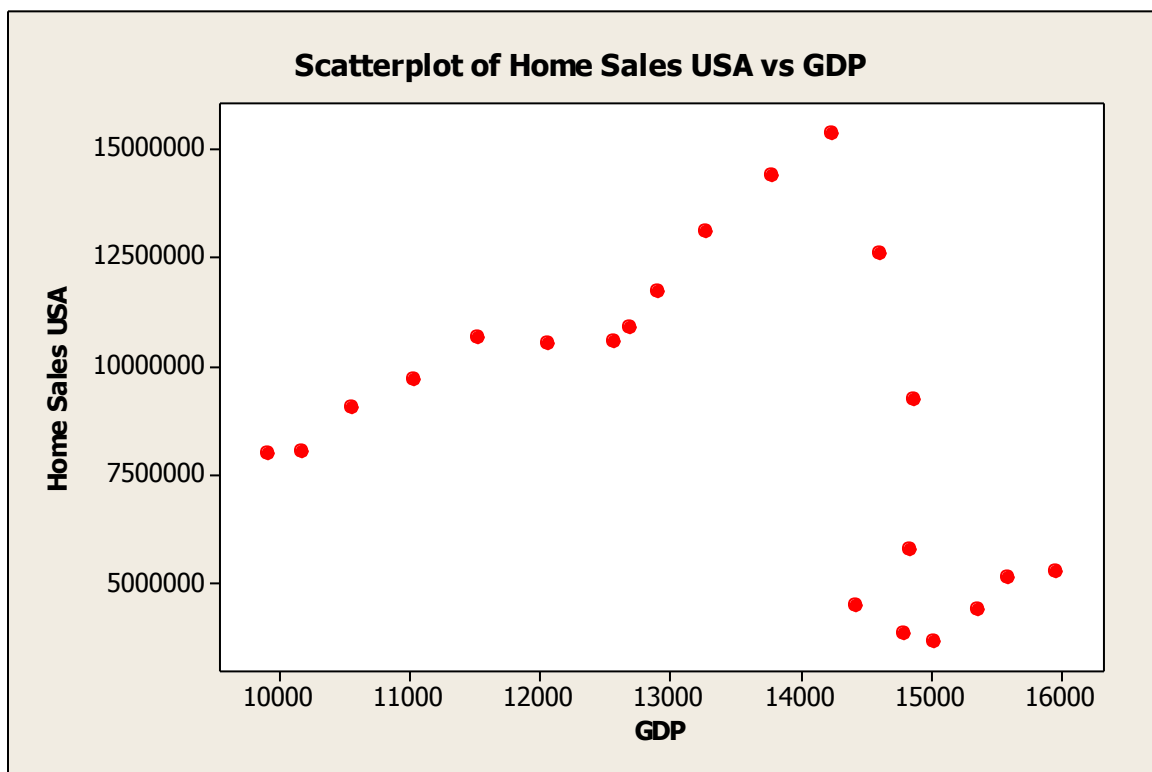


Figure 11 Scatter plot of the Homes sales VS GDP

Pearson correlation of Home Sales USA and GDP = -0.335

P-Value = 0.137

Data analysis. The Pearson value indicates that there is a moderate negative relationship between GDP and the sales of a new single family home in USA. The P- Value indicates that this relationship is not significant statistically.

According to this analysis the GDP should not be included in the regression model. However, the scatter plot indicates that there was an important positive relationship between the GDP and the sales of new single family houses for a period of time. Then this relationship was converted into a negative relationship between the two factors, again, the relationship between the two factors appeared again.

The conversion of the relationship between the two factors is also due to the financial crisis that took place in 2007. Its effect started in 2006 and lasted until 2010. So this factor cannot be excluded as an independent variable. The correlation analysis was conducted again after dividing the set of data into three periods, which have been identified through Population analysis. The following Pearson's values and P-Values were generated through this analysis.

Correlations: Home sales 1, GDP 1

Pearson correlation of Home sales 1 and GDP 1 = 0.920

P-Value = 0

Correlations: Home Sales 2, GDP2

Pearson correlation of Home Sales 2 and GDP2 = -0.028

P-Value = 0.947

Correlations: Home Sales 3, GDP3

Pearson correlation of Home Sales 3 and GDP3 = 0.991

P-Value = 0.009

The Pearson Values for both Periods one and three indicates a very strong positive relationship between the Personal consumption and the dependent factor. The P- Values for both periods one and three indicates that this relationship is highly significant. This analysis indicated that even though there is a conversion in the relationship between the two factors in period two, this factor cannot be excluded as a dependent factor in the model. However, a suitable technique should be used to allow the use of this set of fluctuated data.

