International Journal of Advanced Statistics and Probability, 3 (2) (2015) 138-145
www.sciencepubco.com/index.php/IJASP
©Science Publishing Corporation
doi: 10.14419/ijasp.v3i2.4686
Research Paper

Analysis of quantile regression as alternative to ordinary least squares

Abdullahi Ibrahim*, Abubakar Yahaya

Department of Mathematics, Ahmadu Bello University, Zaria – Nigeria
**Corresponding author E-mail: ibworld82@yahoo.com*

Abstract

In this article, 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 as well as approximate distributions of the associated test statistics for the parameters. Furthermore, we suggest a goodness of fit statistic called the least absolute deviation (LAD) coefficient of determination. 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.

Keywords: *Quantile Regression; Model Validation; Stepwise Regression; Linear Programming.*

1. Introduction

In most regression problems, interest lies in studying the relationship between two or more variables, this is because it is an important aspect in the philosophy of science to study the concept of relationship between varying qualities or events. The classical regression analysis procedures normally focus on the mean; that is to say, the relationship between the response and predictor variable(s) is summarized by the conditional mean of the response for each fixed value of the predictors. The idea of modeling and fitting the conditional-mean is at the core of a broad family of regression-modeling approaches, including the familiar simple linear regression, multiple linear regression models with "heteroskedastic" errors using weighted least squares as well as nonlinear regression models.
Robust estimation refers to the ability of a procedure to produce an estimate that is highly insensitive to model misspecifications. Hence, robust estimates should be good under wide range of possible data generating distributions. In regression context, under the normality assumptions, the errors are assumed to be independent, identically and normally distributed random variable; the OLS is believed to be one of the most efficient among the classical estimation procedures. However, once the normality assumption is dropped, it is possible to find estimation methods that are more efficient than the OLS. Specifically, this is true when the data generating process has fat tails resulting to several outliers. In these cases, the OLS becomes highly unstable and sample dependent because of the quadratic weighting, which makes the procedure very sensitive to outliers.
QR was first introduced by Koenker and Bassett [7] and is intended to offer a comprehensive strategy for completing the regression picture (Koenker, [6]). Unlike OLS, QR does not impose any strict parametric assumptions. 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.
As with multiple linear regressions (MLR), QR has many applications, and was originally developed for statistical use, as the first QR publication was in Econometrics (Koenker and Bassett, [7]) where it was insightfully envisioned a more robust regression approach capable of modeling conditional quantile functions beyond the classical OLS approach to model building. Koenker and Bassett [7] noted that "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".

QR 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. Quantiles are robust in relation to handling outliers (an interested reader should consult Lee [8] for further details). 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 away from the rest of the data points. On the other hand, the effect of such a distant observation on the sample median (middle quartile) is bounded no matter how far the outlying observation is. Other quantiles enjoy similar property as the effect of outlying observations on the t-th sample quantile is bounded and QR inherits this robustness property Cizek [4]; hence QR estimates are reliable in the presence of extreme outliers. Chen [3] also investigated this property in a survey by considering the data of body mass index (BMI) against age for up to 8,280 men over a four year (1999-2002) period.

Some other interesting applications of QR include those in ecological and environmental studies by Cade and Noon [2] in which some prediction intervals were estimated. As noted by Cade and Noon [2], it is extremely difficult to identify, document, and measure every ecological independent variable; as a result, using classical methods such as OLS may prove to be difficult in arriving at a statistically significant model. However, models built using only portions of the response variable distribution may be more useful (Cade and Noon, [2]). Interestingly, Green and Kozek [5] use an approximate QR method to model weather data. These models are approximate because they are formed by applying quantile functions onto parametric models. Parametric weather distributions are modeled over time and regression quantiles are then applied to the models. Five-curve summaries were obtained for the probability distributions of the weather data and the results were quite interesting.

Buhai [1] provided 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 [1] was able to articulate a thorough summary of each.

Bayesian approach to quantile regression is another current area that has attracted a lot of interest as it has shown two obvious advantages over classical inference. First, Bayesian approach does not rely on approximations to asymptotic variances of the estimators thereby leading to nearly exact estimates. Second, it provides estimation and forecasts, which fully take into account parameter uncertainty (Yu and Zhang, [11]). The idea of Bayesian QR has been explored by Yu, Kerm and Moyeed, [10].

