ANALYSIS OF QUANTILE REGRESSION AS ALTERNATIVE TO ORDINARY LEAST SQUARES REGRESSION

BY

IBRAHIM ABDULLAHI
B.Sc. STATISTICS (UNIMAID, 2011)
M.sc/SCI/44167/2012-2013

A THESIS SUBMITTED TO THE SCHOOL OF POSTGRADUATE STUDIES, AHMADU BELLO UNIVERSITY, ZARIA NIGERIA

IN PARTIAL FULFILLMENT FOR THE AWARD OF MASTER OF SCIENCE (M.Sc) DEGREE IN STATISTICS,

DEPARTMENT OF MATHEMATICS
AHMADU BELLO UNIVERSITY, ZARIA
NIGERIA.

JULY, 2015

DECLARATION

I declare that the work in this thesis entitled "**ANALYSIS OF QUANTILE REGRESSION AS ALTERNATIVE TO ORDINARY LEAST SQUARES REGRESSION**" has been carried out by me in the Department of Mathematics.The information derived from literature has been duly acknowledged in the text and a list of references provided. No part of this thesis was previously presented for another degree or diploma at this or any other Institution.

_____ _____

Abdullahi Ibrahim Date

CERTIFACATION

This thesis entitled "ANALYSIS OF QUANTILE REGRESSION AS ALTERNATIVE TO ORDINARY LEAST SQUARES REGRESSION" by Ibrahim, ABDULLAHI (M.Sc/Sci/44167/2012-2013) meets the regulations governing the award of the degree of Master of Science of the Ahmadu Bello University, Zaria and is approved for its contribution to knowledge and literary presentation.

Dr. A. Yahaya _____ _____
Chairman, Supervisory Committee SignatureDate

Prof. O.E. Asiribo _____ _____
Member, Supervisory Committee Signature Date

Name_____ _____
External Examiner Signature Date

Prof. B. Sani_____ _____
Head of Department Signature Date

Prof. Adebayo A. Joshua _____ _____
Dean, School of Postgraduate Studies Signature Date

DEDICATION

This master's thesis is dedicated to everyone who has believed in me throughout my life especially recently and to those who have shown me unconditional and unfaltering love. My parents Mal. Abdullahi and HajiyaGambo have been a constant source of love and support since the day I was born. There is no doubt that their influence has shaped my life tremendously.

I must always thank Allah for providing the great comfort and peace we all need to succeed. Without his presence in my life none of this would be possible.

-

ACKNOWLEDMENT

My profound gratitude goes to Allah the almighty on whom my life solely depends on, who by his grace this dream became a reality. I also acknowledge the work of previous authors and papers cited. My appreciation also goes to my supervisors Dr. A.Yahaya and Prof Asiribo for their relentless effort to make this work a success and their suggestions and strictness has really brought out the best in me, which were highly appreciated.

And I also want to use this medium to appreciate my lecturers especially Mr. C.N Nnamani for his guidance and direction all through this work. My special thanks goes to my one and only my brothers and sisters Prof DauadAbdullahi, Abdulmumini, Usman and HajjaNuratu, for their love and care. Also to my uncles, aunties and my cousins and best friend Prof. Segun (Asiribo). Finally to the whole youths of Nguru I want to say I love you all. To the whole staff of Nguru Local Government Education Authority and to the entire PG (2012-2013) students of mathematics department, I want to say thank you all.

ABSTRACT

In this thesis, we present an alternative to ordinary least squares (OLS) regression based on analytical solution in the Statgraphics software is considered, and this alternative is no other than quantile regression (QR) model. We also present goodness of fit statistic called Quantile regression coefficient of determination as well as heteroskedasticity test statistics for the parameters. The procedure is well presented, illustrated and validated by a numerical example based on publicly available dataset on fuel consumption in miles per gallon in highway driving. Theresults obtained from the analysis in this thesissuggest that sometimes OLS estimates can even be misleading what the true relationship between response variable and covariate as the effects can be very different for different subsections of the sample. Quantile Regression therefore gives a better and more complete view of the relationship among random variables.

TABLE OF CONTENTS

DECLARATION i
CERTIFACATION ii
DEDICATION iii
ACKNOWLEDMENT iv
ABSTRACT v
TABLE OF CONTENTS vi
List of Tables viii
List of Figures ix
Appendix x
CHAPTER ONE 1
INTRODUCTION 1
1.0 Background of the Study 1
1.1 Limitations for the Conditional Mean 2
1.2 Motivation of the Study 4
1.3 Statement of the Problem 5
1.4 Scope of the study 5
1.5 Significance of the study 6
1.6 Aim and Objective(s) of the Study 6
1.7 Limitation of the Study 7
1.8 Statement of Hypotheses 7
CHAPTER TWO 8
LITERATURE REVIEW 8
2.0 Introduction 8
2.1 Multiple Linear Regression 8
2.2 Quantile Regression 10
CHAPTER THREE 16
METHODOLOGY 16
3.0 Introduction 16
3.1 Data Collection 16
3.2 Classical Linear Regression 17
3.3 Estimation of the Parameters in Linear Regression Models 18
3.4 Estimating σ^2 22

3.5 Properties of the Estimators 23
3.6 Coefficient of Multiple Determination R^2 23
3.7 Stepwise Selection 25
3.8 Quantile Regression 25
3.8.1 Computation of Quantile Regression 26
3.8.2 Least Absolute Deviation Regression Goodness of Fit 30
3.8.3 Quantile Regression Goodness of Fit 31
CHAPTER FOUR 33
RESULTS AND DISCUSSION 33
4.0 Introduction 33
4.1 Numerical Illustration and Discussion of the result 33
4.2 Ordinary least squares regression 34
4.3 Heteroskedasticity Test: White 35
4.4 Comparison of OLS and QR as the number of variable increases 36
CHAPTER FIVE 38
SUMMARY CONCLUSIONS AND RECOMMENDATIONS 38
5.0 Introduction 38
5.1 Summary 38
5.2 Conclusion 39
5.3 Recommendations 39
5.4 Contribution to knowledge 39
5.5 Further research 40
REFERENCES 41
APPENDIX A 44

List of Tables

Table 4.1: Correlations matrix…….33

Table 4.2: OLS Regression model……34

Table 4.3: Stepwise regression……..35

Table 4.4: White (Error) test…….35

Table 4.5: Numerical result of the OLS and QR simple linear models……36

Table 4.6: Numerical result of the OLS and QR multiple regression models…….36

List of Figures

Fig 4.1: Residual plot……37

Appendix

Appendix A: Fuel consumption data set of 93 cars was measured on 6 regressors……..44

Appendix B: R Code for Multiple Quantile Regression……46

CHAPTER ONE

INTRODUCTION

1.0 Background of the Study

In most regression problems, interest lies in studying the relationship between two or more variables. Where it is an important aspect in the philosophy of science to study the concept of relationship between varying qualities or events.

The purpose of regression analysis is to expose the relationship between a response variable and predictor variables. In real applications, the response variable cannot be predicted exactly from the predictor variables. Instead, the response for a fixed value of each predictor variable is a random variable. For this reason, we often summarize the behavior of the response for fixed values of the predictors using measures of central tendency. Typical measures of central tendency are the average value (mean), the middle value (median), or the most likely value (mode).

Traditional regression analysis is focused on the mean; that is, we summarize the relationship between the response variable and predictor variables by describing the mean of the response for each fixed value of the predictors, using a function we refer to as the conditional mean of the response. The idea of modeling and fitting the conditional-mean function is at the core of a broad family of regression-modeling approaches, including the familiar simple linear-regression model, multiple regression, models with "heteroscedastic" errors using weighted least squares, and nonlinear regression models.

Conditional-mean models have certain attractive properties. Under ideal conditions, they are capable of providing a complete and parsimonious description of the relationship between the covariates and the response distribution. In addition, using conditional-mean models leads to

estimators (least squares and maximum likelihood) that possess attractive statistical properties that are easy to calculate, and are straightforward to interpret. Such models have been generalized in various ways to allow for "heteroscedastic" errors so that given the predictors, modeling of the conditional mean and conditional scale of the response can be carried out simultaneously. Conditional-mean modeling has been applied widely in the social sciences, particularly in the past half century, and regression modeling of the relationship between a continuous response and covariates via least squares and its generalization is now seen as an essential tool. More recently, models for binary response data, such as Logistic, Probit and Poisson regression models for count data have become increasingly popular in social science research. These approaches fit naturally within the conditional mean modeling framework. While quantitative social-science researchers have applied advanced methods to relax some basic modeling assumptions under the conditional-mean framework, this framework itself is seldom questioned.

The conditional-mean framework has inherent limitations.

1.1 Limitations for the Conditional Mean

- The mean of the errors is zero, i.e. $E(\varepsilon_i)=0$.
- The error is assumed to have equal variance that is, $\text{var}(\varepsilon_i)=\delta^2$ for all value of $X = x$
- The distributions of errors for different value of regressors are independent, denoting the errors in different value by $\varepsilon_i \; and \; \varepsilon_j$, that is, $Cov(\varepsilon_i \varepsilon_j)=0 \quad \forall i \neq j.$
- The probability distribution of errors for all values of *x* is normally distributed by $\varepsilon_i \sim N(0,\sigma^2).$

An alternative to conditional-mean modeling believed to have roots that can be traced to the mid-18th century. This approach can be referred to as conditional median modeling, or simply median regression. It addresses some of the issues mentioned above regarding the choice of a measure of central tendency. The method replaces least-squares estimation with

least-absolute distance estimation. While the least-squares method is simple to implement without high-powered computing capabilities, least-absolute-distance estimation demands significantly greater computing power. It was not until the late 1970s, when computing technology was combined with algorithmic developments such as linear programming, that median-regression modeling via least-absolute-distance estimation became practical.

