https://doi.org/10.56093/JISAS.V79I2.5
Available online at http://isas.org.in/isa/jisas
JOURNAL OF THE INDIAN SOCIETY OF
AGRICULTURAL STATISTICS 79(2) 2025 95–104

Censored Regression Models with Endogeneity: A Survey

Swati Shukla, Shalabh and Subhra Sankar Dhar

Indian Institute of Technology, Kanpur

Received: 02 October, 2025; Revised: 02 November, 2025; Accepted: 06 November, 2025

SUMMARY

This survey article provides a comprehensive overview of methodologies developed to handle endogeneity in censored regression models. We emphasize their theoretical basis, and estimation methodologies. The paper summarizes the benefits and drawbacks of each technique and addresses expansions to nonlinear, panel, and high-dimensional settings, offering researchers practical assistance and identifying options for future methodological improvement in censored models with endogenous regressors.

Keywords: Censored Regression model; Endogeneity; Two-Stage Estimation; Instrumental Variables.

About Professor M.N. Das

Professor MN Das has made significant contributions to the field of design and analysis of experiments, particularly in the areas of incomplete block designs and factorial designs. His book on experimental design has been a valuable resource for students in our country, providing one of the first comprehensive introductions to incomplete designs and their variants. I can attest to the book's impact, having benefited from it both as a student and as a teacher. In fact, it inspired me to delve deeper into the subject and write my own book on experimental design. Professor MN Das's legacy will endure, and his work will continue to guide future generations of researchers and students. Professor MN Das worked in the area of block designs, which are related to linear models. In this context, the present paper considers censoring in a linear regression model, which motivates researchers to explore the development of analysis of variance in different experimental designs when observations are censored.

1. INTRODUCTION

Censored regression models constitute an important class of regression models in which the dependent variable is observed only within a certain threshold. This threshold can be either fixed or random, leading to fixed or random censoring, respectively. Fixed-censoring models have been widely studied in both theoretical and applied econometrics. In these models, parametric estimation techniques typically assume a known distribution for the error term. However, un-unlike classical linear regression, misspecification of the error distribution in censored regression models can result in inconsistent estimates of the structural parameters (see, e.g., Amemiya (1979), Amemiya (1984), Heckman (1976), Heckman (1978), Goldberger (1981), Arabmazar and Schmidt (1982)).

To address this issue, semi parametric approaches have been developed, which avoid fully specifying the error distribution while imposing sufficient restrictions to ensure identification of the structural parameters. A common restriction is the independence of the error term from the regressors, which facilitates the construction of consistent estimators (see, e.g., Powell (1984), Powell (1986a), Powell (1986b), Lee (1992)). Despite their theoretical appeal, these estimators have practical limitations. For instance, the CLAD estimator of Powell (1986 a) involves minimizing a non-convex objective function, and commonly used iterative linear programming methods can

Corresponding author: Shalabh
E-mail address: shalab@iitk.ac.in

converge only to local minima rather than the global minimum (see Buchinsky (1994)). To overcome these challenges, several alternative approaches have been proposed. Notably, Buchinsky and Hahn (1998) and Khan and Powell (2001) suggested subsample selection using nonparametric estimates of selection probabilities, though these methods are limited by the curse of dimensionality. Chernozhukov and Hong (2002) introduced a simpler three-step procedure that is easier to implement, but it relies on parametric assumptions for the selection mechanism, which may result in inconsistent estimates if the selection model is misspecified.

Endogeneity presents an additional challenge in censored regression models, often arising due to omitted variables or measurement errors. Endogeneity leads to biased and inconsistent parameter estimates if not properly addressed. Various methods have been proposed to handle endogeneity in censored regression models. Moreover, many empirical applications involve both censoring and endogeneity. For instance, Chay and Powell (2001) studied changes in black–white earnings inequality using longitudinal Social Security earnings records. In their analysis, education is treated as endogenous in the earnings equation, while a substantial portion of earnings records are censored at the Social Security maximum taxable limit, with censoring rates ranging from 32% to 54%.Similarly, in the estimation of Engel curves, Blundell and Powell (2007) and Chernozhukov *et al.* (2015) highlighted the endogeneity of total household expenditure due to the joint determination of total spending and its components. Censoring arises because a significant fraction of households report zero expenditure on alcohol, affecting the estimation of its budget share.

Over the past decades, several methodologies have been developed to address the joint problem of censoring and endogeneity. Classical approaches, such as the two-step control function method, provide consistent estimates under certain conditions. More recently, robust estimation techniques, including M-estimation, L-estimation, and quantile-based approaches, have been proposed to handle heavy-tailed errors, outliers, and model misspecification. Despite these advances, there is a lack of comprehensive surveys summarizing the theoretical developments, practical implementation, and comparative performance of these methods.

This article aims to provide a systematic overview of censored regression models under endogeneity. We discuss the evolution of estimation strategies, highlight their theoretical properties and limitations, and review applications in empirical research. Finally, we outline open challenges and potential directions for future research.