Conventional regression models have been used in numerous statistical downscaling studies and are the cornerstone of software packages such as SDSM (Wilby et al. [9]). Despite their popularity, these conditional mean models have some limitations. When interest is in the quantiles of the conditional distribution rather than the mean, standard regression models may fail to provide the desired information because the assumption of homogenous variance may not be justified. Also, it is common practice to assume that regression residuals are normally distributed but this may not be a valid assumption, even after application of some normalizing transformation. The non-normality of residuals may not be a serious issue if the only interest is in the mean of the conditional distribution. However, when interest is in the tails of the conditional distribution, the distribution of residuals becomes important. We also note that conventional regression models can be sensitive to outliers. While methods are available to deal with outliers, it is an issue that is often not properly dealt with in practice.

This paper presents a QR approach which provides different estimator for each quantile, hence the main aim of the study is to investigate the robustness of QR as an alternative to OLS, especially when the number of regressors gets larger. The rest of this paper is organized in such a way that, the next section gives a brief background of the development of QR model. In section three, we provide a procedure for finding the estimate of the parameters of the models after which a numerical example and discussion of the results followed. Section four provides some concluding remarks.

2. Materials and methods

The data used in this research was obtained from statgraphics software. The response variable is miles per gallon and the independent variables are Horsepower, Weight (pounds), Width (inches), Length, Engine size and Wheel base (inches). The data set is publicly available at: http://www.csus.edu/indiv/v/velianitis/ds101/schedule.htm. In order to demonstrate the analytical power of QR, we have used statgraphics and Eviews (a complete programming package) to analyze the miles per gallon in highway driving data.

2.1. Multiple linear regressions

The general linear regression model can be expressed as

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i \\ &= \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i \qquad i = 1, 2, \ldots, n \end{aligned} \tag{1}$$

The OLS method chooses the $\beta's$ in equation 1 so that the sum of the squares due to errors, $\sum_{i=1}^{n}\varepsilon_i^2$ is minimized. The model in terms of the observations, Equation 1, may be written in matrix notation as:

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon} \text{ and } E(\mathbf{y}) = \mathbf{X\beta}$$

where:

$$\mathbf{y} = \textit{response vector} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \textit{design matrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}, \quad \boldsymbol{\beta} = \textit{parameter vector} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \text{ and}$$

$$\boldsymbol{\varepsilon} = \textit{error vector} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Now we wish to find the vector of least squares estimators of the parameters, $\hat{\boldsymbol{\beta}}$ that minimizes:

$L = \sum_{i=1}^{n}\varepsilon_i^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X\beta})'(\mathbf{y} - \mathbf{X\beta})$ and this can be obtained by:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{2}$$

Thus, the fitted regression model can be obtained by:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \tag{3}$$

2.2 Quantile regression

The model for linear quantile regression *p*-variables problem is given by:

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

where $\mathbf{y}$ a column vector of responses, $\mathbf{X}$ is the regressor matrix of order $n$xp , $\boldsymbol{\beta}(\theta)$ 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 θ = 0.5. The least absolute estimates $\hat{\beta}$ for the conditional median is obtained as the solution of the minimization problem:

$$\min_{\beta} \sum |\mathbf{y} - \mathbf{X\beta}| \tag{4}$$

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

$$\mathbf{s}_1 = [\mathbf{y} - \mathbf{X\beta}]_+$$
$$\mathbf{s}_2 = [\mathbf{X\beta} - \mathbf{y}]_+$$

The original L_1 problem can be formulated as:

$$\min_{\beta} \left\{ \mathbf{1}'\mathbf{s}_1 + \mathbf{1}'\mathbf{s}_2 \,\middle|\, \mathbf{y} = \mathbf{X\beta} + \mathbf{s}_1 - \mathbf{s}_2, \{\mathbf{s}_1, \mathbf{s}_2\} \in \Re_+^2 \right\}.$$