The median-regression model can be used to achieve the same goal as conditional-mean-regression modeling in order to represent the relationship between the central location of the response and a set of covariates. However, when the distribution is highly skewed, the mean can be challenging to interpret while the median remains highly informative. As a consequence, Conditional-median modeling has the potential to be more useful.

The median is a special quantile, which describes the central location of a distribution. Conditional-median regression is a special case of quantile regression in which the conditional 0.5 quantile is modeled as a function of covariates. More generally, other quantiles can be used to describe non-central positions of a distribution. The quantilenotion generalizes specific terms like quartile, quintile, decile, and percentile. The *p*thquantile denotes that value of the response below which the proportion of the population is *p*. Thus, quantiles can specify any position of a distribution. For example, 2.5% of the population lies below the 0.25th quantile.

Koenker and Bassett (1978) introduced quantile regression, which models conditional quantiles as functions of predictors. The quantile-regression model is a natural extension of the linear-regression model. While the linear-regression model specifies the change in the conditional mean of the dependent variable associated with a change in the covariates, the quantile regression model specifies changes in the conditional quantile. Since any quantile can be used, it is possible to model any predetermined position of the distribution.

Quantile-regression models can be fitted by minimizing a generalized measure of distance using algorithms based on linear programming. As a result, quantile regression is now a practical tool for researchers. Software packages familiar to statistical scientists offer readily accessed commands for fitting quantileregression models.

This research work aims to investigate the robustness of quantile regression as an alternative to least squares regression, especially when the number of regressors increases in the model to a broad audience of social scientists who are interested in modeling both the location and shape of the distribution they wish to study. It is imperative for researchers to utilize the comparison in two parts.

1.2 Motivation of the Study

Standard linear regression techniques summarize the average relationship between a set of regressors and the outcome variable based on the conditional mean function $E(y/x)$. This provides only a partial view of the relationship, as we might be interested in describing the relationship at different points in the conditional distribution of y. Quantile regression provides that capability. Analogous to the conditional mean function of linear regression, we may consider the relationship between the regressors and outcome using the conditional median function $Q_q(y/x)$, where the median is the 50th percentile, or quantile q, of the empirical distribution. The quantile q $\in$ (0; 1) is that y which splits the data into proportions q below and $1-q$ above: $F(y_q) = q \; and \; y_q = F^{-1}(q)$: for the median, q = 0.5.If ε_i is the model prediction error, OLS minimizes $\sum \varepsilon_i^2$. Median regression, also known as least-absolute-deviations (LAD) regression, minimizes $\sum |\varepsilon_i|$. Quantile regression minimizes a sum that gives asymmetric penalties $(1-q)|\varepsilon_i|$ for overprediction and $q|\varepsilon_i|$ for underprediction. Although its computation requires linear programming methods, the quantile regression estimator is asymptotically normally distributed. Median regression is more robust to outliers

than least squares regression, and is semi parametric as it avoids assumptions about the parametric distribution of the error process. Just as regression models conditional moments, such as predictions of the conditional mean function, we may use quantile regression to model conditional quantiles of the joint distribution of *y* and *x*.

1.3 Statement of the Problem

Regression analysis is robust in application for various kind of research, especially when provisions are made to control for problems dealing with heteroskedasticity, due to the violation of OLS assumption $\mathrm{E}(\varepsilon^2)=\sigma^2$ *i.e* $\mathrm{E}(\varepsilon_1^2)\neq \mathrm{E}(\varepsilon_2^2)$. What is the implication of heteroskedasticity?The ordinary least squares (OLS) estimators and regression predictions based on them remain unbiased and consistent.The OLS estimators are no longer the BLUE because they are no longer efficient. As a result, regression predictions will be inefficient as well.Because of the inconsistency of the covariance matrix of the estimated regression coefficients, the tests of hypotheses, that is, t-tests or F-tests, are no longer valid.

The purpose of this study is to introduce a good model that controls the problem of heteroskedasticity, and the newly model is considered as quantile regression model as a robust alternative to ordinary least squares regression when the said assumptions fail to hold.

1.4 Scope of the study

The study, seeks to investigate the behavior or nature of the two regression processes as the number of predictor (explanatory) variables increases. Keeping in mind that quantile regressiondoes not take into account the failure or otherwise of the existing assumptions. Discussion of the two techniques, various significant tests of the importance of independent variables, the reliability of the models and results interpretation in the two techniques will be considered. The study also involves empirical analysis using the two techniques and a comparison of the results will be made to establish the discrepancies in using the two techniques, if any.

1.5 Significance of the study

The study has been found worthy of research due to the fact that QR has been found to give a more holistic view of the effect of the explanatory variables on the response variable at different quantiles. The main significance of the study is that it strongly propose the use of QR as against the OLS regression that was used in the past for examiningthe heteroscedasticity, thereby recommending the use of QR to researchers for the study of heteroscedasticity. Most research involves estimation of the relationship between independent variables and dependent variable on the average. The relationship between dependent and independent variables can be estimated and examined at each quantile of the dependent variable. Inthis research it is our purpose to revisit the least absolute deviation estimation in regression analysis, consider some of its theoretical properties, and consider its implementation from a computational mathematical programming point of view. We also consider goodness of fit statistics as well as approximate distributions of the associate test statistics for the parameters. Furthermore, we suggest a new goodness of fit statistic, called the Quantilecoefficient of determination, which is adapted to the metrics used in LAD estimation. Finally some examples are provided to illustrate the behavior of the procedures in data that include outliers growing recognition of the need for a more flexible, more complete analysis is a driving force in the use of QR in the literature.

1.6 Aim and Objective(s) of the Study

The main aim of the study is to investigatequantile regression as an alternative to least squares regression, especially when the number of regressors increases.

1. To examine the quantile regression andleast squares regression.
2. To compare the models in term of goodness of fit statistic.
3. To recommend a suitable model for regression analysis.

1.7 Limitation of the Study

The study would have focused on conditional, unconditional, mean and quantilesbut for time constraint, the study will focused only on conditional mean and conditional quantiles.

1.8 Statement of Hypotheses

t - test for significance of one coefficient

The t - test is used to determine whether the relationship between *y* andx_j is significant

H_0: $\beta_j = 0$ (i.e. the coefficient β_j is not significantly different from zero).

H_1: $\beta_j \neq 0$ (i.e. the coefficient is significantly different from zero).

F - Test for overall significant of all coefficients

Testing whether the relationship between *y* and all *x*variables is significant

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

H_1 : at least one of $\beta_{j's}$ is different from zero.

CHAPTER TWO

LITERATURE REVIEW

2.0 Introduction

Elementary statistics texts tell us that the method of least squares was first discovered about 1805 (Stigler, 1986). There has been a dispute about who first discovered the method of least squares. It appears that it was discovered independently by Carl Friedrich Gauss (1777-1855) and Adrien Marie Legendre (1752-1833), that Gauss started using it before 1803 (he claimed in about 1795, but there is no corroboration of this earlier date, and that the first account was published by Legendre in 1805, see(Draper & Smith, 1981). Stigler (1986) notes that Sir Francis Galton discovered regression about 1885 in studies of heredity. Any contemporary course in regression analysis today starts with the methods of least squares and its variations.

2.1 Multiple Linear Regression

Multiple linear regression (MLR) is one of the most commonly used data mining techniques, and can provide insightful information in cases where the rigid assumptions associated with MLR are met. The assumptions include:

- ✓ linearity of the coefficients;
- ✓ Normal or Gaussian distribution for the response errors (ε); and
- ✓ The errors (ε)have a common distribution.
- ✓ Equal variance (homoscedasticity).

MLR is a very versatile tool and can be applied to almost any process, system, or area of study. Much has been published regarding this subject, and the following text may be useful to the reader: Kutner*et al* (2004) as well as (Myers, 1990), provide thorough accounts of MLR and will be indispensable for most readers.

A key step in developing an appropriate MLR model is selecting a method of model building and a set of best model criteria. As used in this thesis, stepwise regression is commonly used for model building. Introduced by Efroymson (1960), stepwise regression was intended to be a n automated procedure that selects the most statistically significant variables from a finite pool of independent variables. There are three separate stepwise regression procedures; forward selection, backward elimination and mixed selection. Mixed selection is the most statistically defendable type of stepwise regression, and is a mixture of the forward and backward proceduresKutner*et al.* (2004), Neter*et al*. (1996), and (Draper and Smith, 1981).

A set of best model criteria are commonly used in conjunction with stepwise regression in order to select the optimal model. As cited by Young and Guess (2000) and Young and Huber (2004), multicollinearity and heteroscedasticity can be significant problems when modeling the IB of MDF using industrial data. Young and Guess (2002) used the following best model criteria: maximum Adjusted R^2, parameters (p) Mallow's Cp (Mallow 1973), minimum Akaike's Information Criterion (AIC), Akaike(1974), Variance Inflation Factor (VIF) < 10, significance of independent variables pvalue< 0.10, absence of heteroscedasticity in residuals, $E(\varepsilon) = 0$.

For this thesis, we focus on the aforementioned criteria. We also use a pvalue< 0.05 for significance among the independent variables. The adjusted R^2 statistic, R_a^2, is a better measure of fit for MLR models built with the potential to contain significantly more independent variables than data records. As additional independent variables are added to a regression model, R^2 will always increase regardless of the fit. The R_a^2 statistic only increases if the residual sum of squares decreases (Draper and Smith 1981).The R_a^2 statistic minimizes the risk of, and penalizes for, using too many independent variables. AIC measures the complexity of the model and guards against model bias. VIFs are reported to protect against

multicollinearity, and redundancy in the model. Models with VIF < 10 can be said to be relatively free of these effects (Kutner et al. 2004).