2. MATHEMATICAL FRAMEWORK

Consider a sequence of paired independent and identically distributed (i.i.d.) data points $(y_i, \mathbf{x}_i^*)$, $i = 1, \ldots, n$, where y_i denotes the i^{th} observation on the response variable. The observed variable y_i is related to an unobserved latent variable y_i^* through censoring at a fixed non-random threshold c, such that

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* \geq c \\ c & \text{if } y_i^* \leq c. \end{cases}$$

Thus, y_i is always observed, while the latent outcome y_i^* is only partially observed. The latent variable is modeled as

$$y_i^* = \tilde{\mathbf{x}}_i^\top \boldsymbol{\alpha}_0 + w_i \gamma_0 + \epsilon_i, \tag{2.1}$$

where $\mathbf{x}_i^* = (\tilde{\mathbf{x}}_i, w_i)^\top$ denotes the $(p+1)$-dimensional vector of explanatory variables. Here, $\tilde{\mathbf{x}}_i = (\tilde{x}_{i1}, \ldots, \tilde{x}_{ip})^\top$ is a p-dimensional vector of exogenous regressors, with the first component normalized to one (intercept term). These covariates are assumed to be uncorrelated with the error term ϵ_i. The variable w_i represents an endogenous regressor that is correlated with ϵ_i. For simplicity, we restrict attention to a single endogenous variable, but the methodology and results readily extend to cases with multiple endogenous regressors. The unknown parameter vector associated with the exogenous regressors is denoted by $\boldsymbol{\alpha}_0 = (\alpha_{10}, \ldots, \alpha_{p0})^\top$, while γ_0 represents the coefficient of the endogenous variable. The error term ϵ_i is assumed to be independent of $\tilde{\mathbf{x}}_i$. The censored regression model with left censoring at the fixed threshold c can be expressed as

$$y_i = \max\{c, \tilde{\mathbf{x}}_i^\top \boldsymbol{\alpha}_0 + w_i \gamma_0 + \epsilon_i\}, \quad i = 1, \ldots, n, \tag{2.2}$$

commonly referred to as the Tobit model. Since no distributional assumption is imposed on ϵ_i, this specification is often called a semi-parametric censored regression model. Without loss of generality, the censoring threshold can be normalized to zero ($c = 0$), in which case the model simplifies to

$$y_i = \max\{0, \tilde{\mathbf{x}}_i^\top \boldsymbol{\alpha}_0 + w_i\gamma_0 + \epsilon_i\}, \quad i = 1, \ldots, n. \qquad (2.3)$$

3. ESTIMATION TECHNIQUE

In this section, we will discuss several estimation techniques that were introduced to estimate the parameter of the model defined in (2.3).We start with Nelson and Olson (1978) and Amemiya (1979), who applied two-stage least squares ideas to the censored regression framework. Next, the control function technique by Smith and Blundell (1986) is discussed, which has become a more generally used and computationally feasible solution.

Following that, the semi parametric and distribution-free estimators such as Honoré and Hu (2004) and Blundell and Powell (2007) for semi parametric censored regression are described. Finally, the recent advancements, Chernozhukov *et al.* (2015), Chen (2018) and Kobayashi (2017) discuss the quantile-based techniques and robust estimating frameworks that deal with heavy-tailed errors and model misspecification.

3.1 Classical Approaches

The journey of estimation in censored regression models commenced by Nelson and Olson (1978) and Amemiya (1979), who applied traditional two-stage approaches, such as two-stage least squares (2SLS). Their work provided the groundwork by demonstrating that standard Tobit estimators become inconsistent under endogeneity and proposing instrumental-variable-based adjustments.

Soon after, Smith and Blundell (1986) improved on these methods and introduced a control function approach that addresses endogeneity by including residuals from the first-stage regression in to the Tobit model. This methodology became a cornerstone because it was more adaptable and computationally practical than previous methods. The two techniques are described as follows.

3.1.1 Nelson and Olsen's (1978) and Amemiya's (1979) Estimation Techniques

Nelson and Olson (1978) proposed an instrumental variables approach for censored regression models with endogenous regressors. Note that the instrumental variables a redefined as the set of variables which are highly correlated with the explanatory variables, in limit and uncorrelated with the random errors, in limit. The instrumental variables help in obtaining a consistent estimator of the parameters, see Rao *et al.* (2008).Their estimator replaces the endogenous explanatory variable with its fitted value, obtained from the first-stage regression of w_{i} on the instruments. Specifically, if $\hat{w}_i = \mathbf{z}_i^\top \hat{\boldsymbol{\delta}}_n$ is the predicted value from the least squares regression in (3.1), then the structural Tobit model is estimated as

$$y_i = \max\left\{0, \tilde{\mathbf{x}}_i^\top \boldsymbol{\alpha}_0 + \hat{w}_i\gamma_0 + \epsilon_i\right\}.$$

In this procedure, the standard Tobit maximum likelihood estimator is applied after substituting $\hat{w}_i$ in place of the endogenous regressor.

Building on this framework, Amemiya (1979) introduced an alternative estimator designed to improve efficiency. This method first estimates the reduced-form Tobit model for the endogenous variable and then applies a generalized least squares (GLS) step to recover the structural parameters. By exploiting additional information from the reduced form, Amemiya's estimator achieves greater asymptotic efficiency than the Nelson–Olsen approach. Nevertheless, this estimator does not reach the efficiency of the full information maximum likelihood method. The key limitation is that, unlike models with fully observed dependent variables, maximizing the marginal likelihood of the reduced form in the censored setting does not generally produce asymptotically efficient estimates of the reduced-form parameters. Consequently, while Amemiya's approach improves upon the two-step substitution method of Nelson and Olsen, it still remains less efficient than full maximum likelihood.