The technique of dummy factors that have been used to overcome the fluctuation in the population set of data will handle the fluctuation in the GDP set of data.

In Summery,the correlation analysis indicates that the inflation factor should not be considered as independent factors in the regression model. All the other factors should be included as independent variables in the model.

Three dummy variables were decided to be added to the regression model to handle the fluctuation in the data sets due the financial crisis of 2007.

Stepwise data presentation. To define the most important variables in the model, a step wise analysis has been conducted to automatically identify the useful subset of predictors to be included in the model. Only the factors that have been fulfill the assumption of linearity were included in this step. The three dummy factors of Period 1, Period 2, and period3 has been added the factors to be examined in this step.

Stepwise Regression: Home Sales USA versus GDP, Personal Consumption,

Alpha-to-Enter: 0.15 Alpha-to-Remove: 0.15

Response is Home Sales USA on 8 predictors, with N = 21

Step	1	2	3	4	5
Constant	18390410	13727351	-3015417	127366927	136691887
Unemployment Rate	-1581208	-1070762	-1054758	-136996	
T-Value	-4.87	-2.89	-3.62	-0.49	
P-Value	0.000	0.010	0.002	0.629	
Period1	2787907	6616924	5368999	5415700	
T-Value	2.27	4.52	5.26	5.45	
P-Value	0.036	0.000	0.000	0.000	
Personal Consumption	1628	8800	9343		
T-Value	3.48	5.58	8.47		
P-Value	0.003	0.000	0.000		
Population USA	-0.0566	-0.0609			
T-Value	-4.64	-7.25			
P-Value	0.000	0.000			
S	2441655	2211822	1739120	1170621	1144231
R-Sq	55.56	65.45	79.83	91.40	91.27
R-Sq(adj)	53.22	61.61	76.27	89.25	89.73

The automated step wise regression, recommended a model includes Period 1, Population, Personal consumption as independent variables in the model.

The adjusted R-Square Value for this model is 89.73; This Value indicates that the model explains 89.73 % of the variation of the dependent value of the model. The model could be considered as a valid model if it was proven to fulfill the assumption of multicollinearity between the independent variables. To check the assumption, a normal regression analysis was conducted, it has been named as Model 1.

Step wise regression analysis data presentation/Model 1.

By conducting a regression analysis with the factors have been recommended by the step wise regression (personal consumption, population, and period one). The following equation and results were generated.

Regression Analysis: Home Sales U versus Personal Con, Population U, ...

The regression equation is

$$\text{Home Sales USA} = 1.38\text{E}+08 + 9391 \text{ Personal Consumption} - 736 \text{ Population USA} + 5419252 \text{ Period1}$$

Predictor	Coef	SE Coef	T	P	VIF
Constant	137955853	20482516	6.74	0.000	
Personal Consumption	9391	1096	8.57	0.000	39.075
Population USA	-736.3	100.2	-7.35	0.000	46.965
Period1	5419252	983082	5.51	0.000	3.874

S = 1132741 R-Sq = 91.4% R-Sq(adj) = 89.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	2.33086E+14	7.76953E+13	60.55	0.000
Residual Error	17	2.18127E+13	1.28310E+12		
Total	20	2.54899E+14			

Source	DF	Seq SS
Personal Consumption 1	1	3.17917E+13
Population USA	1	1.62304E+14
Period1	1	3.89907E+13

Regression data analysis / Model 1. As expected, the regression equation explains high amount of variation in the response of the independent variable as it has 89.9 % value for its adjusted square. However the VIF values indicate a multicollinearity between the personal consumption and population as the VIF for these factors are greater than 10. There is a debate about accepted (VIF) s between 10-100, Some researcher believe that VIF Value more than 10 is alarming, while others consider it as an indication about multicollinearity. So, many other attempts have been taken to build regression model that fulfills the assumption of multicollinearity with (VIF) Values less than 10.

In Case of multicollinearity one of the correlated independent factors should be dropped form the model. The adjusted R-Squared Value was used to evaluate the best model.

Model 2 was developed to consider (period one, and personal consumption) in the regression. The Adjusted R-Squared value of this model was (60.3 %).

Model 3 was developed to consider (period one, and population) The Adjusted R-Squared value of this model was (49.4%).

