A Simple Consistent Non-parametric Estimator for the Regression Function in a Truncated Sample

Armando Levy [†]

Department of Economics

North Carolina State University

Raleigh, NC 27695-8110

armando@ncsu.edu

May 1998

revised December 9, 1999

[†]Thanks to Mark Coppejans, Alastair Hall and Dan McFadden for helpful comments.

KEY WORDS: truncated regression, semi-parametric, backfit

Abstract

Much recent work has focused on the estimation of regression functions in samples which are truncated or censored. Much of this work has focused on the estimation of a parametric regression function with an error distribution of unknown form. While these method relax a strong parametric assumption about which we seldom have *a priori* information, they still impose a strong parametric assumption on the regression equation (which is presumably the focus of the analysis). Here we take the other approach. An estimator is proposed for the problem of non-parametric regression when the sample is *truncated* above or below some known threshold of the dependent variable. We specify the error distribution up to a vector of parameters θ while estimating the regression function without assuming a parametric form. A simple "backfit" estimator based on an initial kernel smooth is proposed. We establish consistency results for this estimator when the error distribution is known up to a finite parameter vector and satisfies some regularity conditions. A small monte-carlo study is performed to ascertain the finite sample properties of the estimator. The estimator is found to perform well in our experiment: achieving reasonable average absolute errors relative to the maximum likelihood estimator— especially when truncation is severe.

1 Introduction

Consider the following equation:

$$y = g(\mathbf{x}) + \epsilon \tag{1.1}$$

where y is a quantity of interest, $\mathbf{x}$ a vector of regressor variables, g a continuous function and ϵ is a random draw from distribution F_{θ^*}, where F_{θ^*} is a member of a parametric family indexed by $\theta \in \Theta \subset \mathbb{R}^p$. Moreover, ϵ is independent of $\mathbf{x}$ and both θ^* and g are unknown. It is well known that, given some regularity conditions and a random sample from the process (1.1), g can be consistently estimated by kernel regression, splines, sieves or a variety of other non-parametric methods.

Here we consider estimation of g when data are observed from (1.1) only when $y \in [\alpha, \infty)$ where α is a known threshold. In this case, we will say the data are truncated *below* α. Alternatively we can examine the case where the data are truncated *above* some known threshold α if we observe the pair $(y, \mathbf{x})$ only when $y \in (-\infty, \alpha]$. A common feature of biometric and econometric data is that they are recorded from truncated samples and not random samples. Examples include survival times for subjects after the onset of disease, sales when we wish to estimate demand for *all* (potential) consumers and labor supply for an entire work-eligible population from data on those already working.

Since F_θ belongs to a parametric family, an estimator for g and θ in (1.1) may be called semi-parametric. While there is a significant literature on semi-parametric estimation in censored and/or truncated regression problems, all of it has focused on the related problem of estimation of a parametric regression function in the presence of an *unknown* error distribution. For examples of this approach in biometric literature see

Buckley and James (1982),Tsui et al. (1988), while for econometrics see Powell (1984,1986), Duncan (1986), Fernandez (1986), Horowitz (1986), Ruud (1986), Newey(1986,1991), Cosslett (1991), Ahn and J.L.Powell (1993), Lee (1994a,b) for linear regression and Gallant and Nychka (1987), Ichimura (1993) for nonlinear regression. While these method relax a strong parametric assumption about which we seldom have *a priori* information, they still impose a strong parametric assumption on the regression equation (which is the focus of the analysis). Additionally, since the expectation of the error term in the truncated sample is a smooth, bounded and *monotonically increasing* function, the bias induced by misspecification of the error distribution will be equivalent to differences among functions in a restricted class. On the other hand, estimating a highly nonlinear regression function within a linear class will usually engender quite a large bias.

Here we take the other approach. We specify the error distribution up to a vector of parameters θ while estimating g without assuming a parametric form. The estimator is very simple: requiring a initial kernel smooth followed by calculation of the unique root to a simple fixed point problem. No simulation or grid search is required. In the next section, we describe the estimator and establish uniform consistency of the kernel estimator for a monotonic transformation of the population regression function. The third section establishes a consistency result for the backfit estimator. The fourth section gives the results of a small monte carlo study while section five concludes.

2 A Simple Estimator

Consider the truncated regression problem where we observe pairs $(y_i, \mathbf{x_i}) \in \mathbb{R} \times \mathbb{R}^k$ from iid draws of the process (1.1) *only if* y_i exceeds a known threshold α which we set to zero without loss of generality. We will consider the case of truncation from below in this paper, but the case for truncation from above is essentially the same. Note that the truncated pairs $(y_i, \mathbf{x_i})$ are also iid, though in contrast to (1.1), the distribution of $\epsilon_i = y_i - g(\mathbf{x}_i)$ conditional on $\mathbf{x}_i$ are not identically distributed. Henceforth we direct our analysis to this truncated sample.

