Advances and Applications in Statistics
© 2015 Pushpa Publishing House, Allahabad, India
Published Online: April 2015
http://dx.doi.org/10.17654/ADASFeb2015_133_143
Volume 44, Number 2, 2015, Pages 133-143 **ISSN: 0972-3617**

A PARAMETRIC APPROACH FOR ESTIMATING CONDITIONAL PROBABILITY DISTRIBUTIONS

T. K. Mak and F. Nebebe

Department of Supply Chain and B.T.M.
JMSB, Concordia University
1455 De Maisonneuve West
Montreal, Quebec, H3G 1M8, Canada
e-mail: tak.mak@concordia.ca
Fassil.Nebebe@concordia.ca

Abstract

We consider in this paper a fully parametric approach that directly models the conditional distribution of a response variable as a function of the explanatory variables using a class of probability distributions recently proposed in Mak and Nebebe [4]. A fully parametric approach, unlike traditional quantile regression which fits a quantile regression function for each quantile, yields automatically an estimate of the quantile regression function for any quantile once the parameters in the model have been estimated. Furthermore, statistical inferences on how the conditional distribution varies with the values of the explanatory variables are considerably easier with the proposed parametric modeling than quantile regression. We also studied the relationship between quantile regression and the present approach under linear modeling. A real example using a real estate data set is used to illustrate the proposed methodology.

Received: November 11, 2014; Accepted: January 31, 2015

2010 Mathematics Subject Classification: Primary 62F10; Secondary 62E17.

Keywords and phrases: conditional distributions, non-normal probability distributions, quantile regression.

Communicated by K. K. Azad

1. Introduction

There is little doubt that multiple regression is one of the most widely used statistical tools in analyzing data from a wide variety of areas, such as social sciences, business intelligence, engineering environmental and health sciences. It is very effective in assisting researcher understand the relationship between the response variable and the explanatory variables. This relationship is described in terms of the mean as the central location measure. Such modeling has been extended to the case of scale measure changes, and there has been now a broad literature in heteroscedastic regression models, which has attracted in particular great attention in off-line quality control that aims at choosing settings of the explanatory variables to minimize variability around a targeted value of the quality characteristic.

Despite its popularity, multiple regression models (and its ramification) fail to address many emerging issues in recent research in many areas. The assumption of normality in regression is in many applications unrealistic and the conditional distribution of the response variable for given levels of the explanatory variables may be a skewed one. While multiple regression is effective in studying central location measure such as the mean, it cannot be used to draw inference about non-central location measures, and more generally to describe how the shape of the conditional distribution varies with the explanatory variables. Hao and Naiman [2] discussed the importance in some interesting applications of studying non-central location measures and the shape of the distribution.

To study non-central location measure, Koenker and Bassett [3] introduced quantile regression, modeling a given quantile of interest directly as a function of the explanatory variables. While the computation of the regression parameters, typically involving linear programming for the minimization of a certain distance measure, is more complex than those in multiple regression, it poses little challenge to today’s computing technology. Since the introduction of quantile regression, there has been a rapid growth of its applications in many empirical researches.

The approach of quantile regression fits a regression model for each quantile. It does not model directly the conditional distribution of the response given the values of the explanatory variables. The latter approach was considered briefly in Yu et al. [7] (see also Yu and Zhang [8]) who modeled the error of the regression model using the asymmetric Laplace distribution. Recently, Noufaily and Jones [5] also adopted a fully parametric approach, modeling directly the conditional distribution of the response variable as a generalized gamma distribution with the parameters in the log generalized gamma distribution modeled parametrically as either linear or log linear function of a single explanatory variable. These fully parametric approaches can be useful as pointed out by Noufaily and Jones [5] as, unlike individual quantile regression, it is not necessary to conduct a separate regression analysis for each quantile as any *p*th level quantile regression function can be easily determined once the parameters in the conditional distribution have been estimated. More importantly, statistical inferences concerning various hypotheses about the conditional distributions can be conducted using the familiar maximum likelihood theory and are asymptotically efficient under the assumed model. With the "one regression for each quantile" approach, statistical inferences are considerably less straightforward. While their classes of conditional distributions used in their fully parametric approaches provide some flexibility, there is no theoretical justification and assurance of their validities in any given application. In fact, the asymmetric Laplace distributions have a cusp at its mode and both the Laplace and generalized gamma for modeling positive response variable do not include the normal distribution as a special case (and therefore cannot be used straightforwardly to test normality). Mak and Nebebe [4] proposed a very flexible class of probability distributions and demonstrated its usefulness in a range of statistical problems. Since the class is shown to provide approximation to the distribution of any continuous variable with a continuously differentiable density function, it therefore provides some theoretical justification for its application in the modeling of the conditional distribution of the response variable. This is illustrated with an example in Section 4. In Section 2, we discussed the modeling of the conditional distribution of the response variable with this new class of distributions. We then established the relationship between quantile regression and the proposed conditional distribution model under linear parametric modeling. In Section 3, we suggested two methods of estimating the parameters in the conditional distribution model.