Furthermore, let $\boldsymbol{\beta} = \left[\mathbf{X} - \mathbf{XI} \text{ - } \mathbf{I}\right]$ and $\boldsymbol{\psi} = \begin{bmatrix} [\boldsymbol{\beta}]_+ \\ [-\boldsymbol{\beta}]_+ \\ \mathbf{s}_1 \\ \mathbf{s}_2 \end{bmatrix} = \begin{bmatrix} [\boldsymbol{\beta}]_+ \\ [-\boldsymbol{\beta}]_+ \\ [\mathbf{y} \text{ - } \mathbf{X}\boldsymbol{\beta}]_+ \\ [\mathbf{X}\boldsymbol{\beta} \text{ - } \mathbf{y}]_+ \end{bmatrix}$, $\mathbf{d} = \begin{bmatrix} \mathbf{0}_{[p]} \\ \mathbf{0}_{[p]} \\ \mathbf{1}_{[n]} \\ \mathbf{1}_{[n]} \end{bmatrix}$.

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

$$\begin{aligned} &\underset{\boldsymbol{\psi}}{\textit{minimize}} \ \mathbf{d}'\boldsymbol{\psi} \\ &\textit{subject to} \\ &\qquad \mathbf{B}\boldsymbol{\psi} = \mathbf{y} \\ &\qquad \boldsymbol{\theta} \geq \mathbf{0}. \end{aligned}$$

Therefore, its dual counterpart is:

$$\begin{aligned} &\underset{\mathbf{d}}{\textit{maximize}} \ \mathbf{y}'\mathbf{z} \\ &\textit{subject to} \\ &\qquad \mathbf{B}'\mathbf{z} \leq \mathbf{d}. \end{aligned}$$

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, by a simple position, the above problem can be reformulated as follows:

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

In fact the equality $\mathbf{X}'\mathbf{z} = \mathbf{0}$ can be transformed as follows:

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

The obtained formulation: $\mathbf{X}'\underbrace{\left(\frac{1}{2}\mathbf{z} + \frac{1}{2}\mathbf{1}\right)}_{\eta} = \underbrace{\frac{1}{2}\mathbf{X}'\mathbf{1}}_{\mathrm{b}}$ (5)

permits the expression of the dual problem as follows:

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

The role of 1/2 in equation (4) 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:

$$\min_{\beta(\theta)} \sum_{i=1}^{n} \rho_\theta \left(y_i - \mathbf{x}_i'\boldsymbol{\beta}(\boldsymbol{\theta}) \right)$$

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

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

where $(1 - \theta)$ plays the same role that 1/2 played for the median formulation.

2.3 Quantile regression goodness of fit

It follows that the equivalent of the residual sum of squares is, for each considered quantile θ, 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, we've:

$$Q_\theta(\hat{y}/x) = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta)x \tag{6}$$

The residual absolute sum of weighted differences is the corresponding minimizer

$$\text{RASW}_\theta = \sum_{y_i \geq \hat{\beta}_0(\theta)+\hat{\beta}_1(\theta)x_i} \theta\left|y_i - \hat{\beta}_0(\theta) - \hat{\beta}_1(\theta)x_i\right| + \sum_{y_i < \hat{\beta}_0(\theta)+\hat{\beta}_1(\theta)x_i} (1-\theta)\left|y_i - \hat{\beta}_0(\theta) - \hat{\beta}_1(\theta)x_i\right| \tag{7}$$

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

$$\text{TASW}_\theta = \sum_{y_i \geq \theta} \theta\left|y_i - \hat{\theta}\right| + \sum_{y_i < \theta} (1-\theta)\left|y_i - \hat{\theta}\right| \tag{8}$$

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_θ is always less than TASW_θ, the $pseudo\ R_\theta^2$ 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 of size θ. In practice, for each considered quantile, the corresponding $pseudo\ R_\theta^2$ can be evaluated at a local level, thereby indicating whether the presence of the covariates influences the considered quantile. The $pseudo\ R_\theta^2$ can also be used to assess the model with the best goodness of fit between nested models.

3. Results and discussions

In previous sections, we described the two regression methods considered in this article; while in this section, we conduct an analysis using numerical data obtained from statgraphics software package to investigate a good alternative to ordinary least square. The analysis was performed using Eviews statistical package.

