SHIFT RESTRICTIONS AND SEMIPARAMETRIC ESTIMATION IN ORDERED RESPONSE MODELS

Roger W. Klein[*] *and Robert P. Sherman*[†]

[*]Rutgers University [†]California Institute of Technology

Abstract

We develop a $\sqrt{n}$-consistent and asymptotically normal estimator of the parameters (regression coefficients and threshold points) of a semiparametric ordered response model under the assumption of independence of errors and regressors. The independence assumption implies shift restrictions allowing identification of threshold points up to location and scale. The estimator is useful in various applications, particularly in new product demand forecasting from survey data subject to systematic misreporting. We apply the estimator to assess exaggeration bias in survey data on demand for a new telecommunications service.

KEYWORDS: Ordered response model, semiparametric estimation, shift restrictions, threshold points, survey data, transformation model.

INTRODUCTION

Consider the model $Y^* = \alpha_0 + X'\gamma_0 + U$ where Y^* is a latent response variable, α_0 is an intercept, $X = (X_0, X_1, \ldots, X_k)$ is a vector of nonconstant regressor variables, $\gamma_0 = (\gamma_{00}, \gamma_{01}, \ldots, \gamma_{0k})$ is a vector of slope coefficients, and U is an error term with distribution function F. In an ordered response model, we observe (Y, X) where

$$Y = \sum_{j=0}^{J+1} j\{t_{j-1} < Y^* \leq t_j\} \tag{1}$$

with $J < \infty$ and threshold points t_j satisfying $-\infty = t_{-1} < t_0 < \cdots < t_J < t_{J+1} = \infty$.

Suppose that U is independent of X, with mean zero and positive, finite variance σ_0^2. Further, suppose that the distribution function F is known up to the scale factor σ_0. Then it is possible to identify the regression coefficients and the threshold points up to location and scale. Specifically, the identifiable parameters are γ_0/σ_0 and $(t_j - \alpha_0)/\sigma_0$, $j = 0, 1, \ldots, J$. Maximum likelihood estimators of these parameters are $\sqrt{n}$-consistent and asymptotically normally distributed (Amemiya, 1985, pp.286-295). Common choices for F are the normal and logistic cumulative distribution functions. Under the assumption that F is known up to scale, Klein and Sherman (1997) develop a computationally attractive procedure for estimating these parameters. Their procedure can be used more generally to estimate transformation functions up to location and scale even when the response variable is censored and discrete.

In fully parametric estimation of the ordered response model, if the distribution function F is misspecified, then all the estimators mentioned above can be inconsistent. In our simulations, threshold point estimates seem especially sensitive to this distributional assumption. This leads to consideration of a semiparametric estimation approach.

Suppose, once again, that U and X are independent, but let F be unspecified. Rewrite the latent model as $Y^* = \alpha_0 + \beta_0(X_0 + \mathcal{X}'\theta_0) + U$ where $\beta_0 = \gamma_{00}$, $\mathcal{X} = (X_1, \ldots, X_k)$, and $\theta_0 = (\gamma_{01}, \ldots, \gamma_{0k})/\gamma_{00}$. There exists a number of $\sqrt{n}$-consistent and asymptotically normal semiparametric estimators of θ_0 in the literature. Examples include the estimators of Han (1987), Powell, Stock, and Stoker (1989), Ichimura (1993), and Cavanagh and Sherman (1998). All these estimators cover general single index models, of which the ordered response model in (1) is a special case.

However, to our knowledge, there exists no semiparametric estimator of the threshold points up to location and scale. This paper fills this gap by providing a $\sqrt{n}$-consistent and asymptotically normal estimator of (θ_0, Δ_0), where $\Delta_0 = (t_1 - t_0, \ldots, t_J - t_0)/\beta_0$. We estimate these parameters in two stages. The first stage estimator of θ_0 is an optimization estimator generalizing the binary response estimator of Klein and Spady (1993). The second stage estimator of Δ_0 is a direct estimator involving no optimization, and so is extremely fast to compute.

We are able to identify the threshold points, which act like intercept terms, by

exploiting special properties of single index models with additive errors that are independent of regressors. These properties are called *shift restrictions*. These restrictions enable us not only to identify the threshold points, but also to estimate them at the usual parametric rate.

Estimates of (θ_0, Δ_0) are useful in a variety of applications. For example, in marketing or political survey data where the response variable is an integer rating of a product or candidate, such estimates can be used to target demographic groups that are likely to fall in the various rating groups.

As another example, consider survey data on demand for new products or services where the response variable is projected level of usage. Nonnegative integer quantities Y are commonly reported. Moreover, systematic misreporting of projected usage levels seems to be a common phenomenon in surveys of this type. One way to model this data is with an ordered response model where Y^* represents accurate demand projections. Accurate projections are related to reported projections through the transformation $Y^* = \Lambda(Y)$, where Λ is a strictly increasing function. As we show below, $\Lambda(Y)$ values correspond to threshold points in an ordered response model. Thus, accurate usage and revenue forecasting require estimates of the threshold points as well as the regression parameters.

In fact, work on a survey eliciting projected usage of potential new telecommunications services provided the impetus for this paper. In this survey, respondents were asked to estimate the average number of times each month they would use each service. They reported nonnegative integer values. Moreover, for each service, there was strong evidence suggesting that many respondents systematically exaggerated: it appeared that the more they liked the service, the more they exaggerated their projected usage. We apply our estimator to this data to assess exaggeration bias and compare these results to the fully parametric procedure of Klein and Sherman (1997) mentioned above.

Finally, we note a duality, alluded to above, between ordered response models and transformation models that lets us use the estimator developed here to estimate the transformation function at points of interest without making parametric assumptions about either the error distribution or the transformation function. We also allow the

response variable to be censored and discrete.

Consider the transformation model $\Lambda(Y) = Y^*\{Y^* > c\}$, where Y is a continuous response variable, Λ is a strictly increasing function on the support of Y, $Y^* = \beta_0 V_0 + U$, $V_0 = X_0 + \mathcal{X}'\theta_0$, the error term U is independent of the regressor vector $X = (X_0, \mathcal{X})$, and $c \in [-\infty, \infty)$. Note that $c = -\infty$ corresponds to the standard transformation model. We observe (Y, X) in this model. Suppose we wish to estimate Λ at the points $y_0 < \cdots < y_J$ in the support of Y. Define $Z = j$ if $y_{j-1} < Y \leq y_j$, $j = 0, 1, \ldots, J$, where $y_{-1} = -\infty$. By the strict monotonicity of Λ, there exist points $t_0 < \cdots < t_J$ with $t_j = \Lambda(y_j)$, $j = 0, 1, \ldots, J$, such that

$$Z = \sum_{j=0}^{J+1} j\{t_{j-1} < Y^* \leq t_j\}$$

where $t_{-1} = -\infty$ and $t_{J+1} = \infty$. Thus, the threshold points in this ordered response model are the transformation function values of interest. The estimator developed in this paper can be applied to estimate these quantities. Note that censoring poses no problem for this estimator. We also note that a slight generalization of the model defined above allows Y to take on discrete values. Again, discrete Y values pose no problem for the estimator developed here. This estimator will be $\sqrt{n}$-consistent and asymptotically normal even when Y values are censored and discrete. This is crucial for applications like the one presented in Section 6, where Y is censored and integer-valued. This is also in contrast to the semiparametric estimators of Λ developed by Horowitz (1996), Ye and Duan (1997), and Gørgens and Horowitz (1999) which require a very smooth distribution for Y in order to achieve $\sqrt{n}$-consistency and asymptotic normality.

In the next section, we derive the shift restrictions on which the estimator of (θ_0, Δ_0) is based. We also give a high-level overview of the estimation procedure. In Section 3, we provide a detailed definition of the estimator and discuss the assumptions needed to establish its asymptotic properties. Sections 4 and 5 establish consistency and asymptotic normality. Simulation results are presented in Section 6. In Section 7, we apply the estimator to survey data on projected usage of a new telecommunications service. Section 8 summarizes. Proofs of technical results are collected in an appendix.

Shift Restrictions and Estimators

In this section, we derive the shift restrictions on which the estimator of Δ_0 is based. We also give a high-level description of the estimator of (θ_0, Δ_0). A more detailed description is given in the next section.

Let $S_Y = \{0, 1, \ldots, J\}$ denote the support of Y, and let j and k be points in S_Y. Define the scaled difference

$$\Delta_0(k, j) = (t_k - t_j)/\beta_0 \,.$$

Note $\Delta_0 = (\Delta_0(1,0), \ldots \Delta_0(J,0))$ and $\Delta_0(i,0) = \sum_{j=1}^{i} \Delta_0(j, j-1)$ for $i = 1, \ldots, J$. Thus, we can estimate Δ_0 up to location and scale by estimating $\Delta_0(j, j-1)$ for $j = 1, \ldots, J$ and then forming partial sums.

Suppose the latent response variable underlying the ordered response model has the form $Y^* = \alpha_0 + \beta_0(X_0 + \mathcal{X}'\theta_0) + U$ presented in the last section. Write V_0 for the index $X_0 + \mathcal{X}'\theta_0$. Assume that U and X are independent and V_0 has support $\mathbb{R}$. Write U^* for $U + \alpha_0$ and fix v in $\mathbb{R}$. The independence assumption implies that

$$\begin{aligned} \mathbb{P}\{Y \leq j \mid V_0 = v\} &= \mathbb{P}\{U^* \leq t_j - \beta_0 v\} \\ &= \mathbb{P}\{U^* \leq t_k - \beta_0(v + \Delta_0(k,j))\} \\ &= \mathbb{P}\{Y \leq k \mid V_0 = v + \Delta_0(k,j)\}. \end{aligned}$$

Note that without independence, the first and third equalities above need not hold. For each j in S_Y and v in $\mathbb{R}$, write $P_j(v)$ for the conditional probability function $\mathbb{P}\{Y \leq j \mid V_0 = v\}$. The relation

$$P_j(v) = P_k(v + \Delta_0(k,j)) \tag{2}$$

is called a *shift restriction*.[1] It says that if U and X are independent, then conditional

[1] In a personal communication prior to the writing of this paper, Whitney Newey suggested that this restriction might be exploited to develop an efficient estimator of θ_0 in the ordered response model. Here, the restriction is used to identify Δ_0 up to location and scale.

probability functions are equal up to a location shift given by $\Delta_0(k, j)$. In other words, the quantity $P_j(v)$ carries hidden information about Δ_0 that is revealed through (2). Figure 1 gives a graphical illustration of (2).

We will use (2) to construct a second-stage estimator of Δ_0, given a first-stage estimator of θ_0. We begin by developing the first-stage estimator. Let $\theta = (\theta_1, \ldots, \theta_k)$ and let Θ denote the parameter space for θ_0. Write $V(X, \theta)$ for $X_0 + \mathcal{X}'\theta$ and $P_j(v, \theta)$ for $I\!P\{Y \leq j \mid V(X, \theta) = v\}$. Note that $P_j(v) = P_j(v, \theta_0)$. Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ denote a sample of independent observations from model (1). Let $V_i(\theta) = V(X_i, \theta)$. Define $\hat{\theta}$ to be the maximizer, over Θ, of the quasi-likelihood function

$$\frac{1}{n} \sum_{i=1}^{n} \hat{\tau}(X_i) \sum_{j=0}^{J+1} \{Y_i = j\} \log \left[\hat{P}_j(V_i(\theta), \theta) - \hat{P}_{j-1}(V_i(\theta), \theta) \right] \tag{3}$$

where $\hat{P}_{-1}(\cdot, \theta) \equiv 0$, $\hat{P}_{J+1}(\cdot, \theta) \equiv 1$, $\hat{P}_j(v, \theta)$ is a kernel regression estimator of $P_j(v, \theta)$, $j = 0, 1, \ldots, J$, and the trimming function $\hat{\tau}$ rules out summands associated with poorly estimated conditional probabilities. The kernel estimators and the trimming function will be defined in detail in the next section. This estimator is a extension of the quasi-likelihood estimator of Klein and Spady (1993).