It is well known that methods which are strongly consistent under random sampling become biased in a truncated sample (see Goldberger (1981) and Greene (1997) for the case of OLS). We may view this bias as due to the omitted variable $\mathbb{E}(\epsilon|\epsilon > -g(\mathbf{x}))$ since the mean of y conditional on $\mathbf{x}$ and a positive observation is:

$$\mathbb{E}(y|\mathbf{x}, y > 0) = g(x) + \mathbb{E}(\epsilon|\epsilon > -g(\mathbf{x})) \tag{2.1}$$

where,

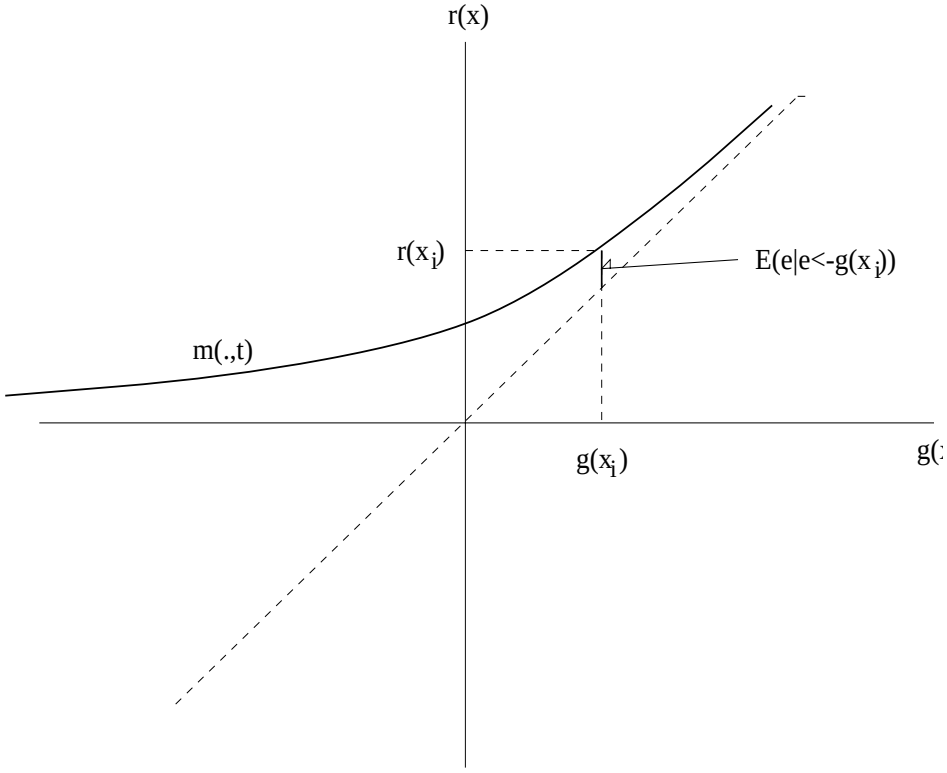
$$\mathbb{E}(\epsilon|\epsilon > -g(\mathbf{x})) = \frac{\int_{-g(\mathbf{x})}^{\infty} u f_{\theta^*}(u) du}{1 - F_{\theta^*}(-g(\mathbf{x}))} \tag{2.2}$$

We will denote the conditional mean of y given $\mathbf{x}$ and inclusion into the truncated sample as:

$$r(\mathbf{x}) = m(g(\mathbf{x}), \theta^*) = \mathbb{E}(y|\mathbf{x}, y > 0) = g(\mathbf{x}) + \mathbb{E}(\epsilon|\epsilon > -g(\mathbf{x})) \tag{2.3}$$

where $r(\mathbf{x})$ gives the conditional mean of y as a function of $\mathbf{x}$ while $m(g(\mathbf{x}), \theta)$ denotes the mean conditional on the value $g(\mathbf{x})$ *and* θ. Since g is continuous and F_θ is absolutely

Figure 1: Relationship Between r(**x**), m(**x**) and g(**x**)



continuous, r will be a continuous function of $\mathbf{x}$ and hence measurable. Additionally if (2.2) is differentiable in $g(\mathbf{x})$ with derivative exceeding -1, $m(.,\theta)$ will be a continuous, positive, monotonically increasing function of $g(\mathbf{x})$ and hence invertible. Our strategy to estimate g is to simply estimate r consistently and then to invert m to "backfit" g. Assume the following:

A1: $g : \mathbb{R}^k \to \mathbb{R}$ is a continuous function, and $\mathbb{E}y1(y > 0) < \infty$ where $1(.)$ is the indicator function.

A2: $\mathbf{x} \in \mathbb{R}^k$ is a random vector, independent of ϵ with an absolutely continuous distribution function and continuous density h.

A3: ϵ is a random variable with an absolutely continuous distribution F^*_θ and an everywhere positive density f_{θ^*}. In order to identify a level for g, we assume $\mathbb{E}\epsilon = 0$.

Let

$$\hat{r}_n(\mathbf{x}) = \frac{\sum_{i=1}^n K(\frac{x-x_i}{\lambda})y_i}{\sum_{i=1}^n K(\frac{x-x_i}{\lambda})}$$

be the Nadaraya-Watson kernel estimator of $r(\mathbf{x})$ based on kernel K and window width λ_n which satisfies the following properties:

A4:

- K is a density on $\mathbb{R}^k$ with an absolutely integrable characteristic function.
- $\lambda_n \to 0$ as $n \to \infty$, and $\sqrt{n}\lambda_n^k \to \infty$.

Theorem 1 below establishes that while kernel regression of y on $\mathbf{x}$ in the truncated sample is biased for g, it is uniformly consistent for the monotonic transform of g: r.