3.1.2 Smith and Blundell's (1986) Estimation Technique

The estimation technique of Smith and Blundell (1986) is a two-step procedure that modifies the Tobit model to account for endogeneity. The first stage involves regressing the endogenous variable on the instruments, while the second stage incorporates the residuals from this regression into the structural Tobit model. Let $z_{1,i}$ denote the i-th observation on the instrumental variable corresponding to the endogenous regressor w_i in model (2.1), for $i = 1, \ldots, n$. In the first stage, w_i is regressed on the instrument $z_{1,i}$ and the exogenous covariates $\tilde{\mathbf{x}}_i$ as follows:

$$w_i = z_{1,i}\delta_{1,0} + \tilde{\mathbf{x}}_i^\top \boldsymbol{\delta}_{2,0} + \vartheta_i = \mathbf{z}_i^\top \boldsymbol{\delta}_0 + \vartheta_i, \quad i = 1, \ldots, n, \tag{3.1}$$

where $\mathbf{z}_i = (z_{1,i}, \tilde{\mathbf{x}}_i^\top)^\top$ is a $(p+1)$-dimensional vector of instruments, $\boldsymbol{\delta}_0 = (\delta_{1,0}, \boldsymbol{\delta}_{2,0})^\top$ is the corresponding parameter vector, and ϑ_i is the first-stage error term. Assuming $\mathbb{E}[\vartheta_i \mid \mathbf{z}_i] = 0$, the parameter vector $\boldsymbol{\delta}_0$ can be estimated by OLS as

$$\hat{\boldsymbol{\delta}}_n = \arg\min_{\delta} \sum_{i=1}^{n} (w_i - \mathbf{z}_i^\top \delta)^2,$$

with residuals $\hat{\vartheta}_i = w_i - \mathbf{z}_i^\top \hat{\boldsymbol{\delta}}$. These residuals $\hat{\vartheta}_i$ capture the endogenous component of w_i that is correlated with the structural error ϵ_i. Suppose that

$$\begin{pmatrix} \vartheta_i \\ \epsilon_i \end{pmatrix} \sim \mathcal{N}_2\left(\mathbf{0}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}\right),$$

so that ϵ_i can be expressed as

$$\epsilon_i = \rho_{10}\vartheta_i + \varepsilon_i,$$

with ε_i independent of ϑ_i. Substituting this decomposition into the structural equation yields the transformed Tobit model

$$y_i = \max\{0, \tilde{\mathbf{x}}_i^\top \boldsymbol{\alpha}_0 + w_i\gamma_0 + \hat{\vartheta}_i\rho_{10} + \eta_i\},$$

where $\eta_i = \varepsilon_i + \rho_{10}(\vartheta_i - \hat{\vartheta}_i)$ is a new error term. Thus, in the second stage, the conditional likelihood is constructed by including the residuals $\hat{\vartheta}_i$ from the first stage as additional regressors. This yields consistent estimates of the parameters. However, as noted by Smith and Blundell(1986), this two-step procedure is generally less efficient than the full maximum likelihood estimator, since the least squares estimates of the reduced-form parameters in the first stage do not exploit the restrictions implied by the structural model.

3.2 Semi parametric and Non parametric Approaches

Honoré and Hu (2004) contributed semi parametric estimation methods that avoided strong parametric distributional assumptions, making estimation more robust to misspecification.

Later, Blundell *et al.* (2007) advanced the field by proposing a semiparametric censored regression estimator that combined control functions with flexible modeling of the error structure. This was a major leap since it relaxed assumptions about normality and homoscedasticity, making it applicable to a wider range of empirical settings. These two techniques are described in detail below.

3.2.1 Honoré and Hu Estimation Technique

For the censored regression model under an independence restriction, Honoré and Powell (1994) constructed moment conditions using a pairwise comparison approach. Building on this, Honoré and Hu (2004) extended the idea to models with endogenous regressors, formulating analogous moment conditions. While these conditions are satisfied at the true parameter values, Honoré and Hu (2004) noted that they do not guarantee uniqueness; that is, the moment conditions may not be sufficient to identify the regression coefficients fully. Although originally motivated by panel data censored regression models with predetermined (but not strictly exogenous) explanatory variables, the same principles and insights extend naturally to cross-sectional models with endogenous regressors, providing a unified perspective across panel and cross-sectional contexts. To formalize the idea, consider a panel censored regression model with individual-specific fixed effects $\tilde{\alpha}_i$, given by

$$y_{it} = \max\{0, \tilde{\mathbf{x}}_{it}^\top + w_{it}\gamma_0 + \tilde{\alpha}_i + \epsilon_{it}\}, i = 1, \ldots, n \text{ and } t = 1, \ldots, T,$$

where ϵ_{it} and ϵ_{is} are i.i.d. conditional on $(\tilde{\mathbf{x}}_{it}, \tilde{\mathbf{x}}_{is}, \mathbf{z}_{it}, \mathbf{z}_{is}, \tilde{\alpha}_i)$ for two distinct time periods t and s. Under the restrictions that $w_{it}\gamma_0 \geq 0$ with probability one, the following auxiliary variables can be defined

$$e_{its}(\boldsymbol{\beta}_0) = \max\{-\tilde{\mathbf{x}}_{it}^\top \boldsymbol{\alpha}_0, -\tilde{\mathbf{x}}_{is}^\top \boldsymbol{\alpha}_0, \tilde{\alpha}_i + \epsilon_{it}\},$$

and similarly,

$$e_{ist}(\boldsymbol{\beta}_0) = \max\{-\tilde{\mathbf{x}}_{it}^{\top}\boldsymbol{\alpha}_0, -\tilde{\mathbf{x}}_{is}^{\top}\boldsymbol{\alpha}_0, \tilde{\alpha}_i + \epsilon_{is}\}.$$

