Quantile Regression due to Skewness and Outliers

Neda Jalali and Manoochehr Babanezhad

Department of Statistics
Faculty of Sciences
Golestan University
Gorgan, Golestan, Iran
m.babanezhad@gu.ac.ir

Abstract

Regression models explore relationship between a response variable and some explanatory variables based often on conditionally mean function. The choice of mean framework is not always appropriate for two reasons. First, when the distribution of explanatory variable may highly be skewed, and second when sever outliers may be observed in the analysis. In contrast, quantile regression, in special case median regression, remains informative in such situations. In this paper, we briefly define quantile regression. We investigate the efficiency of this method by estimating the effect of age on satisfaction score by median regression.

Keywords: Linear regression, Skewness, Outliers, Quantile regression, Median

1 Introduction

Ordinary regression models explore the relationship between a response variable and some explanatory variables by a conditionally mean function,

\[ Y_i = E(Y_i|X_i) + \epsilon_i \]

(e.g., \( E(Y_i|X_i) = \alpha + \beta X_i \)). If regular assumptions such as, uncorrelated random distributed of error term \( \epsilon_i \), with mean zero and constant variance \( \sigma^2 \)
are satisfied, then the least square estimator \( \hat{\beta} \) for \( \beta \) is the best linear unbiased estimator. In some situation where regular assumptions are not met [1, 2], the conditionally mean functions are then poor to characterize the relationship between \( Y \) and \( X \) [3, 4, 5]. In contrast, quantile regression, in special case median regression, which is the extension of classical regression model, might lead to the best and unbiased estimator. In next section, we briefly introduce quantile regression, and in Section 3, we investigate the efficiency of quantile regression by estimating the effect of age on satisfaction score adjusted by sex, education and number of children through median regression. Section 4 is ended by conclusion.

2 Quantile regression

As stated before, quantile regression is constructed by conditionally quantile given one or more explanatory variables. Following Koenker and Basset [1], linear quantile regression can be modelled as,

\[
Q(\tau | X_i) = \alpha + \beta(\tau)X_i
\]

where \( \hat{\beta}(\tau) \) can be estimated by solving:

\[
\hat{\beta}(\tau) = \sum_{i: Y_i \leq \alpha + \beta X_i} \tau(Y_i - \alpha - \beta X_i) + \sum_{i: Y_i > \alpha + \beta X_i} (1 + \tau)(Y_i - \alpha - \beta X_i)
\]

In special case when \( \tau = 0.5 \) which minimizes absolute deviations is median regression [1, 3]. The main advantage of median regression is its ability to estimate the effect of \( X \) without making assumption about error term. In addition, estimating parameters by least absolute deviations ignore the effect of outliers because it merely considers the sign of them not magnitude. In contrast, ordinary least square consider the magnitude of the deviations and do not control the extension of the outliers [5].

3 Case Study

A research was constructed based on a questioner which includes 20 multiple choice questions. Samples were taken from Gorgan population (a city in north of Iran). Using Cochran formula 406 questionnaires were prepared. Every
question was ranked from 1 to 4 and sum of ranks were introduced as a satisfaction criteria for each participant. The aim of the research was to estimate the effect of age on satisfaction score adjusted covariates including sex (man or woman), education, number of children by quantile (median) regression. The model for quantile (median) regression can be written as,

\[ Q_{0.5} = \beta_0 + \beta_1 \text{age} + \beta_2 \text{sex} + \beta_3 \text{education} + \beta_4 \text{children} \]  

(1)

where \( \beta_1 \) is the effect of age. Table I shows parameters estimates and 95 % confidence intervals. For instance, the analysis shows age has significant effect on life satisfaction score. In addition, the analysis shows a 44 years old woman has less than 37.44 % life satisfaction score. In the contrary, this score for a same age man is less than 37.11 %. A histogram of the standardized residuals from median regression and normal fitted density curve shows residuals has normal distribution (Figure 1) with constant variance (Figure 2). It implied that median regression fits the data well.
4 Discussion

The most of applied statistics may be constructed as linear regression model, and associated estimation method often is ordinary least squares. However, we would be able to check the regular assumptions about data (normality) and residuals. In addition, one may be interested in other position parameter instead of mean [4]. Thus quantile regression is more preferable in such situations. Quantile regression has this capability to analysis the whole distribution whereas ordinary regression merely considers the central distribution [1, 4, 5, 6]. In our example, different covariates influence response variable in different quantiles. Unreported analysis displays that quantile regression results for quantiles 0.2 and 0.9 are also doing better than the ordinary regression.
estimators.

References


Received: December, 2010