Turn to estimation of Δ_0 and reparametrize. Let $\delta_0 = (\delta_{01}, \ldots, \delta_{0J})$ where $\delta_{0j} = \Delta_0(j, j-1)$, $j = 1, \ldots, J$. Recall that $\Delta_0 = (\Delta_0(1, 0) \ldots, \Delta_0(J, 0))$ and $\Delta_0(i, 0) = \sum_{j=1}^{i} \delta_{0j}$, $i = 1, \ldots, J$. Thus, an estimator of δ_0 can be used to produce an estimator of Δ_0 by forming partial sums. Write $\hat{V}_i$ for $V(X_i, \hat{\theta})$. From (2), for each i,

$$P_j(\hat{V}_i + \delta_{0j}) = P_{j-1}(\hat{V}_i)\,. \tag{4}$$

Notice that the difference in the arguments of the conditional probability functions in (4) is equal to δ_{0j}. Write $\hat{P}_j(v)$ for $\hat{P}_j(v, \hat{\theta})$. For each i, define $\hat{V}_{ij}$ as a point for which

$$\hat{P}_j(\hat{V}_{ij}) = \hat{P}_{j-1}(\hat{V}_i)\,.$$

A natural estimator of δ_{0j} is $\hat{V}_{ij} - \hat{V}_i$, the difference in the arguments of the estimated conditional probability functions. For a given i, the estimator $\hat{V}_{ij} - \hat{V}_i$ does not converge

to δ_{0j} at a $\sqrt{n}$ rate. To achieve $\sqrt{n}$-consistency and asymptotic normality, we average a subset of these individual estimators. This subset is chosen to exclude individual estimators that perform poorly in small samples. After a relabeling of indices, the estimator of δ_{0j} is given by

$$\hat{\delta}_j = \frac{1}{m}\sum_{i=1}^{m}(\hat{V}_{ij} - \hat{V}_i) \tag{5}$$

where m is the cardinality of the subset described above. The estimator of δ_0 is given by $\hat{\delta} = (\hat{\delta}_1 \ldots, \hat{\delta}_J)$. As with other estimators of this type, a grid search is used to approximate the $\hat{V}_{ij}$'s. In the next section, we describe an algorithm for producing a grid that is fine enough to ensure $\sqrt{n}$-consistency and asymptotic normality but coarse enough to make computation very fast.

Assumptions and Definitions

In this section, we present assumptions and definitions used to establish the asymptotic properties of the estimator of (θ_0, δ_0) introduced in the last section.

Recall that $X = (X_0, \mathcal{X})$, $Y^* = \alpha_0 + \beta_0 V_0 + U$ where $V_0 = V(X, \theta_0) = X_0 + \mathcal{X}'\theta_0$, and (Y, X) denotes a generic observation from model (1). Assumption A1 describes the parameter spaces for θ_0 and δ_0. A2 restates assumptions about the data, and in particular, the key assumption of independence between errors and regressors. A3 gives regularity conditions on the conditional distribution of X_0 given $\mathcal{X}$ and Y. The support condition is used with the trimming scheme (see D5 below) in establishing the limiting distribution of the estimator of θ_0.[2] The smoothness and boundedness conditions help control bias in kernel regression estimators defined below.

A1. Parameter Spaces. θ_0 is an interior point of Θ, a compact subset of $\mathbb{R}^k$. δ_0 is an interior point of D, a compact subset of the strictly positive orthant of $\mathbb{R}^J$.

[2]This support condition is made for analytic convenience and can be dispensed with provided a more complicated trimming scheme is used. For example, if observations are trimmed when estimated densities are small (as is done in Klein and Spady, 1993), then no restrictions on the support are needed. The asymptotic results still hold, but the proofs are more complicated and a degradation in small sample performance can be expected.

A2. Data. $(Y_1, X_1), \ldots, (Y_n, X_n)$ are iid observations from model (1). In addition, each X_i is independent of U_i.

A3. Conditional Densities. Write f for the conditional density of X_0 given $\mathcal{X}$ and Y. On any compact set, f is strictly positive and has bounded derivatives up to order 4.

Recall that $P_j(v, \theta) = I\!P\{Y \leq j \mid V(X, \theta) = v\}$ where $V(X, \theta) = X_0 + \mathcal{X}'\theta$. Write $g_1(\cdot, \theta)$ for the density of $V(X, \theta)$ given $Y \leq j$ and $g_0(\cdot, \theta)$ for the density of $V(X, \theta)$ given $Y > j$, where, for ease of notation, we have suppressed the dependence of these densities on j. Define $p_1 = I\!P\{Y \leq j\}$ and $p_0 = 1 - p_1$. By Bayes' rule,

$$P_j(v, \theta) = p_1 g_1(v, \theta) / [p_1 g_1(v, \theta) + p_0 g_0(v, \theta)] .$$

We develop a kernel regression estimator of $P_j(v, \theta)$ by constructing kernel density estimators of the component densities $g_i(v, \theta)$, $i = 0, 1$.

The kernel density estimators are defined in D1 through D3 using local smoothing techniques for bias reduction developed by Abramson (1982) and Silverman (1986). D1 defines the pilot density estimators used to construct estimated local smoothing parameters in D3. D2 defines smooth functions used in D3 to control the rate at which estimated local smoothing parameters converge to zero. By this means, we allow local windows to open arbitrarily wide, but not too fast. Thus, we reap the benefits of local smoothing on a widening support, and at the same time achieve better small sample performance by preventing local windows from opening too wide in the tails of the conditional distributions. D4 defines the kernel estimator of $P_j(v, \theta)$. D5 defines the trimming function used in first-stage estimation of θ_0.

D1. Pilot Density Estimators. Let K denote a symmetric kernel density function with rth derivatives bounded and integrable, $r = 0, 1, 2, 3, 4$. Fix $\delta \in (0, 1/3)$ and define the pilot window $h_p = n^{-\gamma}$ where $1/10 < \gamma < 1/3(3 + \delta)$.[3] Fix j in

[3]The pilot window h_p in D1 is wider than the second stage window h in D3. This is needed to control asymptotic bias in density estimates (see Lemma 1A through Lemma 5A in the appendix.). This particular range is chosen to make relatively short work of proving that pilot density estimators can be taken as known in asymptotic arguments (see Lemma 4A in the appendix.).

S_Y and recall that $V_k(\theta) = V(X_k, \theta)$. Let $n_1 = \sum_k \{Y_k \leq j\}$ and write $\hat{\sigma}_1(\theta)$ for the sample standard deviation of the $V_k(\theta)\{Y_k \leq j\}$'s. Define the pilot density estimator of $g_1(V_i(\theta), \theta)$ to be

$$\hat{\pi}_{1i}(\theta) = \frac{1}{n_1} \sum_{k \neq i} \{Y_k \leq j\} K([V_i(\theta) - V_k(\theta)]/\hat{\sigma}_1(\theta)h_p)/\hat{\sigma}_1(\theta)h_p \,.$$

Define $\hat{\pi}_{0i}(\theta)$ in like fashion, replacing $\{Y_k \leq j\}$ with $\{Y_k > j\}$. The resulting $\hat{\pi}_{0i}(\theta)$ is a pilot density estimator of $g_0(V_i(\theta), \theta)$.

D2. Smooth Damping. Refer to D1. Write $\hat{m}_1(\theta)$ for the geometric mean of the $\hat{\pi}_{1i}(\theta)$'s. Define estimated local smoothing parameters

$$\hat{l}_{1i}(\theta) = \hat{\pi}_{1i}(\theta)/\hat{m}_1(\theta) \,.$$

Choose $\epsilon \in (0, 1/40 - \delta/20)$ and take $a_{n_1} \propto [\ln n_1]^{-1}$. For $l > 0$, define the damping function

$$\hat{d}_1(l) = 1/[1 + \exp(-n_1^{\epsilon}[l - a_{n_1}])] \,.$$

Note that $\hat{d}_1(l)$ smoothly approximates the indicator function $\{l > a_{n_1}\}$. Define $\hat{d}_0(l)$ analogously.

D3. Locally Smoothed Density Estimators. Refer to D1 and D2. Write $\hat{d}_{1i}(\theta)$ for $\hat{d}_1(\hat{l}_{1i}(\theta))$ and define damped estimated local smoothing parameters

$$\hat{\lambda}_{1i}(\theta) = \hat{\sigma}_1(\theta) \left[\hat{l}_{1i}(\theta)\hat{d}_{1i}(\theta) + a_{n_1}\left[1 - \hat{d}_{1i}(\theta)\right]\right]^{-1/2} .$$

Choose $\alpha \in ((3 + \delta)/20, 1/6)$. Let $h = n^{-\alpha}$. Define

$$\hat{g}_1(v, \theta) = \frac{1}{n_1} \sum_k \{Y_k \leq j\} K([v - V_k(\theta)]/\hat{\lambda}_{1k}(\theta)h)/\hat{\lambda}_{1k}(\theta)h \,.$$

Define $\hat{g}_0(v, \theta)$ analogously, replacing $\{Y_k \leq j\}$ with $\{Y_k > j\}$.

D4. Estimated Conditional Probability Functions. Refer to D3. Let $\hat{p}_i = n_i/n$,

$i = 0, 1$. Define

$$\hat{P}_j(v, \theta) = \hat{p}_1 \hat{g}_1(v, \theta) / \left[\hat{p}_1 \hat{g}_1(v, \theta) + \hat{p}_0 \hat{g}_0(v, \theta)\right] .^4$$

D5. Quasi-Likelihood Trimming. Refer to (3) in Section 2. Fix $q \in (1/2, 1)$ and write $\hat{\xi}_q$ for the sample qth quantile of the $|X_i|$'s, $i = 1, \ldots, n$. For $x \in I\!R^{k+1}$, define the trimming function

$$\hat{\tau}(x) = \{|x| \leq \hat{\xi}_q\} .$$

A3 implies that for each $\theta \in \Theta$, the densities $g_i(V(x, \theta), \theta)$, $i = 0, 1$, which appear in the denominator of $P_j(V(x, \theta), \theta)$, approach zero if and only if $|x|$ approaches infinity. Thus, $\hat{\tau}$ rules out observations associated with poorly estimated tail probabilities.

Next, consider $\hat{\delta} = (\hat{\delta}_1, \ldots, \hat{\delta}_J)$, the second-stage estimator of δ_0.

Recall that $\hat{V}_i = V(X_i, \hat{\theta})$, where $\hat{\theta}$ is the first-stage estimator of θ_0 and $\hat{P}_j(v) = \hat{P}_j(v, \hat{\theta})$. Refer to (5) in Section 2 and the discussion preceding it. As noted there, in order to construct $\hat{\delta}_j$, for each $\hat{V}_i$ in a target set $\mathcal{T}$ we must find a number $\hat{V}_{ij}$ such that $\hat{P}_j(\hat{V}_{ij}) = \hat{P}_{j-1}(\hat{V}_i)$. Since achieving equality in the estimated probabilities is impracticable, we construct a fixed set of grid points $\mathcal{G}$ and approximate each $\hat{V}_{ij}$ with a point $\tilde{V}_{ij}$ from $\mathcal{G}$. That is, for each $\hat{V}_i$ in $\mathcal{T}$, we find a point $\tilde{V}_{ij}$ in $\mathcal{G}$ such that $\hat{P}_j(\hat{V}_{ij})$ is closest to $\hat{P}_{j-1}(\tilde{V}_i)$. The average of the $\tilde{V}_{ij} - \hat{V}_i$'s approximates $\hat{\delta}_j$. The target set $\mathcal{T}$ is designed to rule out individual estimators that perform poorly in small samples. The grid $\mathcal{G}$ is designed to be fine enough to ensure $\sqrt{n}$-consistency and asymptotic normality, but coarse enough to make computation very fast.