It follows that for any function h_1, the condition $\mathbb{E}[h_1(e_{its}(\boldsymbol{\beta}_0) - h_1(e_{ist}(\boldsymbol{\beta}_0)) \mid \tilde{\mathbf{x}}_{is}, \tilde{\mathbf{x}}_{it}, \mathbf{z}_{it}, \mathbf{z}_{is}] = 0$ holds and can be exploited to construct valid moment conditions. Furthermore, if $(\epsilon_{it}, \epsilon_{is})$ is exchangeable conditional on $(\tilde{\mathbf{x}}_{is}, \tilde{\mathbf{x}}_{it}, \mathbf{z}_{it}, \mathbf{z}_{is}, \tilde{\alpha}_i)$, then additional moment conditions are available of the form

$$\mathbb{E}[g(h_1(e_{its}(\hat{\mathbf{a}}_0) - h_1(e_{ist}(\hat{\mathbf{a}}_0))) \mid \tilde{\mathbf{x}}_{is}, \tilde{\mathbf{x}}_{it}, \mathbf{z}_{it}, \mathbf{z}_{is}] = 0$$

for any odd function g. These moment conditions can be used to estimate $\boldsymbol{\beta}$ via the continuously updating Generalized Method of Moments (GMM) framework (Hansen *et al.,* 1996), which provides a practical and flexible estimation strategy in the presence of censoring and endogeneity.

3.2.2 Blundell and Powell (2007) Estimation Technique

Their approach is based on a "distributional exclusion" restriction, which assumes that the conditional distribution of the structural error depends on the regressors and instruments only through a lower-dimensional control variable. This control variable is defined as the residual from the reduced-form regression—that is, the difference between the endogenous regressor and its conditional expectation given the instruments. Formally, the first-stage reduced form can be expressed as $w_i = h(\mathbf{z}_i) + \vartheta_i \; i = 1, \ldots, n,$ where $h(.)$ is an unknown (possibly nonparametric) function of the instruments $\mathbf{z}_i$, and ϑ_i denotes the first-stage error.

This assumption translates into a similar exclusion restriction for the conditional quantiles of the censored dependent variable, motivating a two-stage estimation procedure. In the first stage, the conditional expectation $h(.)$ is estimated non parametrically, and the corresponding residuals are obtained. In the second stage, Blundell and Powell (2007) estimates the structural parameters using a weighted least squares regression of pairwise differences in estimated quantiles of the dependent variable on the corresponding differences in regressors. The estimator is given by

$$(\hat{\mathbf{a}}_n, \hat{\gamma}_n)^{\top} = \left[\sum_{i<j} K_{\vartheta}\left(\frac{\hat{\vartheta}_i - \hat{\vartheta}_j}{h_n}\right)\hat{t}_i\hat{t}_j(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^{\top}\right]^{-1} \times \sum_{i<j} K_{\vartheta}\left(\frac{\hat{\vartheta}_i - \hat{\vartheta}_j}{h_n}\right)\hat{t}_i\hat{t}_j(\mathbf{x}_i - \mathbf{x}_j)(\hat{\mathbf{q}}_i - \hat{\mathbf{q}}_j),$$

where $q_i = Q_{\tau}(y_i \mid \mathbf{x}_i, \mathbf{z}_i) = q_i(\tau)$, $\mathbf{x}_i = (\tilde{\mathbf{x}}_i^{\top}, w_i)$ and $Q_{\tau}(\epsilon_i \mid \mathbf{x}_i, \mathbf{z}_i) = Q_{\tau}(\epsilon_i \mid \vartheta_i) = \lambda_{\tau}(\vartheta_i)$. Here, $K_{\vartheta}(.)$ denotes a kernel function that integrates to one, vanishes outside a compact set, and satisfies the usual regularity conditions. The sequence $\{h_n\}$ is a bandwidth parameter tending to zero at an appropriate rate as $n \to \infty$. The trimming term $\hat{t}_i$ is defined as

$$\hat{t}_j = \omega(\hat{q}_i).1_{(\mathbf{x}_i, \mathbf{z}_i) \in \mathcal{B}}$$

where, $\omega(\hat{q}_i) = 0$ unless $\hat{q}_i > a$ for some $a > 0$. This ensures that only pairs of observations in the uncensored region and within a compact support $\mathcal{B}$ are used in estimation, with sufficiently small differences in their control variables. The estimation uses only those pairs of observations where both estimated quantiles lie in the uncensored region and the difference in the estimated control variables is sufficiently small. Although this approach provides a consistent semi parametric estimator, it relies on several strong assumptions: (i) the endogenous regressors must be continuously distributed, (ii) a valid distributional exclusion restriction must hold, and (iii) the nonparametric first stage must be estimated at an appropriate rate. As a result, the method cannot be directly applied in settings with censored or discrete endogenous variables. This limitation is particularly important in program evaluation studies, where treatment variables are often binary or discrete-valued.

3.3 Quantile-Based Estimation Approaches

Quantile-based approaches represented a substantial advancement in the literature. Chernozhukov and Hong (2002) and Chen (2018), in particular, extended the frame work of Censored Quantile Instrumental Variable (CQIV) estimation, allowing researchers to capture heterogeneous impacts at different points of the outcome distribution rather than focusing simply on mean effects. This method provides a more complete knowledge of the influence of endogenous regressors under censoring, especially when effects vary across quantiles. The two techniques are described as follows.

3.3.1 Chernozhukov (2015) and Chen (2018) EstimationTechnique

Chernozhukov *et al.* (2015) extended the approach of Chernozhukov and Hong (2002) to handle endogeneity within a non-additive triangular system of equations. Despite this generalization, both methods rely on the existence of a control function, which imposes strong structural assumptions on the underlying model. As a result, if the true nature of endogeneity is misspecified, the control function approach may yield in consistent estimates and misleading inferences. Moreover, this approach requires the endogenous variables to be continuously distributed, thereby excluding cases with censored or discrete endogenous variables. This restriction is particularly limiting inprograme valuation studies, where the endogenous treatment variable is often binary or multi-valued but discrete, rendering the control function approach generally inapplicable.

