Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 12, Number 1 (2016), pp. 995-1002

http://www.ripublication.com

Truncated Regression Model and Nonparametric Estimation for Gifted and Talented Education Program

Ferry Safriandi[1]
DPK Al-Azhar University, Kopertis Area 1, Medan, Indonesia.

Liza Setyaning Pertiwi[2]
Polytechnic State Finances, Eka Warni Road, Medan, Indonesia.

Aida Fitriani[3]
University of North Sumatra, Department of Mathematics, Bioteknologi Road, Medan, Indonesia.

Zuhri [4]
STIM Sukma Medan, Kopertis Area 1, Medan, Indonesia.

Abstract

In this paper we consider identification and estimation of a nonparametric location scale model. We first use the truncated data. Then we use truncated regression model. Truncated regression is used to model dependent variables for which some of the observations are not included in the analysis because of the value of the dependent variable. In the latter case we propose a simple estimation procedure based on combining conditional quantile estimators for three distinct quantiles. The new estimator is shown to converge at the optimal nonparametric rate with a limiting normal distribution. A small scale simulation study indicates that the proposed estimation procedure performs well in finite samples. We also present an empirical application on GATE Program using example data test.

Keywords: GATE Program, Language Score, Truncated Regression.

Introduction

The nonparametric location-scale model is usually of the form:

$$y_i = \mu(x_i) + \sigma(x_i)\epsilon_i$$

where x_i is an observed d−dimensional random vector and ϵ_i is an unobserved random variable, distributed independently of x_i, and assumed to be centered around zero in some sense. The functions $\mu(\cdot)$ and $\sigma(\cdot)$ are unknown. In this paper, we consider extending the nonparametric location-scale model to accommodate censored data. The advantage of our nonparametric approach here is that economic theory rarely provides any guidance on functional forms in relationships between variables.

To allow for censoring, we work within the latent dependent variable framework, as is typically done for parametric and semiparametric models. We thus consider a model of the form:

$$y_i^* = \mu(x_i) + \sigma(x_i)\epsilon_i$$
$$y_i = \max(y_i^*, 0)$$

where y_i^* is a latent dependent variable, which is only observed if it exceeds the fixed censoring point, which we assume without loss of generality is 0. We consider identification and estimation of μ(xi) after imposing the location restriction that the median of $\epsilon_i = 0$. We emphasize that our results allow for identification of μ(xi) on the entire support of xi. This is in contrast to identifying and estimating μ(xi) only in the region where it exceeds the censoring point, which could be easily done by extending Powell's(1984) CLAD estimator to a nonparametric setting. One situation is when the data set is heavily censored. In this case, μ(xi) will be less than the censoring point for a large portion of the support of xi, requiring estimation at these points necessary to draw meaningful inference regarding its shape.

Our approach is based on a structural relationship between the conditional median and upper quantiles which holds for observations where μ(xi)≥0. This relationship can be used to motivate an estimator for μ(xi) in the region where it is negative. Our results are thus based on the condition

$$P_X(x_i: \mu(x_i) \geq 0) > 0$$

where PX(·) denotes the probability measure of the random variable x_i.

The paper is organized as follows. The next section explains the key identification condition, and motivates a way to estimate the function μ(·) at each point in the support of x_i. Section 3 introduces the new estimation procedure and establishes the asymptotic properties of this estimator when the identification condition is satisfied. Section 4 considers an extension of the estimation procedure to estimate the distribution of the disturbance term. Section 5 explores the finite sample properties of the estimator through the results of a simulation study. Section 6 presents an empirical application STIFIN test, in which we estimate the survivor function in the region beyond the censoring point. Section 7 concludes by summarizing results.

Censored and Truncated Data: Comparison Definitions

- Y is censored when observe X for all observations, but we only know the true value of Y for a restricted range of observations. Values of Y in a certain range are reported as a single value or there is significant clustering around a value, say 0.
- if y=k or Y>k for all Y =>Y is censored from below or left censored
- if y=k or Y<k for all Y =>Y is censored from above or right censored

We usually think of an uncensored Y, Y^*, the true value oh Y when the censoring mechanism is not applied. We typically have all the observations for {Y,X}, but not {Y^*,X}.