To determine $\mathcal{T}$ and $\mathcal{G}$, we first compute $\hat{P}_{j-1}(\hat{V}_i)$ for each $\hat{V}_i$ in the sample. This defines a set of estimated probabilities with a range in $(0, 1)$. We do the same for the $\hat{P}_j(\hat{V}_i)$'s and then intersect the two ranges. Next, we trim a proportion p of the

[4]One could estimate $P_j(v, \theta)$ with an ordinary kernel regression estimator with local bandwidths or any nonparametric estimator that achieves the required bias reduction. We estimate component densities with their own local bandwidths. As suggested in Silverman (1986), this approach should produce an estimate of the marginal density of $V(X, \theta)$ (and ultimately, an estimate of $P_j(v, \theta)$) with better small sample properties.

smallest and largest of the probability estimates from the set of estimates that lie in the intersection. This determines a new interval of probability estimates. We take $\mathcal{T}$ to be the set of $\hat{V}_i$'s for which $\hat{P}_{j-1}(\hat{V}_i)$ falls in this interval, and $\mathcal{G}$ to be (roughly) the set of $\hat{V}_i$'s for which $\hat{P}_j(\hat{V}_i)$ falls in this interval. This construction guarantees that each $\hat{V}_i$ in $\mathcal{T}$ has a counterpart $\tilde{V}_{ij}$ in $\mathcal{G}$ for which $\hat{P}_j(\tilde{V}_{ij})$ is close to $\hat{P}_{j-1}(\hat{V}_i)$.

The ideas in the last paragraph are captured precisely in D6 and D7 below. Note that D7 augments $\mathcal{G}$ with a few extra points to ensure that the grid is fine enough so that each $\tilde{V}_{ij}$ is within $o_p(n^{-1/2})$ of the corresponding $\hat{V}_{ij}$. Also, note that computing all the $\tilde{V}_{ij}$'s in the manner described above is extremely fast, involving a comparison of at most $|\mathcal{T}||\mathcal{G}|$ pairs of fixed numbers, where $|A|$ denotes the cardinality of the set A. Finally, D8 formally defines the approximator of $\hat{\delta}_j$.

D6. The Target Set. Consider the interval

$$[l, u] = [\max\{\min_i \hat{P}_{j-1}(\hat{V}_i), \min_i \hat{P}_j(\hat{V}_i)\}, \min\{\max_i \hat{P}_{j-1}(\hat{V}_i), \max_i \hat{P}_j(\hat{V}_i)\}] .$$

Fix $p \in (0, 1/2)$. Let $\hat{P}_L$ denote the pth quantile of the estimated probabilities ($\hat{P}_{j-1}(\hat{V}_i)$'s and $\hat{P}_j(\hat{V}_i)$'s) that fall in the interval $[l, u]$. Let $\hat{P}_U$ denote the corresponding $(1-p)$th quantile. Define the target set

$$\mathcal{T} = \{\hat{V}_i : \hat{P}_L \leq \hat{P}_{j-1}(\hat{V}_i) \leq \hat{P}_U\} .$$

D7. The Grid. Refer to D6. Define $\mathcal{R} = \{\hat{V}_i : \hat{P}_L \leq \hat{P}_j(\hat{V}_i) \leq \hat{P}_U\}$. Let $\hat{V}_L$ denote the largest estimated index value smaller than the smallest estimated index value in $\mathcal{R}$. Let $\hat{V}_U$ denote the smallest estimated index value larger than the largest estimated index value in $\mathcal{R}$. Fix $\gamma > 1/2$ and define a grid of points

$$\mathcal{P} = \{\hat{V}_L + [\hat{V}_U - \hat{V}_L]j/n^\gamma, j = 0, 1, 2, \ldots, n^\gamma\} .$$

Notice that $\mathcal{P}$ contains the points $\hat{V}_L$ and $\hat{V}_U$ and that the distance between adjacent points in $\mathcal{P}$ is $o_p(n^{-1/2})$. Define the grid

$$\mathcal{G} = \mathcal{R} \cup \mathcal{P} .$$

All the points in $\mathcal{G}$ are contained in the interval $[\hat{V}_L, \hat{V}_U]$. Therefore, $\mathcal{G}$ is a finer partition of $[\hat{V}_L, \hat{V}_U]$ than $\mathcal{P}$. Deduce that adjacent points in $\mathcal{G}$ are a distance $o_p(n^{-1/2})$ apart, uniformly over $\mathcal{G}$. Also, note that for each $\hat{V}_i$ in $\mathcal{T}$, there exists a point $\hat{V}_{ij}$ in $[\hat{V}_L, \hat{V}_U]$ such that $\hat{P}_j(\hat{V}_{ij}) = \hat{P}_{j-1}(\hat{V}_i)$ This follows from the construction of $\mathcal{T}$ and $\mathcal{G}$ and the continuity of $\hat{P}_j(v)$ as a function of v.

D8. The Approximator of $\hat{\delta}_j$. Refer to D6 and D7. For each $\hat{V}_i$ in $\mathcal{T}$, define $\tilde{V}_{ij}$ to be a point in $\mathcal{G}$ for which $\hat{P}_j(\tilde{V}_{ij})$ is closest to $\hat{P}_{j-1}(\hat{V}_i)$. Write $\hat{t}(\cdot)$ for the indicator function of the convex hull of $\mathcal{T}$ and $\hat{\rho}$ for $\sum_{i=1}^n \hat{t}(\hat{V}_i)/n$. Define

$$\hat{d}_j = \frac{1}{n\hat{\rho}} \sum_{i=1}^{n} \hat{t}(\hat{V}_i)(\tilde{V}_{ij} - \hat{V}_i)\,.$$

Note that $\hat{d}_j$ is just the average of the $\tilde{V}_{ij} - \hat{V}_i$'s for $\hat{V}_i$'s in $\mathcal{T}$.

Consistency

Consistency of $\hat{\theta}$ follows from standard uniform convergence and identification results similar to those established in Klein and Spady (1993). The details are given in the appendix. We get the following result:

Theorem 1. *If A1 through A3 hold, then $|\hat{\theta} - \theta_0| = o_p(1)$ as $n \to \infty$.*

Refer to D6 through D8 in Section 3. Recall that for each $\hat{V}_i$ in the target set $\mathcal{T}$ there is a $\tilde{V}_{ij}$ in the grid set $\mathcal{G}$ such that $\hat{P}_j(\tilde{V}_{ij})$ is closest to $\hat{P}_{j-1}(\hat{V}_i)$. Further, there exists a point $\hat{V}_{ij}$ in the interval $[\hat{V}_L, \hat{V}_U]$ containing $\mathcal{G}$ such that $\hat{P}_j(\hat{V}_{ij}) = \hat{P}_{j-1}(\hat{V}_i)$. Note that

$$\hat{\delta}_j = \frac{1}{n\hat{\rho}} \sum_{i=1}^{n} \hat{t}(\hat{V}_i)(\hat{V}_{ij} - \hat{V}_i)\,.$$

The asymptotic properties of $\hat{\delta}_j$ are relatively easy to establish. The following asymptotic equivalence result allows us to infer the first-order asymptotic properties of $\hat{d}_j$ from those of $\hat{\delta}_j$. The proof is given in the appendix.

Lemma 1. *If A1 through A3 hold, then $\sqrt{n}(\hat{d}_j - \hat{\delta}_j) = o_p(1)$ as $n \to \infty$.*

By Lemma 1, $\hat{d}_j$ is consistent if $\hat{\delta}_j$ is consistent. The following representation, established in the appendix, will be useful in proving consistency of $\hat{\delta}_j$.

LEMMA 2. *If A1 through A3 hold, then* $\hat{\delta}_j - \delta_{0j}$ *equals*

$$\frac{1}{n\hat{\rho}} \sum_{i=1}^{n} \hat{t}(\hat{V}_i)([\hat{P}_{j-1}(\hat{V}_i) - P_{j-1}(V_i)] - [\hat{P}_j(\hat{V}_i + \delta_{0j}) - P_j(V_i + \delta_{0j})])/\hat{P}'_j(V_{ij}^+)$$

where V_{ij}^+ *is between* $\hat{V}_{ij}$ *and* $\hat{V}_i + \delta_{0j}$.

The next result follows directly from Lemma 1, Lemma 2, and uniform convergence results for estimated probability functions and their derivatives established in the appendix.

THEOREM 2. *If A1 through A3 hold, then* $|\hat{d}_j - \delta_{0j}| = o_p(1)$ *as* $n \to \infty$.

Asymptotic Normality

As with consistency, $\sqrt{n}$-consistency and asymptotic normality of $\hat{\theta}$ follows in a straightforward manner from standard uniform convergence arguments that allow us to replace estimated probability functions in (3) with their estimands. Standard Taylor expansion arguments are then applied to prove $\sqrt{n}$-consistency and asymptotic normality. Let $Q(\theta)$ denote the expected value of the criterion function in (3) after estimated probability functions have been replaced by their estimands. Let $H(\theta)$ denote the hessian of $Q(\theta)$. A proof of the following result is given in the appendix. The symbol $\Longrightarrow$ denotes convergence in distribution.

THEOREM 3. *If A1 through A3 hold, then, as* $n \to \infty$,

$$\sqrt{n}(\hat{\theta} - \theta_0) \Longrightarrow N(0, -[H(\theta_0)]^{-1}) .$$

The rest of this section is devoted to giving an overview of the argument for $\sqrt{n}$-consistency and asymptotic normality of $\hat{d}_j$.

We begin with a useful characterization. Refer to D6 and D8 and note that as $n \to \infty$ the indicator function of the convex hull of $\mathcal{T}$ converges in probability to the indicator function $t(v) = \{a \leq v \leq b\}$ where a and b are real numbers satisfying $-\infty < a < b < \infty$. Also, note that $\hat{\rho}$ converges in probability to $\rho = P\{a \leq V_0 \leq b\}$ as $n \to \infty$.

LEMMA 3. *If A1 through A3 hold, then as* $n \to \infty$, $\hat{d}_j - \delta_{0j}$ *equals*

$$[D_{j-1}(\theta_0) - D_j(\theta_0)] + [M_{j-1}(\theta_0) - M_j(\theta_0)]\,(\hat{\theta} - \theta_0) + o_p(n^{-1/2})$$

where $D_k(\theta)$ *equals*

$$\frac{1}{n\rho}\sum_{i=1}^{n} t(V_i)[\hat{P}_k(V(X_i,\theta) + \Delta_0(k, j-1), \theta) - P_k(V(X_i,\theta) + \Delta_0(k, j-1), \theta)]/P_j'(V_i + \delta_{0j})$$

and $M_k(\theta) = I\!E \nabla_\theta D_k(\theta)$.

This characterization, which is proved in the appendix, follows from Lemma 1, Lemma 2, condition (2), and standard uniform convergence arguments.

In the appendix, we show that $D_{j-1}(\theta_0) - D_j(\theta_0)$ behaves asymptotically like a U-statistic. By a standard projection argument, this can be written as a average of iid random variables plus a negligible remainder. From Theorem 3, we see that $[M_{j-1}(\theta_0) - M_j(\theta_0)]\,(\hat{\theta} - \theta_0)$, which accounts for estimation uncertainty, has a similar characterization. These characterizations are given in Lemma 4 below.

Let $g(\cdot)$ denote the marginal density of V_0. For each v in $I\!R$ and k in S_Y, write $w_k(v)$ for $g(v)/[g(v + \Delta_0(k, j-1))P_k'(v)]$. Write $\tilde{G}(\theta)$ for the gradient of the quasi-likelihood function in (3) with all hats removed.

LEMMA 4. *If A1 through A3 hold, then as* $n \to \infty$

(i) $D_k(\theta_0) = \mathcal{A}_{1k}(\theta_0) + o_p(n^{-1/2})$ *where*

$$\mathcal{A}_{1k}(\theta_0) = \frac{1}{n\rho}\sum_{i=1}^{n} t(V_i - \Delta_0(k, j-1))[\{Y_i \leq k\} - P_k(V_i)]w_k(V_i - \Delta_0(k, j-1))$$

(ii) $M_k(\theta_0)(\hat{\theta} - \theta_0) = \mathcal{A}_{2k}(\theta_0) + o_p(n^{-1/2})$ *where*

$$\mathcal{A}_{2k}(\theta_0) = -M_k(\theta_0)\,[H(\theta_0)]^{-1}\,\tilde{G}(\theta_0)\,.$$

We now give a formal statement of the asymptotic normality result, proved in the appendix. Write $\mathcal{A}_j(\theta_0)$ for $\mathcal{A}_{1j}(\theta_0) + \mathcal{A}_{2j}(\theta_0)$, γ_j for $\sqrt{n}(\mathcal{A}_j(\theta_0) - \mathcal{A}_{j-1}(\theta_0))$, and σ_j^2 for $I\!E\gamma_j^2$.

