International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Comparison between Tobit and Interval Censored Regression Model in STIFIN Test and GPA Prediction

Eka Rusdiana[1], Nilayani[2], Sri Pra Viana Elina[3], Liza Setyaning Pertiwi[4]

Department of Mathematics, University of North Sumatera, Indonesia

Abstract: ***In this paper we consider identification and estimation of a censored nonparametric location scale model. We first show that in the case where the location function is strictly less than the (fixed) censoring point for all values in the support of the explanatory variables, then the location function is not identified anywhere. In contrast, if the location function is greater or equal to the censoring point with positive probability, then the location function is identified on the entire support, including the region where the location function is below the censoring point. 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 STIFIN Test and GPA prediction using example data test. The survival curve for benefit receipt based on our new estimator closely matches the Kaplan-Meier estimate in the non-censored region and is relatively flat past the censoring point. We find that incorrect distributional assumptions can significantly bias the results for estimates past the censoring point.***

Keywords: Censored Regression, Tobit Estimator, Interval Regression.

1. 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:

$$\begin{aligned} y_i^* &= \mu(x_i) + \sigma(x_i)\epsilon_i \\ y_i &= \max(y_i^*, 0) \end{aligned}$$

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.

2. Estimation Procedure and Asymptotic Properties

2.1 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 , $j \neq i$, we let $x_j \in C_n(x_i)$ 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 $\alpha \in (0, 1)$ 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 $\rho_\alpha(\cdot) \equiv \alpha|\cdot| + (2\alpha - 1)(\cdot)I[\cdot < 0]$.

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:

$$\hat{c}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} \tau(x_i) w(\hat{q}_{0.5}(x_i)) \cdot \frac{(\hat{q}_{\alpha_1}(x_i) - \hat{q}_{0.5}^{(p)}(x_i))}{(\hat{q}_{\alpha_2}(x_i) - \hat{q}_{\alpha_1}(x_i))}}{\frac{1}{n}\sum_{i=1}^{n} \tau(x_i) w(\hat{q}_{0.5}(x_i))}$$

$$\hat{c}_2 = \frac{\frac{1}{n}\sum_{i=1}^{n} \tau(x_i) w(\hat{q}_{0.5}(x_i)) \cdot \frac{(\hat{q}_{\alpha_2}(x_i) - \hat{q}_{0.5}^{(p)}(x_i))}{(\hat{q}_{\alpha_2}(x_i) - \hat{q}_{\alpha_1}(x_i))}}{\frac{1}{n}\sum_{i=1}^{n} \tau(x_i) w(\hat{q}_{0.5}(x_i))}$$

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):

$$\hat{\mu}(x) = \hat{c}_2\hat{q}_{\alpha_1}(x) - \hat{c}_1\hat{q}_{\alpha_2}(x)$$

3. Estimating the Distribution of ϵi

As mentioned in Section 2, the distribution of the random variable ϵ_i is identified for all quantiles exceeding $\alpha_0 \equiv \inf\{\alpha: \sup_{x\in X} q_\alpha(x) > 0\}$. 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 $c0.75 - c0.25 \equiv 1$. We also assume without loss of generality that $\alpha 0 \leq 0.25$. To estimate c_α for any $\alpha \geq \alpha_0$, we let $\alpha_- = \min(\alpha, 0.5)$ and define our estimator as

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

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).

4. Monte Carlo Results

In this section the finite sample properties of the proposed estimator are explored by way of a small scale simulation study. We simulated from designs of the form:

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

where xi was a random variable distributed uniformly between -1 and 1, ϵ_i was distributed standard normal, and the scale function σ(xi) was set to $e^{0.15xi}$. We considered four different functional forms for μ(xi) in our study:

1. $\mu(x) = x$
2. $\mu(x) = x^2 - C_1$
3. $\mu(x) = 0.5 \cdot x^3$
4. $\mu(x) = e^x - C_2$

where the constants C1, C2 were chosen so that the censoring level was 50%, as it was for the other two designs. We adopted the following data-driven method to select the quantile pair. For a given point x, we note that the estimator requires that $q_{\alpha 1}(x)$, $q_{\alpha 2}(x)$ both be strictly positive for identification, requiring that the quantiles be sufficiently close to 1. On the other hand, efficiency concerns would suggest that the quantiles not be at the extreme, as the quantile regression estimator becomes imprecise. We thus let the probability of being censored, or the "propensity score" (see Rosenbaum and Rudin(1983)) govern the choice of quantiles for estimating the function μ(·) at the point x. Letting di denote an indicator function which takes the value 1 if an observation is uncensored, we note that

$$1 - E[d_i | x_i = x] = F_\epsilon\left(\frac{-\mu(x)}{\sigma(x)}\right)$$

where $F_\epsilon(\cdot)$ denotes the c.d.f. of ϵ_i. Letting $\alpha^* = F_\epsilon\left(\frac{-\mu(x)}{\sigma(x)}\right)$, we note that

$$\begin{aligned} q_{\alpha^*}(x) &= \max(\mu(x) + c_{\alpha^*}\sigma(x), 0) \\ &= \max(\mu(x) + \frac{-\mu(x)}{\sigma(x)}\sigma(x), 0) \\ &= 0 \end{aligned}$$