To address some of these limitations, Chen (2018) studied sequential estimation methods for censored quantile regressions. These methods have been shown to offer efficiency advantages: when the model is correctly specified, estimators based on the independence restriction are generally more efficient than those relying solely on quantile restrictions. Extensive numerical results illustrating this efficiency gain can be found in Honoré and Powell (1994) and Honoré and Hu (2004).

3.3.2 Chen and Wang (2020) Estimation Technique

Chen and Wang (2020) introduced moment-based estimators for censored outcome models in the presence of endogenous regressor, relying on independence and quantile restrictions, respectively. Under the full independence assumption on ϵ_i, conditional on $(\tilde{\mathbf{x}}_i, \mathbf{z}_i)$, we can construct valid moment conditions. For each quantile $\tau \in (\tau_l, \tau_u)$, let $c_0(\tau)$ denote the τ-th quantile of ϵ_i. The corresponding moment conditions are

$$\mathbb{E}[1_{\tilde{\mathbf{x}}_i^\top \boldsymbol{\alpha}+c>0}(1_{y_i-\mathbf{x}_i^\top \boldsymbol{\theta}<c} - \tau)(\tilde{\mathbf{x}}^\top, \mathbf{z}_i^\top)^\top] = 0, \quad \tau \in [\tau_l, \tau_u],. \tag{3.2}$$

Here $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \gamma)^\top$ and $\mathbf{x}_i = (\tilde{\mathbf{x}}_i, \mathbf{z}_i)^\top$. To handle heavy censoring, top-quantile moment conditions are imposed

$$\mathbb{E}[\left(1_{\tilde{\mathbf{x}}_i^\top \boldsymbol{\alpha}+c>0} - \tau\right)(\tilde{\mathbf{x}}^\top, \mathbf{z}_i^\top)^\top I_{\mathbf{x}^o}(\mathbf{x}_i)] = 0, \quad \tau \geq \tau_u^0, \tag{3.3}$$

for some top quantiles τ_u^0 and $I_{\mathbf{x}^o}(\mathbf{x}_i) = 1_{(\mathbf{x}_i \in \mathbf{x}^o)}$ restricts the sample to a bounded support to handle extreme censoring. The semi parametric estimation proceeds as follows:

1. For a given γ, estimate $\boldsymbol{\alpha}$ via pairwise comparison:

$$\widehat{\boldsymbol{\alpha}}(\gamma) = \arg\min_{\hat{\mathbf{a}}} \mathbb{S}_n(\boldsymbol{\alpha}, \gamma),$$

$$\mathbb{S}_n(\boldsymbol{\alpha}, \gamma) = \frac{1}{\binom{n}{2}} \sum_{i<j} s\left(y_i(\gamma), y_j(\gamma), (\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j)' \boldsymbol{\alpha}\right), \tag{3.4}$$

where $y_i(\gamma) = \max\{y_i - w_i\gamma, \gamma^*\} - \gamma^*$, $\gamma^* = \max\{-\gamma, 0\}$, and $s(.)$ is the convex loss function of Honoré and Powell (1994) .

2. Estimate the quantiles of residuals ϵ_i using the Kaplan Meier estimator:

$$\hat{c}(\widehat{\boldsymbol{\alpha}}(\gamma), \gamma, \tau) = \inf\{t : \hat{F}(\widehat{\boldsymbol{\alpha}}(\gamma), \gamma, t) \geq \tau\}. \tag{3.5}$$

3. Define the objective function over all quantiles:

$$\mathbb{G}_n(\boldsymbol{\alpha}, \gamma, c) = \int_{\tau_l}^{\tau_u^0} \| \mathbb{G}_{1n}(\boldsymbol{\alpha}, \gamma, c(\tau), \tau) \| \, d\tau + \int_{\tau_u^0}^{\tau_u} \| \mathbb{G}_{2n}(\boldsymbol{\alpha}, \gamma, c(\tau), \tau) \| \, d\tau, \tag{3.6}$$

where $\mathbb{G}_{1n}$ and $\mathbb{G}_{2n}$ are the sample analogs of the moment conditions defined by

$$\mathbb{G}_{1n}(\boldsymbol{\alpha}(\gamma), \gamma, c(\tau)) = \frac{1}{n} \sum_{i=1}^{n} 1_{\left(\tilde{\mathbf{x}}_i^\top \boldsymbol{\alpha} + \min\{\gamma, 0\} + c > 0\right)} \left(1_{\left(y_i - \mathbf{x}_i^\top \boldsymbol{\theta} < c\right)} - \tau\right) \mathbf{z}_i,$$

$$\mathbb{G}_{2n}(\boldsymbol{\alpha}(\gamma), \gamma, c(\tau)) = \frac{1}{n} \sum_{i=1}^{n} \left(1_{\left(y_i - \mathbf{x}_i^\top \boldsymbol{\theta} < c\right)} - \tau\right) \mathbf{z}_i I_{\mathbf{x}^o}(\mathbf{x}_i). \tag{3.7}$$

4. Estimate γ_0 and $\boldsymbol{\alpha}_0$ as:

$$\hat{\gamma} = \arg\min_{\gamma} \mathbb{G}_n(\widehat{\boldsymbol{\alpha}}(\gamma), \gamma, \hat{c}(\tau)), \quad \widehat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\alpha}}(\hat{\gamma}). \tag{3.8}$$

This procedure achieves the identification and consistent estimation of (α_0, γ_0) through independence-based quantile moment conditions. Although point identification follows from general assumptions on the distribution of observables, these conditions are often difficult to validate in practice. Building on this, Chesher *et al.* (2023) introduced an alternative framework for censored out come models with endogenous regressors under instrumental variable restrictions. The focus is on Tobit-type left censoring at zero, with extensions to stochastic censoring. In contrast to control function methods, this approach does not impose explicit structure on the process generating endogenous regressors. Consequently, the model may deliver either partial or point identification depending on instrument strength and the support of exogenous variables. The paper characterizes identified sets and develops inference procedures for scalar functions of partially identified parameters. An application to UK household tobacco expenditures demonstrates how the method allows inference on the effect of endogenous total expenditure under both minimal and stronger Gaussian assumptions and contrasts these results with those obtained from fully specified triangular models.