Theroem 1 *Given* **A1-A4**, *then for every* $\delta \in (0, sup_{x \in R^k} h(\mathbf{x})]$,

$$sup_{\{\mathbf{x} \in \mathbb{R}^k : h(\mathbf{x}) > \delta\}} |\hat{r}_n(\mathbf{x}) - r(\mathbf{x})| \to_p 0$$

The proof is standard and given in the appendix. Bounds for the rates of convergence can be found in Bierens (1983).

Given our estimate of r, we may recover an estimate for the population regression function g by inverting the definition (2.3) with respect to a consistent estimate of θ^*. In order to insure this, we must guarantee that $m(g(\mathbf{x}), \theta)$ is monotonically increasing for all values of θ.

In order to identify g from r, we add the following conditions:

A5:

- i. $\theta^* \in \Theta$ where Θ is a compact subset of $\mathbb{R}^p$ and
- ii. If $\theta \neq \theta^*$ then $\frac{f_\theta(y - g(\mathbf{x}))}{1 - F_\theta(-g(\mathbf{x}))} \neq \frac{f_{\theta^*}(y - g(\mathbf{x}))}{1 - F_{\theta^*}(-g(\mathbf{x}))}$.
- iii. $\frac{f_\theta(y - g(\mathbf{x}))}{1 - F_\theta(-g(\mathbf{x}))}$ is continuous at each $\theta \in \Theta$ with probability one.
- iv. There is open set U $\in \mathbb{R}$ such that $r(\mathbf{x}) \in U$ with probability one and

  $sup_{u \in U} \frac{f_\theta(u)}{[1 - F_\theta(u)]^2} \int_u^\infty (\epsilon - u) f_\theta(\epsilon) d\epsilon < 1 \ \forall \theta \in \Theta$
- v. $\mathbb{E}[sup_{\theta \in \Theta} |log[\frac{f_\theta(y - g(\mathbf{x}))}{1 - F_\theta(-g(\mathbf{x}))}]|] < \infty$

Assumption **A5i** is made to insure uniform convergence in θ of the sample likelihood. When $log(f_\theta)$ is a concave density we may relax this requirement to Θ a convex set. Assumption **A5ii.** is an identification condition, while **A5iii.** is a continuity condition. **A5iv.** is necessary and sufficient for strict monotinicity of m for any $\theta \in \Theta$, and **A5v.** is a standard moment restriction. Note that if θ^* were known, $\mathbb{E}(\epsilon | \epsilon > u)$ would be a known function and we could trivially define $\hat{g}_n(\mathbf{x})$ as the inverse of $\hat{r}_n(\mathbf{x}) = m(\hat{g}_n(\mathbf{x}), \theta^*)$ by finding the root $u = \hat{g}_n(\mathbf{x})$ of:

$$\hat{r}_n(\mathbf{x}) = u + \mathbb{E}(\epsilon | \epsilon > -u) = 0$$

and we would have uniform consistency of $\hat{g_n}$ from uniform consistency of $\hat{r}_n$ given Lipschitz continuity of the inverse. On the other hand, if g were known θ^* could be estimated efficiently from the residuals $\epsilon_i = y_i - g(\mathbf{x})$ by maximum likelihood.

Given our initial kernel smooth of $r(\mathbf{x})$, θ defines $\hat{g}_n$, and any estimate of $\hat{g}_n$ will generate an estimate of θ^* through the estimated residuals $\hat{\epsilon_i} = y_i - \hat{g}_n(\mathbf{x_i})$. We will employ an EM-

type algorithm to alternate between estimation of θ^* given $\hat{g}_n$ and updating $\hat{g}_n$ through inversion of (2.3) given θ. We compute our estimator $\hat{g}_n$ by the following algorithm, which we initialize with $\hat{r_n}$, $\mathbb{E}_\theta$ refers to expectation with respect to F_θ.

Algorithm for the Backfit Estimator

- *Step 1:* An estimate $\hat{g_0}$ of g is used to form residuals $\hat{\epsilon}_i = y_i - \hat{g_0}(\mathbf{x}_i), i = 1 \ldots, n$.
- *Step 2:* The estimate $\hat{\theta}_{ML}$ is formed by maximum likelihood based on the residuals $\hat{\epsilon}_i$,
- *Step 3:* We find $\hat{g_1}$ as the root of $g(\mathbf{x}) - \hat{r}_n(\mathbf{x}) - \mathbb{E}_{\hat{\theta}_{ML}}(\epsilon | \epsilon > -g(\mathbf{x})) = 0$
- *Step 4:* Go to step 1 and repeat until $\hat{g}_1 = \hat{g}_0$.

Upon convergence, the algorithm finds a solution $\hat{\theta}_n$ to:

$$\hat{r}_n = m^{-1}_{\hat{\theta}_n}(\hat{r}_n) - \mathbb{E}_{\hat{\theta}_n}(\epsilon | \epsilon > -m^{-1}_{\hat{\theta}_n}(\hat{r}_n)) \tag{2.4}$$

and we estimate $\hat{g}_n = m^{-1}_{\hat{\theta}_n}(\hat{r}_n)$.

3 Consistency

The first main result establishes that the truth satisfies (2.4) asymptotically.