2. Quantile Regression and Conditional Distribution Modeling

2.1. Quantile regression and conditional distribution modeling

Let Y be the response variable and X be a vector of q explanatory variables. In quantile regression, it is assumed that the conditional pth level quantile $y_{p|x}$ of Y given $X = x$, is a function $y_p(x, \alpha_p)$ of x involving a vector of unknown parameter α_p that depends on p. If the purpose is just to estimate the quantile regression function $y_p(x, \alpha_p)$ for one or several values of p, it is neither necessary to specify the distributional form of the conditional distribution of Y given X, nor to model α_p since a separate quantile regression can be conducted for each p to estimate the α_p. Specifically, suppose that a sample of n independent observations (X_i, Y_i), $i = 1, ..., n$, is obtained. Then for any given p, a distribution free estimate of α_p can be obtained by minimizing with respect to a,

$$\sum \rho(Y_i - y_p(X_i, a)), \tag{1}$$

where $\rho(z) = z(p - I(z < 0))$ and I denotes the indicator function. This procedure thus yields an estimator of the quantile regression function via the minimization of (1) for each p. It, however, provides only estimates for selected p but not explicitly the conditional distribution of Y given X, and does not take into consideration the possible form of the functional dependence of α_p on p. Provided that the conditional distribution of Y given X can be approximately specified parametrically, a potentially better approach is to model the parameters of the conditional distribution as a function of x and use asymptotically optimal method of estimation such as the maximum likelihood to estimate the entire conditional distribution. Suppose that the conditional distribution of Y given X is a parametric distribution involving a vector parameter β. Denote by $Q_{Y|x}(p; \beta)$ the conditional quantile function (and therefore a function of p) of the conditional distribution of Y given $X = x$. Here β varies with x and is therefore a function $\beta(x, \gamma)$ of x involving an unknown vector parameter γ. Under the present parametric modeling, $y_{p|x} = Q_{Y|x}(p; \beta(x, \gamma))$ so that $y_{p|x}$ can be estimated for any x and p once γ has been estimated. In contrast, with the quantile regression approach, it is necessary to fit and estimate the regression function $y_p(x, \alpha_p)$ for each p.

Mak and Nebebe [4] introduced a class of probability distributions defined through their quantile functions. They showed that for the case of a continuous distribution with support $(-\infty, \infty)$, its quantile function can always be approximated by $p_k(N(p))$, where $p_k(u) = \beta_0 + \beta_1 u + \cdots + \beta_k u^k$ is a polynomial of degree k, and $N(p) = \Phi^{-1}(p)$ is the quantile function of the standard normal distribution. We consider in this paper the model

$$Q_{Y|x}(p; \beta) = p_k(N(p)) \tag{2}$$

so that $\beta = (\beta_0, \beta_1, ..., \beta_k)'$ which is a function $\beta(x, \gamma)$ of x with an unknown parameter γ. We consider in the next section the special case of linear modeling.

2.2. Linear models

Theorem 1 below asserts that, under model (2), β is linear in x if and only if the quantile regression function is also linear in x.