As noted by Kutner*et al*. (2004), model validation is the final step in the regression modeling-building process. Kutner*et al*. (2004) point to three main methods associated with model validation, as follows: collection of new data to validate the current model and its predictability, comparison of current results with other theoretical values, empirical and simulation results and use of a cross-validation sample to validate and assess the predictive power of the current model.

For this thesis, we use the comparison of OLS and QR model using the both MPG in highway driving results. A plethora of statistics is available to aid in assessing the predictive power of regression models. Another common model validation statistic is the classic coefficient of determination, or R^2, statistic. This value is also computed for the withheld sample, and provides some insight into the predictability of the model. By definition, higher R^2 values are preferred, i.e., the R^2 statistic indicates the amount of variation explained by the regression model.

2.2 Quantile Regression

Quantile Regression (QR) was introduced by Koenker and Bassett (1978) and is intended to offer a comprehensive strategy for completing the regression picture (Koenker, 2005). QR does not impose any strict parametric assumptions Koenker (2005) Response data in the tails, or outer quantiles, of a distribution may behave differently than data in the inner quantiles of the distribution in response to the predictor variables. Traditionally, MLR is used to study causality between independent variables and the central tendency of a response variable as measured by the mean or average, with an important goal of making useful predictions of the response variable. However, several stringent aforementioned assumptions must be met in order for a MLR model to perform well. In contrast to MLR, As Mosteller and Tukey (1977)

noted in their influential text, as cited by Koenker(2005): "…the regression curve gives a grand summary for the averages of the distributions corresponding to the set of Xs…and so regression often gives a rather incomplete picture. Just as the mean gives an incomplete picture of a single distribution, so the regression curve gives a correspondingly incomplete picture for a set of distributions."

As with MLR, QR has many applications, and was originally developed for statistical use, as the first QR publication was in Econometrical (Koenker and Bassett 1978). Koenker and Bassett (1978) insightfully envisioned a more robust regression approach capable of modeling conditional quantile functions beyond the classic MLR least squares approach to model building. Koenker and Bassett (1978) note, "estimators are suggested, which have comparable efficiency to least squares for Gaussian linear models while substantially outperforming the least-squares estimator over a wide class of non-Gaussian error distributions".

Researches began applying these new techniques to the Management Science field of study and introduce an adaptive filtering procedure for exploring regression quantiles. These models are used as part of their Quantile Estimation Procedure (QEP) and are utilized to signal preventative actions and therefore avoid undesirable system states (Gorr and Hsu 1985).

Young and Easterling (1994)investigate QR as applied to reliability data analysis. Typically, in reliability applications, the practitioner is most interested in the outer quantiles of distributions being studied, i.e., products that have an extremely short or long lifespan. Young and Easterling (1994) use QR techniques to explore the outer quantiles of sensitivity test distributions. Various sample sizes are examined, as well as quantiles, and the effect of assuming different specified models is noted. Young and Easterling (1994) find that QR

provides better models for their data when quantiles are estimated as a function of specific model parameters as opposed to tests developed in order to estimate a specific quantile.

Buchinsky (1998)provides a basic guide for empirical research, focusing on cross section applications, while summarizing the most significant aspects of QR and filling in some noted literature gaps. Several alternative covariance matrix estimators are presented and Buchinsky (1998) eloquently discusses useful procedures for testing QR models for homoskedasticity and symmetry of the error distribution. A generous empirical example is presented using data obtained from a current population survey, and the paper concludes with a brief discussion on the application of censored QR models.

In 1999, Koenker and Machado introduce goodness-of-fit procedures for QR .These statistics are quite similar to the R^2 statistic applied to classical regression techniques.Various inference processes designed to assess the adequacy of the regression model are presented (Koenker and Machado 1999). These authors also illustrated their findings using empirical economic growth models, hypothetical examples, and conclude with Monte Carlo evidence.

The idea of computing regression quantiles with the use of a conditional quantile function is further articulated by (Koenker and Hallock, 2001). Koenker and Hallock (2001)discuss the undeniable link between quantile and the "operations of ordering and sorting the sample observations that are usually used to define them" (Koenker and Hallock 2001). The innate symmetry of the absolute value function ensures that there is the same number of observations both below and above the median (Koenker and Hallock 2001). Koenker and Hallock (2001) note there is high demand for more specialize QR models in the finance industry.

In their useful text, Fitzenberger*et al.* (2002)discussed the practical application of QR as compared to the methodology of least-squares regression. They note the important MLR

assumption of constant error, and insightfully articulate a useful example pertaining to wage distributions, acknowledging the importance of proper distribution modeling.

Recently, censored regression models have received substantial attention in economic literature, both theoretical and applied (Honore*et al.* 2002). Honore*et al.* (2002) note that, to date, most estimation procedures for panel data models or cross-sectional models are constructed using fixed censoring. Honore*et al.* (2002) suggest a new procedure for adapting these fixed censoring models to perhaps more applicable random censoring models.

Some other interesting applications include those in ecological and environmental studies (Cade and Noon 2003). As noted by Cade and Noon (2003), it is extremely difficult to identify, document, and measure every ecological independent variable. As a result, using classical MLR methods and others, it is sometimes impossible to arrive at a statistically significant model. However, models built using only portions of the response variable distribution may be more useful (Cade and Noon 2003). Cade and Noon (2003) explore various ecological QR applications, and thoughtfully estimate prediction intervals.

Interestingly, Green and Kozek (2003)use an approximate QR method to model weather data. These models are approximate because they are formed by applying quantile functions onto parametric models (Green and Kozek 2003). Parametric weather distributions are modeled as they vary over time, and regression quantiles are then applied to the models (Green and Kozek 2003). Five-curve summaries are obtained for the probability distributions of the weather data and the results are quite interesting.

Buhai (2004)provides an introduction to QR, discussing basic models and interpretations as well as computational and theoretical aspects of the algorithm, by concentrating only on two applications of QR which are: survival analysis and recursive structural equation models. Buhai (2004) was able to articulate a thorough summary of each.

Quantile regressions are very useful for visualizing changes in the conditional distribution of longitudinal data sets over time (e.gheteroscedasticity in the data set). When the mean regressions are in fact significantly different, the quantile regression can suggest which part of the conditional distributions differ, (Karlsson, 2006).

Quantile regression also goes beyond the location shift model to determine the effect of covariates on the shape and scale of the entire response distribution. The spacing of the quantile lines indicates whether the distribution is skewed to the right or left. Quantile are robust in regards to outliers, (Lee, 2005)r. Sensitivity of an estimator to departures from its distributional assumptions is another important issue. The sample mean can be adversely affected even by a single observation if it is sufficiently far from the rest of the data points. On the other hand, the effect of such a distant observation on the sample median is bounded no matter how far the outlying observation is. Other quantiles enjoy similar property the effect of outlying observations on the *t*-th sample quantile is bounded. Quantile regression inherits this robustness property. Cizek (2003). Thus quantile regression estimates are reliable in the presence of extreme outliers. Chen (2004) also investigated this robustness property, by considering the data of body mass index (BMI) against age for 8,280 men from a four year (1999-2002) survey.

Chernozhukov e*tal.* (2013)established the properties of the quantile and distribution regression estimators and also investigate that quantiles estimate differs across quantile.

Koenker (2005), most panel data approaches to date proceed in a quantile-by-quantile fashion, and include individual dummies as additional covariates in the quantile regression. As shown by some recent work, however, this fixed-effects approach faces special challenges when applied to quantile regression. Galvaoand Montes, (2012)develop the large-N, T analysis of the fixed-effects quantile regression estimator, and show that it may suffer from

large asymptotic biases. Rosen (2010)shows that the fixed-effects model for a single quantile is not point-identified.

Geraci and Bottai (2007)considered a random-effects approach for a single quantile assuming that the outcome variable is distributed as an asymmetric Laplace (for given τ) conditional on covariates and individual effects, and used a Monte Carlo EM algorithm for the computation of the Maximum Likelihood estimator. In addition, our approach is related to Abrevaya and Dahl (2008), who considered a correlated random-effects model to study the effects of smoking and prenatal care on birth weight. Their approach mimics control function approaches used in linear panel models.

Although many ordinary least square regression and quantile regression applications have been explored, and utilized in practice, the literature does not yet support the analysis ofquantile regression analysis as alternative to ordinary least squares as the number of regressors increase.

CHAPTER THREE

METHODOLOGY

3.0Introduction

In this chapter we introduce the process by which we analyze data to provide insight into the phenomenon under investigation rather than a prescription for final decision, which depends on the aim and objectives of the research.

Research methodology is the process or methods used to carry out a research or study. This refers to the method used to collect data or information to be used for the purpose of research. Two Statistical models are employed in this study. The first is an ordinary least squares estimation (OLS) and the second is a quantile regression.

3.1 Data Collection

The sources of data of any research are either primary or secondary or both. Primary Data: are those data, which are collected by the investigator himself for the purpose of a specific inquiry or study. Such data are original in character and are mostly generated by surveys conducted by individuals or research institutions, while Secondary Data: When an investigator uses data, which have already been collected by others, such data are called "Secondary Data". Such data are primary data for the agency that collected them, and become secondary for someone else who uses these data for his own purposes. For this research, secondary method of data collection is used.The data used in this research comes from http://www.csus.edu/indiv/v/velianitis/ds101/schedule.htm

Variables in the Model

The independent variables in the model include:

a. Horse power (maximum)

b. Weight (pounds)

c. Width (inches)

d. Length (inches)

e. Engine size (liters)

f. Wheel base (inches)

The dependent variable of the model is miles per gallon in highway driving.

3.2 Classical Linear Regression

The first-order simple linear regression model is (Draper and Smith 1981),

$$y_i=\beta_0+\beta_1 x_i+\varepsilon_i \qquad (3.1)$$

Where:

y_i is the value of the response variable in the ithobservation,

β_0 is the intercept parameter,

β_1 is a slope parameter,

x_i is the value of the independent variable in the ith observation,

ε_i is a random error term of the ith observation with mean $E(\varepsilon)=0$ and variance $\text{var}(\varepsilon_i)=\sigma^2$, with the error terms being independent and identically distributed, i = 1,…, *n*.