Theroem 2 *Given* **A1-A5**, $\hat{\theta}_{ML}(\theta^*, r) \rightarrow_p \theta^*$.

proof: At the truth θ^* and r, $\hat{\theta}_{ML}$ is the maximum likelihood estimator based on the censored sample $\epsilon_i = y_i - g(\mathbf{x}_i) = y_i - m^{-1}_{\theta^*}(r(\mathbf{x}_i)), i = 1 \ldots, n$ Continuity of f_θ and g implies continuity and hence measureablility and separability of $Q(y, \mathbf{x}, \theta) = log[\frac{f_\theta(y - g(\mathbf{x}))}{1 - F_\theta(-g(\mathbf{x}))}]$

for each $\theta \in \Theta$. Furthermore since $Q(y, \mathbf{x}, \theta)$ is dominated by **A5v.** and continuous with probability one by **A5iii.**, we have by lemma 1 of Tauchen (1985):

$$\frac{1}{n}\sum_{i=1}^{n} Q(y_i, \mathbf{x}_i, \theta) \to \mathbb{E}\{Q(y, \mathbf{x}, \theta)\}$$

almost surely, uniformly in θ. Moreover $\mathbb{E}\{Q(y, \mathbf{x}, \theta)\}$ is continuous in θ. The identification and dominance conditions **A5ii.** and **A5v.** ensure $\mathbb{E}\{Q(y, \mathbf{x}, \theta)\}$ is uniquely maximized at θ^* by the information inequality. The result follows by theorem 2.1 of Newey and McFadden (1994). □

Combined with theorem 1, this result suggests that an estimate based on $\hat{r}_n$ might be consistent as well if there are no other solutions to (2.4). The uniform convergence of $\hat{r_n}$ to r will be sufficient as long as some Lipschitz conditions hold. If there is a sequence of solutions which converges to a limit, theorem 3 establishes that the limit is consistent for θ^*. Before stating the result we add another identification condition:

A6: For all $\theta \in \Theta$,

$$\frac{1}{n}\sum_{i=1}^{n}[A_i(\theta) + B_i(\theta)]$$

has a probability limit of full rank, where

$$\begin{aligned} a_{j,k} = f_\theta^{-1}(\hat{\epsilon_i})\{\frac{\partial^2 f(\hat{\epsilon_i})}{\partial\theta_k\theta_j}\frac{-\partial m_\theta^{-1}(\hat{r_n}(\mathbf{x_i}))}{\partial\theta_j} + \frac{\partial f_\theta(\hat{\epsilon_i})}{\partial\theta_j}\frac{-\partial^2 m_\theta^{-1}(\hat{r_n}(\mathbf{x_i}))}{\partial_k\partial_j}\} \\ + \frac{\partial f_\theta(\hat{\epsilon_i})}{\partial\theta_j}\frac{-\partial m_\theta^{-1}(\hat{r_n}(\mathbf{x_i}))}{\partial\theta_j} f_\theta(\hat{\epsilon_i})^{-2}\{\frac{\partial f_\theta(\hat{\epsilon_i})}{\partial\theta_j}\frac{-\partial m_\theta^{-1}(\hat{r_n}(\mathbf{x_i}))}{\partial\theta_j}\} \end{aligned}$$

and,

$$
\begin{aligned}
b_{j,k} &= f_\theta(-m_\theta^{-1}(\hat{r_n}(\mathbf{x_i})))[1 - F_\theta(-m_\theta^{-1}(\hat{r_n}(\mathbf{x_i})))]^{-1}\frac{\partial^2 m_\theta^{-1}(\hat{r_n}(\mathbf{x_i}))}{\partial\theta_k\partial\theta_j} \\
&+ \frac{\partial f_\theta}{\partial\theta_k}\frac{-\partial m_\theta^{-1}(\hat{r_n}(\mathbf{x_i}))}{\partial\theta_k}\frac{-\partial m_\theta^{-1}(\hat{r_n}(\mathbf{x_i}))}{\partial\theta_j}[1 - f_\theta(-m_\theta^{-1}(\hat{r_n}(\mathbf{x_i})))]^{-1} \\
&+ f_\theta(-m_\theta^{-1}(\hat{r_n}(\mathbf{x_i})))^2\frac{-\partial m_\theta^{-1}(\hat{r_n}(\mathbf{x_i}))}{\partial\theta_j}\frac{-\partial m_\theta^{-1}(\hat{r_n}(\mathbf{x_i}))}{\partial\theta_k}[1 - F_\theta(-m_\theta^{-1}(\hat{r_n}(\mathbf{x_i})))]^{-2}
\end{aligned}
$$

are elements from the jth row and ith column from the pxp matrices $A_i(\theta)$ and $B_i(\theta)$ respectively.

While rather complicated, **A6** is verifiable given some parametric family f_θ as total differentiation of (2.3) with respect to θ reveals:

$$\frac{\partial m^{-1}(r(x),\tilde{\theta})}{\partial\theta} = \frac{-\frac{\partial\mathbb{E}_\theta(\epsilon|\epsilon>-u)}{\partial\theta}|_{u=m^{-1}[r(x),\tilde{\theta}]}}{1+\frac{\partial\mathbb{E}_\theta(\epsilon|\epsilon>-u)}{\partial u}|_{u=m^{-1}[r(x),\tilde{\theta}]}}$$

Hence given an initial smooth $\hat{r_n}$, a parametric family for the error term and Θ, simulation and the law of large numbers could serve to check **A6**. Given this additional condition and some Lipschitz restrictions, we can establish θ^* as the only limit to the backfit estimator.