Model 2 and Model 3 are totally not acceptable due to the low value of the adjusted R- Squared. Another attempt were carried to identify an independent factor that can improve the model for better R-squared with better VIF Values.

Model 4 was developed to consider adding the unemployment rate to Model 3. (Period one, population and unemployment rate). The unemployment rate has been chosen to be added as it were included in the first four steps of the step wise process and has been dropped just by the last step of the step wise regression. Model 4 had an R-squared value of (70.3 %). Model 5 was developed to add the unemployment rate to Model 2. (Period one, personal consumption, and unemployment rate) The adjusted R-Squared value was (76.3). As this model for the best R-Squared value so far with acceptable (VIF) another attempt was carried out to improve it. Model 6 was built to include period two and three to Model 5. The software excluded period three due the fact that it is highly correlated with other variable according to Minitab. The following regression equation and data were generated for Model 6.

Model 6 Data Presentation

Regression Analysis: Home Sales U versus Personal Con, Unemployment,

* Period3 is highly correlated with other X variables

* Period3 has been removed from the equation.

The regression equation is

Home Sales USA = - 8579583 + 1904 Personal Consumption
- 895621 Unemployment Rate + 9187261 Period1 + 2890891 Period2

Predictor	Coef	SE	Coef	T	P	VIF
Constant	-8579583		4902413	-1.75	0.099	
Personal Consumption	1904.4		408.9	4.66	0.000	3.215
Unemployment Rate	-895621		253270	-3.54	0.003	1.672
Period1	9187261		1549869	5.93	0.000	5.684
Period2	2890891		1044576	2.77	0.014	1.913

S = 1474192 R-Sq = 86.4% R-Sq(adj) = 82.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	2.20127E+14	5.50317E+13	25.32	0.000
Residual Error	16	3.47719E+13	2.17324E+12		
Total	20	2.54899E+14			

Source	DF	Seq SS
Personal Consumption	1	3.17917E+13
Unemployment Rate	1	1.09883E+14
Period1	1	6.18066E+13
Period2	1	1.66453E+13

The (VIF) values in Model 6 are all less than (10). There is no doubt that multicollinearity assumption has been fulfilled.

An evaluation for the test statistics were required to choose one model to be used as a robust tool to predict the market demand. The following table presents these statistics.

Table 4 Test Statistic for Model 1 & 6

The test statistic	Model 1	Model 6
Adjusted r- Squared	89.9 %	82.9 %

The Adjusted R-Squared of Model 1 are greater than these Values of Model 6. Model 1 should to be chosen according to this analysis. However, Model one has a probability of multicollinearity.

By considering the fact that Model 6 has an adjusted R- Squared that exceeded 80% and it is free of multicollinearity. A final decision has been made to use Model 6 as a tool to predict the demand, the model provide following equation to be used.

Interpretation of the results.

Home Sales= -8579583+1904 Personal consumption -895621 unemployment rate +9187261 Period 1+2890891 Period 2

Predictor	Coef	SE Coef	T	P	VIF
Constant	-8579583	4902413	-1.75	0.099	
Personal Consumption	1904.4	408.9	4.66	0.000	3.215
Unemployment Rate	-895621	253270	-3.54	0.003	1.672
Period1	9187261	1549869	5.93	0.000	5.684
Period2	2890891	1044576	2.77	0.014	1.913

S = 1474192 R-Sq = 86.4% R-Sq(adj) = 82.9%

Constant Coefficient interpretation. The coefficient in Model 6 is - 8579583. This number is meaningless. However, the constant coefficient is extremely important in any regression model as it guarantees that the residuals of the model have a mean of zero.

Factor coefficient interpretation. As stated earlier, the coefficients tell how strongly each independent factor in the model is associated with its dependent factor. The size of the coefficient for each independent variable gives the size of the effect on the dependent variable, and the sign of the coefficient (positive or negative) gives the direction of the effect. The following table presents the coefficients of the model.

Table 5 Coefficients of Model 6

Factor	coefficient
Unemployment rate	- 895621
Period 1	+ 9187261
Personal consumption	+ 1904
Period 2	2890891

- The raise of one unit of the unemployment rate decrease the demand of the single new family houses with (895621 house)
- The raise of one unit of the personal consumption raises the demand of the single new family houses with (1904 house)
- As period 1 has just the two Values o (1 and 0) the coefficient of this factor indicates that when this value decrease from (1) to (0) a decrease of (9187261 Houses) will occur in the demand.
- As period 2 has just the two Values o (1 and 0) the coefficient of this factor indicates that when this value decrease from (1) to (0) a decrease of (2890891 Houses) will occur in the demand.