Most practitioners use multiple linear regression (MLR) first-order models of the form:

$$y_i=\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k+\varepsilon_i \qquad (3.2)$$

where:

y_i is the value of the response variable in the ithobservation,

β_0 is the intercept parameter,

β_k is the slope parameter associated with the kthvariable,

x_{ki} is the kth independent variable associated with the ithobservation,

ε_i is a random error term of the ith observation.

For the ordinary least squares to give the best linear unbiased estimates, certain assumptions must hold:

Mean $E(\varepsilon)=0$

Homoscedasticity - $\text{var}(\varepsilon_i)=\sigma^2$

No autocorrelation – $\text{cov}(\varepsilon_i\varepsilon_j)=0 \;\; i\neq j$

The question is, if one or more of these assumptions are violated, what will become of the result obtained from OLS method, as it is interesting to know that some data from fields including biostatistics, econometrics, survival analysis etc. violate the assumption of homoscedasticity? On the average has never been a satisfactory statement with which to conclude a study on heterogeneous populations. Characterizations of the conditional mean constitute only limited aspect possible more extensive changes involving the entire distribution.

3.3 Estimation of the Parameters in Linear Regression Models

The method of least squares is typically used to estimate the regression coefficients in a multiple linear regression model. suppose that $n>k$ observations on the response variable are available, say $y_1,y_2,y_3,\ldots,y_n$. Along with each observed response y_i, we will have an observation on each regressor variable and let x_{ij} denote the ith observation or level of variable x_j.We assume that the error term $\boldsymbol{\varepsilon}$ in the model has $E(\varepsilon)=0$, $\text{var}(\varepsilon_i)=\sigma^2$ and $\text{cov}(\varepsilon_i\varepsilon_j)=0 \;\; i\neq j$. We may write the model equation (Equation 2) in terms of the observations.

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\ldots+\beta_kx_{ik}+\varepsilon_i \qquad (3.3)$$
$$=\beta_0+\sum_{j=1}^{k}\beta_jx_{ij}+\varepsilon_i \qquad i=1,2,\ldots,n$$

The method of least squares chooses the $\beta's$ in Equation 3 so that the sum of the squares of the errors, ε_i is minimized. The least squares function is

$$L = \sum_{i=1}^{n} \varepsilon_i^2 \qquad (3.4)$$
$$= \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2$$

The function L is to be minimized with respect to $\beta_0, \beta_1, \ldots, \beta_k$. the least squares estimators, say $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$, must satisfy

$$\left. \frac{\partial L}{\partial \beta_0} \right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) = 0 \qquad (3.5)$$

and

$$\left. \frac{\partial L}{\partial \beta_j} \right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) x_{ij} = 0 \qquad (3.6)$$

Simplifying Equation 5 and 6, we obtain

$$\begin{aligned}
n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} &= \sum_{i=1}^{n} y_i \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1} x_{i2} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{i1} x_{ik} &= \sum_{i=1}^{n} x_{i1} y_i \\
\vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots & \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ik} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{ik} x_{i2} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 &= \sum_{i=1}^{n} x_{ik} y_i
\end{aligned} \qquad (3.7)$$

These equations are called the least squares normal equations. Note that there are $p = k + 1$ normal equations, one for each of the unknown regression coefficients. The solution to the normal equations will be the least estimators of the regression coefficients $\beta_0, \beta_1, \ldots, \beta_k$.

It is simpler to solve the normal equations if they are expressed in matrix notation. We now give a matrix development of the equation that parallels the development of Equation 3.7. The model in terms of the observations, Equation 3.3, may be written in matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$$

$$\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$$

where:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \text{ and } \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

- $\mathbf{X}$ is called design matrix.
- $\mathbf{y}$ is called response vector.
- $\boldsymbol{\beta}$ is the parameter vector.
- $\boldsymbol{\varepsilon}$ is the error vector.

We wish to find the vector least squares estimators, $\hat{\boldsymbol{\beta}}$, that minimizes

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Note that L may be expressed as

$$\begin{aligned} L &= \mathbf{y}'\mathbf{y} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \end{aligned} \qquad (3.8)$$

because $\boldsymbol{\beta}'\mathbf{X}'\mathbf{y}$ is a $(1\text{x}1)$ matrix, or a scalar, and its transpose $(\boldsymbol{\beta}'\mathbf{X}'\mathbf{y})' = \mathbf{y}'\mathbf{X}\boldsymbol{\beta}$ is the same scalar. The least squares estimators must satisfy

$$\left.\frac{\partial L}{\partial \boldsymbol{\beta}}\right|_{\hat{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0}$$

which simplifies to

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} \qquad (3.9)$$

Equation 3.9 is the matrix form of the least squares normal equations. It is identical to Equation 3.7. To solve the normal equations, multiply both side of the Equation 3.9 by the inverse of $\mathbf{X}'\mathbf{X}$. Thus, the least square estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \qquad (3.10)$$

It is easy to see that the matrix form of the normal equations is identical to the scalar form. Writing out Equation 9 in detail, we obtain

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots \sum_{i=1}^{n} x_{i1}x_{ik} \\ \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \sum_{i=1}^{n} x_{ik}x_{i2} & \cdots \sum_{i=1}^{n} x_{ik}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1`}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \vdots \\ \sum x_{ik}y_i \end{bmatrix}$$

If the indicated matrix multiplication is performed, the scalar form of the normal equations (i.e., Equation 3.7) will result. In this form it is easy to see that $\mathbf{X}'\mathbf{X}$ is a $(p \text{ x } p)$ symmetric matrix and $\mathbf{X}'\mathbf{y}$ is a $(p \text{ x } 1)$ column vector. Note the special structure of the $\mathbf{X}'\mathbf{X}$ matrix. The diagonal element of $\mathbf{X}'\mathbf{X}$ are the sum of squares of the elements in the columns of $\mathbf{X}$, and the off diagonal elements are the sum of cross products of the elements in the columns of $\mathbf{X}$. Furthermore, note that the elements of $\mathbf{X}'\mathbf{y}$ are the sums of cross products of the columns of $\mathbf{X}$ and the observations $\{y_i\}$.

The fitted regression model is

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \qquad (3.11)$$

In scalar notation, the fitted model is

$$\hat{y}_i = \hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_j x_{ij} \qquad \text{i} = 1,2,\ldots,\text{n}$$

The difference between the actual observation y_i and the corresponding fitted value $\hat{y}_i$, is the residual say $\varepsilon_i = y_i - \hat{y}_i$. The $(n \text{ x } 1)$ vector of residuals is denoted by

$$\boldsymbol{\varepsilon} = \mathbf{y} - \hat{\mathbf{y}} \qquad (3.12)$$

3.4 Estimating σ^2

It is also usually necessary to estimate σ^2. To develop an estimator of this parameter, consider the sum of squares of the residuals, say

$$\begin{aligned} SS_E &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} \varepsilon_i^2 \\ &= \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} \end{aligned}$$

substituting $\boldsymbol{\varepsilon} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$, we have

$$\begin{aligned} SS_E &= \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) \\ &= \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{y}'\mathbf{y} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \end{aligned}$$

Because $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$, this last equation becomes

$$SS_E = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \qquad (3.13)$$

Equation 3.12 is called the error or residual sum of squares, and it has n-p degree of freedom associated with it. It can be show that

$$E(SS_E) = \sigma^2(n - p)$$

so an unbiased estimator of σ^2 is given by

$$\hat{\sigma}^2 = \frac{SS_E}{n-p} \qquad (3.14)$$

3.5 Properties of the Estimators

The mothed of least squares produces an unbiased estimator of the parameter $\boldsymbol{\beta}$ in the linear regression model. This may be easily demonstrated by taking the expected value of $\hat{\boldsymbol{\beta}}$ as follows:

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\right] \\ &= E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})\right] \\ &= \mathrm{E}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\right] \\ &= \boldsymbol{\beta} \end{aligned}$$

because $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{I}$. Thus, $\hat{\boldsymbol{\beta}}$ is an unbiased estimmator of $\boldsymbol{\beta}$. The variance property of $\hat{\boldsymbol{\beta}}$ is expressed in the covariance matrix:

$$Cov(\hat{\boldsymbol{\beta}}) \equiv E\left\{\left[\hat{\boldsymbol{\beta}} - \mathbf{E}(\hat{\boldsymbol{\beta}})\right]\left[\hat{\boldsymbol{\beta}} - \mathbf{E}(\hat{\boldsymbol{\beta}})\right]'\right\} \qquad (3.15)$$

which is just a symmetric matrix whose ith main diagonal element is the variance of the individual regression coefficient $\hat{\boldsymbol{\beta}}_i$ and whose (ij)th element is the covariance between $\hat{\boldsymbol{\beta}}_i$ and $\hat{\boldsymbol{\beta}}_j$. The covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$Cov(\hat{\boldsymbol{\beta}}) = \sigma^2 (X'X)^{-1} \qquad (3.16)$$

3.6 Coefficient of Multiple Determination R^2

The coefficient of multiple determination R^2, is the square of the correlation, coefficient (r). Its value varies from zero to one. It has advantage over the correlation coefficient in that it may be interpreted directly as the proportion of variance in the response variable (y) or

predictor variable (x) that can be accounted for by the regression equation. For example an R^2=0.76 Implies 76% of the variance in the response variable can be explained by the regression equation (model), while the other 24% is unexplained.

$$R^2 = 1 - \frac{SS_R}{SS_T} \qquad (3.17)$$

where:

$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$SS_E = \sum_{i=1}^{n} (y_i - \hat{y})^2$$

ADJUSTED R^2_{adj}

In a multiple linear regression model R^2_{adj} measures the proportion of the variation in the dependent variable accounted for by the potentially significant explanatory variables in the multiple regression model. Unlike R^2, R^2_{adj} allows for the degrees of freedom associated with the sum of squares. Therefore, even though the residual sum of squares decreases or remains the same as the new explanatory variables are added, the residual variance does not. For this reason, R^2_{adj} is generally considered to be a more accurate goodness-of-fit measure than R^2. Points to note when using adjusted R square.

$$R^2_{adj} = 1 - \frac{SS_E / n - p}{SS_T / n - 1}$$

Where:

SS_E = sum of squared residuals (error)

n = number of observations

p = number of parameters

SS_T = total sum of squares

If it is significantly lower than R square, this means some explanatory variables are missing without them; the variation in the variable is not fully measured. When comparing two models with this measure, make sure you use the same dependent variable.