Theroem 3 *Given* **A1 - A6**, $f_\theta(u), \frac{\partial f_\theta(u)}{\partial\theta}, m_\theta^{-1}(u), \frac{\partial m_\theta^{-1}(u)}{\partial\theta}$ *all Lipschitz in* u *for* $u \in U$ *in some neighborhood of* θ^**, and sequence of solutions* $\{\hat{\theta}_n\}$ *converging to* θ_0*:*

$$\theta_0 = \theta^* \text{ and,}$$

$$\hat{g}_n(\mathbf{x}) \to_p g(\mathbf{x})$$

proof:

The estimator $\hat{g_n}$ uniquely solves for all $\mathbf{x}$ in the support of the empirical distribution:

$$\hat{g_n}(\mathbf{x_i}) = \hat{r_n}(\mathbf{x_i}) - E_{\hat{\theta}_n}(\epsilon|\epsilon > -\hat{g_n}(\mathbf{x_i})) = m_{\hat{\theta}_n}^{-1}(\hat{r}_n(\mathbf{x_i})) \text{ for } i = 1,\ldots,n$$

where $\hat{\theta}_n$ is the maximum likelihood estimator based on $\hat{\epsilon}_i = y_i - \hat{g}_n(\mathbf{x}_i) = y_i - m^{-1}_{\hat{\theta}_n}(\hat{r}_n(x_i))$.

A Taylor's expansion of the score equations around θ^* gives:

$$S_n(\hat{\theta}_n) = 0 = S_n(\theta^*) + \nabla S_n(\bar{\theta})[\hat{\theta}_n - \theta^*]$$

where $\bar{\theta} = \lambda_n \theta^* + (1 - \lambda_n)\hat{\theta}$ for some $\lambda_n \in [0, 1]$, and

$$S_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\{\frac{\partial f_\theta(\hat{\epsilon_i})}{\partial \theta}\frac{\partial m^{-1}_\theta(\hat{r}_n(x))}{\partial \theta} f_\theta(\hat{\epsilon_i})^{-1} + f_\theta(-m^{-1}_\theta(\hat{r}_n(x)))\frac{\partial m^{-1}_\theta(\hat{r}_n(x))}{\partial \theta}[1 - F_\theta(-m^{-1}_\theta(\hat{r}_n(x)))]^{-1}\}$$

$$\nabla S_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} A_i(\theta) + B_i(\theta)$$

Lipschitz where A_i and B_i are as given in **A6**. By theorem 1, $\hat{r_n} \rightarrow_p r$ uniformly for all $\mathbf{x} : h(\mathbf{x}) > \delta\}$, for any $\delta \in (0, sup_{x \in R^k} h(\mathbf{x})]$. By the Lipschitz conditions, we may replace $\hat{r_n}$ by r in the equation for $S_n(\theta)$ above, since the two expressions differ by an average of terms which can be made arbitrarily small. Furthermore $S_n(\theta^*) \rightarrow_p 0$, since this is now the weighted score evaluated at the truth as $n \rightarrow \infty$. Hence $\hat{\theta}_n \rightarrow_p \theta^*$ so long as the probability limit of $\nabla S_n(\bar{\theta})$ is full rank which is given by **A6**. Hence $\hat{\theta} \rightarrow_p \theta^*$, and $\hat{g}(\mathbf{x}) \rightarrow_p g(\mathbf{x})$ by the Mann-Wald Theorem.

□

While we are unable to entirely rule out spurious solutions to (2.4), theorem 3 guarantees that any convergent sequence of solutions is the correct one. Moreover theorem 3 may be relaxed of the conditions if we replace $\hat{r_n}$ with r.

4 Finite Sample Performance

In order to judge the finite sample performance of our estimator, a monte carlo simulation was performed. The simulation was based on the following equation:

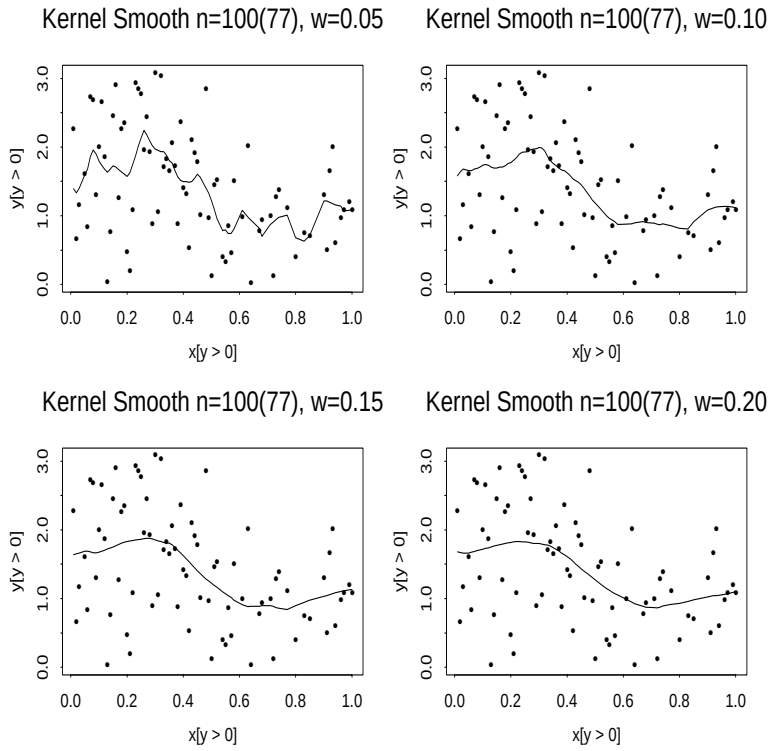
$$y = g(x) + \epsilon = \delta + \beta sin(\gamma x) + \epsilon \tag{4.1}$$