T - Value interpretation. As stated earlier, T-statistic is simply the coefficient divided by its standard error. The following table presents the T-Values of the model.

Table 6 T- Values of Model 6

Factor	T- value
Constant	-1.75
Unemployment rate	- 3.54
Period 1	5.33
Personal consumption	5.23
Period 2	2.77

The absolute value of all of T-Values for all the independent factors is greater than two. That's indicates that these factors are significant in the model. Although the T-value of the constant is less than two, the constant could not be removed as it defines the slope of the regression.

P-Value interpretation. As explained earlier, the P-value help in rejecting the null hypothesis of this test which is that the independent factor has no effect on the dependent factor. And explains how significant each variable in the model is. The following table presents the T-Values of the model.

Table 7 P- Values of Model 6

Factor	P- value
Constant	0.099
Unemployment rate	0.003
Period 1	0.000
Personal consumption	0.000
Period 2	0.014

The P-Values of all the independent factors are smaller than 0.05. These values lead the researcher to reject the null hypothesis, and indicate that these factors are significant in the model. Although the P-value of the constant is greater than 0.05, and indicates that the constant is not significant in the model. The constant could not be removed from the model.

R square interpretation. In model 6, the model explains 86.4 % of the variation, that's covers a considerable amount of the variation.

R adjusted square interpretation. The adjusted R-square is 82.9 %, that's indicates that the model explains 82.9 % of the variation in the respond of the dependent factor.

Analysis of variance.

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	2.20127E+14	5.50317E+13	25.32	0.000
Residual Error	16	3.47719E+13	2.17324E+12		
Total	20	2.54899E+14			

F- Value interpretation: It compares the selected model with a regression model has no predictor is also known as an intercept-only model.

The hypotheses for the F-test of the overall significance are as follows:

Null hypothesis: The fit of the intercept-only model and the selected model are equal.

Alternative hypothesis: The fit of the intercept-only model is significantly reduced compared to selected model.

The P value for the F-test of overall significance test is less than the test significance level (0.05), the null-hypothesis has been rejected and a conclusion has been drawn that the model provides a better fit than the intercept-only model.

Sum of Square interpretation.

Source	DF	Seq SS
Personal Consumption	1	3.17917E+13
Unemployment Rate	1	1.09883E+14
Period1	1	6.18066E+13
Period2	1	1.66453E+13

The Sum of square values indicates that the unemployment rate is the most effective factor in the model, followed by period 1, then the personal consumption, while period 2 is the least effective factor in the model.

Plot interpretation.