3.7 Stepwise Selection

Stepwise regression is a combination of the forward and backward selection techniques. It was very popular at one time, stepwise regression is a modification of the forward selection so that after each step in which a variable was added, and all candidate regressor variables in the model are checked to see if their significance has been reduced below the specified tolerance level. If a non-significant variable is found, it is removed from the model.

Stepwise regression requires two significance levels: one for adding variables and one for removing variables. The cutoff probability for adding variables should be less than the cutoff probability for removing variables so that the procedure does not get into an infinite loop. After careful considerations, the above mentioned procedures for variable selection has been found to mainly base its selection criterion on the zero order correlation between the regressor(s) and the response variable. Which implies that the above mentioned variable selection process does not take into account the correlation within the regressors themselves (multicollinearity) which leads us to the idea of QR.

3.8 Quantile Regression

The median regression also called least absolute deviation regression minimizes $\sum_i |\varepsilon_i|$

Buchinsky (1998) describes the general quantile regression model as follows:

$$y = x_i' \beta_\theta + \varepsilon_{\theta i}, \quad i = 1,2,3,\ldots n \qquad (3.18)$$

where:

y_i Denotes the dependent variable and the θthquantile (0<θ<1) of the conditional distribution of y_i is a linear function of a *k*x*1* vector of explanatory variables, x_i and an unknown error term $\varepsilon_{\theta i}$; β_θ is the unknown vector of regression parameters associated with the θth percentile. The conditional quantile function can be expressed as $Q_\theta(y_i / x_i) = x'\beta$.

The quantile regression minimizes a sum that gives the asymmetric penalties $\theta|\varepsilon|$ for underprediction and $(1-\theta)|\varepsilon|$ for overprediction. The quantile regression estimator $\hat{\beta}_\theta$ minimizes over β_θ the objective function

$$Q(\beta_\theta) = \min_{\beta_\theta} \left\{ \sum_{y_i \geq x'\beta_\theta} \theta |y_i - x'\beta_\theta| + \sum_{y_i \leq x'\beta_\theta} (1-\theta)|y_i - x'\beta_\theta| \right\} \qquad (3.19)$$

The study considers threequantile regressions at the 25th, 50th and 75th quantiles. Where θ is called the regression quantile, $0 < \theta < 1$.

Selecting in equation (3.19) θ = 1/2 gives perhaps the best known L-estimator, called the least absolute deviation (LAD). This method, as the name suggests, minimizes not the sum of squared deviations but the sum of absolute values of the deviations. Consequently it does not put excessive weight on highly deviating observations like the least squares does, and hence produces more robust estimators with respect to outliers.

3.8.1 Computation of Quantile Regression

Wagner (1959) proved that the least absolute deviation criterion can be formulated as a linear programming technique and then solved efficiently exploiting proper methods and algorithms. Koenker and Basset (1978) pointed out how conditional quantiles could be estimated by an optimization function minimizing a sum of weighted absolute deviations, using weights as asymmetric functions of the quantiles. The linear programming formulation

of the problem was therefore natural, allowing to exploit the same methods for estimating the conditional quantiles, and offering researchers and practitioners a tool for looking inside the whole conditional distribution apart from its center.

The linear programming formulation of the QR problem.

The two -variables problem

Let us first consider a two-variable problem. Searching the best line approximating the data scatter according to the L_1 criterion corresponds to the solution of the following minimization problem:

$$\sum_{i=1}^{n} |\beta_0 + \beta_1 x_i - y_i|. \qquad (3.20)$$

Although this cost function is not linear, a simple trick allows us to make it linear, at the price of introducing extra variables. In fact, an objective function or constraints involving absolute values can be handled via linear programming by introducing extra variables or extra constraints.

The resulting linear program is:

$$\begin{aligned} &\min imize && \sum_{i=1}^{n} \varepsilon_i \\ &subject \quad to && \varepsilon_i \geq \beta_0 + \beta_1 x_i - y_i \qquad i = 1,2,3,\ldots,n \\ & && \varepsilon_i \geq -(\beta_0 + \beta_1 x_i - y_i) \qquad i = 1,2,3,\ldots,n \end{aligned}$$

Each ε_i is an auxiliary variable standing for the error at the ith point. The constraints guarantee that: $\varepsilon_i \geq \max\{\beta_0 + \beta_1 x_i - y_i, -(\beta_0 + \beta_1 x_i - y_i)\} = |\beta_0 + \beta_1 x_i - y_i|$.

In an optimal solution, each of these inequalities has to be satisfied with equality; otherwise, we could decrease the corresponding ε_i: the optimal solution thus yields a line minimizing the Expression (3.19). Hence, to solve the L_1 regression problem, it suffices to solve the equivalent linear programming problem.

The *p*-variables problem

The model for linear quantile regression is given by:

$\mathbf{y} = \mathbf{X}\beta(\theta) + \boldsymbol{\varepsilon}$,

Where $\mathbf{y}$ is a column vector of responses, $\mathbf{X}$ is the regressor matrix of order $n$xp , $\beta(\theta)$ is the vector of p unknown parameters for the generic conditional quantile θ and $\boldsymbol{\varepsilon}$ is the vector of n unknown errors. The simpler notation $\boldsymbol{\beta}$ will be used to refer to the conditional median case having $\theta = 0.5$.

The least absolute estimates $\hat{\beta}$ for the conditional median is obtained as the solution of the minimization problem:

$$\min_{\beta} \sum |\mathbf{y} - \mathbf{X}\boldsymbol{\beta}| \qquad (3.21)$$

Let us denote $[x]_+$ as the non-negative part of x. By posing:

$\mathbf{s_1} = [\mathbf{y} - \mathbf{X}\boldsymbol{\beta}]_+$

$\mathbf{s_2} = [\mathbf{X}\boldsymbol{\beta} - \mathbf{y}]_+$

the original L_1 problem can be formulated as:

$$\min_{\boldsymbol{\beta}} \{\mathbf{1}'\mathbf{s_1} + \mathbf{1}'\mathbf{s_2} | \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{s}_1 - \mathbf{s_2}, \{\mathbf{s}_1, \mathbf{s_2}\} \in \Re_+^2\}.$$

Furthermore, let:

$$\mathbf{B} = [\mathbf{X} - \mathbf{X}\mathbf{I} - \mathbf{I}],$$

and

$$\psi = \begin{bmatrix} [\boldsymbol{\beta}]_+ \\ [-\boldsymbol{\beta}]_+ \\ \mathbf{s_1} \\ \mathbf{s_2} \end{bmatrix}$$

$$\psi = \begin{bmatrix} [\boldsymbol{\beta}]_+ \\ [-\boldsymbol{\beta}]_+ \\ [\mathbf{y} - \mathbf{X}\boldsymbol{\beta}]_+ \\ [\mathbf{X}\boldsymbol{\beta} - \mathbf{y}]_+ \end{bmatrix}$$

$$\mathbf{d} = \begin{bmatrix} \mathbf{0}_{[\mathbf{p}]} \\ \mathbf{0}_{[\mathbf{p}]} \\ \mathbf{1}_{[n]} \\ \mathbf{1}_{[\mathbf{n}]} \end{bmatrix}.$$

Such reformulation of the problem leads to a standard linear programming problem. The primal formulation of such a problem (equational form) is:

$$\begin{array}{ll} \underset{\psi}{\min imize} & \mathbf{d}'\psi \\ subject\ to & \mathbf{B}\psi = \mathbf{y} \\ & \boldsymbol{\theta} \geq \mathbf{0}. \end{array}$$

Therefore, its dual counterpart is:

$$\begin{array}{ll} \underset{d}{\max imize} & \mathbf{y}'\mathbf{z} \\ subject\ to & \mathbf{B}'\mathbf{z} \leq \mathbf{d}. \end{array}$$

Bearing in mind the main result of linear programming, that is the theorem for which the solutions of such a minimization problem have to be searched in the vertices of the simplex[1], by a simple position, the above problem can be reformulated as follows:

$$\max_{\mathbf{z}} \left\{ \mathbf{y}'\mathbf{z} \middle| \mathbf{X}'\mathbf{z} = \mathbf{0}, \mathbf{z} \in [-1,+1]^n \right\}.$$

In fact the equality:

$$\mathbf{X}'\mathbf{z} = \mathbf{0}$$

can be transformed as follows:

$$\begin{array}{ll} \frac{1}{2}\mathbf{X}'\mathbf{z} = \mathbf{0} & by\ multiplying\ \ by\ \frac{1}{2} \\ \frac{1}{2}\mathbf{X}'\mathbf{z} + \frac{\mathbf{1}}{\mathbf{2}}\mathbf{X}'\mathbf{1} = \frac{\mathbf{1}}{\mathbf{2}}\mathbf{X}'\mathbf{1} & by\ adding\ \frac{1}{2}\mathbf{X}'\mathbf{1}. \end{array}$$

The obtained formulation:

$$\mathbf{X}'\underbrace{\left(\frac{1}{2}\mathbf{z} + \frac{\mathbf{1}}{\mathbf{2}}\mathbf{1}\right)}_{\eta} = \underbrace{\frac{1}{2}\mathbf{X}'\mathbf{1}}_{\mathbf{b}} \qquad (3.22)$$

permits the expression of the dual problem as follows:

$$\max_{\mathbf{J}} \left\{ \mathbf{y}'\mathbf{J} \middle| \mathbf{X}'\mathbf{J} = \mathbf{b}, \mathbf{J} \in [0,1]^n \right\}.$$

The role of 1/2 in Equation (3.21) is seemingly neutral, but it is the key to the generalization to the other conditional quantiles. In fact, the minimization problem for the conditional median, becomes for the generic θ-th conditional quantile:

$$\min_{\beta(\theta)} \sum \rho_\theta \left(y_i - \mathbf{x}'\beta(\theta)\right)$$

A similar set of steps leads to the following dual formulation for the generic quantileregression problem:

$$\max_{\mathbf{z}} \left\{ \mathbf{y}'\mathbf{z} \middle| \mathbf{X}'\mathbf{z} = (\mathbf{1} - \theta)\mathbf{X}'\mathbf{1}, \mathbf{z} \in [0,1]^n \right\},$$

where*(*1 − *θ)* plays the same role that 1/2 played for the median formulation.