3.4 Recent Developments and Extensions

The recent past has witnessed some novel approaches, such as the Genya Estimation Technique, which provide alternative frameworks to jointly tackle censoring and endogeneity in a more flexible and robust manner. Along side these, significant progress was made in the development of robust estimation methods aimed at reducing the impact of outliers, accommodating heavy-tailed error distributions, and addressing model misspecification. Collectively, these advancements have enhanced the reliability and applicability of inference in censored regression models under practical data scenarios. two such estimation techniques are presented below.

3.4.1 Genya Estimation Technique

Kobayashi (2017) introduces the *p*-th Tobit quantile regression models with endogenous variables in a Bayesian framework. In this framework, the first-stage regression of the endogenous variable on the exogenous variables incorporates the assumption that the τ-th quantile of the error term is zero. The residual from this regression is then included in the *p*-th quantile regression model, ensuring that the *p*-th conditional quantile of the new error term is zero.To account for the unknown distribution of the first-stage regression error, both parametric and semi parametric approaches are employed around the zeroτ-th quantile assumption. Since the value of τ is generally unknown, it is treated as an additional parameter and estimated from the data. The effectiveness of the proposed models is illustrated using simulated data and applied to real data on the labour supply of married women.

3.4.2 Robust Estimation Technique

Shukla *et al.* (2023) proposed a two-step estimation procedure for censored regression models with endogenous regressors using the control function approach. In the first step, the endogenous regressor is modeled via ordinary least squares under standard error assumptions. The residuals from this regression, denoted as $\hat{\vartheta}_i$, are then included in the second stage. The second stage employs an M-estimation technique to estimate the model parameters. Specifically, the M-estimator $\hat{\boldsymbol{\theta}}_n$ is defined as the solution to the estimating equation

$$\sum_{i=1}^{n}\psi\Big(y_i - \max\{0, \hat{\mathbf{x}}_i^\top\boldsymbol{\theta}\}\Big) = 0,$$

where $\psi(.)$ is the derivative of a general loss function $\rho(.)$, $\hat{\mathbf{x}}_i = (\tilde{\mathbf{x}}_i, w_i, \hat{\vartheta}_i)^\top$, and $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \gamma, \rho_1)^\top$. Under suitable regularity conditions defined in Shukla *et al.* (2023), the M-estimator is consistent and asymptotically normal, i.e.,

$$\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0,$$

and

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}\big(0, \Sigma_M\big),$$

where $\boldsymbol{\theta}_0$ is the true parameter vector and Σ_M is the asymptotic covariance matrix in Theorem~3.2 from Shukla *et al.* (2023).

Building on this framework, Shukla *et al.* (2024) introduced a robust alternative based on L-estimation. The first step of the procedure remains the same, while the second step replaces M-estimation with an L-estimator defined as

$$\pi_n = \int_0^1 \hat{\boldsymbol{\theta}}_n(\tau) J(\tau) d\tau,$$

where $\hat{\boldsymbol{\theta}}_n(\tau)$ is the censored quantile regression estimator at quantile τ, and $J(\tau)$ is a smooth weight function defined on a compact subinterval of $(0,1)$. The

L-estimator also satisfies consistency and asymptotic normality under standard conditions, i.e.,

$$\hat{\boldsymbol{\pi}}_n \xrightarrow{p} \boldsymbol{\pi}_0,$$

and

$$\sqrt{n}(\hat{\boldsymbol{\pi}}_n - \boldsymbol{\pi}_0) \xrightarrow{d} \mathcal{N}\left(0, \Sigma_L\right),$$

where Σ_L is the corresponding asymptotic covariance matrix defined in Theorem 3.4 from Shukla *et al.* (2024).These results extend naturally to nonlinear censored regression models with endogenous regressors. If the non linear regression function $\tilde{g}(\bullet)$ is Lipschitz continuous in both parameters and regressors, then both M- and L-estimators retain their consistency property (see, in Shukla *et al.* (2025)). In addition to these theoretical guarantees, the performance of the proposed methods has been evaluated through Monte Carlo simulations and an empirical application using the "Affair" dataset, demonstrating both finite-sample efficiency and robustness in practical contexts. In summary, the M- and L-estimation procedures offer complementary strengths: both ensure asymptotic reliability, simulations validate their finite-sample behavior, and real data applications confirm their practical relevance.

4. APPLICATIONS

In this section, we provide the application of Mroz data set available in wooldridge package (available at https://github.com/swati-1602/CRLM-Real-data.git) to demonstrate the application of our proposed estimators. This data set focuses on the labor force participation of U.S. women and includes data for 753 individuals. The response variable of interest is the total number of hours worked per week. Notably, 325 out of the 753 women reported not working any hours, rendering the dependent variable left-censored at zero. This implies that the censoring probability (CP) is of approximately 0.43.

In this data set, the explanatory variables include a woman's years of education (*educ*), years of experience (*exper*) and its squared term (*expersq*), age (*age*), the number of young children under six years of age (*kidslt6*), and non-wife household income (*nwifeinc*). Non-wife household income is treated as an endogenous variable due to its potential correlation with unobserved household preferences that may influence the wife's labour force participation.