3.1. Results on OLS and QR as the number of regressors increases

Table 1: Comparison for Simple OLS and QR Methods

	OLS		Q(0.25)		Q(0.50)		Q(0.75)	
Parameter	Estimate	P-Values	Estimate	P-Values	Estimate	P-Values	Estimate	P-Values
Constant	79.0963	0.0000	73.9091	0.0000	73.2857	0.0000	75.5714	0.0000
Wheelbase	-0.4811	0.0000	-0.4545	0.0000	-0.4286	0.0000	-0.4286	0.0000
Important Regression Statistics								
	R^2 = 0.3787		*pseudo* R^2 =	0.1822	*pseudo* R^2 =	0.2295	*pseudo* R^2 =	0.2772
	R^2-Adj = 0.3719		R^2-Adj =	0.1732	R^2-Adj =	0.2210	R^2-Adj =	0.2693

Dependent variable: MPG Highway

From table 1 the output of the results of fitting a linear model to describe the relationship between MPG and wheelbase. In determining whether the model can be simplified, notice that the highest P-value on the independent variables is 0.0000, belonging to Wheelbase. Since the P-value is less than 0.05, that term is statistically significant at the 95% confidence level. Consequently, we may not want to remove any variable from the model, despite this our quantile regression differs with each of its coefficient estimates showing significance variation with each quantile. This suggests

that a benefit exists to utilizing quantile regression to examine the impact of our independent variable on our dependent variable MPG. From our model explaining MPG through wheelbase, the impact of wheelbase on MPG showed a decrease in MPG per unit increase in wheelbase. The OLS R^2 statistic indicates that the model as fitted explains 37.78% of the variability in MPG Highway and $R_{0.75}$-squared statistic indicates that the model as fitted explains only 27.72% of the variability in MPG Highway.

Table 2: Comparison for Multiple OLS and QR Methods

	OLS		Q(0.25)		Q(0.50)		Q(0.75)	
Parameter	Estimate	P-Values	Estimate	P-Values	Estimate	P-Values	Estimate	P-Values
Constant	68.6614	0.0000	65.9155	0.0000	57.6544	0.0000	63.3398	0.0000
Wheelbase	-0.3218	0.0000	-0.3383	0.0001	-0.2234	0.0018	-0.2471	0.0003
Horsepower	-0.0426	0.0000	-0.0303	0.0002	-0.0375	0.0000	-0.0483	0.0000
Important Regression Statistics								
	R^2 = 0.5124		Pseudo R^2 =	0.2773	Pseudo R^2 =	0.3161	Pseudo R^2 =	0.3528
	R^2-Adj = 0.5016		R^2-Adj =	0.2612	R^2-Adj =	0.3009	R^2-Adj =	0.3384

Dependent variable: MPG Highway

From table 2 the results of fitting a MLR and QR model to describe the relationship between MPG Highway and 2 independent variables. In determining whether the model can be simplified, notice that the highest P-value on the independent variables is 0.0000, belonging to wheelbase and horsepower. Since the P-value is less than 0.05, that term is statistically significant at the 95% confidence level. Consequently, we probably don't want to remove any variables from the model, despite this our quantile regression differs with each of its coefficient estimates showing significance varying with each quantile. This suggests that a benefit exists to utilizing quantile regression to examine the impact of our independent variable on our dependent variable MPG.

From our model explaining MPG through wheelbase, the impact of wheelbase on MPG showed a decrease in MPG per unit increase in wheelbase while horsepower will remain constant and the impact of horsepower on MPG also show a decrease in MPG per unit increase in horsepower while wheelbase will remain constant.

The OLS R^2 statistic indicates that the model as fitted explains 51.24% of the variability in MPG Highway. The adjusted R^2 statistic, which is more suitable for comparing models with different numbers of independent variables, is 50.16%. The $R_{0.75}$-squared statistic indicates that the model as fitted explains 35.28% 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 33.84%.