3.8.2 Least Absolute Deviation Regression Goodness of Fit

From the user's view point an important question is how to evaluate the successfulness of the estimated model. In the least squares context the most used measure of goodness of fit is the coefficient of determination, R^2 It measures the closeness of the fit to the observed values. In the least squares sense the fit is perfect if $R^2 = 1$. The other end of the range is zero, meaning no explanatory power of the regressors. The coefficient of determination can be defined as the square of the correlation between observed and fitted values.

We suggest to alleviate the arbitrary nature of R^2 in the LAD estimation by defining the coefficient of determination rather as

$$G = 1 - \frac{\sum_{i=1}^{n} \left| y_i - x_i' \tilde{\beta} \right|}{\sum_{i=1}^{n} \left| y_i - md_y \right|} \qquad (3.23)$$

In the definition $\tilde{\beta}$ is the LAD estimator of the coefficient vector, β , and *mdy*is the median of *y*. Like R^2 (3.17) assumes a1so va1ues between zero and one, the former indicating no regression and the 1atter a perfect fit. These properties are due to the following re1ation,

$$\min_{\beta} \sum_{i=1}^{n} \left| y_i - x_i'\beta \right| \le \sum_{i=1}^{n} \left| y_i - md_y \right| \qquad (3.24)$$

where the equa1ity ho1ds if $x_i'\beta = md_y$ for all i = 1, ... , *n.* It may be noted that (3.23) is an analog to the R^2 , because

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - x_i' \tilde{\beta}\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (3.25)$$

3.8.3 Quantile Regression Goodness of Fit

The assessment of goodness of fit for the QR model exploits the general idea leading to the typical R^2 goodness of fit index in classical regression analysis (Koenker and Machado 1999).

Among the different available formulations of the R^2 index, the formulation expressed in terms of the complement to 1 of the ratio between the residual sum of squares and the total sum of squares of the dependent variable (Gujarati 2003) allows us to derive the corresponding index for QR. The latter can be obtained by taking into account that QR differs from OLS regression because it is based on the minimization of an absolute weighted sum (not an unweighted sum of squares as in OLS). In addition, this approach aims to provide an estimation of the quantile conditional distribution of the dependent variable (not an expected conditional distribution as in OLS). It follows that the equivalent of the residual sum of squares is, for each considered quantileθ, the residual absolute sum of weighted differences between the observed dependent variable and the estimated quantile conditional distribution. For the simplest regression model with one explanatory variable

$$Q_\theta(\hat{y}/x) = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta)x \qquad (3.26)$$

The residual absolute sum of weighted differences is the corresponding minimizer

$$\text{RASW}_\theta = \sum_{y_i \geq \hat{\beta}_0(\theta)+\hat{\beta}_1(\theta)x_i} \theta\left|y_i - \hat{\beta}_0(\theta) - \hat{\beta}_1(\theta)x_i\right| + \sum_{y_i < \hat{\beta}_0(\theta)+\hat{\beta}_1(\theta)x_i} (1-\theta)\left|y_i - \hat{\beta}_0(\theta) - \hat{\beta}_1(\theta)x_i\right| \qquad (3.27)$$

the equivalent of the total sum of squares of the dependent variable is, for each considered quantileθ, the total absolute sum of weighted differences between the observed dependent variable and the estimated quantile:

$$\text{TASW}_{\theta} = \sum_{y_i \geq \theta} \theta \left| y_i - \hat{\theta} \right| + \sum_{y_i < \theta} (1-\theta) \left| y_i - \hat{\theta} \right| \qquad (3.28)$$

The obtained *pseudo* R^2 can be considered as an index comparing the residual absolute sum of weighted differences using the selected model with the residual absolute sum of weighted differences using a model with only the intercept. The obtained *pseudo* R^2 can be computed as follows:

$$pseudo R_{\theta}^2 = 1 - \frac{\text{RASW}_{\theta}}{\text{TASW}_{\theta}}$$

As $RASW_{\theta}$ is always less than $TASW_{\theta}$, the *pseudo* R_{θ}^2 ranges between 0 and 1. It is worth noting that the index cannot be considered a measure of the goodness of fit of the whole model because it is related to a given quantile. In practice, for each considered quantile, the corresponding *pseudo* R_{θ}^2 can be evaluated at a local level, thereby indicating whether the presence of the covariates influences the considered quantile.

The *pseudo* R_{θ}^2 can also be used to assess the model with the best goodness of fit between nested models. If two nested models are considered, the comparison between the corresponding *pseudo* R^2 can be used to evaluate whether it is advantageous to move from a restricted to a more complex model. Any eventual improvement can be quantified through the so-called *relative* R^2. This is derived from a different interpretation of the *pseudo* R_{θ}^2, where $RASW_{\theta}$ and $TASW_{\theta}$ represent the residual absolute sum of weighted differences between the observed dependent variable and the estimated quantile conditional distribution, respectively, in the more complex model and in the simplest one (Hao and Naiman 2007).

CHAPTER FOUR

RESULTS AND DISCUSSION

4.0 Introduction

This chapter consists of the results obtained from regression analysis using OLS and QR techniques. Correlation and Stepwise regression was also examined. Criteria used for the goodness of fit of the model is coefficient of determination. All test of significance were conducted at 5% level using a statistical software package Eview, R and Statgraphics.

4.1 Numerical Illustration and Discussion of the result

We have assumed simple linear and multiple regression models for simplicity. The results of the OLS regression as well as the results from the QR are shown in Tables

Table 4.1:Correlations matrix

	MPG Highway	Weight	Wheelbase	Width	Horsepower	Engine Size	Length
MPG Highway		-0.8107	-0.6154	-0.6404	-0.6190	-0.6268	-0.5429
		(93)	(93)	(93)	(93)	(93)	(93)
		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Weight	-0.8107		0.8719	0.8750	0.7388	0.8451	0.8063
	(93)		(93)	(93)	(93)	(93)	(93)
	0.0000		0.0000	0.0000	0.0000	0.0000	0.0000
Wheelbase	-0.6154	0.8719		0.8072	0.4869	0.7325	0.8237
	(93)	(93)		(93)	(93)	(93)	(93)
	0.0000	0.0000		0.0000	0.0000	0.0000	0.0000
Width	-0.6404	0.8750	0.8072		0.6444	0.8671	0.8221
	(93)	(93)	(93)		(93)	(93)	(93)
	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
Horsepower	-0.6190	0.7388	0.4869	0.6444		0.7321	0.5509
	(93)	(93)	(93)	(93)		(93)	(93)
	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000
Engine Size	-0.6268	0.8451	0.7325	0.8671	0.7321		0.7803
	(93)	(93)	(93)	(93)	(93)		(93)
	0.0000	0.0000	0.0000	0.0000	0.0000		0.0000
Length	-0.5429	0.8063	0.8237	0.8221	0.5509	0.7803	
	(93)	(93)	(93)	(93)	(93)	(93)	
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	

Correlation (Sample Size) P-Value

From table 4.1 this table shows Pearson product moment correlations between each pair of variables. These correlation coefficients range between -1 and +1 and measure the strength of the linear relationship between the variables. Also shown in parentheses is the number of

pairs of data values used to compute each coefficient. The third number in each location of the table is a P-value which tests the statistical significance of the estimated correlations. P-values below 0.05 indicate statistically significant non-zero correlations at the 95.0% confidence level.

4.2 Ordinary least squares regression

Table 4.2:OLS Regression model

		Standard	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	18.9878	12.4967	1.51943	0.1323
Weight	-0.0129731	0.00171341	-7.5715	0.0000
Wheelbase	0.257333	0.115077	2.23619	0.0279
Width	0.159445	0.204413	0.780014	0.4375
Length	0.0549857	0.0429656	1.27976	0.2041
Horsepower	0.0066876	0.0105224	0.635559	0.5268
Engine Size	0.418763	0.691919	0.60522	0.5466

From table 4.2 the output shows the results of fitting a multiple linear regression model to describe the relationship between MPG Highway and 6 independent variables. The equation of the fitted model is

$$mpg = 18.9878 - 0.0129731\,weight + 0.257333\,wheelbase + 0.159445\,width + 0.0549857\,length + 0.0066876\,horsepower + 0.418763\,engine\,size$$

In determining whether the model can be simplified, notice that the highest P-value on the independent variables is 0.5466, belonging to Engine Size. Since the P-value is greater or equal to 0.05, that term is not statistically significant at the 95.0% or higher confidence level. Consequently, we should consider removing Engine Size from the model and performed stepwise regression.