- Y is truncated when we only observe X for observations where Y would not be censored. We do not have a full sample for {Y,X}, we exclude observations based on characteristics of Y.

Estimation Procedure and Asymptotic Properties

Estimation Procedure:

In this section we consider estimation of the function μ(·). Our procedure will be based on our identification results in the previous section, and involves nonparametric quantile regression at different quantiles and different points in the support of the regressors. Our asymptotic arguments are based on the local polynomial estimator for conditional quantile functions introduced in Chaudhuri(1991a,b). For expositional ease, we only describe this nonparametric estimator for a polynomial of degree 0, and refer readers to Chaudhuri(1991a,b), Chaudhuri et al.(1997), Chen and Khan(2000,2001), and Khan(2001) for the additional notation involved for polynomials of arbitrary degree.

First, we assume the regressor vector xi can be partitioned as (xi^{ds},x^c) where the d_{ds}–dimensional vector xi^{ds} is discretely distributed, and the d_c-dimensional vector xi^c is continuously distributed.

We let Cn(x_i) denote the cell of observation x_i and let h_n denote the sequence of bandwidths which govern the size of the cell. For some observation x_j, j ≠ i, we let x_j ϵ $C_n(x_i)$ denote that $x_j^{(ds)}$=$x_i^{(ds)}$ and $x_j^{©}$ lies in the d_c-dimensional cube centered at x_i^c with side length $2h_n$.

Let I[·] be an indicator function, taking the value 1 if its argument is true, and 0 otherwise. Our estimator of the conditional α^{th} quantile function at a point x_i for any α ϵ (0, 1) involves α-quantile regression (see Koenker and Bassett (1978)) on observations which lie in the defined cells of x_i. Specifically, let θ minimize:

$$\sum_{j=1}^{n} I[x_j \in C_n(\,x_i)]\rho_\alpha(y_j-\theta)$$

Where

$$\rho_\alpha(.) \equiv \alpha|.| + (2\alpha-1)(.)I[.<0]$$

Our estimation procedure will be based on a random sample of n observations of the vector $(y_i,x_i^{'})^{'}$ and involves applying the local polynomial estimator at three stages. Throughout our description, ˆ· will denote estimated values.

1) Local Constant Estimation of the Conditional Median Function.

In the first stage, we estimate the conditional median at each point in the sample, using a polynomial of degree 0. We will let h_{1n} denote the bandwidth sequence used in this stage. Following the terminology of Fan(1992), we refer to this as a local constant estimator, and denote the estimated values by ˆq0.5(xi). Recalling that our identification result is based on observations for which the median function is positive, we assigns weights to these estimated values using a weighting function, denoted by w(·). Essentially, w(·) assigns 0 weight to observations in the sample for which the estimated value of the median function is 0, and assigns positive weight for estimated values which are positive.

2) Weighted Average Estimation of the Disturbance Quantiles

In the second stage, the unknown quantiles $c_{\alpha 1}$, $c_{\alpha 2}$ are estimated (up to the scalar constant _c) by a weighted average of local polynomial estimators of the quantile functions for the higher quantiles α1, α2. In this stage, we use a polynomial of degree k, and denote the second stage bandwidth sequence by h_{2n}.

We let ˆc1, ˆc2 denote the estimators of the unknown constants $\frac{c_{\alpha 1}}{\Delta c}, \frac{c_{\alpha 2}}{\Delta c}$ and define them
as:

$$\widehat{c_1} = \frac{\frac{1}{n}\sum_{i=1}^{n} \tau(x_i) w\left(\widehat{q_{0.5}}\,(x_i)\right) \cdot \frac{\left(\hat{q}_{\alpha 1}(x_i) - \hat{q}_{0.5}^{(p)}(x_i)\right)}{\left(\hat{q}_{\alpha 2}(x_i) - \hat{q}_{\alpha 1}(x_i)\right)}}{\frac{1}{n}\sum_{i=1}^{n} \tau(x_i) w\left(\widehat{q_{0.5}}\,(x_i)\right)}$$