Table 3: Comparison for Multiple OLS and QR Methods

	OLS		Q(0.25)		Q(0.50)		Q(0.75)	
Parameter	Estimate	P-Values	Estimate	P-Values	Estimate	P-Values	Estimate	P-Values
Constant	26.2699	0.0009	31.3873	0.0042	29.8301	0.0030	28.3395	0.0084
Wheelbase	0.3563	0.0012	0.2256	0.1199	0.2678	0.0530	0.3396	0.0287
Horsepower	0.0116	0.2526	-0.0107	0.2657	0.0070	0.5668	-0.0015	0.9190
Weight	-0.0117	0.0000	-0.0095	0.0000	-0.0096	0.0000	-0.0108	0.0000
Important Regression Statistics								
	R^2 = 0.6965		*pseudo* R^2 =	0.4136	*pseudo* R^2 =	0.4422	*pseudo* R^2 =	0.4792
	R^2-Adj = 0.6863		R^2-Adj =	0.3938	R^2-Adj =	0.4234	R^2-Adj =	0.4617

Dependent variable: MPG Highway

From table 3 the results of fitting a MLR and QR model to describe the relationship between MPG Highway and 3 independent variables. The independent variable horsepower, showed insignificance in our OLS regression or in any of our quantile regressions. Since the P-value in the other variables is less than 0.05, that term is statistically significant at the 95% confidence level. Consequently, we probably want to remove horsepower variables from OLS model.

Two of our independent variables Wheelbase and Horsepower showed no significance in our QR model. Interestingly, only weight showed significant impact across the QR.

The OLS R^2 statistic indicates that the model as fitted explains 69.65% of the variability in MPG Highway. The adjusted R^2 statistic, which is more suitable for comparing models with different numbers of independent variables, is 68.63%.

The $R_{0.75}$-squared statistic indicates that the model as fitted explains 47.92% 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 46.17%.

Table 4: Comparison for Multiple OLS and QR Methods

	OLS		Q(0.25)		Q(0.50)		Q(0.75)	
Parameter	Estimate	P-Values	Estimate	P-Values	Estimate	P-Values	Estimate	P-Values
Constant	22.7061	0.0040	33.0546	0.0017	26.24442	0.0252	27.9184	0.0417
Wheelbase	0.4243	0.0002	0.1879	0.1811	0.339352	0.0441	0.3899	0.0417
Horsepower	-0.0051	0.6857	-0.0024	0.8597	-0.011886	0.4807	-0.0137	0.5169
Weight	-0.0102	0.0000	-0.0072	0.0006	-0.007924	0.0003	-0.0088	0.0001
Passengers	-1.1291	0.0341	-0.5714	0.2657	-1.236038	0.0605	-1.7158	0.0070
Important Regression Statistics								
	R^2 = 0.7117		*pseudo* R^2 =	0.4201	*pseudo* R^2 =	0.4590	*pseudo* R^2 =	0.5174
	R^2-Adj = 0.6986		R^2-Adj =	0.3938	R^2-Adj =	0.4344	R^2-Adj =	0.4955

Dependent variable: MPG Highway

From table 4 the results of fitting OLS and QR model to describe the relationship between MPG and 4 independent variables (Wheelbase, Horsepower, Weight and Passengers).
Three of our independent variables Wheelbase, Weight and Passengers showed significance in our OLS regression and probably horsepower should be removed after fitting stepwise regression.
Three of our independent variables Wheelbase, Horsepower and Passengers showed no significance in our QR model. Interestingly, we have one independent variable (weight) that exists to utilizing quantile regression to examine the impact of our independent variable on our dependent variable MPG. The OLS R^2 statistic indicates that the model as fitted explains 71.17% of the variability in MPG Highway. The adjusted R^2 statistic, which is more suitable for comparing models with different numbers of independent variables, is 69.86%.
The $R_{0.75}$-squared statistic indicates that the model as fitted explains 51.74% of the variability in MPG Highway. The adjusted R^2 statistic, which is more suitable for comparing models with different numbers of independent variables, is 49.55%.

3.2. Result on stepwise model

Table 5: Variable Selection Model

Parameter	Estimate	Standard Error	t-Statistic	P-Value
CONSTANT	22.1484	7.5291	2.9417	0.0042
Weight	-0.0107	0.0011	-10.0767	0.0000
Wheelbase	0.4322	0.1064	4.0624	0.0001
Passengers	-0.9975	0.4105	-2.4299	0.0171

Dependent variable: MPG Highway

From Table 5 the output shows the results of fitting a MLR model to describe the relationship between MPG Highway and 4 independent variables. The equation of the fitted model is

$$\text{MPG Highway} = 22.1484 - 0.0107 * \text{Weight} + 0.4322 * \text{Wheelbase} - 0.9975 * \text{Passengers}$$

The R^2 statistic indicates that the model as fitted explains 71.12% of the variability in MPG Highway. The adjusted R^2 statistic, which is more suitable for comparing models with different number of independent variables is 70.14%.