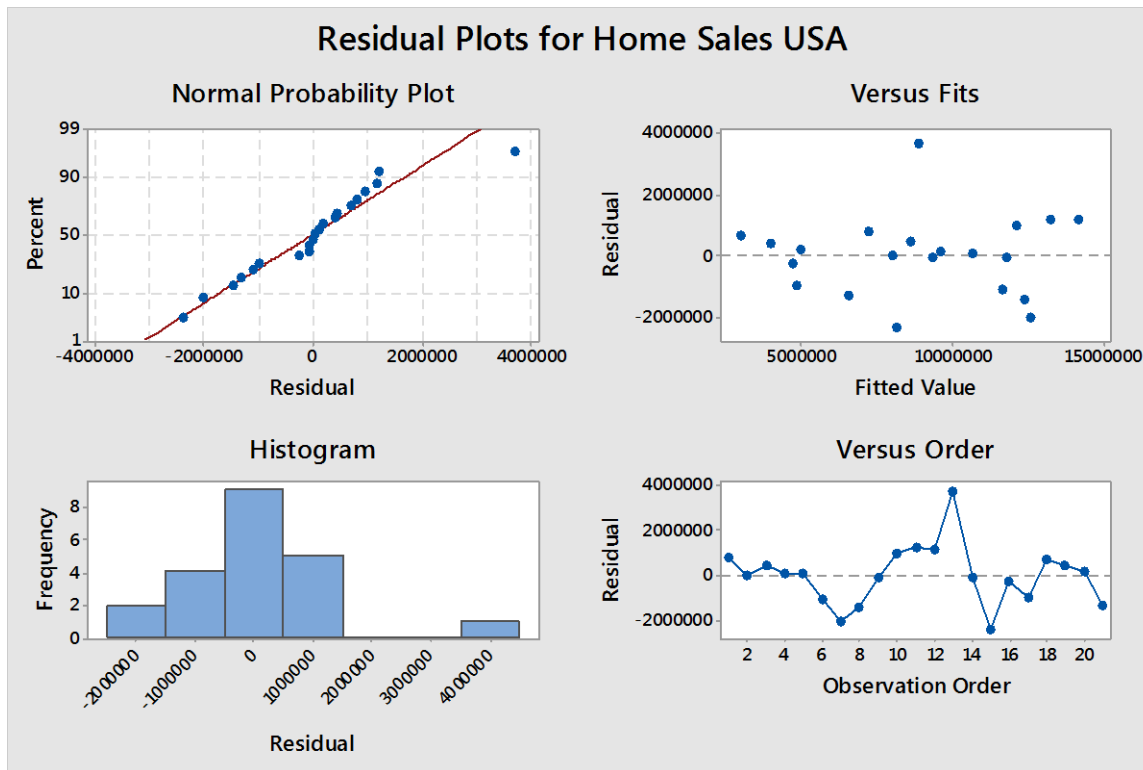


Figure 12 Residual Plots for Home Sales USA

The plot interpretation satisfies the assumption equal variance or homoscedasticity as the residuals plot versus fitted has a random pattern.

The plot interpretation indicated that the most of the residuals are normally distributed, however, an outlier is noticed and it is observation (16) the year 0f 2006 when the effect of the

financial crisis began. To verify the assumption of normality, the predicted values of the market demand were calculated and a normality test was conducted. The results found that residuals are normally distributed.

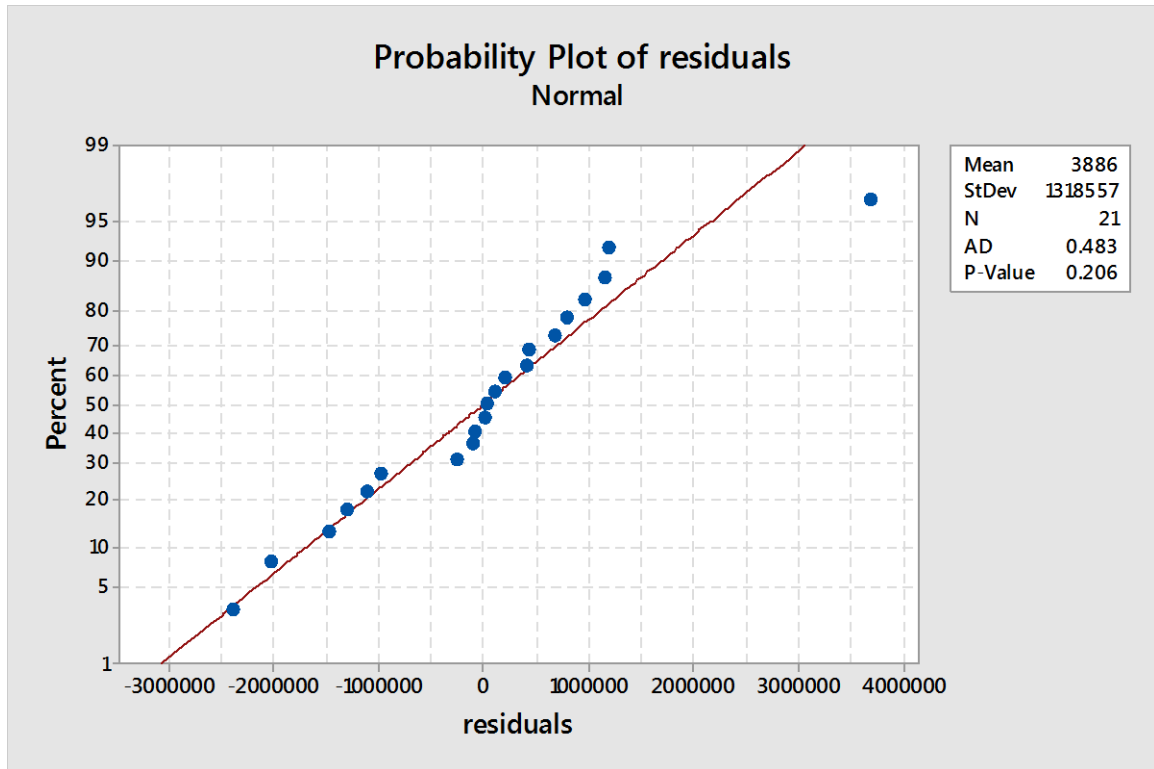


Figure 13 Probability Plot of Residuals

Model Validation. The predicted values were plotted versus the actual values to visualize the difference between the two values. The following graph was generated.