Theorem 1. *If the conditional distribution of Y given X is specified by the conditional quantile function* (2), *then* $\beta(x, \gamma) = \gamma x$ *if and only if* $y_{p|x} = x'\alpha_p$ *for any given p.*

Proof. We first prove the sufficient condition. For any integer $r \geq k + 1$, consider any r values $0 < p_j < 1,\ j = 1, \ldots, r.$ Let $Y_p = (y_{p_1 | x}, \ldots, y_{p_r | x})'$. We then have

$$Y_p = Gx,$$

where the ith row of the matrix G is the transpose of α_{p_i}. Since the conditional quantile function is given by (2), we have also

$$Y_p = M\beta,$$

where

$$M = \begin{pmatrix} 1 \ N(p_1) \cdots N^k(p_1) \\ \vdots \\ 1 \ N(p_r) \cdots N^k(p_r) \end{pmatrix}.$$

Consequently,

$$Gx = M\beta \tag{3}$$

so that $\beta = (M'M)^{-1}M'Gx = \gamma x,$ where $\gamma = (M'M)^{-1}M'G.$

The necessary condition is also true since

$$y_{p|x} = Q_{Y|x}(p; \beta(x, \gamma)) = p_k(N(p)) = \beta' m = x'\gamma' m = x'\alpha_p,$$

where $m = (1, N(p), \ldots, N^k(p))'$ and $\alpha_p = \gamma' m.$

3. Estimation of Conditional Distributions

3.1. Estimation via quantile regression

Theorem 1 establishes the connection between quantile regression and conditional distribution modeling. Furthermore, from (3), we have $Gx = M\gamma x$ for all x so that

$$M\gamma = G.$$

Thus an estimate of γ can be obtained by first estimating G. Now each row of G is the transpose of α_p which can be estimated using quantile regression (Chen [1]). Consequently, an estimator $\hat{G}$ of the entire matrix G can be obtained from quantile regression. Let $\varepsilon = vec(\hat{G}) - vec(G)$. Then

$$vec(\hat{G}) = vec(G) + \varepsilon = (I \otimes M)vec(\gamma) + \varepsilon. \tag{4}$$

Let V be the covariance matrix of $vec(\hat{G})$ or equivalently of ε. Note that (4) is of the form of a linear regression model with design matrix $\Lambda = I \otimes M$ and a general covariance matrix V for the error term ε. Consequently, $vec(\gamma)$ can be estimated by

$$(\Lambda' V^{-1} \Lambda)^{-1} \Lambda' V^{-1} vec(\hat{G}).$$

In computing the estimate, the matrix V can be replaced by an estimate obtained based on asymptotic results or the bootstrap. Such estimates are generally available in statistical software.

3.2. Maximum likelihood estimation

It can be easily shown (Mak and Nebebe [4]) that the p.d.f. of the conditional distribution of Y given $X = x$ is given by

$$f(y \mid x, \gamma) = \frac{\psi(N(F(y \mid x, \beta)))}{p_k'(N(F(y \mid x, \beta)))}, \tag{5}$$

where ψ is the p.d.f. of the standard normal distribution and $F(y \mid x, \gamma)$ is the conditional cumulative distribution function which is an inverse function of (2) with of course $\beta = \gamma x$. Thus the asymptotically optimal maximum likelihood estimator can be obtained from maximizing the log likelihood

$$\sum\{\ln(\psi(N(F(Y_i \mid X_i, \beta)))) - \ln(p_k'(N(F(Y_i \mid X_i, \beta))))\}. \tag{6}$$

Note that the function $F(y \mid x, \beta)$ can only be obtained numerically by inverting the function in (2). The conditional quantile function as defined by (2) provides a close approximation to the true conditional quantile function for p inside an interval $(\varepsilon, 1-\varepsilon)$ for a small ε (which can be made arbitrary small by increasing the order of the polynomial). This is reflected in the fact that the conditional density in (5) may be negative for some extreme values of Y for certain polynomials. Thus it is necessary to modify the likelihood function for those values using a certain kind of penalized function (Mak and Nebebe [4]).

