The Korean Journal of Applied Statistics (2012) DOI: http://dx.doi.org/10.5351/KJAS.2012.25.1.197
25*(1), 197–203*

Pointwise Estimation of Density of Heteroscedastistic Response in Regression

Jihoon Hyun[1] · Siwon Kim[2] · Sungdong Lee[3] · Wookjae Byun[4] · Mi Kyoung Son[5] · Choongrak Kim[6]

[1]Korea Science Academy of KAIST; [2]Korea Science Academy of KAIST
[3]Korea Science Academy of KAIST; [4]Korea Science Academy of KAIST
[5]Department of Statistics, Pusan National University
[6]Department of Statistics, Pusan National University

(Received November 24, 2011; Revised January 16, 2012; Accepted January 16, 2012)

Abstract

In fitting a regression model, we often encounter data sets which do not follow Gaussian distribution and/or do not have equal variance. In this case estimation of the conditional density of a response variable at a given design point is hardly solved by a standard least squares method. To solve this problem, we propose a simple method to estimate the distribution of the fitted vales under heteroscedasticity using the idea of quantile regression and the histogram techniques. Application of this method to a real data sets is given.

Keywords: Conditional distribution function, heteroscedasticity, histogram, quantile regression.

1. Introduction

Regression analysis is a statistical technique to investigate and model the relationship between variables. In particular, the aim of regression analysis is to construct mathematical models that describe or explain relationships that may exist between variables. Applications of regression are numerous and occur in almost every field such as engineering, the physical and chemical sciences, economics, management, life and biological sciences, and the social sciences. For this reason, regression analysis has been one of the most important statistical technique for many decades.

After fitting a regression model, we are interested quite often in estimating the confidence region for the fitted values at a specific design point. Both in the parametric and the nonparametric regression, Gaussian assumptions are allowed to the error terms. When the homoscedastistic assumptions on the variance of the error terms are doubtful, heteroscedastistic assumptions are given in addition to the Gaussian assumptions. However, it is very hard to reflect the heteroscedastistic assumptions in evaluating confidence regions for the fitted values. In general, this problem shares the same

This research was supported by KSA of KAIST 2011 R&E program and Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (2010-0003699).
[6]Corresponding author: Professor, Department of Statistics, Pusan National University, Jangjeon 2-dong, Geumjeong-gu, Busan 609-735, Korea. E-mail: crkim@pusan.ac.kr

difficulty to estimate the conditional distribution of the response variable given the covariates, and many studies are done in this area. Among them, Copas (1995), Hjort and Jones (1996), Loader (1996), Hall and Presnell (1997), and Hall *et al.* (1999) gave excellent results.

In this article, we propose a simple method to estimate the distribution of the fitted vales under heteroscedasticity using the idea of quantile regression (Koenker and Bassett, 1978). Quantile regression is very robust technique in regression and gives many desirable properties when error terms are far away from the Gaussian assumption. For simplicity, we consider a one-dimensional covariate case. First, we evaluate quantile regression estimate for various values of quantile, and using this result we make a smooth version of density using the histogram technique and the spline.

The rest of the paper is organized as follows. In Section 2, we review the quantile regression. In Section 3 we propose the method to estimate the conditional density function using the quantile regression estimates, and smoothing technique based on the histogram. In Section 4, three dimension graphs for the distribution function based on a real data set using R package for visualization of the conditional distribution function. Concluding remarks are given in Section 5.

2. Quantile Regression

The method of quantile regression estimation, first introduced by Koenker and Bassett (1978), is contrary to the least squares method estimating conditional expectation, and is concerned about estimating quantiles instead of mean. Quantile regression can give a more complete assessment of covariate effects at a properly chosen set of quantiles. See, for example, Koenker and Bassett (1978), Portnoy and Koenker (1997), Yu *et al.* (2003), and Koenker (2005), among others. Because of these theoretical advantages, quantile regression is being used in various fields. Cole and Green (1992) and Haegerty and Pepe (1999) used quantile regression to create reference charts in the medical field and Hendricks and Konenker (1992) and Koenker and Hallock (2001) used quantile regression to build economic models among others.