Table 4.3: Stepwise regression

		Standard	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	26.3169	7.08803	3.71286	0.0004
Weight	-0.0110749	0.00111771	-9.90852	0.0000
Length	0.0816996	0.0389896	2.09542	0.0390
Wheelbase	0.210046	0.100849	2.08278	0.0401

From the table 4.3 the output shows the results of fitting a stepwise regression model to describe the relationship between MPG Highway and 6 independent variables. The equation of the fitted model is

$$mpg = 26.3167 - 0.0110749\, weight + 0.0816996\, length + 0.210046\, wheelbase$$

Since the P-value in the table is less than 0.05, there is a statistically significant relationship between the variables at the 95.0% confidence level. The R-Squared statistic indicates that the model as fitted explains 70.6502% of the variability in MPG Highway. The adjusted R-squared statistic, which is more suitable for comparing models with different numbers of independent variables, is 69.6609%.

In determining whether the model can be simplified, we notice that the highest P-value on the independent variables is 0.0401, belonging to Wheelbase. Since the P-value is less than 0.05, that term is statistically significant at the 95.0% confidence level. Consequently, we probably don't want to remove any variables from the model.

4.3 Heteroskedasticity Test: White

Table 4.4: White (Error) test

F-statistic	3.283434	Prob. F	0.0000
Obs*R-squared	53.65804	Prob. Chi-Square	0.0017
Scaled explained SS	85.32316	Prob. Chi-Square	0.0000
Important Regression Statistics			
R^2	57.6968		
Adjusted R^2	40.1247		

4.4 Comparison of OLS and QR as the number of variable increases

Table 4.5: Numerical result of the OLS and QRsimple linear models

	OLS		QR(0.25)		QR(0.50)		QR((0.75)	
Parameter	*Estimate*	*P-Value*	*Estimate*	*P-Value*	*Estimate*	*P-Value*	*Estimate*	*P-Value*
CONSTANT	51.6014	0.0000	47.19070	0.0000	48.54839	0.0000	55.08696	0.0000
Weight	-0.00732706	0.0000	-0.006512	0.0000	-0.006452	0.0000	-0.007826	0.0000
Important Regression Statistics								
	R^2	65.7166	R_θ^2	38.5070	R_θ^2	41.6138	R_θ^2	44.868
	Adjusted R^2	65.3399	Adjusted R_θ^2	37.8312	Adjusted R_θ^2	40.9722	Adjusted R_θ^2	44.262

From table 4.5 the output of the results of fitting an OLS regression as well as the result from the QR are shown. Estimated coefficient for the weightshow a negative and significant impact of these variable on MPG. However, white test for heteroskedasticity indicates that heteroskedasticity is present in the data.This leads to OLS estimates being inefficient.Furthermore, quantile regression has the advantage that it does not assume normally distributed errors for the estimation of the coefficients, as OLS does and allows coefficients to change for different sub-sets of the sample.

Table 4.6: Numerical result of the OLS and QRmultiple regression models

	OLS		QR(0.25)		QR(0.50)		QR((0.75)	
Parameter	*Estimate*	*P-Value*	*Estimate*	*P-Value*	*Estimate*	*P-Value*	*Estimate*	*P-Value*
CONSTANT	37.5218	0.0000	23.04633	0.0000	29.15135	0.0000	43.15210	0.0000
Weight	-0.00963	0.0000	-0.009112	0.0000	-0.008969	0.0000	-0.009709	0.0000
Length	0.115526	0.0019	0.176062	0.0000	0.146861	0.0000	0.093851	0.0202
Important Regression Statistics								
	R^2	69.2197	R_θ^2	47.8174	R_θ^2	47.8445	R_θ^2	47.726
	Adjusted R^2	68.5357	Adjusted R_θ^2	46.6578	Adjusted R_θ^2	46.6855	Adjusted R_θ^2	46.565

From table 4.6 the output of the results of fitting aMLR as well as the result from the QRs are shownas the number of regressor increase. Estimated coefficient for the weight show a negative and significant impact of these variable on MPG while length have significantly positive effects on MPG. However, white test for heteroskedasticity indicates that heteroskedasticity is present in the data.This also leads to OLS estimates being inefficient.Furthermore, quantile regression has the advantage that it does not assume

normally distributed errors for the estimation of the coefficients, as OLS does and allows coefficients to change for different sub-sets of the sample.

QR improve the efficiency of the estimators compared to OLS and allow analysing independently. They may allow detecting significant effects from variables whose coefficients may have been dismissed for not appearing to be statistically different from zero in a mean based model such as OLS.

The high $R^2$of the OLS could be interpreted as evidence of bias due to omitted variable; quantile method is also a good alternative when this type of bias is suspected. Nevertheless, the pseudo R^2which is the relevant goodness of fit measure in quantile regression show an improvement on the fit of the model compared to OLS. Furthermore, the model appears to explain better the MPG for the higher quantiles than for lower quantiles

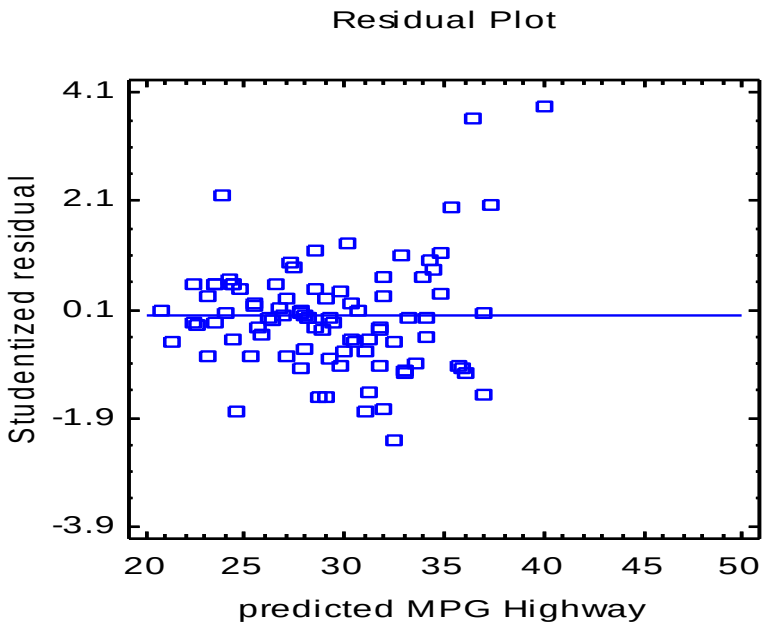


Fig. 4.1: Residual plot

CHAPTER FIVE

SUMMARY CONCLUSIONS AND RECOMMENDATIONS

5.0 Introduction

In this chapter, we present summary, conclusion and recommendations based on the results obtained in the preceding chapter.

5.1 Summary

Our primary goal in this work as initially stated in our objectives is to investigate the robustness of quantile regression as an alternative to least squares regression, especially when the number of regressors increases. This thesis presented a general overview of the quantile regression method, consisting of a non-technical introduction to the basic model and its crucial features and of a short review of two major applications. We have seen that quantile regression offers an extension of univariatequantile estimation to estimation of conditional quantile functions and that it complements the established mean regression methods by adding more flexibility in the estimation sand more robustness particularly in non-Gaussian distribution settings. The covariate effects are allowed to influence location, scale and shape of the response distribution unlike conventional techniques which usually investigated location-shift paradigms. Furthermore, by focusing on local parts of the conditional distribution, quantile regression methods offer a useful deconstruction of conditional mean models.

Effort are made to model miles per gallon in highway driving using the quantile regressions approach, showing that OLS estimation is not always an appropriate method to analyze miles per gallon in highway driving. The two independent variables have been found to have an influence on the miles per gallon in highway driving. We suggest that researchers retain their list of independent variables, even if those variables are not significantly associated with the dependent (response) variable at the bivariate level, until they examine their multiple regression results for any evidence of heteroskedasticity.

QR is an invaluable tool for facing heteroskedasticity, and provides a method for modeling the rates of change in the response variable at multiple points of the distribution when such rates of change are different. It is, however, also useful in the case of homogeneous regression models outside of the classical normal regression model, and in the case where the error independence assumption is violated, as no parametric distribution assumption is required for the error distribution.

5.2 Conclusion

Quantile regression is offering a comprehensive strategy for completing the regression picture as it goes beyond this primary goal of determining only the conditional mean, and enables one to pose the question of relationship between the response variable and covariate at any quantile of the conditional distribution function. Quantile regression overcomes various problems that OLS is confronted with frequently; error terms are not constant across a distribution, thereby violating the axiom of homoscedasticity. Also, by focusing on the mean as a measure of location, information about the tails of a distribution is lost. As indicate in the data of miles per gallon in highway driving.

5.3 Recommendations

From the analysis and evaluation of the results via preceding discussions in these study so far, the following recommendations are proffered.

- ✓ The performance is stable, and robust against common deviations from the model assumptions.
- ✓ The model should trigger reviews rather than automatic disallowances. The researcher used QR as a tool in guiding policymakers toward sound policy decisions rather than as the final determinant of policy decisions.

5.4 Contribution to knowledge

- Ability to bring to limelight the advantage of quantile regression in the data analysis

- This research has also help to employ the pseudo R^2 to identify or determine the presence of outliers in the model

5.5 Further research

Based on publicly available dataset on fuel consumption in miles per gallon in highway driving, QR model performs better compared to the OLS methods. QR can also be applied when rigid assumptions associated with OLS hold.

REFERENCES

Abreyaya, J., and Dahl, C. (2008). The Effects of Birth Inputs on Birthweight. *Journal of Business & Economic Statistics*, 5(2), 379-397.

Akaike, H. (1974). Factor analysis and AIC. *Pschychometrika, 52*, 317 - 332.