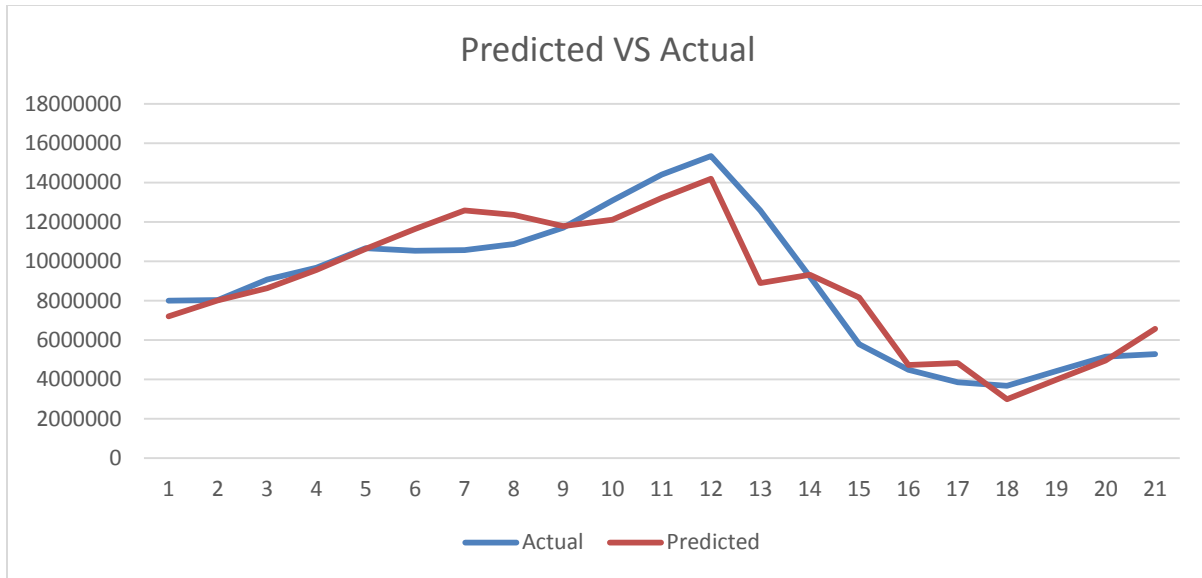


Figure 14 Home Sales Predicted VS Actual

The graph indicate acceptable variation at the most points of observation.

Summary

This chapter presented the in depth analysis was conducted to sort the data by assessing the linearity assumption, and the different model developed to predict the defendant factor. It presented all the analysis and the logic about choosing the final model. It interpreted the element of the chosen model and validate its output.

Chapter V

Results, Conclusion, and Recommendations

Introduction

This chapter presents the results of the study, summarize the conclusion, and provide recommendation.

Results

The goal of the study was to predict the demand of new single family homes in the USA by building a regression model that relate the demand to some economic and social factors. The study began with identifying the possible economic and social factors that may affect the demand, the data were collected from reliable data, the assumptions of the regression analysis were checked, and the regression model was built to answer the questions formulated at the beginning stage of the search.

1- What factors do affect the demand of building new houses?

Despite of assuming that each of Population, GDP, Mortgage rate, inflation rate, unemployment rate, and personal consumption are factors that may affect the demand of new single family houses. The study reveals the fact that unemployment rate and personal consumption are the only factors that could be considered as important factors to affect the demand of the sales. The study also forced Period one which is the period between 1994- 1006 and period 2 (2007-2010) which is the period of the financial crisis and its affects as independent factors that affect the demand.

2- How strong is the relationship between these factors and the market?

There is a moderate relationship between each of inflation, mortgage rate and Market demand.

Under normal conditions, there is a very strong relationship between GDP, Unemployment Rate, Population, Personal consumption and Market demand.

3- How the variation of these factors will affect the demand of building new houses?

The increase of one unit in the personal consumption increases the demand with (+ 1904) units.

The increase of one unit in the unemployment rate decreases the demand with (895621) units.

The demand in period one raises the demand in that period with (9187261) units. That factor has no effect in period two and three.

The demand in period two raises the demand in that period with (2890891) unit. Period two factor has no effect in period one and three.

The demand in period three is not affected with period one and period two factors. So the only two factors that affect the demand are the unemployment rate and the personal consumption.