4. Example

Shmulei et al. [6] considered data on 506 housing tracts in the Boston areas. The response variable is the median value of owner-occupied homes (in thousands). As an illustrative example, we consider the predictor LSTAT, the percentage of lower status of the population. Our preliminary analysis shows that median home values has a nonlinear relationship with LSTAT, but is approximately linear in $x = 1/\sqrt{LSTAT}$. Furthermore, the median home values seem to have been truncated at 50, and we used only observations with $y < 50$ in our analysis (the sample size reduces to 491). For $k = 3$, we have:

$$Q_{Y|x}(p, \beta) = \beta_0 + \beta_1 N(p) + \beta_2\{N(p)\}^2 + \beta_3\{N(p)\}^3.$$

Here we consider linear modeling so that $\beta_i = \gamma_{i0} + \gamma_{i1}x$ for $i = 0, 1, 2, 3$. Note that for the case of the median, we have $Q_{Y|x}(0.5, \beta) = \beta_0 = \gamma_{00} + \gamma_{01}x$, which is therefore the median regression function. Some important special cases deserve some attentions. Consider the following hierarchical models:

1. When β_2 and β_3 are identically zero (or equivalently, $\gamma_{21} = \gamma_{31} = 0$) the model reduces to the traditional regression model with normally distributed errors and a heterogeneous variance function.

2. For location-scale models, we have $(y - \beta_0)/\beta_1$ is identically distributed. Consequently, β_2 and β_3 are of the forms: $\beta_2 = \gamma_2\beta_1$ and $\beta_3 = \gamma_3\beta_1$.

3. All the γ_{ij} are free parameters and the conditional distributions may have different shapes.

Maximum likelihood estimates of the unknown parameters in Models 1 to 3 above can be obtained by maximizing (6) with respect to the unknown parameters in each model. Furthermore, the standard likelihood ratio test can be applied to test a sub-model against a larger model.

Table 1 gives the estimates of the unknown parameters for Models 1 to 3 above, as well as the values of the likelihood test statistic for testing Model 1 against Model 2, and Model 2 against Model 3. It is seen that Model 1 is rejected, and we conclude that the conditional distribution is non-normal. However, Model 2 cannot be rejected and there is no strong evidence against the location-scale distribution. In fact additional calculations show that Models 2 and 3 yield very similar conditional distributions. Based on Model 2, the quantile regression functions $Q_{Y\,|\,X}(p, \beta)$, as a function of *LSTAT* is plotted in Figure 1 for $p = 0.01, 0.25, 0.5, 0.75$ and 0.99. The p.d.f. of the location-scale distribution of Model 2 is drawn in Figure 2, which appears to be right-skewed.

Table 1. Estimates and log likelihoods of Models 1 to 3

Model parameter estimates		$-2\log(\lambda)$	Change in $-2\log(\lambda)$
Model 1	$\beta_0(x) = -0.386 + 70.171x$ $\beta_1(x) = 2.118 + 7.218x$	540.853	
Model 2	$\beta_0(x) = -0.119 + 67.70x$ $\beta_1(x) = 1.921 + 6.07x$ $\beta_2 = 0.09531\beta_1$ $\beta_3 = 0.03183\beta_1$	526.691	14.162
Model 3	$\beta_0(x) = 0.055 + 67.362x$ $\beta_1(x) = 1.609 + 8.288x$ $\beta_2(x) = -0.641 + 3.861x$ $\beta_3(x) = -0.046 + 0.673x$	526.473	0.218

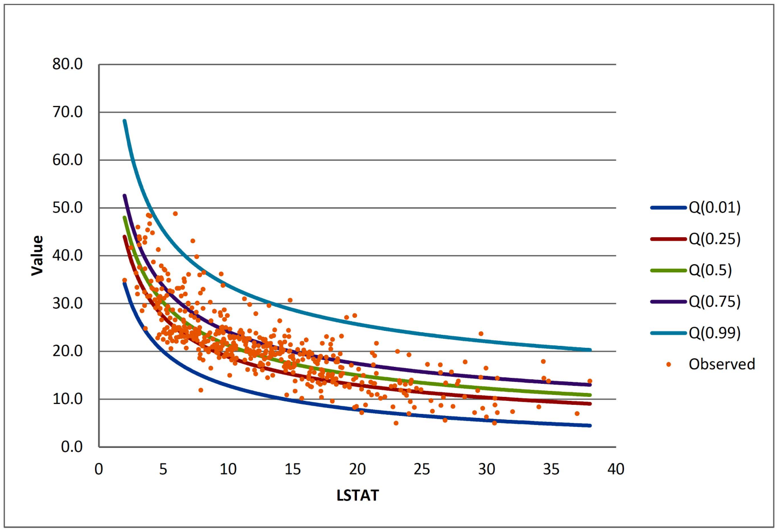

Figure 1. Quantile functions.