with threshold $\alpha = 0$ where x is uniformly distributed on the unit interval and ϵ is a standard normal deviate. The parameter values were set to $\beta = 1$ and $\gamma = 2\pi$ while δ was varied to achieve different expected levels of truncation (10%, 25% and 50%). Five hundred replications were performed at samples sizes of 100, 250, 500, 1000 and 5000. Bandwidths were set to 0.15, 0.10, 0.08, 0.08 and 0.06 for sample sizes 100, 250, 500, 1000 and 5000 respectively (bandwidths were increased slightly for the 50% truncation level). Each bandwidth was chosen subjectively by the author from examination of kernel smoothes for the first monte carlo sample at the various bandwidths. Plots of four kernel smoothes of the first sample of size 100 at 25% truncation are shown in figure 4. The Epanechnikov kernel, $K(u) = \frac{3}{4}(1 - u^2)I(|u| \leq 1)$ was employed.

Equation (4.1) obviously satisfies **A1-A3** while the Epanechnikov kernel satisfies **A4**. As $\theta = \sigma \in (0, \infty)$ in the present setting, assumptions **A5i** and **A5v** require σ be bounded from above and away from zero. **A5iv** corresponds to $\frac{\partial}{\partial u}[\sigma \frac{\phi(u/\sigma)}{\Phi(u/\sigma)}] > -1$ where ϕ and Φ denote the standard normal density and distribution respectively.

As a basis for comparison, we calculated three estimates for g: Our backfit estimator, our backfit estimator with known variance, and the maximum likelihood estimator initialized at the true parameter values. The backfit estimate with known variance will capture the rate bound achievable from kernel regression, while our MLE gives a sense for the overall

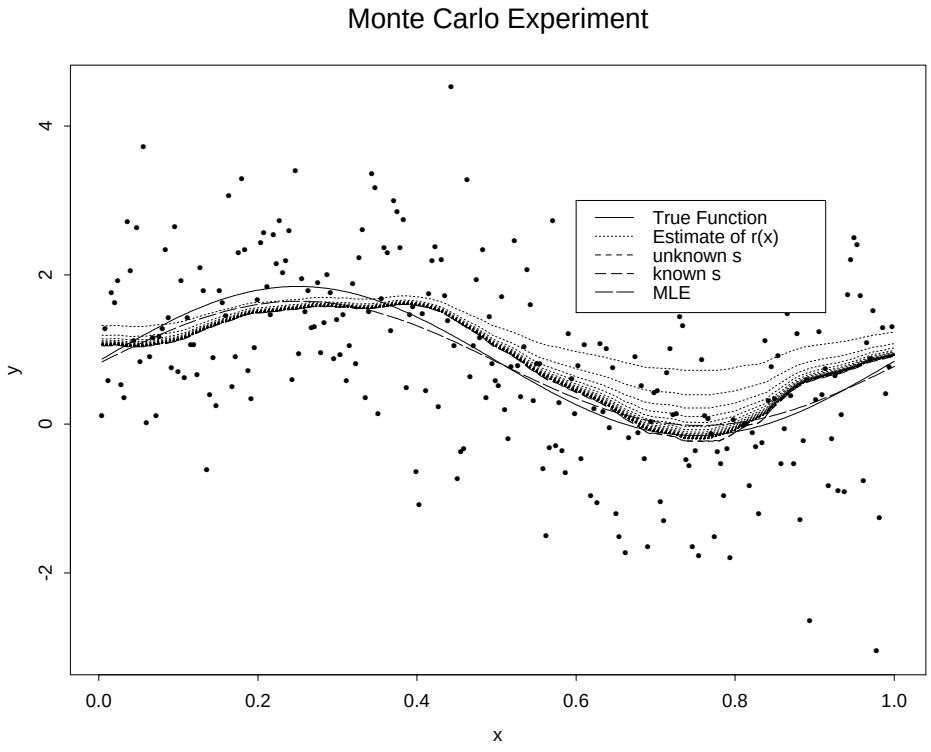
Figure 2: Four Bandwidths for the First sample



relative efficiency of our estimator. Since our function is C^{∞} the initial kernel smooth is bounded below, but may obtain, a $\sqrt{n}$ rate of convergence (Stone (1982)).

Figure 4 graphs the three estimators along with the true regression function for the first monte carlo sample at $n = 250$ and a 25% truncation rate. Iterations of the algorithm result in a "bending" of the curve towards the origin as the bias term is updated as the algorithm finds a convergent value of θ. The MLE estimate and the backfit with true $\theta = \theta^*$ are also displayed for comparison. A feature of the backfit estimator which becomes apparent is its tendency to exaggerate variations in $\hat{r}_n$ which suggests optimal rates for window widths may be slower than under conventional kernel regression.

Figure 3: Examples of the Three Estimators



In order to capture a global measure of convergence, the average absolute errors where computed and the results are given in the table 1 below.