4- How can we predict the demand of building new houses according to the prediction of the values of these factors?

As this study has been performed in period three, both period one and two will have a value of zero. The only two factors to determine the demand are the unemployment rate and the personal consumption values. The projected values of the unemployment rate and the personal consumption for any year in period 3 should be obtained to be used according to the following equation to forecast the demand of that given year.

Home Sales = $-8579583 + 1904 \text{ Personal consumption} - 895621 \text{ unemployment rate}$

Conclusion

The sales of new single family houses in the USA are mainly affected by the unemployment rate and the personal consumption in the USA. The sales are negatively affected by the raise of the unemployment rate, while it positively affected by the personal consumption. The study was able to predict the market demand using the historical data of the unemployment rate and the personal consumption with an acceptable level of accuracy.

Recommendations

This study was able to build a model that explains 82.9% of the variation in the demand. Companies can use this model for the strategic planning for the company to predict the demand on the middle term (3-5) years. However the average forecasting of two periods will provide better output on the short term of forecasting (yearly demand). Therefore, it is highly recommended to keep an eye on the output of the average of two period forecasting on the short term of forecasting.

As the model of this study explained 82.9 % of the variation, it is believed that adding other independent variables may improve this percentage. For future work, it is recommended to repeat the study to consider evaluating the average of the income and the number of vaccines as possible variables that may affect the demand.

Building the model on (2015) forced the use of dummy variables to handle the fluctuation in the data sets due to the financial crises period as there were not enough data points to do the regression analysis in period three from (2010-2014). It is recommended to repeat the study when more points of data are available after 2010, ten points of data are the minimum recommended amount of points to repeat the study, and it is the year of 2020.

References

- [1] Amadeo, K. (2015, July). Personal consumption expenditures price index. Retrieved from <http://useconomy.about.com/od/inflationfaq/p/Personal-Consumption-Expenditures.htm>
- [2] Annual inflation rate in the United States from 1990 to 2014 (n.d). *Statista Website*. Retrieved from <http://www.statista.com/statistics/191077/inflation-rate-in-the-usa-since-1990/>
- [3] Assumptions of multiple linear regression. (2015). *Statistic solution*. Retrieved from <http://www.statisticssolutions.com/assumptions-of-multiple-linear-regression/>
- [4] Balance, D. (2012). Assumptions in multiple regression. Retrieved from http://www.dianneballanceportfolio.com/uploads/1/2/8/2/12825938/assumptions_in_multiple_regression.pdf
- [5] Belsky, E. S., Drew, R. B., & McCue, D. (2007). *Projecting the Underlying Demand for New Housing Units: Inferences from the Past, Assumptions about the Future*. Joint Center for Housing Studies, Graduate School of Design [and] John F. Kennedy School of Government, Harvard University.
- [6] Boyd, T (2014). *Property Market Analysis The key to looking forward*.
- [7] Chen, W. (2013, June). What is the difference between R-squared and Adjusted R-squared? *Quora Website*. Retrieved from <https://www.quora.com/What-is-the-difference-between-R-squared-and-Adjusted-R-squared>
- [8] Dallal, G, E. (2000), How to Read the output from multiple linear regression analysis. Retrieved from <http://www.jerrydallal.com/lhsp/regout.htm>

- [9] Easton, V.J & McColl, J, H. (n.d). *Statistic Glossary Website*. Retrieved from http://www.stats.gla.ac.uk/steps/glossary/hypothesis_testing.html#hypothtest
- [10] Eppli, M. J & Childs, M. J. (2001). *A descriptive analysis of U.S. housing demand for the 1990s*.
- [11] Freddie Mac, (n.d). 15-Year fixed rate mortgage average in the United States© [MORTGAGE15US]. *Economic Research Website*. Retrieved from <https://research.stlouisfed.org/fred2/series/MORTGAGE15US/>
- [12] Finding the pearson correlation. (2014, June). *Ableau Website*. Retrieved from <http://kb.tableau.com/articles/knowledgebase/statistics-finding-correlation>
- [13] Frost, J. (2015, June). What is the F-test of overall significance in regression Analysis? *The Minitab Blog Website*. Retrieved from <http://blog.minitab.com/blog/adventures-in-statistics/what-is-the-f-test-of-overall-significance-in-regression-analysis>
- [14] Frost, J. (2013, May). Regression Analysis: How do I interpret R-squared and assess the goodness-of-fit? *The Minitab Blog Website*. Retrieved from <http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit>
- [15] Frost J. (2014, June) What Is the difference between linear and nonlinear equations in regression analysis? *The Minitab Blog Website*. Retrieved from <http://blog.minitab.com/blog/adventures-in-statistics/what-is-the-difference-between-linear-and-nonlinear-equations-in-regression-analysis>
- [16] Hoerl, R. (2008, June). The reality of residual analysis. *Quality Progress Website*. Retrieved from <http://asq.org/quality-progress/2008/06/statistics-roundtable/the-reality-of-residual-analysis.html>