$$\widehat{c_2} = \frac{\frac{1}{n}\sum_{i=1}^{n} \tau(x_i) w\left(\widehat{q_{0.5}}\,(x_i)\right) \cdot \frac{(\hat{q}_{\alpha 2}(x_i) - \hat{q}_{0.5}^{(p)}(x_i))}{(\hat{q}_{\alpha 2}(x_i) - \hat{q}_{\alpha 1}(x_i))}}{\frac{1}{n}\sum_{i=1}^{n} \tau(x_i) w(\widehat{q_{0.5}}\,(x_i))}$$

where $\tau(x_i)$ is a trimming function, whose support, denoted by X_τ, is a compact set which lies strictly in the interior of X. The trimming function serves to eliminate“boundary effects” that arise in nonparametric estimation. We use the superscript (p) to distinguish the estimator of the median function in this stage from that in the first stage.

3) Local Polynomial Estimation at the Point of Interest

Letting x denote the point at which the function μ(·) is to be estimated at, we combine the local polynomial estimator, with polynomial order k and bandwidth sequence h_{3n}, of the conditional quantile function at x using quantiles α1, α2, with the estimator of the unknown disturbance quantiles, to yield the estimator of μ(x):

$$\hat{\mu}(x) = \hat{c}_2 \hat{q}_{\alpha 1}(x) - \hat{c}_1 \hat{q}_{\alpha 2}(x)$$

Estimating the Distribution of ϵ_i

As mentioned in Section 2, the distribution of the random variable ϵ_i is identified for all quantiles exceeding $\alpha_0 \equiv \inf\{\alpha: \sup_{x\epsilon X} q_\alpha(x) > 0\}$. In this section we consider estimation of these quantiles, and the asymptotic properties of the estimator. Estimating the distribution of ϵ_i is of interest for two reasons. First, the econometrician may be interested in estimating the entire model, which would require estimators of σ(xi) and the distribution of ϵ_i as well as of μ(xi). Second, the estimator can be used to construct tests of various parametric forms of the distribution of ϵ_i, and the results of these tests could then be used to adopt a (local) likelihood approach to estimating the function μ(xi).

Before proceeding, we note that the distribution of ϵ_i is only identified up to scale, and we impose the scale normalization that c0.75 − c0.25 ≡ 1. We also assume without loss of generality that α0 ≤ 0.25. To estimate c_α for any $\alpha \geq \alpha_0$, we let $\alpha_- = \min(\alpha, 0.5)$ and define our estimator as

$$\widehat{c_\alpha} = \frac{\frac{1}{n}\sum_{i=1}^{n} \tau(x_i) w(\widehat{q_{\alpha-}}(x_i)) \cdot (\hat{q}_\alpha(x_i) - \hat{q}_{0.5}^{(p)}(x_i))}{\frac{1}{n}\sum_{i=1}^{n} \tau(x_i) w(\widehat{q_{\alpha-}}(x_i)) \cdot (\hat{q}_{0.75}(x_i) - \hat{q}_{0.25}(x_i))}$$

The proposed estimator, which involves averaging nonparametric estimators, will converge at the parametric (√n) rate and have a limiting normal distribution, as can be rigorously shown using similar arguments found in Chen and Khan(1999b).

Truncated Regression

- Data truncation is (B-1): the truncation is based on the y-variable.
- We have the following regression satisfies all CLM assumptions:
  $y_i = x_i'\beta + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$
- ➢ we sample only if $y_i < c_i$
- ➢ Observations dropped if $y_i \geq c_i$ by design.
- ➢ We know the exact value of c_i for each person.
- Given the normality assumption for ε_i, ML is easy to apply.