Consider a simple linear model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

where y_i is a response variable, x_i is a covariate, β_0 and β_1 are unknown regression coefficients, and ϵ_i is a identically and independently distributed error with mean 0 and variance σ^2. The least squares estimation(LSE) of $\boldsymbol{\beta}$ is obtained by minimizing the quadratic loss function $r(u) = u^2/2$, *i.e.*, given $\{x_i, y_i\}$, the LSE is obtained by minimizing

$$\sum_{i=1}^{n} r(y_i - \beta_0 - \beta_1 x_i) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

Therefore, the LSE is concerned with the estimation of the conditional expectation $E(Y|X = x)$. However, median quantile regression estimates the conditional median of Y given $X = x$, and the corresponding loss function is $|u|/2$. The resulting estimator is called the least absolute deviation(LAD) estimator, because it minimizes

$$\sum_{i=1}^{n} r(y_i - \beta_0 - \beta_1 x_i) = \frac{1}{2}\sum_{i=1}^{n}|y_i - \beta_0 - \beta_1 x_i|.$$

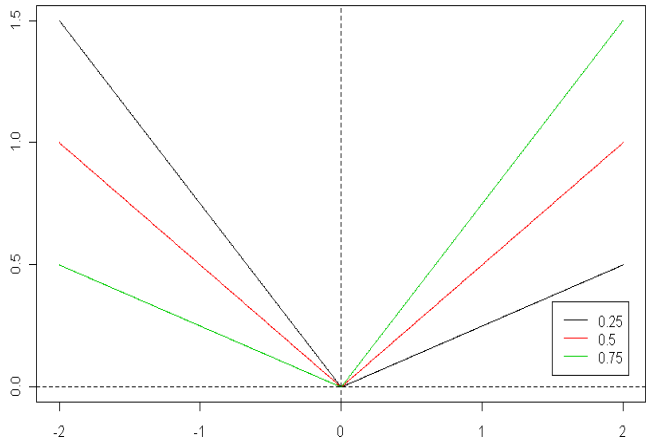

Figure 2.1. Check function ($\tau = 0.25, 0.5, 0.75$)

Note that

$$\begin{aligned}\rho_{0.5}(u) &= 0.5|u| \\ &= 0.5uI_{[0,\infty)}(u) - (1-0.5)uI_{(-\infty,0)}(u),\end{aligned}$$

where $I(\cdot)$ is an indicator function. By replacing 0.5 by τ, $100\tau\%$ quantile regression $q_\tau(x)$ at x can be defined as the value of θ that minimizes

$$E[\rho_\tau(Y-\theta)|X=x].$$

Here,

$$\rho_\tau(u) = \tau uI_{[0,\infty)}(u) - (1-\tau)uI_{(-\infty,0)}(u)$$

is called the check function (see Figure 2.1), and it can also be written as $\rho_\tau(u) = u(\tau - I(u<0))$. According to Figure 2.1, check function has a weight of τ in the case of a positive value and a weight of $(1-\tau)$ in the case of a negative value. Therefore, the $100\times\tau\%$ estimate of regression coefficients in the linear quantile regression at is given by

$$\hat{\beta}_\tau = \arg\min_{\boldsymbol{\beta}\in R^p} \sum_{i=1}^{n} \rho_\tau\left(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}\right)$$

and the corresponding fitted value is

$$\hat{y}_{\tau,i} = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_\tau.$$

The parametric quantile regression is extended to the nonparametric quantile regression by Yu and Jones (1998) and Hall *et al.* (1999). Yu and Jones (1998) used a double-kernel approach and Hall *et al.* (1999) proposed an adjusted version of Nadaraya-Watson estimator. In the nonparametric quantile regression, the bandwidth selection is important. To solve this problem, Yu and Jones (1998) adopted the idea of asymptotic mean squared error criterion proposed by Ruppert *et al.* (1995).

3. Estimation of Density Function of Response Variable

3.1. Estimation in the simple linear regression

Before we propose estimation technique via a quantile regression, we first consider a prediction problem for Y given X in the simple linear regression case. Let $Y(x)$ be the Y given $X = x$. Then

$$Y(x) \sim N\left(\beta_0 + \beta_1 x, \sigma^2\right).$$

Let $\hat{Y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$, then we it is easy to show that

$$\frac{Y(x) - \hat{Y}(x)}{s\sqrt{1 + \frac{1}{n} + \frac{(x - \bar{X})^2}{S_{xx}}}} \sim t(n-2),$$

where

$$s = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2},$$

$$S_{xx} = \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2.$$

Therefore, $Y(x)$ is represented briefly as follows :

$$Y(x) \xrightarrow{D} N\left(\hat{\beta}_0 + \hat{\beta}_1 x, s\sqrt{1 + \frac{1}{n} + \frac{\left(x - \bar{X}\right)^2}{S_{xx}}}\right).$$

Using this result we can obtain an approximate density of the response variable at x, however, it is available under very restrictive assumptions such as normality with equal variance for all x. Therefore, this approach is not realistic for application to the real data sets.