THEOREM 4. *If A1 through A3 hold, then, as* $n \to \infty$,

$$\sqrt{n}(\hat{d}_j - \delta_{0j}) \Longrightarrow N(0, \sigma_j^2)\,.$$

Recall that $\Delta_0 = (\Delta_0(1,0)\ldots, \Delta_0(J,0))$ and that for each $i = 1, \ldots, J$, $\Delta_0(i,0) = \sum_{j=1}^{i} \Delta_0(j, j-1)$. Write $\hat{d}$ for $(\hat{d}_1, \ldots \hat{d}_J)$ and $\hat{\Delta}$ for the vector whose ith component equals $\sum_{j=1}^{i} \hat{d}_j$, $i = 1, \ldots, J$. It is easy to show that $\sqrt{n}(\hat{d} - \delta_0)$ converges in distribution to a normal random vector with mean zero and $J \times J$ covariance matrix Ω with ijth entry equal to $I\!E\gamma_i\gamma_j$.

Write C for the $J \times J$ matrix with zeros above the main diagonal and ones elsewhere. Note that $C\hat{d} = \hat{\Delta}$ and $C\delta_0 = \Delta_0$. It follows that $\sqrt{n}(\hat{\Delta} - \Delta_0)$ converges in distribution to a normal random vector with mean zero and covariance matrix $C'\Omega C$.

SIMULATION RESULTS

This section explores through simulations some aspects of the finite sample performance of the two-stage estimator of (θ_0, Δ_0) in a semiparametric ordered response model.

The latent variable model in the simulations has the form

$$Y^* = \alpha_0 + \beta_0(X_0 + \theta_0 X_1) + U$$

where Y^* is the latent response variable, X_0 and X_1 are explanatory variables, α_0, β_0,

and θ_0 are unknown parameters, and U is an error term, independent of $X = (X_0, X_1)$. We observe (Y, X) where

$$Y = 0\{-\infty < Y^* \leq t_0\} + 1\{t_0 < Y^* \leq t_1\} + 2\{Y^* > t_1\}\,.$$

In each simulation, $\alpha_0 = \beta_0 = \theta_0 = 1$. Since $\beta_0 = 1$, $\Delta_0 = t_1 - t_0$. The objects of estimation are the slope parameter θ_0 and the threshold point difference Δ_0.

The object of the simulations is to study the accuracy and precision of the two-stage estimator of (θ_0, Δ_0) under different error distributions, different index distributions, different threshold points and threshold point differences, and different sample sizes. The error distributions are either normal (N) or nonnormal (NN), the index distributions are either symmetric (S) or asymmetric (A), the threshold points are either symmetrically or asymmetrically placed around the median of the Y^* distribution, and the sample sizes are 500 or 1000.

More specifically, the error term U has either a standard normal distribution (skewness = 0, kurtosis = 3) or a $\chi^2(4)$ distribution (skewness $\approx$ 1.5, kurtosis $\approx$ 6), standardized to have zero mean and unit variance. In all designs, X_1 is the difference between two $\chi^2(1)$ variates, standardized to have mean zero and unit variance. The index distribution is either symmetric (X_2 is standard normal) or asymmetric (X_2 is standardized $\chi^2(1)$). The threshold points t_1 and t_2 are either (i) the 50th and 80th, (ii) the 35th and 65th, (iii) the 60th and 70th, or (iv) the 45th and 55th percentiles of the Y^* distribution. When U is standard normal and the index is symmetric, the value of the threshold difference Δ_0 is approximately 1.38 in (i), 1.28 in (ii), .47 in (iii), and .40 in (iv). When U is $\chi^2(4)$ and the index is symmetric, Δ_0 is approximately 1.23 in (i), 1.06 in (ii), .39 in (iii) and .34 in (iv). When U is standard normal and the index is asymmetric, Δ_0 is approximately 1.36 in (i), 1.15 in (ii), .42 in (iii) and .39 in (iv). Finally, when U is $\chi^2(4)$ and the index is asymmetric, Δ_0 is approximately 1.34 in (i), .92 in (ii), .41 in (iii) and .29 in (iv).

A word about scaling. In order to compare slope and threshold difference estimates on the same scale across all designs, we divided all threshold difference estimates by their true values before computing bias and variance estimates. So, for example, all

bias calculations for slope and threshold difference estimates alike are made relative to the true value of unity.

Figure 2 presents results on threshold point difference estimation under all designs. The first row presents results for the model with nonnormal errors and asymmetric index (NNA), for the 4 threshold point designs. Note that the threshold point differences, Δ_0, decrease from left to right. The second row presents results for the model with normal errors and asymmetric index (NA). The third row corresponds to the model with nonnormal errors and symmetric index (NNS) while the fourth row corresponds to the model with normal errors and symmetric index (NS). By comparison, the columns fix the threshold points and vary the error and index distributions.

In order to better understand Figure 2, focus on the scatterplot in the first row and second column of the figure, corresponding to model NNA, 35–65. This scatterplot represents summary statistics for the second-stage estimator of the scaled threshold difference $\Delta_0/.92$ where $\Delta_0 = t_1 - t_0 = .92$. Solid points represent summary statistics for the estimator. As a visual aid, points are connected with a solid line. For example, the solid line closest to the zero line represents standard error (SE) estimates, based on 100 replications, for the second-stage estimator of $\Delta_0/.92$ for sample sizes of 500 and 1000. The solid line above the SE line represents the corresponding root-mean-squared-error ($RMSE$) estimates. Finally, the solid line above the $RMSE$ line represents the corresponding standard error plus absolute bias ($SE + |BIAS|$) estimates. These three lines can never cross because of the relation

$$SE \leq RMSE \leq SE + |BIAS| .$$

Note that absolute bias can be inferred from the difference between the highest and lowest solid lines.

From Figure 2, we see that measures of accuracy and precision are somewhat insensitive to design changes, with the exception of model NNA, 45–55. Absolute bias is quite small for most designs, even for samples of size 500. Note that within rows, standard error estimates tend to increase somewhat as the difference Δ_0 decreases. Figure 3 gives results for slope parameter estimates for the same models that were

estimated to produce Figure 2. The results are qualitatively the same as those for the threshold point difference estimates.

In a slightly different format, Table 1 below presents the results of another set of simulations for the models described above when $n = 1000$. The results of these simulations are qualitatively the same as those of the other simulations.

Table 1 ($n = 1000$)

	50-80		35-65		60-70		45-55	
NNA	$\hat{\theta}$	$\frac{\hat{\Delta}}{\Delta_0}$	$\hat{\theta}$	$\frac{\hat{\Delta}}{\Delta_0}$	$\hat{\theta}$	$\frac{\hat{\Delta}}{\Delta_0}$	$\hat{\theta}$	$\frac{\hat{\Delta}}{\Delta_0}$
BIAS	.021	-.022	.133	-.067	.011	-.055	.081	-.226
SE	.051	.092	.169	.168	.026	.203	.165	.304
RMSE	.055	.094	.215	.181	.028	.210	.184	.379
NA								
BIAS	-.001	.005	.040	.017	-.017	.082	.008	.075
SE	.087	.084	.181	.094	.120	.151	.161	.147
RMSE	.087	.086	.186	.096	.121	.172	.161	.165
NNS								
BIAS	-.005	-.014	-.026	-.010	-.023	.021	-.002	.028
SE	.026	.082	.032	.112	.035	.157	.022	.197
RMSE	.027	.083	.041	.112	.042	.158	.022	.199
NS								
BIAS	-.006	.037	-.006	-.023	.008	.053	.019	.017
SE	.084	.091	.063	.068	.090	.136	.076	.102
RMSE	.084	.098	.063	.072	.091	.146	.079	.103

An Application

In this section, we consider an application requiring estimation of threshold points in an ordered response model to accurately estimate level of demand for a potential new telecommunications service. We apply the semiparametric estimator developed in this paper and compare results with those based on the parametric *Orbit* estimator for the ordered response model introduced in Klein and Sherman (1997). The *Orbit* estimator requires the error distribution in the model to be normally distributed and will generally be inconsistent if the normality assumption does not hold.

We begin with a description of the salient features of the data. A more detailed description can be found in Klein and Sherman (1997).

In a survey on demand for a potential new video service, 922 respondents were asked to estimate the average number of times per month they would use the service at a given charge for each use. They reported nonnegative integer values, with about 20% zeros. The data consist of these reported quantities, the per usage prices which varied across respondents, and other relevant explanatory variables such as income, etc.[5] In examining the reported quantities, we found that median and lower levels of demand roughly agreed with what one might expect for this type of service. However, beyond the median level of demand, it seemed reasonable to infer that respondents were exaggerating in a systematic manner: the higher the level of reported demand, the greater the exaggeration. For example, at the upper extreme, one respondent reported average monthly usage 20 times greater than the median level of usage.

Let $Y^* = \alpha_0 + \beta_0 V_0 + U$ denote accurate projected usage, where $V_0 = X_0 + \mathcal{X}'\theta_0$ with X_0 the price variable and $\mathcal{X}$ the other relevant explanatory variables, and β_0 and θ_0 are unknown parameters. We assume that U and $X = (X_0, \mathcal{X})$ are independent. We observe (Y, X) where Y is reported projected usage and assume the ordered response model

$$Y = \sum_{j=0}^{J+1} j\{t_{j-1} < Y^* \leq t_j\}\,.$$

where $t_j = \Lambda(y_j)$ for a function Λ that is strictly increasing on S_Y, the support of Y. We see that we must be able to estimate the threshold points to accurately estimate future usage of this service.

Up to this point, the only constraint placed on Λ is strict monotonicity. For example, the function $\Lambda(y_j) = y_j$, which characterizes accurate reporting, is allowed. Similarly, strictly increasing functions that capture understatement, exaggeration, or any mixture of these possibilities are also allowed. In any case, Λ can be interpreted as a correcting function quantifying the nature and extent of deviations from accurate reporting.

As mentioned above, if we can estimate Λ at each support point of interest, then

[5]Due to the proprietary nature of the data, we are not at liberty to identify either the service or any explanatory variables that may help identify the service.

we can correct misreporting and accurately forecast demand for the new service at these points. However, without extra information, Λ can only be estimated up to location and scale at points of interest. Nonetheless, it is easy to show that Λ itself can be estimated provided it is known at 2 (or more) points. In our application, it is natural to assume that $\Lambda(0) = 0$ and $\Lambda(3) = 3$ where 3 is the median of the reported quantities. See Klein and Sherman (1997) for a detailed discussion of these "safety point" assumptions. Under these assumptions, we can estimate Λ itself at points of interest.

We begin by constructing the two-stage estimator $(\hat{\theta}, \hat{d})$. Table 2 displays the observed support points of Y together with the corresponding sample frequencies.

Table 2

Y	Frequency	Y	Frequency	Y	Frequency
0	167	7	10	16	1
1	100	8	34	20	3
2	185	9	1	25	3
3	110	10	37	30	1
4	124	12	11	40	1
5	87	13	1	60	1
6	35	15	10		

We obtain $\hat{\theta}$ by partitioning the Y values into 3 regions: $(-\infty, 0]$, $(0, 3]$, and $(3, \infty)$. We choose this particular partition to facilitate comparison with the parametric *Orbit* estimates from Klein and Sherman (1997) where the same partition is used to get a first-stage estimator of θ_0. After coding the elements of these regions as $0, 1, 2$, respectively, we estimate $\hat{\theta}$ as prescribed in Section 2.