Thus if one knew the propensity score value, identification would require that α^* be a lower bound for the choice of quantile pair. The propensity score can be easily estimated using kernel methods, suggesting an estimator of α^*:

$$\hat{\alpha}^* = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n} d_i K_h(x_i^{(c)} - x^{(c)}) I[x_i^{(d)} = x^{(d)}]}{\frac{1}{n}\sum_{i=1}^{n} K_h(x_i^{(c)} - x^{(c)}) I[x_i^{(d)} = x^{(d)}]}$$

where $K_h(\cdot) = h^{-d_c} K(\frac{\cdot}{h})$ where h is a bandwidth sequence, and $K(\cdot)$ is a kernel function.

Our proposed choice of quantile pair takes into account this lower bound as well as the efficiency loss of estimating quantiles at the extreme. We set:

$$\alpha_1 = \frac{2\hat{\alpha}^* + 1}{3} \qquad \alpha_2 = \frac{2 + \hat{\alpha}^*}{3}$$

which divides the interval [ˆα^*, 1] into three equal spaces. In implementing this procedure in the Monte Carlo study, the propensity scores were estimated using a normal kernel function and a bandwidth of $n^{-1/5}$.

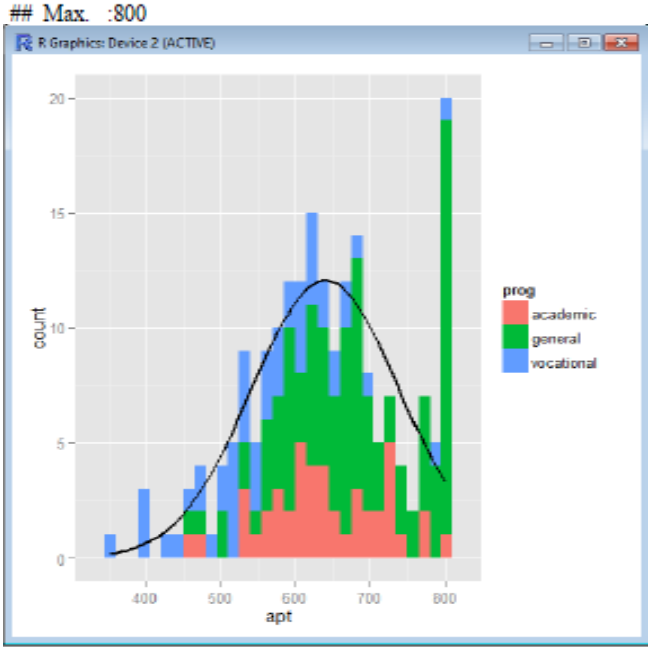
For the quantile estimators, a local constant was fit in the first stage, using a bandwidth of $n^{-1/5}$, and a local linear estimator was used in the second and third stages, using a bandwidth of the form $kn^{-1/5}$. The constant k was selected using the "rule of thumb" approach detailed on page 202 in Fan and Gijbels(1996).

5. Case Studies

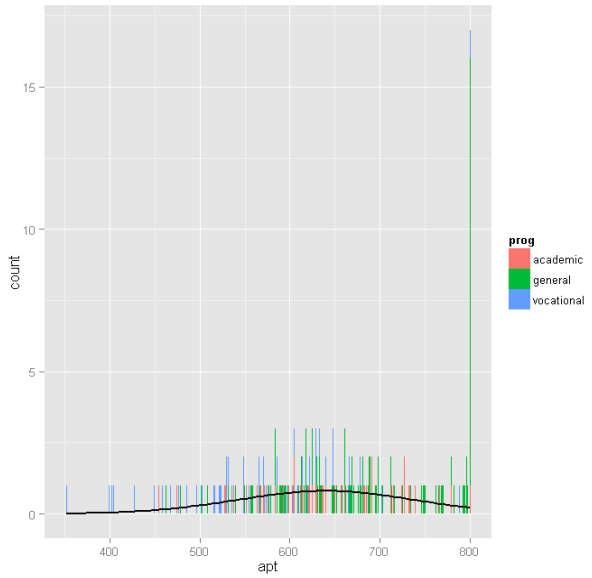
5.1 Application Tobit regression to STIFIN Test