To address this endogeneity, the husband's years of education (*huseduc*) is employed as an instrumental variable. It is assumed that while this variable does not directly affect the wife's decision to participate in the labour force, it influences both the husband's income and thenon-wife household income. The variable *hours* represents the total number of hours a woman worked in a year.

In the context of this study, an exogenous variable is one that is uncorrelated with the error term. In contrast, an endogenous variable is correlated with the error term, typically due to omitted variables, simultaneity, or measurement error. This correlation can lead to biased and inconsistent parameter estimates if ignored. To demonstrate the impact of endogeneity, two analyses are conducted:

1. **Exogenous Case:** A censored regression model (Tobit model) is estimated assuming all regressors are exogenous. The model parameters are obtained by minimizing the absolute loss function using the M-estimation framework, also known as the CLAD (Censored Least Absolute Deviations) method Powell (1984).
2. **Endogenous Case (Control Function Approach):** To account for endogeneity in the *nwifeinc* variable, a control function approach is adopted.
    - In the first stage, the potentially endogenous regressor (*nwifeinc*is regressed on all exogenous variables and an instrumental variable (*huseduc*).
    - The residuals from this first-stage regression are obtained and included as an additional regressor in the censored regression model.
    - The model is then re-estimated using the least absolute loss-based M-estimation method.

The performance of both models is compared using the Bootstrap Mean Squared Error (BMSE) of the parameter estimates. The results show that the BMSE values under the control function approach are slightly smaller than those obtained under the exogeneity assumption, indicating an improvement in estimation efficiency and bias reduction when endogeneity is properly addressed. This application thus highlights that ignoring endogeneity in censored models can lead to biased inference, whereas incorporating a control function correction within the M-estimation (CLAD-type) framework provides more reliable parameter estimates.

Table 1. Parameter Estimates and Bootstrap Mean Squared Error (BMSE) for Exogenous and Endogenous Case

Estimator	Variable	Estimate	BMSE				
			B=100	B=200	B=300	B=400	B=500
CLAD	age	-0.0646	0.5113	0.396	0.3593	0.3325	0.2959
	educ	-0.0100	0.4840	0.3710	0.3348	0.3132	0.2736
	exper	0.0253	0.4694	0.3579	0.3220	0.3038	0.2623
	expersq	0.1437	0.4390	0.3325	0.2977	0.2908	0.2428
	kidslt6	-0.0530	0.5050	0.3902	0.3536	0.3279	0.2906
	kidsge6	-0.0075	0.4828	0.3700	0.3338	0.3124	0.2727
	nwifeinc	-0.0107	0.4843	0.3713	0.3350	0.3134	0.2738
IVCLAD	age	-0.4872	0.2081	0.187	0.1688	0.1636	0.1540
	educ	0.3316	0.1914	0.1857	0.1644	0.1543	0.1261
	exper	-0.2086	0.1269	0.0837	0.0801	0.0741	0.0627
	expersq	2.0238	0.0699	0.0677	0.0638	0.0633	0.0522
	kidslt6	-0.3742	0.0285	0.0282	0.0274	0.0226	0.0217
	kidsge6	-0.5716	0.2252	0.2239	0.2238	0.2533	0.2528
	nwifeinc	-0.3465	0.0937	0.0800	0.0795	0.0794	0.0754
	res	-0.375	0.2071	0.2005	0.1674	0.1609	0.1532

5. AVAILABLE SOFTWARE FOR ESTIMATION OF CENSORED REGRESSION MODELS

In practice, several statistical software packages provide functions and routines for estimating censored regression models, thereby facilitating empirical applications and replication of research findings. Among these, the censReg package (Henningsen (2011)) is one of the most widely used in R. It estimates Tobit and censored regression models using maximum likelihood estimation (MLE) for cross-sectional data, and also supports random-effects MLE for panel data through Gauss–Hermite quadrature. The package accommodates both left- and right-censoring and provides tools for model diagnostics and inference.

Similarly, the AER package (Kleiber and Zeileis (2008, 2009)) offers the function tobit() for estimating Tobit models, while the cenmle() function from the NADA package serves as a front-end interface to the survreg() function in the survival package. The tobit() functionin the VGAM package, on the other hand, performs estimation using its own maximum likelihood routine, rather than relying on survreg(). Additionally, the survival package provides the survreg() function, which can be adapted to fit censored regression models within the broader framework of survival analysis.

For models that involve endogeneity, the sampleselection package provides both two-step and maximum likelihood estimators for handling sample selection and endogenous censoring models. When the focus extends to semiparametric or quantile-based estimation, the quantreg package becomes particularly useful. It includes the function crq(), which is suitable for estimating quantile regression models with either random or fixed censoring. The quantreg package offers flexible computational algorithms for quantile regression, and with appropriate modifications, these methods can be effectively adapted to censored data settings. Moreover, the function MCMCtobit from the MCMC package uses the Bayesian Markov Chain Monte Carlo (MCMC) method to estimate censored regression models.

In the present study, the quantile loss (L-estimation) approach has been implemented using user-defined optimization routines in R, while M-estimation has been carried out using the optim() function in R. These implementations enable the estimation of parameters in censored regression models under endogeneity. The proposed approach complements the existing software ecosystem by providing a robust alternative to traditional likelihood-based methods, particularly in cases where the error distribution is non-normal or heavy-tailed. The R codes of the Real data study is available at https://github.com/swati-1602/CRM.git.

At present, there is no R package that directly supports both M-estimation and L-estimation techniques for censored regression models. Hence, the development of an efficient algorithm and its incorporation in to an R package for handling such estimation methods could form a valuable direction for future research.

6. CONCLUSION

This survey article provides an overview of the development of various methodologies available in the literature to handle endogeneity in censored regression models. The main contribution is to compile the theoretical background and estimation techniques.