3.2. Estimation using quantile regression

In real data, the error terms hardly follows Gaussian distribution with equal variance. To meet more realistic situations, we apply nonparametric quantile regression. To do this, we first evaluate quantile estimates at different values of τ. In this paper, we used quantile estimates at $\tau = 0.01, 0.10, 0.25, 0.50, 0.75, 0.90, 0.99$. For a given x, let $\hat{y}_{0.01}, \hat{y}_{0.10}, \hat{y}_{0.25}, \hat{y}_{0.50}, \hat{y}_{0.75}, \hat{y}_{0.90}, \hat{y}_{0.99}$ be fitted values obtained at each τ. Then, we have the following approximate probabilities for the response variable.

$$\begin{aligned}
P(Y \le \hat{y}_{0.01}) &= 0.01,\\
P(\hat{y}_{0.01} \le Y \le \hat{y}_{0.10}) &= 0.09,\\
P(\hat{y}_{0.10} \le Y \le \hat{y}_{0.25}) &= 0.15,\\
P(\hat{y}_{0.25} \le Y \le \hat{y}_{0.50}) &= 0.25,\\
P(\hat{y}_{0.50} \le Y \le \hat{y}_{0.75}) &= 0.25,\\
P(\hat{y}_{0.75} \le Y \le \hat{y}_{0.90}) &= 0.15,\\
P(\hat{y}_{0.90} \le Y \le \hat{y}_{0.99}) &= 0.09,\\
P(Y \ge \hat{y}_{0.99}) &= 0.01.
\end{aligned}$$

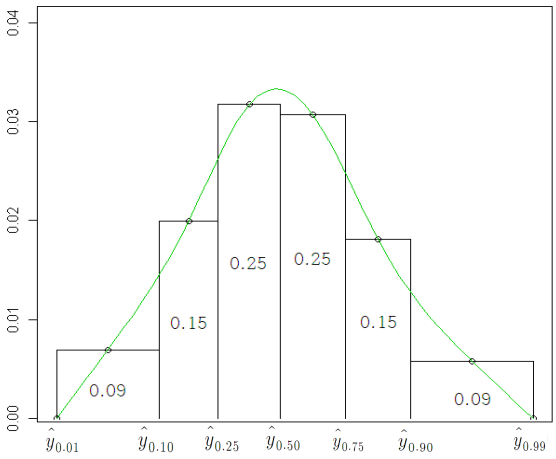

Figure 3.1. Estimation of conditional distribution function (cubic interpolation of histogram based on quantile estimates)

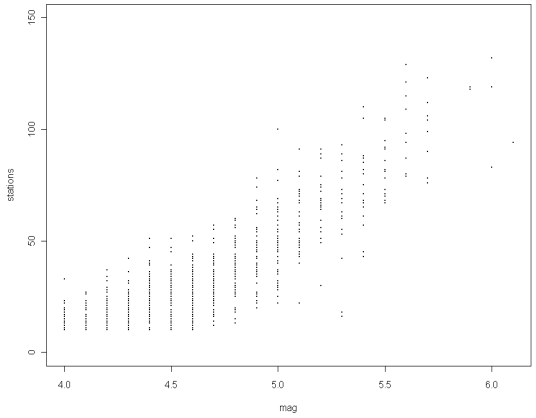

Figure 4.1. Scatter plot for earthquakes data

To obtain the conditional distribution function using seven quantile estimators, we apply the histogram method. We calculate the heights that satisfy the above equations and associates each point smoothly using cubic interpolation. These methods can be easily understood from Figure 3.1. We propose the conditional distribution function as a smoothing curve obtained by this technique.

4. Example

As an illustrative example, we use the Locations of Earthquakes in Fiji Island data set in R package (available by typing "quakes{datasets}"; see Figure 4.1 for the scatter plot). We see that dispersion of response values increases as the covariate increases. We fit the data using a local linear quantile estimation at different values of τ given in Figure 4.2. The bandwidth was chosen by the method of Yu and Jones (1998). Finally, a three-dimensional plot for the density function of the response is given in Figure 4.3.