Consider the situation in which we have a measure of academic aptitude (scaled 200-800) which we want to model using reading and math test scores, as well as, the type of program the student is enrolled in (academic, general, or vocational). The problem here is that students who answer all questions on the academic aptitude test correctly receive a score of 800, even though it is likely that these students are not "truly" equal in aptitude. The same is true of students who answer all of the questions incorrectly. All such students would have a score of 200, although they may not all be of equal aptitude.

```
##        id          read            math          prog
## Min.   :  1.0   Min.   :28.0   Min.   :33.0   academic  : 45
## 1st Qu.: 50.8   1st Qu.:44.0   1st Qu.:45.0   general   :105
## Median :100.5   Median :50.0   Median :52.0   vocational: 50
## Mean   :100.5   Mean   :52.2   Mean   :52.6
## 3rd Qu.:150.2   3rd Qu.:60.0   3rd Qu.:59.0
## Max.   :200.0   Max.   :76.0   Max.   :75.0
##       apt
## Min.   :352
## 1st Qu.:576
## Median :633
## Mean   :640
## 3rd Qu.:705
## Max.   :800
```



Looking at the above histogram, we can see the censoring in the values of **apt**, that is, there are far more cases with scores of 750 to 800 than one would expect looking at the rest of the distribution. Below is an alternative histogram that further highlights the excess of cases where **apt**=800. In the histogram below, the **breaks** option produces a histogram where each unique value of **apt** has its own bar (by setting breaks equal to a vector containing values from the minimum of **apt** to the maximum of **apt**). Because **apt** is continuous, most values of **apt** are unique in the dataset, although close to the center of the distribution there are a few values of **apt** that have two or three cases. The spike on the far right of the histogram is the bar for cases where **apt**=800, the height of this bar relative to all the others clearly shows the excess number of cases with this value.



Next we'll explore the bivariate relationships in our dataset.
read math apt
read 1.0000 0.6623 0.6451
math 0.6623 1.0000 0.7333
apt 0.6451 0.7333 1.0000

```
                                LR chi2(4)      =  188.97
                                Prob > chi2     =  0.0000
Log likelihood = -1041.0629     Pseudo R2       =  0.0832
------------------------------------------------------------------------------
         apt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        read |   2.697939   .618798     4.36   0.000     1.477582    3.918296
        math |   5.914485   .7098063    8.33   0.000     4.514647    7.314323
             |
        prog |
          2  |  -12.71476   12.40629   -1.02   0.307    -37.18173    11.7522
          3  |  -46.1439    13.72401   -3.36   0.001    -73.2096    -19.07821
             |
       _cons |    209.566   32.77154    6.39   0.000     144.9359    274.1961
-------------+----------------------------------------------------------------
       /sigma |   65.67672   3.481272                      58.81116    72.54228
------------------------------------------------------------------------------
  Obs. summary:     0  left-censored observations
                  183     uncensored observations
                   17 right-censored observations at apt>=800
```

5.2 Application Interval Regression to GPA Prediction

We wish to predict GPA from teacher ratings of effort, writing test scores and the type of program in which the student was enrolled (vocational, general or academic). The measure of GPA is a self-report response to the following item:
Select the category that best represents your overall GPA.
0.0 to 2.0
2.0 to 2.5
2.5 to 3.0
3.0 to 3.4
3.4 to 3.8
3.8 to 4.0

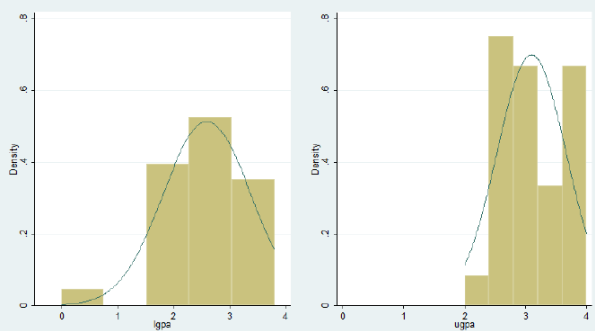
And These are the Dataset:

	lgpa	ugpa
1.	2.5	3
2.	3.4	3.8
3.	2.5	3
4.	0	2
5.	3	3.4
6.	3.4	3.8
7.	3.8	4
8.	2	2.5
9.	3	3.4
10.	3.4	3.8
11.	2	2.5
12.	2	2.5
13.	2	2.5
14.	2.5	3
15.	2.5	3
16.	2.5	3
17.	3.4	3.8
18.	2.5	3
19.	2	2.5
20.	3	3.4
21.	3.4	3.8
22.	3.8	4
23.	2	2.5
24.	3	3.4
25.	3.4	3.8
26.	2	2.5
27.	2	2.5
28.	2	2.5
29.	2.5	3
30.	2.5	3

Note that there are two GPA responses for each observation, **lgpa** for the lower end of the interval and **ugpa** for the upper end.