The paper comprehends the advantages, disadvantages, and constraints of these estimation techniques. Such a development will facilitate further research in the area of censored regression with endogeneity and extend help to address its further extension to nonlinear, panel, and high-dimensional settings.

ACKNOWLEDGEMENT

The authors are grateful for the comments from the referee and editor, which improved the exposition of the paper. The author, Subhra Sankar Dhar, gratefully acknowledges his core research grant, CRG/2022/001489, from the Government of India.

REFERENCES

Amemiya, T. (1979). The estimation of a simultaneous-equation to bit model. *International Economic Review*, **20(1)**,169-181.

Amemiya, T. (1984). To bit models: A survey. *Journal of Econometrics*, **24(1)**, 3-61.

Arabmazar, A.and Schmidt, P.(1982). An investigation of the robustness of the tobit estimator tonon-normality. *Econometrica*, **50(4)**, 1055-1063.

Blundell, R., Chen, X., and Kristensen, D. (2007). Semi-nonparametri civestimationofshapeinvariant engel curves. *Econometrica*, **75(6)**, 1613-1669.

Blundell, R.and Powell, J.L. (2007). Censored regression quantiles with endogenous regressors. *Journal of Econometrics*, **141(1)**, 65-83. Semiparametric methods in econometrics.

Buchinsky, M. (1994). Changes in the U.S. wagestructure 1963-1987: Application of quantile regression. *Econometrica*, **62(2)**, 405-458.

Buchinsky, M. and Hahn, J. (1998). An alternative estimator for the censored quantile regression model. *Econometrica*, **66(3)**, 653-671.

Chay, K.Y. and Powell, J.L. (2001). Semi parametric censored regression models. *Journal of Economic Perspectives*, **15(4)**, 29-42.

Chen, S. (2018). Sequential estimation of censored quantile regression models. *Journal of Econometrics*, **207(1)**, 30-52.

Chen, S. and Wang, Q. (2020). Semi parametric estimation of a censored regression model with endogeneity. *J. Econometrics*, **215(1)**, 239-256.

Chernozhukov, V., Fern´andez-Val, I., and Kowalski, A.E. (2015). Quantile regression with censoring and endogeneity. *Journal of Econometrics*, **186(1)**, 201-221.

Chernozhukov, V. and Hong, H. (2002). Three-step censored quantile regression and extramarital affairs. *Journal of the American Statistical Association*, **97(459)**, 872-882.

Chesher, A., Kim, D., and Rosen, A.M. (2023). Iv methods for tobit models. *Journal of Econometrics*, **235(2)**, 1700-1724.

Goldberger, A.S. (1981). Linear regression after selection. *Journal of Econometrics*, **15(3)**, 357- 366.

Hansen, L.P., Heaton, J., and Yaron, A. (1996). Finite-sample properties of some alternative gmm estimators. *Journal of Business Economic Statistics*, **14(3)**, 262-280.

Heckman, J.J. (1976). The common structure of statistical models of truncation,sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, **5(4)**, 475-492.

Heckman, J.J. (1978). Dummy endogenous variables in a simultaneou sequation system. *Econometrica*, **46(4)**, 931-959.

Henningsen, A. (2011). *censReg: Censored Regression (Tobit) Models*. R package version 0.5.

Honoré, B.E. and Hu,L. (2004). Estimation of cross sectional and panel data censored regression models with endogeneity. *Journal of Econometrics*, **122(2)**, 293-316.

Honoré, B.E. and Powell, J.L. (1994). Pairwise difference estimators of censored and truncated regression models. *Journal of Econometrics*, **64(1)**, 241-278.

Khan, S. and Powell, J.L. (2001). Two-step estimation of semi parametric censored regression models. *Journal of Econometrics*, **103(1)**, 73-110. Studies in estimation and testing.

Kleiber, C. and Zeileis, A. (2008). *Applied Econometrics with R*. Springer, New York.

Kleiber, C. and Zeileis, A.(2009). *AER: Applied Econometrics with R*.R package version1.1.

Kobayashi, G. (2017). Bayesian endogenous tobit quantile regression. *Bayesian Analysis*, **12(1)**, 161-191.

Lee, M.-J. (1992). Winsorized mean estimator for censored regression. *Econometric Theory*, **8(3)**, 368-382.

Nelson, F. and Olson, L. (1978). Specification and estimation of a simultaneous-equationmodel with limited dependent variables. *International Economic Review*, **19(3)**, 695-709.

Powell, J.L. (1984). Least absolute deviations estimation for the censored regression model. *Journal of Econometrics*, **25(3)**, 303-325.

Powell, J.L. (1986a). Censored regression quantiles. *Journal of Econometrics*, **32(1)**, 143-155.

Powell, J.L. (1986b). Symmetrically trimmed least square sestimation for tobit models. *Econometrica*, **54(6)**, 1435-1460.

Rao, C., Toutenburg, H., Shalabh, and Heumann, C. (2008). *Linear Models and Generalizations-Least Squares and Alternatives*. Springer.

Shukla, S., Dhar, S. S., and Shalabh (2023). M-estimation in censored regression model using instrumental variables under endogeneity, (Under review).

Shukla, S., Dhar, S. S., and Shalabh (2024). L-estimation in instrumental variables regression for censored data in presence of endogeneity and dependent errors, (Under review).

Shukla, S., Dhar, S.S., and Shalabh (2025). Applications of nonlinear tobit model sunder endogeneity.*Sankhya B* (Accepted for publication)

Smith, R. J. and Blundell, R. W. (1986). An exogeneity test for a simultaneous equation tobit model with an application to labors upply. *Econometrica*, **54(3)**, 679-685.