Given $\hat{\theta}$, we estimate $\Lambda(y)$ at the points $y = 0, 1, 2, 3, 4, 5, 6, 9$ and 12.[6] Specifically, we take $y_i = i$, $i = 0, 1, \ldots, 6$, $y_7 = 9$, and $y_8 = 12$. Then, for $j = 1, \ldots, 8$, we use $\hat{d}_j$ to estimate $\delta_{0j} = \Delta_0(j, j-1) = [\Lambda(y_j) - \Lambda(y_{j-1})]/\beta_0$, where β_0 is the coefficient of the price variable in the model. We then use $\sum_{j=1}^{i} \hat{d}_j$ to estimate $\sum_{j=1}^{i} \delta_{0j} = [\Lambda(y_i) - \Lambda(0)]/\beta_0$, obtaining estimates of $\Lambda(y_i)$ up to location and scale at $i = 1, 2, \ldots, 8$. However, since

[6]Note that observed Y values greater than 12 constitute approximately 2% of the sample of size 922. A reliable estimate of Λ cannot be obtained this far out in the tail of the distribution of Y with a sample of this size. Extrapolation procedures based on the estimates of $\Lambda(0), \Lambda(1), \ldots, \Lambda(12)$ can be developed to obtain estimates of Λ at the higher reported quantities. However, we do not do so here.

$\Lambda(0) = 0$, these estimates of $\Lambda(y_i)$ are actually consistent up to scale. Then, using the fact that $\Lambda(3) = 3$, we get that $\hat{\beta} = 3/\hat{\Delta}(3, 0)$ is a consistent estimate of β_0. It follows that $\hat{\beta} \sum_{j=1}^{i} \hat{d}_j$ is a consistent estimate of $\Lambda(y_i)$, $i = 1, 2, \ldots, 8$.

Figure 4 displays the semiparametric estimate of the reporting function Λ^{-1} superimposed over the parametric *Orbit* estimate of Λ^{-1}, which was obtained in a similar fashion. In other words, we plot y_i verses $\hat{\Lambda}(y_i)$, $i = 0, 1, \ldots, 8$, for both procedures. Solid points represent the semiparametric estimates while circles represent the *Orbit* estimates. A reference line corresponding to accurate reporting is also displayed, as well as brackets representing ± 2 estimated asymptotic standard deviations from the semiparametric estimates.

Both procedures tell a similar story, though there are differences. Both procedures suggest that there is roughly accurate reporting at lower levels of demand, but exaggeration at higher levels. However, for reported quantities between 4 and 12 (about 37% of the sample), the semiparametric procedure indicates less exaggeration, perhaps even some understatement. The differences between the two procedures may be attributable to an error distribution with a heavier than normal right tail. Since the parametric *Orbit* procedure assumes a normal error distribution, it will mistakenly treat accurate reported quantities that are high due to heavier than normal tails as exaggerated responses. However, given the magnitude of the estimated standard deviations, we would need a much larger sample size to investigate this possibility.

Summary

This paper develops a $\sqrt{n}$-consistent and asymptotically normal estimator of regression parameters and threshold points (up to location and scale) in a semiparametric ordered response model. To our knowledge, this is the first estimator of the threshold points in this model. Identification of the threshold points, which act like intercept terms, is achieved by exploiting shift restrictions inherent to single index models with additive errors when errors and regressors are independent.

Estimation proceeds in two stages. Regression parameters are estimated in a first

stage optimization of a quasi-likelihood function, and are used to estimate the threshold points in a second stage involving no optimization. Second-stage estimation is extremely fast, computationally.

Because of a duality between ordered response models and transformation models, the estimators developed in this paper can be used to estimate transformation functions at points of interest without assuming parametric forms for either error distributions or transformation functions. Censoring and discrete responses are also allowed, in contrast to previously available semiparametric techniques. We apply the new estimator to survey data (involving discrete responses and censoring) on demand for a new telecommunications service to assess exaggeration bias.

APPENDIX

Intermediate Results

In this subsection, we prove intermediate results used to obtain the main results in the text. Throughout, we maintain assumptions A1 through A3 and definitions D1 through D8 given in section 2.

Recall that $\hat{V}_i = V_i(\hat{\theta})$ and $V_i = V_i(\theta_0)$. Refer to D1 and D3. Write $\hat{\lambda}_{1\theta}$ for the $n \times 1$ vector with ith component $\hat{\lambda}_{1i}(\theta)$. Write $f_1(v,\theta)$ for $p_1 g_1(v,\theta)$ and estimate $f_1(v,\theta)$ with

$$\hat{f}_1(v,\theta,\hat{\lambda}_{1\theta}) = \frac{1}{n}\sum_k \{Y_k \leq j\} K([v - V_k(\theta)]/\hat{\lambda}_{1k}(\theta)h)/\hat{\lambda}_{1k}(\theta)h\,.$$

In like fashion, define $\hat{f}_0(v,\theta,\hat{\lambda}_{1\theta})$ to be an estimator of $f_0(v,\theta) = p_0 g_0(v,\theta)$ by replacing $\{Y \leq j\}$ with $\{Y > j\}$ in the definitions above. Finally, note that the conditional probability function $P_j(v,\theta)$ equals $f_1(v,\theta)/g(v,\theta)$ where $g(v,\theta) = f_1(v,\theta) + f_0(v,\theta)$. The corresponding estimator is $\hat{P}_j(v,\theta) = \hat{f}_1(v,\theta,\hat{\lambda}_{1\theta})/\hat{g}(v,\theta)$ where $\hat{g}(v,\theta) = \hat{f}_1(v,\theta,\hat{\lambda}_{1\theta}) + \hat{f}_0(v,\theta,\hat{\lambda}_{1\theta})$.

We now establish rates at which $\hat{f}_1(v,\theta,\hat{\lambda}_{1\theta})$ and its derivatives with respect to v and θ converge to the corresponding population quantities. The same rates hold for the estimator of $f_0(v,\theta)$ and its derivatives. In Lemma 1A, we obtain rates that are uniform in v and θ. Here, we are not concerned with bias reduction resulting from local smoothing. Subsequently, when we examine estimated densities evaluated

at true parameter values, we will show that estimated local smoothing results in the appropriate bias reduction.

Let ∇_i denote the partial derivative operator with respect to the ith argument of an operand. Also, write ∇_i^k for the kth partial derivative with respect to the ith argument. Interpret ∇_i^0 as the identity operator. Also, write $\mathbf{1}$ as an $n \times 1$ vector of ones.

Lemma 1A (Estimated Densities and Derivatives): For v and θ in compact sets, $k = 0, 1$, and $\delta > 0$,

(1) $\sup_{v,\theta}|\hat{f}_1(v,\theta,\hat{\lambda}_{1\theta}) - f_1(v,\theta)| = O_p(h^2)$

(2) $\sup_{v,\theta}|\nabla_1[\hat{f}_1(v,\theta,\hat{\lambda}_{1\theta}) - f_1(v,\theta)]| = O_p(n^{-1/2}h^{-2})$

(3) $\sup_{v,\theta}|\nabla_2\nabla_1^k[\hat{f}_1(v,\theta,\hat{\lambda}_{1\theta}) - f_1(v,\theta)]| = O_p(n^{-1/2}h^{-(k+2)})$.

PROOF: Write $\hat{\pi}_{1j}(\theta)$ for $\hat{\pi}_{1j}(V_j(\theta),\theta)$ and define $\hat{\eta}(\theta) \equiv [\hat{\sigma}_1(\theta), \hat{m}_1(\theta)]$, where $\hat{m}_1(\theta)$ is the geometric mean of the $\hat{\pi}_{1j}(\theta)$'s. Write the jth component of $\hat{\lambda}_{1\theta}$, the estimated local smoothing vector, as $\hat{\lambda}_{1j}(\theta) \equiv \lambda_1(\hat{\pi}_{1j}(\theta), \hat{\eta}(\theta))$. Write $\mu_1(v,\theta)$ for the expectation of $\hat{\pi}_{1j}(v,\theta)$ and define $\bar{\pi}_{1j}(\theta) \equiv \mu_1(V_j(\theta),\theta)$. With $\bar{m}_1(\theta)$ as the geometric mean of the $\bar{\pi}_{1j}(\theta)'s$ and with $m_1(\theta)$ as its probability limit, let $\eta(\theta) \equiv [\sigma_1(\theta), m_1(\theta)]$. Finally, write $\hat{\lambda}_{1\theta}^*$ and $\bar{\lambda}_{1\theta}$ for the n x 1 vectors of smoothing parameters with jth components:

$$\begin{aligned}\hat{\lambda}_{1j}^* &\equiv \lambda_1(\bar{\pi}_{1j}(\theta), \hat{\eta}(\theta)) \\ \bar{\lambda}_{1j} &\equiv \lambda_1(\bar{\pi}_{1j}(\theta), \eta(\theta)).\end{aligned}$$

Then, to establish (1), write

$$\begin{aligned}&|\hat{f}_1(v,\theta,\hat{\lambda}_{1\theta}) - f_1(v,\theta)| \leq |T_1| + |T_2| + |T_3| + |T_4|, \\ &T_1 = \hat{f}_1(v,\theta,\hat{\lambda}_{1\theta}) - \hat{f}_1(v,\theta,\hat{\lambda}_{1\theta}^*) \\ &T_2 = \hat{f}_1(v,\theta,\hat{\lambda}_{1\theta}^*) - \hat{f}_1(v,\theta,\bar{\lambda}_{1\theta}) \\ &T_3 = \hat{f}_1(v,\theta,\bar{\lambda}_{1\theta}) - E(\hat{f}_1(v,\theta,\bar{\lambda}_{1\theta})) \\ &T_4 = E(\hat{f}_1(v,\theta,\bar{\lambda}_{1\theta})) - f_1(v,\theta)\end{aligned}$$

Start with T_1. Write $\hat{f}_1(v,\theta,\hat{\lambda}_{1\theta})$ as $n^{-1}\sum_j \kappa(v,\theta,\hat{\pi}_{1j}(\theta))$. By a Taylor series expansion

about $\bar{\pi}_{1j}(\theta)$,

$$\begin{aligned} |T_1| &= n^{-1}|\sum_j [\hat{\pi}_{1j}(\theta) - \bar{\pi}_{1j}(\theta)]\nabla_3\kappa(v,\theta,\pi_{1j}^+(\theta))| \\ &\leq \sup_j |\hat{\pi}_{1j}(\theta) - \bar{\pi}_{1j}(\theta)| n^{-1}\sum_j |\nabla_3\kappa(v,\theta,\pi_{1j}^+(\theta))| \end{aligned}$$

From Klein and Spady (1993, Lemma 1), with h_p as the pilot window:

$$\sup_{j,\theta} |\hat{\pi}_{1j}(\theta) - \bar{\pi}_{1j}(\theta)| = O_p([\sqrt{n}h_p]^{-1}),$$

the uniform rate at which a kernel density estimator converges to its expectation. It can be shown that

$$\sup_{v,\theta} n^{-1}\sum_j |\nabla_3\kappa(v,\theta,\pi_{1j}^+(\theta))| = O_p(n_1^{\varepsilon}(\ln n_1)^{3/2}),$$

which implies from the choice of windows and ϵ that

$$\sup_{v,\theta} |T_1| = O_p(n_1^{\varepsilon}(\ln n_1)^{3/2}/(h_p\sqrt{n})) = O_p(h^2)\,.$$

The argument for T_2 is essentially the same as that above for T_1. From Klein and Spady (1993, Lemma 1):

$$\sup_{v,\theta} |T_3| = O_p(1/(h\sqrt{n})) = O_p(h^2).$$

Finally, from a Taylor expansion in h, $|T_4|$ (the bias term) is uniformly $O(h^2)$.