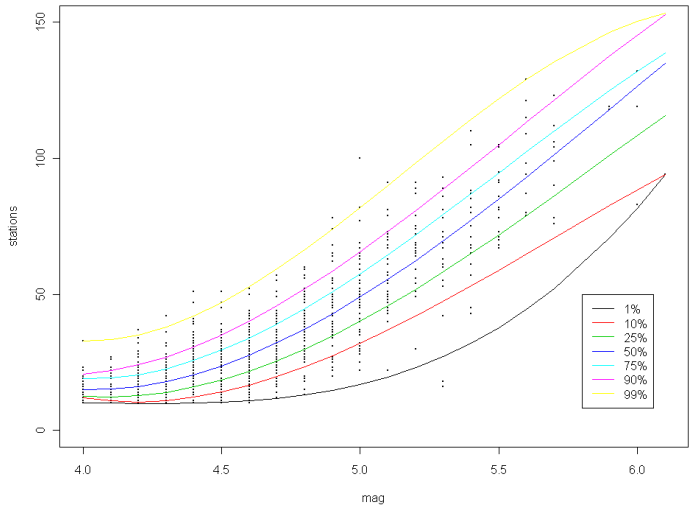

Figure 4.2. Nonparametric quantile regression fit in earthquakes data

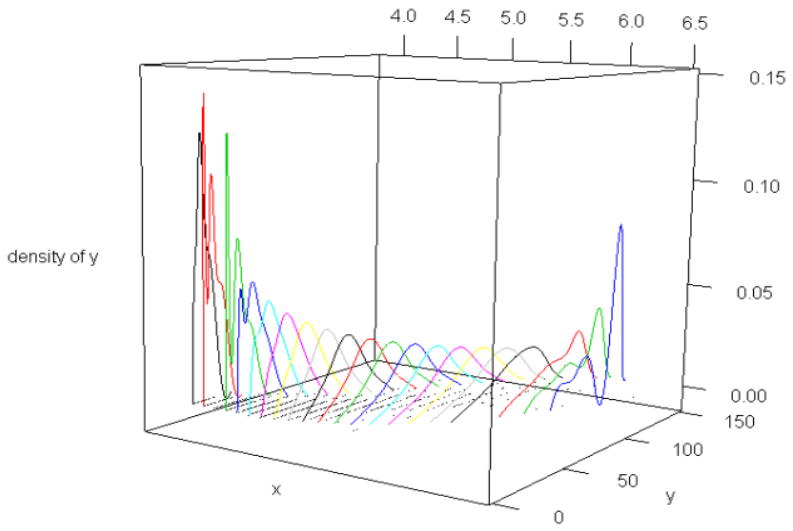

Figure 4.3. 3D plot for densities of response varuable using R package 'rgl' in earthquakes data

5. Concluding Remarks

In estimating confidence region for the fitted values at a specific design point in regression, Gaussian assumptions are allowed to the error terms both in the parametric and the nonparametric regression. When the homoscedastistic assumptions on the variance of the error terms are doubtful, heteroscedastistic assumptions are given in addition to the Gaussian assumptions.

In this paper, we proposed a simple method to estimate the distribution of the fitted vales under heteroscedasticity using the idea of quantile regression. To do this, we evaluated nonparametric

quantile regression estimate for various values of quantile, and using this result we made a smooth version of density using the histogram technique and the spline. For easier visualization, a 3D plot for densities at design points of interest was given. This method would be very useful in recognizing the quantile fits and the corresponding densities of response variable.

References

Cole, T. J. and Green, P. J. (1992). Smoothing reference centile curves: The LMS method and penalized likelihood, *Statistics in Medicine*, **11**, 1305–1319.

Copas, J. B. (1995). Local likelihood based on kernel censoring, *Journal of the Royal Statistical Society, Series B*, **57**, 221–235.

Hall, P. and Presnell, B. (1997). *Intentionally Biased Bootstrap Methods*, unpublished manuscript.

Hall, P., Wolff, R. C. and Yao, Q. (1999). Methods for estimating a conditional distribution, *Journal of the American Statistical Association*, **94**, 154–163.

Heagerty, P. J. and Pepe, M. S. (1999). Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in US children, *Journal of the Royal Statistical Society (Applied Statistics)*, **48**, 533–551.

Hendricks, W. and Koenker, R. (1992). Hierarchical spline models for conditional quantiles and the demand for electricity, *Journal of the American Statistical Association*, **87**, 58–68.

Hjort, N. L. and Jones, M. C. (1996). Locally parametric nonparametric density estimation, *Annals of Statistics*, **24**, 1619–1647.

Koenker, R. (2005). *Quantile Regression*, Cambridge, U.K., Cambridge University Press.

Koenker, R. and Bassett, G. (1978). Asymptotic theory of least absolute error regression, *Journal of the American Statistical Association*, **73**, 618–622.

Koenker, R. and Hallock, K. F. (2001). Quantile regression, *The Journal of Economic Perspectives*, **15**, 143–156.

Loader, C. R. (1996). Local likelihood density estimation, *Annals of Statistics*, **24**, 1602–1618.

Portnoy, S. and Koenker, R. (1997). The Gaussian hare and the Laplacian tortoise: Computability of squared-error versus absolute-error estimators (with discussion), *Statistical Science*, **12**, 279–300.

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Ruppert, D., Sheather, S. J. and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression, *Journal of the American Statistical Association*, **90**, 1257–1270.

Yu, K. and Jones, M. C. (1998). Local linear regression quantile estimation, *Journal of the American Statistical Association*, **93**, 228–238.

Yu, K., Lu, Z. and Stander, J. (2003). Quantile regression: Applications and current research areas, *Journal of the Royal Statistical Society*, **52**, 331–350.