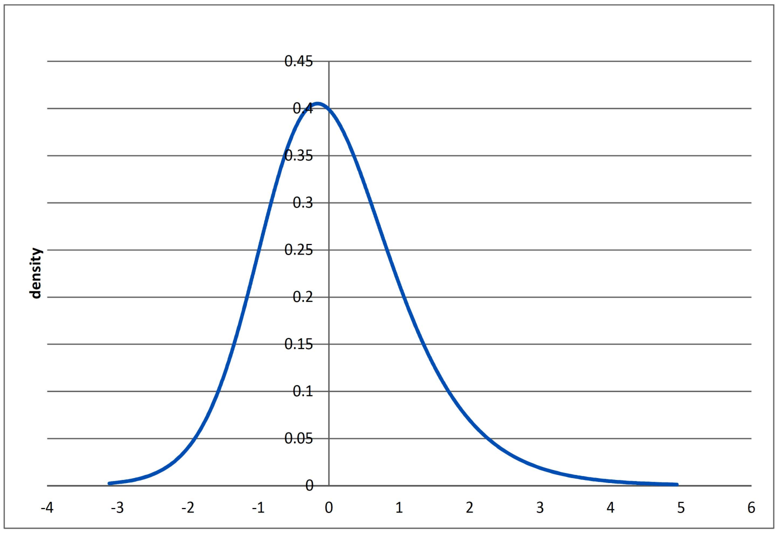

Figure 2. Density of location-scale distribution.

5. Conclusions

In this paper, the conditional distribution of the response variable is modeled directly as a function of the explanatory variables. Unlike quantile regression, once the unknown parameters in the model have been estimated, any quantile regression function can be obtained without the need for the estimation of additional regression models. Such properties are useful if the objective is to study the entire conditional distribution. However, this advantage is made possible at the expense of a somewhat larger collective model. For the case of linear modeling, the number of unknown parameters is $(q+1)(k+1)$, where q is the number of explanatory variables. For estimating a single quantile regression function using quantile regression, the number of unknown parameters is $q+1$. Thus if the purpose is just to obtain a few quantile regression function estimates, quantile regression is a more direct and preferred approach. For the case where the number of data points is sufficiently large, the proposed modeling is capable of giving a more comprehensive analysis of the conditional distribution. Furthermore, statistical inferences are considerably simpler using a fully parametric approach discussed in the present paper, using standard asymptotic theory of general maximum likelihood estimation.

References

[1] C. Chen, An introduction to quantile regression and the QUANTREG procedure, SAS Institute, paper 213-30.

[2] L. Hao and D. Naiman, Quantile regression, Quantitative applications in the social sciences series 149 (2007).

[3] R. Koenker and G. Bassett, Regression quantiles, Econometrica 46 (1978), 33-50.

[4] T. K. Mak and F. Nebebe, On a general class of probability distributions and its applications, Comput. Statist. 28 (2013), 2211-2230.

[5] A. Noufaily and M. C. Jones, Parametric quantile regression based on the generalized gamma distribution, J. R. Stat. Soc. Ser. C. Appl. Stat. 62 (2013), 723-740.

[6] G. Shmulei, N. Patel and P. Bruce, Data Mining for Business Intelligence, John Wiley and Sons, 2007.

[7] K. Yu, Z. Lu and J. Stander, Quantile regression: application and current research areas, The Statistician 52 (2003), 331-350.

[8] K. Yu and J. Zhang, A three-parameter asymmetric Laplace distribution and its extension, Comm. Statist. Theory Methods 34 (2005), 1867-1879.