Table 1: Monte Carlo Results

Average Absolute Error									
	Backfit			Backfit with known σ			MLE		
Expected % Truncated	10%	25%	50%	10%	25%	50%	10%	25%	50%
n=100	0.230	0.309[a]	0.574[a]	0.223	0.298	0.544[a]	0.180	0.263	0.560[b]
n=250	0.162	0.228	0.411[a]	0.158	0.219	0.358[a]	0.116	0.162	0.362
n=500	0.123	0.178	0.339[a]	0.124	0.169	0.299[a]	0.079	0.117	0.242
n=1000	0.095	0.130	0.247	0.093	0.121	0.209	0.056	0.086	0.184
n=5000	0.051	0.068	0.146	0.049	0.064	0.126	0.026	0.038	0.097

[a] Some estimates failed due to underflow.

[b] Two samples failed to converge.

Initially, we notice average errors vary (as expected) positively with the level of truncation and negatively with sample size for all three estimators. While increasing the truncation rate for a fixed sample effectively reduces the sample size, higher truncation rates have the effect of unbalancing the distribution of x— filtering very little data when g is high and filtering heavily when g is low. For example, a sample of 250 with 10% truncation is very close to a sample of 500 with 50% truncation in terms of expected sample size, but for all three estimators the latter sample has errors approximately twice the size as the former.

From examination of table 1, we see that the backfit estimator produces errors which are comparable to the MLE, but increasing in relative terms as the sample increases: from 2.5% larger for sample size 100 and 50% truncation up to 96% larger for a sample size of 5000

and 10% truncation. Moreover, the backfit estimator performs relatively better to MLE as conditions deteriorate. The backfit estimator gains ground in terms of relative average error as the sample decreases, and it performs better as truncation becomes more severe.

An interesting question is to examine the effects of estimating θ to error rates. The backfit estimator converges to the rate of the kernel estimator with known variance, indicating that estimation of σ has no effect on the precision of the backfit estimator asymptotically (although there is still a sizable difference at the highest truncation rate). Since the rate of the first stage kernel smooth is never faster than $\sqrt{n}$, we expect the limiting factor in precision to be the initial estimate of r.

While the algorithm to calculate the solution to the fixed point problem in (2.3) generally converged in a finite number of steps, calculation of the estimator failed in some samples due to computer underflow when the initial kernel smooth was near the boundary and $\hat{\sigma}^2$ is relatively small. This occurred in our experiment because inverting $r(x)$ necessitates inverting $\frac{\phi(u/\sigma)}{\Phi(u/\sigma)}$ near zero. Although this is not a problem asymptotically (see **A5iv**), it can be in finite samples. Here, this was only a problem for small samples with high levels of truncation. Failures accounted for 31.4% of the samples of size 100 and 50% truncation. they accounted for 11%, 1.6% and 0.8% for the 250(50%), 500(50%) and 100(25%) sample size-truncation cohorts respectively. Over-smoothing should help to alleviate this in applications.

5 Conclusion and Extensions

In this paper, we have developed a non-parametric estimator for the regression function in a truncated a sample when the error distribution is known up to a parameter vector. We have

established consistency results for the backfit estimator and found that it performs well in finite samples, even (and especially in terms of relative efficiency) in the presence of severe truncation. Our backfit estimator is easy to implement, relatively inexpensive with respect to computer resources and achieves reasonable average absolute errors in our monte carlo experiments without any optimization of window width. Higher relative efficiency should be achievable through cross-validation or some other optimizing procedure.

This work is a first step towards exploring a useful model for inference in truncated samples. While confidence bands may be easily calculated with a parametric bootstrap, asymptotic results for the backfit estimator would be of interest. In addition, the issues of bandwidth selection and optimal rates of convergence for the backfit estimator have not been touched upon. While the results developed here used a kernel smooth as the initial estimate for the truncated regression function any consistent non-parametric method robust to heteroskedasticity will do. For example, a series estimator would allow the addition of linear control variables within a restricted generalized additive model in applications where the dimensionality of $\mathbf{x}$ is high.

Appendix

Before moving to prove theorem 1, we establish two helpful lemmas. Define

$$a_n(x, \lambda) = \frac{1}{n} \sum_{i=1}^{n} y_i K(\frac{x - x_i}{\lambda})$$

as the numerator of our kernel estimator $\hat{r_n}$.

Lemma 1 *Given* **A1,A3,A4**,

$$\mathbb{E}sup_{\mathbf{x}\in R^k}|a_n(\mathbf{x},\lambda)-\mathbb{E}a_n(\mathbf{x},\lambda)| = O(n^{-1/2})$$

proof:

Since K has an absolutely integrable characteristic function $\beta(t) = \int_{\mathbb{R}^k} exp(i\mathbf{t}'\mathbf{x})K(\mathbf{x})dx$ we may invert it and substitute,

$$a_n(x,\lambda) = \frac{1}{n}(\frac{1}{2\pi})^k \sum_{i=1}^{n} y_i \int exp(-it'[\frac{x-x_i}{\lambda}])\beta(t)dt$$

and,

$$a_n(x,\lambda) - Ea_n(x,\lambda) = \lambda^k(\frac{1}{2\pi})^k\{\int \frac{1}{n}\sum_{i=1}^{n} y_i exp(x-x_i)\beta(\lambda t)dt - E\int \frac{1}{n}\sum_{i=1}^{n} y_i exp(x-x_i)\beta(\lambda t)dt\}$$

and with a change of variable,

$$\begin{aligned}
&\mathbb{E}\sup_{\mathbf{x}\in R^k}|a_n(\mathbf{x},\lambda)-\mathbb{E}a_n(\mathbf{x},\lambda)| \\
&\leq \lambda^k(\frac{1}{2\pi})^k \int_{R^k} E|\frac{1}{n}\sum_{i=1}^{n}[y_i exp(it'x_i) - Ey_i exp(it'x_i)]||\beta(\lambda t)|dt \\
&= \lambda^k(\frac{1}{2\pi})^k \int_{R^k} w_n(t)|\beta(\lambda t)|dt,
\end{aligned}$$

By Liapounov's inequality,

$$w_n(t) \leq var(\frac{1}{n}\sum y_i cos(t'x_i)) + var(\frac{1}{n}\sum y_i cos(t'x_i))$$

and independence with bounded second moments for $y_i I(y_i > 0)$ gives the result.