Buchinsky, M. (1998). Recent advances in quantile regression models: a practical guideline for empirical research. *The Journal of Human Resources, 33*(1), 88-126.

Buhai, S. (2004). Quantile regressions: overview and selected applications. *Unpublished manuscript*, Rotterdam Tinbergan Institute and Erasmus University.

Cade, B., and Noon., B. (2003). A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment., 1*(8), 412-420.

Chen, C. (2004). An introduction to Quantile Regression and the Quantreg Procedure.*SUGI*,30, 213-230.

Chernozhukov, V., Fernandez-Val, I., and Melly, B. (2013). Quantile and Probability Curves. *Econometrica*, 19(2), 2205-2268.

Cizek, P. (2003). *Quantile regression in XploRe Guide.* (Z. H. W. Hardle, Ed.) Berlin: Springer.

Draper, N. R., & Smith, H. (1981). *Applied Regression Analysis* (2nd ed.). John wiley and sons.

Efroymson, M. (1960). *Multiple Regression.* (A. a. Ralston, Ed.) New York, NY: John Wiley and Sons, Inc.

Fitzenberger, B., Koenker, R., & (editors), J. M. (2002). *Economic Applications of Quantile Regression.* New York, NY: Physica-Verlag Heidelberg.

Galvao, K., and Montes, R. (2012). Asymptotics for quantile regression model with different effect. *Journal of Econometrics*, 5, 76-91.

Geraci, M., and Bottai, M. (2007). Quantile Regression for Longitudinal Data Using the Asymmetric Laplace Distribution. *Biostatistics*, 15, 140-154.

Gorr, W.L., and C. Hsu. (1985). An adaptive filtering procedure for estimating regression Quantiles. *Management Science*.31(8), 1019-1029

Green, H., and Kozek, A. (2003). Modeling weather data by approximate regression. *Quantiles Anziam, 44*, 229-249.

Gujarati DN (2003).*Basic Econometrics*, International Edition. McGrawHill

Hao L, N. D. (2007). *AppliedQuantile Regression*(3nd ed.). John Wiley and Sons.

Honore, B., J. Powell, and S. Khan. (2002). Quantile regression under random censoring. *Journal of Econometrics. 109(1):67-109.*

Karlsson, S. (2006). Estimation and inference for Quantile Regression of Longitudinal Data, With Applications in Biostatistics. *International joural of social science* , 8, 18-36.

Koenker, R. (2005). Quantile Regression. New York, NY: Cambridge University Press.

Koenker, R. and G. Bassett, Jr. (1978.). Regression quantiles. *Journal ofEconometrica*, 46(1), 33-50

Koenker, R., and Hallock, K. (2001). Quantile regression. *Journal of Economic Perspective, 15*(4), 143-156.

Koenker, R., and Machado, J. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association, 94*, 1296-1310.

Kutner, M., Nachtsheim, C., and Neter, J. (2004). *Applied Linear Regression Models. 4th Ed. McGraw-Hill Irwin, Inc Boston, MA.*

Lee. (2005). Quantile robust to outlier. *JournalofStatistics and numerical method*, 4, 35-57.

Mallow, C. (1973). Some comments on Cp. *Technometrics, 15*, 661 - 675.

Mosteller, F. and J. Tukey. (1977). *Data Analysis and Regression: A Second Course in Statistics*. Addison-Wesley: Reading, MA.

Myers, R. (1990). *Classical and Modern Regression with Applications.* Boston, MA: PWS-Kent Publishing Company.

Neter, J., and M.H. Kutner, C. N. (1996). *Applied Linear Regression Models* (3rd ed.). Chicago, IL: Irwin, Inc.

Rosen. (2010). Set Identification via Quantile Restrictions in Short Panels. *Journal of Econometrics*, 23,127-137.

Smith, R. L. (1992). Suppressor variables in multiple regression/correlation. *Educational and Psychological Measurement*, 5, 17-29.

Stigler, S. (1986). *The History of Statistics: The Measurement of Uncertainty* . Cambridge: The Belknap Press of Harvard University Press.

Young, L., and Easterling, R. (1994). Estimation of extreme quantiles based on sensitivity tests: a comparative study. *Technometrics, 36*(1), 48-60.

Wagner HM.(1959) Linear programming techniques for regression analysis. *Journal of the American Statistical Association* 54, 206–212

Young, T., and Guess, F. (2002). Developing and mining higher quality information in automated relational databases for forest product manufacture. *International Journal of Reliability and Application, 3*(4), 155-164.

Young, T.M. and C.W. Huber. (2004). Prodictive Modeling of the physical properties of wood composite using genetic algorithms with considerations for distributed data fusion. *Proceedings of the 38th international particleboard/ Composite Materials Symposium*. Washington: Washington State Univ., Pullman, WA. 145-153.

APPENDIX A

Fuel consumption data set of 93 cars was measured on 6 regressors. The response variable is miles per gallon.

MPG	Weight	Length	Wheelbase	Width	Engine size	Horsepower
31	2705	177	102	68	1.8	140
25	3560	195	115	71	3.2	200
26	3375	180	102	67	2.8	172
26	3405	193	106	70	2.8	172
30	3640	186	109	69	3.5	208
31	2880	189	105	69	2.2	110
28	3470	200	111	74	3.8	170
25	4105	216	116	78	5.7	180
27	3495	198	108	73	3.8	170
25	3620	206	114	73	4.9	200
25	3935	204	111	74	4.6	295
36	2490	182	101	66	2.2	110
34	2785	184	103	68	2.2	110
28	3240	193	101	74	3.4	160
29	3195	198	108	71	2.2	110
23	3715	178	110	74	3.8	170
20	4025	194	111	78	4.3	165
26	3910	214	116	77	5	170
25	3380	179	96	74	5.7	300
28	3515	203	113	74	3.3	153
28	3085	183	104	68	3	141
26	3570	203	110	69	3.3	147
33	2270	174	98	66	1.5	92
29	2670	172	97	67	2.2	93
27	2970	181	104	68	2.5	100
21	3705	175	112	72	3	142
27	3080	192	105	69	2.5	100
24	3805	180	97	72	3	300
33	2295	174	98	66	1.5	92
28	3490	202	113	74	3.5	214
33	1845	141	90	63	1.3	63
30	2530	171	98	67	1.8	127
27	2690	177	100	68	2.3	96
29	2850	180	101	68	2.3	105
30	2710	179	103	70	2	115
20	3735	176	119	72	3	145
30	3325	192	106	71	3	140
26	3950	212	114	78	4.6	190
50	1695	151	93	63	1	55
36	2475	164	97	67	1.6	90
31	2865	175	100	70	2.3	160
46	2350	173	103	67	1.5	102
31	3040	185	107	67	2.2	140
33	2345	168	94	63	1.5	81

MPG	Weight	Length	Wheelbase	Width	Engine size	Horsepower
29	2620	172	98	66	1.8	124
34	2285	166	94	64	1.5	92
27	2885	184	104	69	2	128
22	4000	200	113	72	4.5	278
24	3510	188	103	70	3	185
23	3515	191	106	71	3	225
26	3695	205	109	73	3.8	160
26	4055	219	117	77	4.6	210
37	2325	164	97	66	1.6	82
36	2440	172	98	66	1.8	103
34	2970	184	103	69	2.5	164
24	3735	190	110	72	3	155
25	2895	169	96	69	1.3	255
29	2920	175	105	67	2.3	130
25	3525	187	110	69	3.2	217
26	2450	166	95	65	1.6	100
26	3610	199	113	73	3.8	140
33	2295	172	98	67	1.5	92
24	3730	190	107	70	3	202
33	2545	170	96	66	1.6	110
30	3050	181	103	67	2.4	150
23	4100	190	112	74	3	151
26	3200	188	104	69	3	160
31	2910	188	103	67	2.3	155
31	2890	190	105	70	2.2	110
23	3715	194	110	74	3.8	170
28	3470	201	111	74	3.8	170
30	2640	173	97	67	1.8	92
41	2350	177	99	66	1.6	74
31	2575	181	101	66	2	110
28	3240	196	101	75	3.4	160
27	3450	195	108	72	3.4	200
28	3495	177	111	74	3.8	170
26	2775	184	99	67	2.1	140
38	2495	176	102	68	1.9	85
37	2045	146	90	60	1.2	73
30	2490	175	97	65	1.8	90
30	3085	179	102	67	2.2	130
43	1965	161	93	63	1.3	70
37	2055	162	94	65	1.5	82
32	2950	174	99	69	2.2	135
29	3030	188	103	70	2.2	130
22	3785	187	113	71	2.4	138
33	2240	163	93	63	1.8	81
21	3960	187	115	72	2.5	109
30	2985	180	103	67	2	134
25	2810	159	97	66	2.8	178
28	2985	190	104	67	2.3	114

APPENDIX B

R Code for Multiple Quantile Regression:

*First, we create a function in R, then open it up to edit;

```
function(){
```

*Here, we declare variable names (i.e., x is the first column of the data table “test”);

```
x=test[,1]
y=test[,2]
z=test[,3]
```

*When plotting more than one variable, use the “+” sign;

```
plot(x+z,y)
points(x+z,y,cex=.5,col="blue")
```

*We declare the percentiles of interest using the “taus” command.

```
taus<- c(.05,.1,.25,.75,.9,.95)
f <- coef(rq((y)~(x+z),tau=taus))
yy<- cbind(1,x+z)%*%f
for(i in 1:length(taus)){
lines(x+z,yy[,i],col = "gray")
}
```

*Below, we add the multiple linear regression line and the quantile regression lines to the plot;

```
abline(lm(y~x+z),col="red",lty = 2)
abline(rq(y~x+z),col="blue")
```

*Lastly, we ask for a summary of the quantile output;

```
summary(rq(y~x+z, ci=FALSE, tau=taus))
}
```