4. Main results

The following table shows the comparison between the quantiles models generated

Table 6: Comparison QR (0.25), QR (0.50) and QR (0.75) Regression Methods

	Q(0.25)		Q(0.50)		Q(0.75)	
Parameter	Estimate	P-Values	Estimate	P-Values	Estimate	P-Values
Constant	47.1907	0.0000	48.5484	0.0000	55.0870	0.0000
Weight	-0.0065	0.0000	-0.0065	0.0000	-0.0078	0.0000
	pseudo R^2 = 0.3851		*pseudo* R^2 = 0.4161		*pseudo* R^2 = 0.4487	
	R^2-Adj = 0.3783		R^2-Adj = 0.4097		R^2-Adj = 0.4487	

Dependent variable: MPG Highway
From table 6 we found the impact of weight across the QR model with each of its coefficient estimates showing significance varying with each quantile. This suggests that a benefit exists to utilizing quantile regression to examine the impact of our independent variable on our dependent variable MPG.

The equations of the fitted QR models are therefore:

$$\text{MPG Highway } (0.25) = 47.1907 - 0.0065 * \text{Weight}$$

$$\text{MPG Highway } (0.50) = 48.5484 - 0.0064 * \text{Weight}$$

$$\text{MPG Highway } (0.75) = 55.0870 - 0.0078 * \text{Weight}$$

The $R_{0.75}$-Squared statistic indicates that the model as fitted explains 44.87% of the variability in MPG Highway. The adjusted $R_{0.75}$-squared statistic, which is more suitable for comparing ssmodels with different numbers of independent variables, is 44.87%. In determining whether the model can be simplified, notice the p-value of 0.0000, belonging to Weight, which is less than 0.05 indicating that weight is statistically significant at 95% confidence level.

5. Conclusion

QR 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. QR overcomes various problems that OLS is confronted with; especially the fact that error terms are not constant across distribution, thereby violating vital assumption of homoscedasticity. Also, by focusing on the mean as a measure of location, information about the tails of a distribution is lost as indicated in the data of miles per gallon in highway driving.

References

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

[2] Cade, B. S. and Noon, B. R., "A gentle introduction to quantile regression for ecologists", Frontiers in Ecology and the Environment, 1(8), (2003), pp: 412-420. http://dx.doi.org/10.1890/1540-9295(2003)001[0412:AGITQR]2.0.CO;23.

[3] Chen, C., An introduction to Quantile Regression and the Quantreg Procedure. Sugi. (2004).

[4] Cizek, P., Quantile Regression, in "XploRe Application Guide", *edited. by* W. Härdle, Z. Hlavka, and S. Klinke, chap. 1, Springer, Berlin, pp: 19 – 48, (2003).

[5] Green, H. M. and Kozek, A.S., Modeling weather data by approximate regression. Quantiles. Anziam. 44:C229-C248, (2003).

[6] Koenker, R., Quantile Regression. New York, NY: Cambridge University Press, (2005). http://dx.doi.org/10.1017/CBO9780511754098.

[7] Koenker, R. and G. Bassett, G. Jr., "Regression quantiles", Econometrica, Vol. 1, (1978), pp: 33-50. http://dx.doi.org/10.2307/1913643.

[8] Lee, "Quantile robust to outlier". Statistics and numerical methods, (2005), pp: 35-57.

[9] Wilby, R. L., Dawson, C. W., and Barrow, E. M., "SDSM - A decision support tool for the assessment of regional climate change impacts". Environmental Modelling and Software, 17(2), (2002), pp: 147-159. http://dx.doi.org/10.1016/S1364-8152(01)00060-3.

[10] Yu, K. and Moyeed, R. A., "Bayesian Quantile Regression", Statistics and Probability Letters, 54, (2001), pp: 437 - 447. http://dx.doi.org/10.1016/S0167-7152(01)00124-9.

[11] Yu, K., Kerm, P.V. and Zhang, J., "Bayesian Quantile Regression: An Application to the wage Distribution in 1990s Britain". *IRISS Working Paper 2004-10, CEPS/INSTEAD, Differdange, G. -D. Luxembourg.* (2004).