□

Lemma 2 *Let f be a bounded, uniformly continuous real function on $\mathbb{R}^k$. For every density K, on R^k we have,*

$$lim_{\lambda \to 0} sup_{x \in R^k} | \int \lambda^{-k} f(z) K(\frac{x-z}{\lambda}) dz - f(x)| = 0$$

The proof of lemma 2 is standard and left to the reader.

A.1 Proof of Theorem 1

proof: the supremum is measurable by **A3** (for measureability of r) and the fact that the set $\{x \in R^k : h(x) > \delta\}$ is compact combined with Lemma 1 of Jennrich (1969). Let $f(x) = g(x)h(x)$. Since f is bounded and uniformly continuous, by **A4** and lemma 2:

$$lim_{n \to \infty} sup_{x \in R^k} | E \frac{1}{n} \sum \lambda^{-k} y_i K(\frac{x-x_i}{\lambda}) - r(x)h(x)| = 0$$

combining this with lemma 1, we have:

$$lim_{n \to \infty} E sup_{x \in R^k} | \frac{1}{n} \sum \lambda^{-k} y_i K(\frac{x-x_i}{\lambda}) - r(x)h(x)| = 0 \qquad \text{(A.1)}$$

and replacing y_i by 1,

$$lim_{n \to \infty} E sup_{x \in R^k} | \frac{1}{n} \lambda^{-k} y_i K(\frac{x-x_i}{\lambda}) - h(x)| = 0 \qquad \text{(A.2)}$$

(A.1) and (A.2) combined with Chebeshev's inequality gives the result.

□

References

H. Ahn and J.L.Powell. Semiparametric estimation of censored selection models. *Journal of Econometrics*, 58:3–29, 1993.

S. Berry, J. Levinsohn, and A. Pakes. Automobile prices in market equilibrium. *Econometrica*, 63:841–890, 1995.

H. J. Bierens. Uniform consistancy of kernel estimators of a regression function under generlaized conditions. *Journal of the American Statistical Association*, 78:699–707, 1983.

J. Buckley and I. James. Linear regression with censored data. *Biometrika*, 66:429–436, 1982.

S. R. Cosslett. Semiparametric estimation of a regression model with sample selectivity. In *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, pages 175–197. Cambridge University Press, 1991.

G. Duncan. A semi-parametric censored regression estimator. *Journal of Econometrics*, 32:5–34, 1986.

L. Fernandez. Non-parametric maximum likelihood estimation of censored regression models. *Journal of Econometrics*, 32:35–58, 1986.

A.R. Gallant and D.W. Nychka. Semi-nonparametric maximum likelihood estimation. *Econometrica*, 55:363–390, 1987.

A. S. Goldberger. Linear regression after selection. *Journal of Econometrics*, 15:357–366, 1981.

W. H. Greene. *Econometric Analysis*. Prentice-Hall, 3rd edition, 1997.

J. L. Horowitz. A distribution-free least squares estimator for censored linear regression

model. *Journal of Econometrics*, 32:59–84, 1986.

H. Ichimura. Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, 58:71–120, 1993.

R. I. Jennrich. Asymptotic properties of non-linear least squares estimators. *Annals of Mathematical Statistics*, 40:633–43, 1969.

L. F. Lee. Semiparametric two-stage estimation of sample selection models subject to tobit-type selection rules. *Journal of Econometrics*, 61:305–344, 1994a.

L. F. Lee. Semiparametric instrumental variable estimation of simultaneous equation sample selection models. *Journal of Econometrics*, 63:305–344, 1994b.

W. K. Newey and D. McFadden. Large sample estimation and hypothesis testing. In *Handbook of Econometrics: Volume IV*, pages 2113–2241. Elsevier Science, 1994.

W. K. Newey. Linear instrumental variables estimation of limited dependent variables models with endogenous explanatory variables. *Journal of Econometrics*, 32:127–141, 1986.

W. K. Newey. Efficient estimation of tobit models under conditional symmetry. In *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, pages 291–336. Cambridge University Press, 1991.

J. L. Powell. Least absolute deviations estimation for the censored regression model. *Journal of Econometrics*, 25:303–325, 1984.

J. L. Powell. Symmetrically trimmed least squares estimation for tobit models. *Econometrica*, 54:1435–1460, 1986.

P. Ruud. Consistent estimation of limited dependent variable despite misspecification of

distribution. *Journal of Econometrics*, 32:157–187, 1986.

C. J. Stone. Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, 10:1040–53, 1982.

G. Tauchen. Diagnostic testing and evaluation of maximum likelihood models. *Journal of Econometrics*, 30:415–443, 1985.

K-L. Tsui, N. Jewell, and F. J. Wu. A nonparametric approach to the truncated regression problem. *Journal of the American Statistical Association*, 83:785–792, 1988.