- [17] Hutchings, D. M. & Andrus, D. L. (2006). *An Analysis of the Accounting Methods Used by Builders Producing 25 or Fewer Homes per Year*. [18] Jiang, H & Liu, C (2014). A panel vector error correction approach to forecasting demand in regional construction markets. *Construction Management and Economics*, 32(12), 1205-12121. doi: 10.1080/01446193.2014.977800.
- [19] Keefer, A. (2015). The effect of inflation on housing prices. *Global Post Website*. Retrieved from <http://everydaylife.globalpost.com/effect-inflation-housing-prices-33139.html>
- [20] Mayerhauser, N & Reinsdorf, M. (2007, September). Housing services in the national economic accounts. Retrieved from <http://www.bea.gov/papers/pdf/RIPfactsheet.pdf>
- [21] Pearson product-moment correlation. (n.d). *Leard Statistics Website*. Retrieved from <https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php>
- [22] Pettinger, T. (2013, November). Factors that affect the housing market. *Economics Help Website*. Retrieved from <http://www.economicshelp.org/blog/377/housing/factors-that-affect-the-housing-market/>
- [23] Prince, M. (2014, May). Understanding construction market forecasts- how they can help with business planning. *Competitive Advantage Website*. Retrieved from <http://www.cadvantage.co.uk/understanding-construction-market-forecasts-how-they-can-help-with-business-planning/>
- [24] Reading and using STATA output. (n.d). Retrieved from <http://web.mit.edu/course/17/17.846/OldFiles/www/Readout.html>
- [25] Suri, A. (2006). *A linear regression model for predicting market demand tool for engineering managers and top management*.

- [26] Types of regression analysis. (2015). Retrieved from <http://support.minitab.com/en-us/minitab/17/topic-library/modeling-statistics/regression-and-correlation/basics/types-of-regression-analyses/>
- [27] Unemployment rate in the United States from 1990 to 2014 (n.d.). Statista Website. Retrieved from <http://www.statista.com/statistics/193290/unemployment-rate-in-the-usa-since-1990/>
- [28] US. Bureau of Economic Analysis, *Real Gross Domestic Product* [A191RL1A225NBEA], retrieved from FRED, Federal Reserve Bank of St. Louis <https://research.stlouisfed.org/fred2/series/A191RL1A225NBEA/>
- [29] US. Bureau of Economic Analysis, *Real Personal Consumption Expenditures* [DPCERL1A225NBEA], retrieved from FRED, Federal Reserve Bank of St. Louis <https://research.stlouisfed.org/fred2/series/DPCERL1A225NBEA/>
- [30] US. Bureau of the Census, *New One Family Houses Sold: United States* [HSN1F], retrieved from FRED, Federal Reserve Bank of St. Louis <https://research.stlouisfed.org/fred2/series/HSN1F/>
- [31] US. Bureau of the Census, *Total Population: All Ages including Armed Forces Overseas* [POP], Retrieved from FRED, Federal Reserve Bank of St. Louis <https://research.stlouisfed.org/fred2/series/POP/>
- [32] What is a test statistic? (n.d). Minitab Test Support Website. Retrieved from <http://support.minitab.com/en-us/minitab-express/1/help-and-how-to/basic-statistics/inference/supporting-topics/basics/what-is-a-test-statistic/>
- [33] Where Americans spend their money. (2015). Retrieved from <http://money.cnn.com/interactive/news/economy/us-spending/>

- [34] Zainun, N, R, I, & Eftekhari, M (2010). Forecasting low- cost housing demand in an urban in Malaysia using artificial neural networks: Batu Pahat, Johor. *WIT Transactions On Ecology And the Environment* 142, no. The Sustainable world: 54-58 Scopus®, EBSCO host (accessed November2, 2015).
- [35] Zaiontz, C. (2015, Dec). Null and alternative hypothesis. *Real Statistics Using Excel Website*. Retrieved from <http://www.real-statistics.com/hypothesis-testing/null-hypothesis/>