To establish (2), use the same decomposition as above and write:

$$|\nabla_v^k[\hat{f}_1(v,\theta,\hat{\lambda}_{1\theta}) - f_1(v,\theta)]| \leq |\nabla_v^k T_1| + |\nabla_v^k T_2| + |\nabla_v^k T_3| + |\nabla_v^k T_4|.$$

The proof is similar to that for (1) above. (3) also follows similarly. *QED.*

The next lemma provides results for density estimators with known local smoothing parameters. Lemmas 4A shows that local smoothing parameters may indeed be taken as known. Referring to the proof of Lemma 1A, write $\lambda_{1\theta_0}$ as a vector of known local

smoothing parameters with jth component: $\lambda_{1j} \equiv \lambda_1 (\pi_{1j} (\theta_0), \eta(\theta_0))$.

Lemma 2A (Known Local Smoothing): For v in a compact set,

(1) $\sup_v |\hat{f}_1(v, \theta_0, \lambda_{1\theta_0}) - E[\hat{f}_1(v, \theta_0, \lambda_{1\theta_0})]| = O_p(n^{-1/2}h^{-1})$

(2) $\sup_v |E[\hat{f}_1(v, \theta_0, \lambda_{1\theta_0})] - f_1(v, \theta_0)| = o_p(n^{-1/2})$

PROOF: The proof of (1) follows the same argument as that used for T_3 in Lemma 1A. The proof for (2) follows from Abramson[1982], Silverman [1986], window restrictions, and the properties of the smooth trimming function. Namely, for v in a compact set and d_1 the damping function defined in D2,

$$\nabla_v^m d_1 (l(v)) = \begin{array}{cc} 1 + o\left(n^{-1/2}\right) & m = 0 \\ o\left(n^{-1/2}\right) & m = 1, 2, 3, 4. \end{array}$$

To establish (2), expand $E[\hat{f}_1(v, \theta_0, \lambda_{1\theta_0})]$ in a Taylor Series in h about 0. Since $d_1(\cdot) = 1 + o(n^{-1/2})$ and $\nabla_v^m d_1(\cdot) = o(n^{-1/2})$ for $m \geq 1$, we obtain:

$$E[\hat{f}_1(v, \theta_0, \lambda_{1\theta_0})] = f_1(v, \theta_0) + \sum_{k=1}^{4} h^k C_k + o(n^{-1/2}),$$

where C_k is evaluated at $h = 0$ for $k \leq 3$ and C_4 is evaluated at $h^+ \in [0, h]$. From symmetry of the kernel, $C_1 = 0 = C_3$. From Abramson[1982] and Silverman[1986], as a consequence of local smoothing, $C_2 = 0$. The lemma now follows as $h^4 C_4 = o(n^{-1/2})$. *QED*.

Lemma 4A uses Lemma 2A to show that estimated smoothing and trimming parameters may be taken as known. Lemma 3A provides an intermediate result used to prove Lemma 4A.

Lemma 3A (Marginal Expectations): Write $\hat{l}(w)$ for $\hat{\lambda}_1(w, \theta_0)^{-1}$, G for the distribution function for $V_0\{Y \leq y\}$, and P for $I\!P\{Y \leq y\}$. Define the "marginal expectation" of $\hat{f}_1(v, \theta_0, \lambda_{1\theta_0})$ as

$$ME(h) = P \int [\hat{l}(w)/h] K([\hat{l}(w)/h][v - w]) dG(w).$$

For v in a compact set, we have

$$\sup_v |ME(h) - f_1(v, \theta_0)| = o(n^{-1/2}).$$

PROOF: As in the proof to Lemma 2A, expand $ME(h)$ in a Taylor series about $h = 0$. In this expansion, there will be terms involving the trimming function d_1 (see D2) and its derivatives evaluated at v, where v is in a compact set. For such terms, it follows from a Taylor series expansion about $\pi_1(v, \theta_0)$ and properties of the trimming function that:

$$\begin{aligned} \nabla_v^k d_1(\hat{\pi}_1(v, \theta_0)) &= \nabla_v^k d_1(l(v)) + o_p(n^{-1/2}) \\ &= \begin{cases} 1 + o_p(n^{-1/2}), k = 0 \\ o_p(n^{-1/2}), k = 1, 2. \end{cases} \end{aligned}$$

Using these properties of the trimming function, we get

$$ME(h) - f_1(v, \theta_0) = \sum_{k=1}^{4} h^k \hat{C}_k.$$

From symmetry of the kernel, $\hat{C}_1 = 0 = \hat{C}_3$. Under the assumed restrictions on the window,

$$h^2 \hat{C}_2 = h^2 C_2 + o_p(n^{-1/2}).$$

As in the proof to Lemma 2A, $C_2 = 0$. Finally, from window conditions, it can be shown that $h^4 \hat{C}_4 = o_p(n^{-1/2})$, which completes the argument. QED.

Using the above lemma and properties of the trimming function, we can now show that estimated pilot densities and estimated smoothing parameters may be taken as known.

Lemma 4A (Treating Estimated Local Smoothing as Known): For v in a compact set,

$$\sup_v |\hat{f}_1(v, \theta_0, \hat{\lambda}_{1\theta_0}) - \hat{f}_1(v, \theta_0, \lambda_{1\theta_0})| = o_p(n^{-1/2}).$$

PROOF: With $D = \hat{f}_1(v, \theta_0, \hat{\lambda}_{1\theta_0}) - \hat{f}_1(v, \theta_0, \lambda_{1\theta_0})$, from Lemmas 2A and 3A,

$$\begin{aligned} |D| &= |D_1 - D_2| + o_p(n^{-1/2}), \\ D_1 &= \hat{f}_1(v, \theta_0, \hat{\lambda}_{1\theta_0}) - ME \\ D_2 &= \hat{f}_1(v, \theta_0, \lambda_{1\theta_0}) - E(\hat{f}_1(v, \theta_0, \lambda_{1\theta_0})) \,. \end{aligned}$$

With G_n as the empirical CDF:

$$\hat{f}_1(v, \theta_0, \hat{\lambda}_{1\theta_0}) = \int [\hat{l}(w)/h] K([\hat{l}(w)/h][v - w]) dG_n(w)$$

Therefore, from the definition of ME,

$$D_1 = \int [\hat{l}(w)/h] K([\hat{l}(w)/h][v - w])[dG_n(w) - dG(w)] \,.$$

Similarly,

$$D_2 = \int [l(w)/h] K([l(w)/h][v - w])[dG_n(w) - dG(w)].$$

Substituting these expressions into D above,

$$\begin{aligned} |D| &= |\int [\hat{d}(w) - d(w)][dG_n(w) - dG(w)]| + o_p(n^{-1/2}), \\ \hat{d}(w) &= [\hat{l}(w)/h] K([\hat{l}(w)/h][v - w]) \\ d(w) &= [l(w)/h] K([l(w)/h][v - w]) \,. \end{aligned}$$

To show that this difference in differences is $o_p(n^{-1/2})$, integrate by parts and factor $\sup_w |G_n(w) - G(w)| = O_p(n^{-1/2})$ outside of the integral. Then, since $\int |\nabla_w [\hat{d}(w) - d(w)]| = o_p(1)$, the result follows.[7] *QED.*

Lemma 4A lets us take estimated smoothing parameters as known. As a result, we will be able to make use of the low order of the bias in the corresponding density estimator in Lemma 2A. This low order bias result will be useful in and of itself in

[7] Nadaraya[1965] initially used a similar integration by parts argument to show that a kernel density estimator converges to its expectation. Such an argument shows that D_1 and D_2 each converge to zero. To obtain the present result, we exploit the fact that these differences are also converging to each other.

establishing a U-statistic result below and also in being able to use a relatively fast convergence rate on estimated probability functions.

Next, we provide convergence results for semiparametric probability functions and their derivatives.

Lemma 5A (Estimated Probability Functions): Choose $\delta \in (0, 1/3)$. Select the second-stage window $h = n^{-\alpha}$ where $3/(20+\delta) < \alpha < 1/6$. Select the pilot window $h_p = h^r$, $r = [2/(3+\delta)]\alpha$. For v and θ in compact sets and $k = 0, 1$,

(1) $\sup_{v,\theta}|\hat{P}_k(v,\theta) - P_k(v,\theta)| = O_p(h^2)$

(2) $\sup_{v,\theta}\nabla_v|\hat{P}_k(v,\theta) - P_k(v,\theta)| = O_p(n^{-1/2}h^{-2}))$

(3) $\sup_{v,\theta}\nabla_\theta\nabla_v^k|\hat{P}_k(v,\theta) - P_k(v,\theta)| = O_p(n^{-1/2}h^{-(k+2)})$

Denote $\hat{P}_k^*(v,\theta_0)$ as the estimated semiparametric probability function with all estimated smoothing parameters replaced by their probability limits. Then, for the windows selected:

(4) $\sup_v|\hat{P}_k^*(v,\theta_0) - P_k(v,\theta_0)| = O_p(n^{-1/2}/h)$.

PROOF: With densities assumed to be bounded away from zero on compact sets, the proof for (1) through (3) of this lemma follows immediately from Lemma 1A. To establish (4), note that from Lemma 2A the bias in density estimates with known local smoothing is $O(h^4)$. Consequently, density estimates converge uniformly to the truth at the uniform rate for which the estimator converges to its expectation, which from Lemma 1A is $O_p(n^{-1/2}/h.)$. Again, with densities bounded away from zero on compact sets, (4) now follows. $QED.$

MAIN RESULTS

In this subsection, we provide proofs of the main results in the text, using the results established in the last subsection.

Proof of Theorem 1 (Consistency of $\hat{\theta}$): Recall from Section 2 that $\hat{\theta} = \text{argmax}_\theta\,\hat{Q}(\theta)$ where $\hat{Q}(\theta)$ is the criterion function in (3). Noting that $\text{plim}\hat{\tau}(X_i) = \tau(X_i)$ does not depend on θ, from Lemma 5A: $\sup_\theta|\hat{Q}(\theta) - \tilde{Q}(\theta)| \to 0$ in probability

where $\tilde{Q}(\theta)$ is obtained from $\hat{Q}(\theta)$ by replacing all estimated functions with their probability limits. From standard arguments and the trimming function, $\tilde{Q}(\theta)$ converges uniformly to its expectation. From A3, this expectation is uniquely maximized at θ_0. $QED.$

Proof of Lemma 1: Since $\hat{d}_j - \hat{\delta}_j = -\frac{1}{n\hat{\rho}} \sum_{i=1}^{n} \hat{t}(\hat{V}_i)(\hat{V}_{ij} - \tilde{V}_{ij})$, it is enough to show that uniformly over $\tilde{V}_{ij}$ in $\mathcal{G}$, $|\tilde{V}_{ij} - \hat{V}_{ij}| = o_p(n^{-1/2})$ as $n \to \infty$.
Taylor expand $\hat{P}_j(\tilde{V}_{ij})$ about $\hat{V}_{ij}$ and rearrange terms to get

$$|\tilde{V}_{ij} - \hat{V}_{ij}| = |\hat{P}_j(\tilde{V}_{ij}) - \hat{P}_j(\hat{V}_{ij})| / |\hat{P}'_j(V_{ij}^{+})| \tag{6}$$

where V_{ij}^{+} is between $\tilde{V}_{ij}$ and $\hat{V}_{ij}$.
By definition of $\hat{V}_{ij}$, $\hat{P}_j(\hat{V}_{ij}) = \hat{P}_{j-1}(\hat{V}_i)$. Let V_{ij}^{*} denote an element of $\mathcal{G}$ that is closest to $\hat{V}_{ij}$. By definition of $\tilde{V}_{ij}$, $|\hat{P}_j(\tilde{V}_{ij}) - \hat{P}_{j-1}(\hat{V}_i)| \leq |\hat{P}_j(V_{ij}^{*}) - \hat{P}_{j-1}(\hat{V}_i)|$. Put these last two facts together with (6) to get

$$|\tilde{V}_{ij} - \hat{V}_{ij}| \leq |\hat{P}_j(V_{ij}^{*}) - \hat{P}_j(\hat{V}_{ij})| / |\hat{P}'_j(V_{ij}^{+})| \,.$$