Variable	Obs	Mean	Std. Dev.	Min	Max
lgpa	30	2.6	.7754865	0	3.8
ugpa	30	3.096667	.5708332	2	4
write	30	113.8333	49.94278	50	205
rating	30	57.53333	8.303441	48	72

type	lgpa	ugpa
vocational	8	8
	1.75	2.4375
	.7071068	.1767767
general	10	10
	2.78	3.24
	.3852849	.3373096
academic	12	12
	3.016667	3.416667
	.6336522	.5474458
Total	30	30
	2.6	3.096667
	.7754865	.5708332




```
Interval regression                             Number of obs   =30
                                                LR chi2(4)      =37.24
Log likelihood = -33.128905                     Prob > chi2     =0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95%    ConInterval]
-------------+----------------------------------------------------------------
       write |   .0052847   .0016921     3.12   0.002     .0019683    .0086012
      rating |   .0133143   .0091197     1.46   0.144    -.0045601    .0311886
             |
        type |
           2 |    .374853     .19275     1.94   0.052      -.00293    .7526359
           3 |   .7097467   .1668399     4.25   0.000     .3827466    1.036747
             |
       _cons |   1.103863   .4452887     2.48   0.013     .2311137    1.976613
-------------+----------------------------------------------------------------
    /lnsigma |  -1.237263   .1596419    -7.75   0.000    -1.550155   -.9243703
-------------+----------------------------------------------------------------
       sigma |   .2901775   .0463245                      .2122151    .3967812
------------------------------------------------------------------------------

  Observation summary:         0  left-censored observations
                               0     uncensored observations
                               0 right-censored observations
                              30       interval observations
```

Finally, a summary of the observations is given. In this dataset, no observations are left- or right-censored, no observations are uncensored, and all 30 observations are interval censored.

6. Conclusion

a) In the output above, the first thing we see is the call, this is R reminding us what the model we ran was, what options we specified, etc.
   - The table labeled coefficients gives the coefficients, their standard errors, and the z-statistic. No p-values are included in the summary table, but we show how to calculate them below. Tobit regression coefficients are interpreted in the similar manner to OLS regression coefficients; however, the linear effect is on the uncensored latent variable, not the observed outcome.
   - For a one unit increase in **read**, there is a **2.6981** point increase in the predicted value of **apt**.
   - A one unit increase in **math** is associated with a **5.9146** unit increase in the predicted value of **apt**.
   - The terms for **prog** have a slightly different interpretation. The predicted value of **apt** is **-46.1419** points lower for students in a vocational program than for students in an academic program.
   - The coefficient labeled "(Intercept):1" is the intercept or constant for the model.
   - The coefficient labeled "(Intercept):2" is an ancillary statistic. If we exponentiate this value, we get a statistic that is analogous to the square root of the residual variance in OLS regression. The value of **65.6773** can compared to the standard deviation of academic aptitude which was 99.21, a substantial reduction.

b) The final log likelihood, -1041.0629, is shown toward the bottom of the output, it can be used in comparisons of nested models.
   - For a one unit increase in **read**, there is a 2.7 point increase in the predicted value of **apt**.
   - A one unit increase in **math** is associated with a 5.91 unit increase in the predicted value of **apt**.
   - The terms for **prog** have a slightly different interpretation. The predicted value of **apt** is 46.14 points lower for students in a vocational program (**prog**=3) than for students in an academic program (**prog**=1).

The tobit model, also called a censored regression model, is designed to estimate linear relationships between variables when there is either left- or right-censoring in the dependent variable (also known as censoring from below and above, respectively). Censoring from above takes place when cases with a value at or above some threshold, all take on the value of that threshold, so that the true value might be equal to the threshold, but it might also be higher. In the case of censoring from below, values those that fall at or below some threshold are censored.

Interval regression is used to model outcomes that have interval censoring. In other words, you know the ordered category into which each observation falls, but you do not know the exact value of the observation. Interval regression is a generalization of censored regression.

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