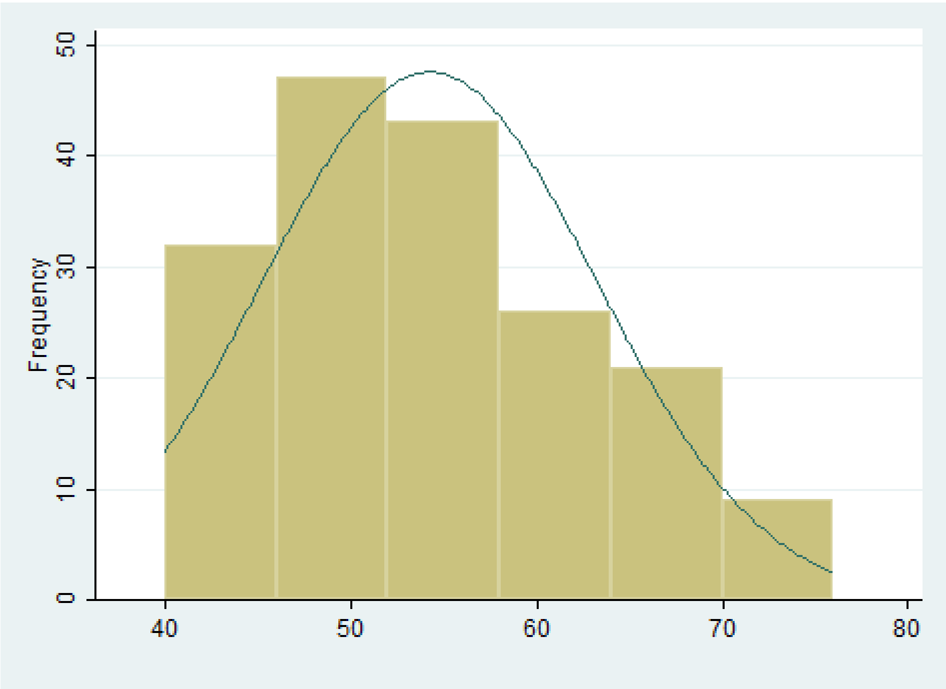
Application to GATE Program

A study of students in a special GATE (gifted and talented education) program wishes to model achievement as a function of language skills and the type of program in which the student is currently enrolled. A major concern is that students are required to have a minimum achievement score of 40 to enter the special program. Thus, the sample is truncated at an achievement score of 40.

Variable	Obs	Mean	Std. Dev.	Min	Max
achiv	178	54.23596	8.96323	41	76
langscore	178	54.01124	8.944896	31	67

Summary for variables: achiv
by categories of: prog (type of program)

prog | N mean sd
---------+-----------------------------
general | 40 51.575 7.97074
academic | 101 56.89109 9.018759
vocation | 37 49.86486 7.276912
---------+-----------------------------
Total | 178 54.23596 8.96323



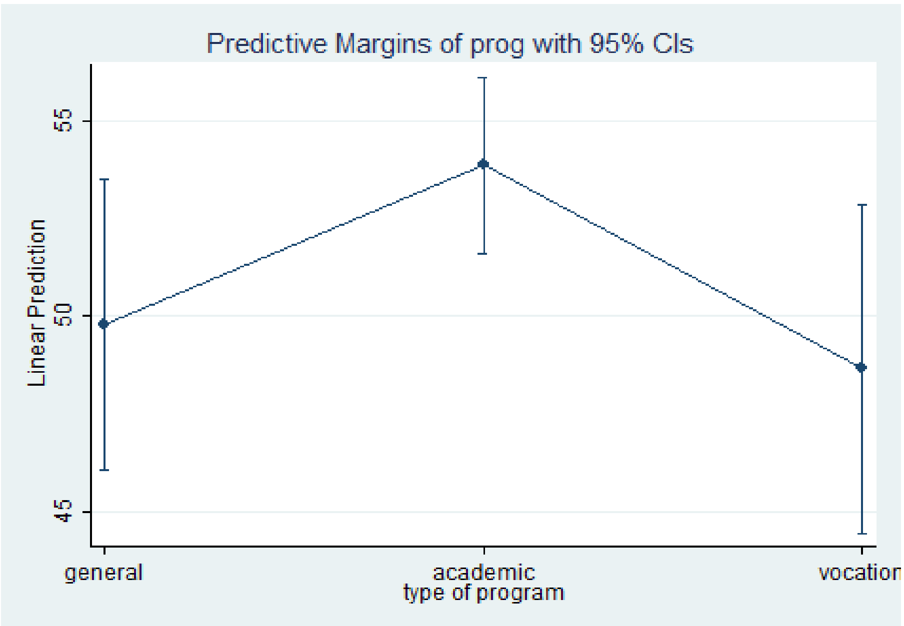
type of |
program | Freq. Percent Cum.
------------+-----------------------------------
general | 40 22.47 22.47
academic | 101 56.74 79.21
vocation | 37 20.79 100.00
------------+-----------------------------------
Total | 178 100.00
Fitting full model:

Iteration 0: log likelihood = -598.11669
Iteration 1: log likelihood = -591.68374
Iteration 2: log likelihood = -591.31208
Iteration 3: log likelihood = -591.30981
Iteration 4: log likelihood = -591.30981

```
Truncated regression
Limit:   lower =        40                  Number of obs =    178
         upper =      +inf                  Wald chi2(3)  =  54.76
Log likelihood = -591.30981                    Prob > chi2   = 0.0000

------------------------------------------------------------------------------
      achiv |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  langscore |   .7125775   .1144719     6.22   0.000     .4882168    .9369383
            |
       prog |
         2  |   4.065219   2.054938     1.98   0.048     .0376131    8.092824
         3  |  -1.135863   2.669961    -0.43   0.671    -6.368891    4.097165
            |
      _cons |   11.30152   6.772731     1.67   0.095     -1.97279    24.57583
-------------+----------------------------------------------------------------
     /sigma |   8.755315    .666803    13.13   0.000     7.448405    10.06222
------------------------------------------------------------------------------
Predictive margins                              Number of obs   =        178
Model VCE    : OIM

Expression   : Linear prediction, predict()

------------------------------------------------------------------------------
            |            Delta-method
            |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       prog |
         1  |   49.78871   1.897166    26.24   0.000     46.07034    53.50709
         2  |   53.85393   1.150041    46.83   0.000     51.59989    56.10797
         3  |   48.65285   2.140489    22.73   0.000     44.45757    52.84813
------------------------------------------------------------------------------
```

In the table above, we can see that the expected mean of **avchiv** for the first level of **prog** is approximately 49.79; the expected mean for level 2 of **prog** is 53.85; the expected mean for the third level of **prog** is 48.65.

Marginsplot



Conclusion

In the output above, we can see that the expected mean of **avchiv** for the first level of **prog** is approximately 49.79; the expected mean for level 2 of **prog** is 53.85; the expected mean for the third level of **prog** is 48.65.

References

[1] Chaudhuri, P. (1991a) "Nonparametric Quantile Regression", Annals of Statistics, 19, 760-777.

[2] Chaudhuri, P. (1991b) "Global Nonparametric Estimation of Conditional Quantiles and their Derivatives", Journal of Multivariate Analysis, 39, 246-269.

[3] Chaudhuri, P., K. Doksum, and A. Samarov (1997) "On Average Derivative Quantile Regression", Annals of Statistics, 25, 715-744.

[4] Chen, S., Dahl. G. B., and S. Khan (2002) "Nonparametric Identification and Estimation of a Censored Regression Model with an Application to Unemployment Insurance Receipt", Center For Labor Economics University of California, Berkeley Working Paper no.54

[5] Chen, S. and S. Khan (2000) "Estimation of Censored Regression Models in the Presence of Nonparametric Multiplicative Heteroskedasticity", Journal of Econometrics, 98, 283-316.

[6] Chen, S. and S. Khan (2001) "Semiparametric Estimation of a Partially Linear Censored Regression Model", Econometric Theory, 17, 567-590.

[7] Fan, J. (1992) "Design-adaptive Nonparametric Regression", Journal of the American Statistical Association, 87, 998-1004.

[8] Fan, J. and I. Gijbels (1996) Local Polynomial Modelling and its Applications, New York: Chapman and Hall.

[9] Khan, S. (2001) "Two Stage Rank Estimation of Quantile Index Models", Journal of Econometrics, 100, 319-355.

[10] Koenker, R. and G.S. Bassett Jr. (1978) "Regression Quantiles", Econometrica, 46, 33-50.

[11] Powell, J.L. (1984) "Least Absolute Deviations Estimation for the Censored Regression Model", Journal of Econometrics, 25, 303-325.

[12] Rosenbaum, P.R. and D.B. Rubin (1983) "The Central Role of the Propensity Score in Observational Studies for Causal Effects", Biometrica, 70, 41-55