By a Taylor expansion of $\hat{P}_j(V_{ij}^{*})$ about $\hat{V}_{ij}$

$$|\tilde{V}_{ij} - \hat{V}_{ij}| \leq |P'_j(V_{ij}^{++}) / \hat{P}'_j(V_{ij}^{+})| |V_{ij}^{*} - \hat{V}_{ij}|$$

where V_{ij}^{++} is between V_{ij}^{*} and $\hat{V}_{ij}$. From Lemma 5A, estimated probability derivatives converge uniformly to the corresponding true probability derivatives. Since both $\tilde{V}_{ij}$ and $\hat{V}_{ij}$ are elements of $\mathcal{G}$, so are V_{ij}^{+} and V_{ij}^{++}. Deduce that $1/\hat{P}'(V_{ij}^{+}) = O_p(1)$. From Lemma 5A and A3, $\hat{P}'(V_{ij}^{++}) = O_p(1)$. Recall from D7 that $\hat{V}_{ij}$ is contained in the interval $[\hat{V}_L, \hat{V}_U]$ where $\hat{V}_L$ and $\hat{V}_U$ are elements of $\mathcal{G}$. Since adjacent elements of $\mathcal{G}$ are a distance $o_p(n^{-1/2})$ apart uniformly over $\mathcal{G}$, it follows from the definition of V_{ij}^{*} that as $n \to \infty$, $|V_{ij}^{*} - \hat{V}_{ij}| = o_p(n^{-1/2})$ uniformly over $\mathcal{G}$. This proves the result. $QED.$

Proof of Lemma 2: For $\hat{V}_i$ in $\mathcal{T}$, write $\hat{\delta}_{ij}$ for $\hat{V}_{ij} - \hat{V}_i$. Taylor expand $\hat{P}_j(\hat{V}_i + \hat{\delta}_{ij})$

about $\hat{V}_i + \delta_{0j}$ to get

$$[\hat{\delta}_{ij} - \delta_{0j}]\hat{P}'_j(V_{ij}^+) = \hat{P}_j(\hat{V}_i + \hat{\delta}_{ij}) - \hat{P}_j(\hat{V}_i + \delta_{0j})$$

where V_{ij}^+ is between $\hat{V}_{ij} = \hat{V}_i + \hat{\delta}_{ij}$ and $\hat{V}_i + \delta_{0j}$. Since $\hat{V}_{ij} = \hat{V}_i + \hat{\delta}_{ij}$, by definition of $\hat{V}_{ij}$, $\hat{P}_j(\hat{V}_i + \hat{\delta}_{ij}) = \hat{P}_{j-1}(\hat{V}_i)$. By (2), $P_j(V_i + \delta_{0j}) = P_{j-1}(V_i)$. Substitute these identities into the last expression to get

$$\begin{aligned}
[\hat{\delta}_{ij} - \delta_{0j}]\hat{P}'_j(V_{ij}^+) &= \hat{P}_j(\hat{V}_i + \delta_{ij}) - \hat{P}_j(\hat{V}_i + \delta_{0j}) \\
&= \hat{P}_{j-1}(\hat{V}) - \hat{P}_j(V_i + \delta_{0j}) \\
&= \hat{P}_{j-1}(\hat{V}_i) - P_{j-1}(V_i) - [\hat{P}_j(\hat{V}_i + \delta_{0j}) - P_j(V_i + \delta_{0j})]\,.
\end{aligned}$$

Divide through by $\hat{P}'_j(V_{ij}^+)$ and sum to complete the proof. $QED.$

Proof of Theorem 2 (Consistency of $\hat{d}_j$): By Lemma 1, it is sufficient to prove consistency of $\hat{\delta}_j$. From Lemma 2, $\hat{\delta}_j - \delta_{0j}$ equals

$$\frac{1}{n\hat{\rho}} \sum_{i=1}^{n} \hat{t}(\hat{V}_i) \left[\hat{P}_{j-1}(\hat{V}_i) - P_{j-1}(V_i) - \left[\hat{P}_j(\hat{V}_i + \delta_{0j}) - P_j(V_i + \delta_{0j})\right]\right] / \hat{P}'_j(V_{ij}^+)$$

where V_{ij}^+ is between $\hat{V}_{ij}$ and $\hat{V}_i + \delta_{0j}$. Argue as in the proof of Lemma 1 to get that $1/\hat{P}'_j(V_{ij}^+)$ is uniformly bounded in probability. Also, note that $\hat{\rho}$ converges in probability to $\rho > 0$. The result now follows directly from Lemma 5A. $QED.$

Proof of Theorem 3 (Asymptotic Normality of $\hat{\theta}$) : Referring to the proof of Theorem 1, from a Taylor Series Expansion: $\sqrt{n}(\hat{\theta} - \theta_0) = -\hat{H}^{-1}(\theta^+)\sqrt{n}\hat{G}(\theta_0)$, where, with θ^+ between $\hat{\theta}$ and θ_0,

$$\begin{aligned}
\hat{H}(\theta^+) &= \nabla^2_\theta \hat{Q}(\theta^+) \\
\hat{G}(\theta_0) &= \nabla_\theta \hat{Q}(\theta_0)\,.
\end{aligned}$$

From Lemma 5A, $\hat{H}(\theta)$ converges in probability, uniformly in θ, to $\tilde{H}(\theta)$, the hessian for $\tilde{Q}(\theta)$. From standard arguments, $\tilde{H}(\theta)$ converges in probability, uniformly in θ, to $H(\theta)$. Deduce from this and the consistency of $\hat{\theta}$, that $\hat{H}(\theta^+)$

converges in probability to $H(\theta_0)$. Let $\hat{P}_{ij} = \hat{P}_j(V_i(\theta_0), \theta_0)$ and $Y_{ij} = \{Y_i = j\}$. Then, the normalized gradient, $\sqrt{n}\hat{G}(\theta_0)$, is given as:

$$\begin{aligned} & \sqrt{n}\sum_i \frac{1}{n}\hat{\tau}(X_i)[\frac{Y_{i0}}{\hat{P}_{i0}} - \frac{Y_{i1}-Y_{i0}}{\hat{P}_{i1}-\hat{P}_{i0}}]\nabla_\theta \hat{P}_{i0} + ... \\ = \quad & \sqrt{n}\sum_i \frac{1}{n}\hat{\tau}(X_i)[\frac{Y_{i0}-\hat{P}_{i0}}{\hat{P}_{i0}} - \frac{(Y_{i1}-Y_{i0})-(\hat{P}_{i1}-\hat{P}_{i0})}{\hat{P}_{i1}-\hat{P}_{i0}}]\nabla_\theta \hat{P}_{i0} + ... \end{aligned}$$

As the above two terms and all remaining terms have the same structure, it will suffice to analyze the first term above. With $\hat{w}_{ij} = \hat{\tau}(X_i)\nabla_\theta \hat{P}_{i0}/\hat{P}_{i0}$, we may write this term as:

$$\begin{aligned} T_1 - T_2 &= \sqrt{n}\sum_{i=1}^n \frac{1}{n}\hat{\tau}(X_i)[Y_{ij} - \hat{P}_{j0}]\hat{w}_{ij} \\ T_1 &= \sqrt{n}\sum_{i=1}^n \frac{1}{n}\hat{\tau}(X_i)[Y_{ij} - P_{j0}]\hat{w}_{ij} \\ T_2 &= \sqrt{n}\sum_{i=1}^n \frac{1}{n}\hat{\tau}(X_i)[\hat{P}_{j0} - P_{j0}]\hat{w}_{ij}\,. \end{aligned}$$

As in Klein and Spady(1993):

$$T_1 = \sqrt{n}\sum_{i=1}^n \frac{1}{n}[Y_{ij} - P_{j0}]\tau(X_i)\nabla_\theta P_{i0}/P_{i0} + o_p(1)$$

and $T_2 = o_p(1)$.

All other terms behave similarly. Thus, $\sqrt{n}\hat{G}(\theta_0) = \sqrt{n}\tilde{G}(\theta_0) + o_p(1)$ where $\tilde{G}(\theta)$ is the gradient of $\tilde{Q}(\theta)$. Deduce that $\sqrt{n}(\hat{\theta} - \theta_0) = -H(\theta_0)^{-1}\sqrt{n}\tilde{G}(\theta_0) + o_p(1)$. Since the information equality holds here, Theorem 3 is now immediate. QED.

Proof of Lemma 3: Recall that V_{ij}^+ is a point between $\hat{V}_{ij} = \hat{V}_i + \hat{\delta}_{ij}$ and $\hat{V}_i + \delta_{0j}$. For convenience, write V_{ij}^+ as $\hat{V}_i + \delta_{ij}^+$, so that δ_{ij}^+ is a point between $\hat{\delta}_{ij}$ and δ_{0j}. Next, refer to the consistency argument above and note that $\hat{t}(\cdot) = \hat{t}^2(\cdot)$. Thus, we have that

$$\hat{\delta}_{ij} - \delta_{0j} \quad = \quad \frac{1}{n\hat{\rho}}\sum_{i=1}^n \hat{t}(\hat{V}_i)\hat{\varepsilon}(\hat{V}_i)[\hat{t}(\hat{V}_i)/\nabla \hat{P}_j(\hat{V}_i + \delta_{ij}^+)],$$

$$\hat{\varepsilon}(\hat{V}_i) \quad = \quad \hat{P}_{j-1}(\hat{V}_i) - P_{j-1}(V_i) - [\hat{P}_j(\hat{V}_i + \delta_{0j}) - P_j(V_i + \delta_{0j})].$$

From D6 and D7 we see that $\hat{t}(v)$ can be written as the indicator function $\{\hat{V}_L < v < \hat{V}_U\}$. From Lemma 5A, $\hat{t}(v)\hat{\varepsilon}(v) = O_p(n^{-1/2}/h)$ uniformly in v and θ. Next, let $\hat{t}^*(w) = \{\hat{V}_L + \delta_{ij}^+ < w < \hat{V}_U + \delta_{ij}^+\}$ and note that $\hat{t}^*(v + \delta_{ij}^+) = \hat{t}(v)$. Then, from (7) and Lemma 5A, $\hat{t}^*(w)/\hat{P}_j'(w) = O_p(1)$ uniformly in v and θ. Therefore, since $\hat{\rho} = O_p(1)$,

$$\hat{\delta}_{ij} - \delta_{0j} = O_p(n^{-1/2}/h)\,.$$

Having obtained a (conservative) rate on $\hat{\delta}_{ij} - \delta_{0j}$, proceed by denoting V_L and V_U as the probability limits of $\hat{V}_L$ and $\hat{V}_U$, respectively, and define the indicator: $t(v) = \{V_L < v < V_U\}$. Then, with ρ as the probability limit of $\hat{\rho}$, we can write $\sqrt{n}(\hat{\delta}_{ij} - \delta_{0j}) = \sqrt{n}(\hat{d}_j - \delta_{0j}) + o_p(1)$ as:

$$\frac{\sqrt{n}}{\hat{\rho}} \sum_{i=1}^{n} \frac{1}{n} \hat{t}(\hat{V}_i)\hat{\varepsilon}(\hat{V}_i)[1/\hat{P}_j'(\hat{V}_i + \delta_{ij}^+)] = \frac{1}{\rho} \sum_{k=1}^{5} T_k + o_p(1),$$

where

$$\begin{aligned}
T_1 &= \sqrt{n} \sum_{i=1}^{n} \frac{1}{n} \hat{t}(\hat{V}_i)\hat{\varepsilon}(\hat{V}_i)[1/\hat{P}_j'(\hat{V}_i + \delta_{ij}^+) - 1/P_j'(\hat{V}_i + \delta_{ij}^+)] \\
T_2 &= \sqrt{n} \sum_{i=1}^{n} \frac{1}{n} \hat{t}(\hat{V}_i)\hat{\varepsilon}(\hat{V}_i)[1/P_j'(\hat{V}_i + \delta_{ij}^+) - 1/P_j'(\hat{V}_i + \delta_{0j})] \\
T_3 &= \sqrt{n} \sum_{i=1}^{n} \frac{1}{n} [\hat{t}(\hat{V}_i) - t(\hat{V}_i)]\hat{t}(\hat{V}_i)\hat{\varepsilon}(\hat{V}_i)[1/P_j'(\hat{V}_i + \delta_{0j})] \\
T_4 &= \sqrt{n} \sum_{i=1}^{n} \frac{1}{n} t(\hat{V}_i)[\hat{t}(\hat{V}_i) - t(\hat{V}_i)]\hat{\varepsilon}(\hat{V}_i)[1/P_j'(\hat{V}_i + \delta_{0j})] \\
T_5 &= \sqrt{n} \sum_{i=1}^{n} \frac{1}{n} t(\hat{V}_i)\hat{\varepsilon}(\hat{V}_i)[1/P_j'(\hat{V}_i + \delta_{0j})]
\end{aligned}$$

We will show that the first four terms above vanish in probability while the fifth is close in probability to the form given in the lemma. Beginning with T_1, with $\hat{t}^*$ defined as above and θ in a compact neighborhood of θ_0,

$$|T_1| \leq \sqrt{n} \sup_{v,\theta} |\hat{t}(v)\hat{\varepsilon}(v)| \sup_{v,\theta} |\hat{t}^*(w)/[1/\hat{P}_j'(w) - 1/P_j'(w)]|,$$

which is $o_p(1)$ from Lemma 5A. Turning to T_2:

$$|T_2| \leq \sqrt{n}\sup_{v,\theta}|\hat{t}(v)\hat{\varepsilon}(v)| \sum_{i=1}^{n} \frac{1}{n}|\hat{t}(\hat{V}_i)/[1/P_j'(\hat{V}_i + \delta_{ij}^+) - 1/P_j'(\hat{V}_i + \delta_{0j})]| \,.$$

The first term in the above product is $O_p(1/h)$ from Lemma 5A. Since δ_{ij}^+ is between $\hat{\delta}_{ij}$ and δ_{0j}], and since from (8), $\hat{\delta}_{ij} - \delta_{0j} = O_p(n^{-1/2}/h)$, $|T_2| = o_p(1)$ from the choice of h. For T_3:

$$|T_3| \leq \sqrt{n}\sup_{v,\theta}|[\hat{t}(v) - t(v)]\hat{t}(v)|\sup_{v,\theta}|\hat{t}(v)\hat{\varepsilon}(v)/P_j'(v + \delta_{0j})| \,.$$

From Lemma 5A, the second term in the above product is $O_p(n^{-1/2}/h)$. From Lemmas A1 and A2 in Klein (1993), which are based on an inequality due to Jim Powell, the first term is much smaller than $o_p(h)$. Therefore, $|T_3| = o_p(1)$. A similar argument shows that $|T_4| = o_p(1)$.

Finally, turn to T_5 and let $\hat{\epsilon}_i(\theta) = \hat{\epsilon}(V_i(\theta))$. We get that

$$\begin{aligned} T_5 &= \sqrt{n}\sum_{i=1}^{n} \frac{1}{n} t(\hat{V}_i)t(V_i)\hat{\varepsilon}_i(\hat{\theta})[1/P_j'(\hat{V}_i + \delta_{0j})] \\ &\quad + \sqrt{n}\sum_{i=1}^{n} \frac{1}{n}[t(\hat{V}_i) - t(V_i)]t(\hat{V}_i)\hat{\varepsilon}_i(\hat{\theta})[1/P_j'(\hat{V}_i + \delta_{0j})] \,. \end{aligned}$$

Employing arguments similar to those above, the second component of T_5 vanishes in probability. Therefore, from a Taylor series expansion of the first component of T_5 about θ_0:

$$\begin{aligned} T_5 &= \sqrt{n}\sum_{i=1}^{n} \frac{1}{n} t(\hat{V}_i)t(V_i)\hat{\varepsilon}_i(\theta_0)/P_j'(V_i + \delta_{0j}) \\ &\quad + \left[\sum_{i=1}^{n} \frac{1}{n} t(\hat{V}_i)t(V_i)\nabla_\theta[\hat{\varepsilon}_i(\theta^+)/P_j'(V_i(\theta^+))]\right]\sqrt{n}[\hat{\theta} - \theta_0] \end{aligned}$$

where θ^+ is between $\hat{\theta}$ and θ_0. Employing arguments similar to those used above, in the first component we may replace $t(\hat{V}_i)t(V_i)$ with $t^2(V_i) = t(V_i)$. This component now has the form given in Lemma 3. Employing Lemma 5A and then replacing $t(\hat{V}_i)t(V_i)$ with $t(V_i)$ as above, it can be shown that the second compo-

nent above also has the form given in Lemma 3. *QED.*

Proof of Lemma 4 (A U-statistic result): For compactness, write $\hat{f}_1(v)$ for $\hat{f}_1(v, \theta_0, \hat{\lambda}_{1\theta_0})$ and $\hat{g}(v)$ for $\hat{g}(v, \theta_0)$. From Lemma 3, with probability tending to one as $n \to \infty$, $D_k(\theta_0)$ equals

$$n^{-1/2} \sum_{i=1}^{n} t(V_i) \left[\frac{\hat{f}_1(V_i + \Delta_0(k, j-1))}{\hat{g}(V_i + \Delta_0(k, j-1))} - P_k(V_i + \Delta_0(k, j-1)) \right] / P_j'(V_i + \delta_{0j}) \,.$$

Lemma 4A lets us take all estimated local smoothing parameters in $D_k(\theta_0)$ as known. To further simplify this expression, we first show that it is unchanged (up to $o_p(1)$) if the summand is multiplied by $\hat{g}/g$. The resulting expression is a U-statistic since it is a linear combination of kernel density estimators. Define $\Delta_0 = \Delta_0(k, j-1)$, $V_i^* = V_i + \Delta_0$, $\delta_i = 1/P_j'(V_i)$, and $\bar{\delta} = \sum_{i=1}^{n} t(V_i)|\delta_i|/n$. Then

$$\begin{aligned} & \left| n^{-1/2} \sum_{i=1}^{n} t(V_i) \left[\frac{\hat{f}_1(V_i^*)}{\hat{g}(V_i^*)} - P_k(V_i^*) \right] \delta_i \left[\frac{\hat{g}(V_i^*)}{g(V_i^*)} - 1 \right] \right| \\ \leq \; & n^{1/2} \sup_v t(v) \left| \frac{\hat{f}_1(v - \Delta_0)}{\hat{g}(v - \Delta_0)} - P_k(v - \Delta_0) \right| \bar{\delta} \sup_v t(v) \left| \frac{\hat{g}(v - \Delta_0)}{g(v - \Delta_0)} - 1 \right| , \end{aligned}$$

which from Lemma 5A converges to zero in probability. Next, define

$$w_k(V_i) = g(V_i)[P_j'(V_i + \delta_{0j}) g(V_i^*)]^{-1} \,.$$

It follows from the last inequality that

$$\begin{aligned} & D_k(\theta_0) = U_n + o_p(n^{-1/2}) \text{ where} \\ & U_n = n^{-1} \textstyle\sum_{i=1}^{n} t(V_i)[\hat{f}_1(V_i^*) - P_k(V_i^*)] \hat{g}(V_i^*) w_k(V_i)/g(V_i) \,. \end{aligned}$$

To write this expression as a U-Statistic, for $d = 0, 1$, let

$$K_{ij}^d = \{Y_i \leq j\}^d \{Y_i > j\}^{1-d} \frac{1}{h \lambda_d(V_j)} K\left(\frac{1}{h \lambda_d(V_j)} [V_i^* - V_j] \right) .$$

Next, with $Z_i = (Y_i, X_i)$, define

$$\rho_k(Z_i, Z_j) = \frac{1}{g(V_i)} t(V_i) w_k(V_i)[K_{ij}^1 - (K_{ij}^1 + K_{ij}^0) P_k(V_i^*)] \,.$$

Then

$$U_n = \frac{2}{n(n-1)} \sum_{j>i} [\rho_k(Z_i, Z_j) + \rho_k(Z_j, Z_i)]/2 \,.$$

Before using standard projection arguments to analyze this U-Statistic, we note several useful properties of the function ρ_k. Define

$$\bar{\rho}_k(Z_i) = [\{Y_i \leq j\} - P_k(V_i)] t(V_i - \Delta_0(k, j-1)) w(V_i - \Delta_0(k, j-1)) \,.$$

Then the following properties hold:

(1) $I\!E \rho_k(Z_i, Z_j)|X_i, Y_i) = \bar{\rho}_k(Z_i) + h^2 \varepsilon_i$, where $\varepsilon_i = O_p(1)$

(2) $I\!E \rho_k(Z_i, Z_j)|X_i, Y_i) = o(n^{-1/2})$

(3) $I\!E \rho_k(Z_i, Z_j)^2 = o(n)$

From (2), $I\!E \rho_k(Z_i, Z_j) = o(n^{-1/2})$. It then follows from (3)[8] that

$$U_n = n^{-1} \sum_{i=1}^{n} I\!E(\rho_k(Z_j, Z_i)|X_i, Y_i) + o_p(n^{-1/2}).$$

From iterated expectations and (2), $I\!E \rho_k(Z_i, Z_j) = I\!E \rho_k(Z_j, Z_i) = o(n^{-1/2})$. Then, since $I\!E \bar{\rho}_k(Z_i) = 0$, $I\!E h^2 \epsilon_i = o(n^{-1/2})$. It now follows from (1) that

$$U_n = n^{-1} \sum_{i=1}^{n} \bar{\rho}_k(Z_i) + o_p(n^{-1/2}).$$

This proves the result. *QED.*

REFERENCES

Abramson, I. S. (1982): "On Bandwidth Variation in Kernel Estimates – A Square Root Law," *Annals of Statistics*, 10, 1217–1223.

[8] See Serfling (1980) and Powell, Stock, and Stoker (1989).

AMEMIYA, T. (1985): *Advanced Econometrics*. Harvard Univ. Press, Cambridge, Mass.

CAVANAGH, C. L. AND R. P. SHERMAN (1998): "Rank Estimators for Monotonic Index Models," *Journal of Econometrics*, 84, 351–381.

HAN, A. K. (1987): "Non-parametric Analysis of a Generalized Regression Model," *Journal of Econometrics*, 35, 303–316.

GØRGENS, T. AND J. HOROWITZ(1999):"Semiparametric Estimation of a Censored Regression Model with an Unknown Transformation of the Dependent Variable," *Journal of Econometrics*, 90, 155–191.

HOROWITZ, J. L. (1996): "Semiparametric Estimation of a Regression Model with an Unknown Transformation of the Dependent Variable," *Econometrica*, 64, 103–138.

ICHIMURA, H. (1993): "Semiparametric Least Squares (SLS) and Weighted Least Squares Estimation of Single Index Models," *Journal of Econometrics*, 58, 71–120.

KLEIN, R. (1993): "Specification Tests for Binary Choice Models Based on Index Quantiles," Journal of Econometrics, 59, 343–375.

KLEIN, R. W. AND R. H. SPADY (1993): "An Efficient Semiparametric Estimator for Binary Response Models," *Econometrica*, 61, 387–432.

KLEIN, R. W. AND R. P. SHERMAN (1997): "Estimating New Product Demand from Biased Survey Data," *Journal of Econometrics*, 76, 53–76.

NADARAYA, E. A. (1965): "On Non-Parametric Estimates of Density Functions and Regression Curves," *Theor. Prob. Appl*, 10, 196–190.

POWELL, J. L., STOCK, J. H., AND T. M. STOKER (1989): "Semiparametric Estimation of Weighted Average Derivatives," *Econometrica*, 57, 1403–1430.

SERFLING, R. J. (1980): *Approximation Theorems of Mathematical Statistics*, Wiley, New York.

SILVERMAN, B. (1986): *Density Estimation*, Chapman and Hall, London.

YE, J. AND N. DUAN(1997):"Nonparametric $\sqrt{n}$-Consistent Estimation for the General Transformation Models,"*Annals of Statistics*, 25, 2682–2717.