Irregular Identification, Support Conditions and Inverse Weight Estimation*

Shakeeb Khan[†] and **Elie Tamer**[‡]

August 2007

ABSTRACT. Inverse weighted estimators are commonly used in econometrics (and statistics). Some examples include: 1) The binary response model under mean restriction introduced by Lewbel(1997) and further generalized to cover endogeneity and selection. The estimator in this class of models is weighted by the density of a special regressor. 2) The censored regression model under mean restrictions where the estimator is inversely weighted by the censoring probability (Koul, Susarla and Van Ryzin (1981)). 3) The treatment effect model under exogenous selection where the resulting estimator is one that is weighted by a variant of the propensity score. We show that *point identification* of the parameters of interest in these models often requires support conditions on the observed covariates that essentially guarantee "enough variation." Under general conditions, these support restrictions, necessary for point identification, drive the weights to take arbitrary large values which we show creates difficulties for regular estimation of the parameters of interest. Generically, these models, similar to well known "identified at infinity" models, lead to estimators that converge at slower than the parametric rate, since essentially, to ensure point identification, one requires some variables to take values on sets with arbitrarily small probabilities, or *thin sets*. For the examples above, we derive rates of convergence under different tail conditions, analogous to Andrews and Schafgans(1998) illustrating the link between these rates and support conditions.

JEL Classification: C14, C25, C13.

Key Words: Irregular Identification, thin sets, rates of convergence.

[*] We thank J. Powell and seminar participants at Berkeley, Duke, Stanford, and Wisconsin, as well as conference participants at the 2007 NASM at Duke University, for helpful comments.

[†] Department of Economics, Duke University, 213 Social Sciences Building, Durham, NC 27708. Phone: (919) 660-1873.

[‡] Department of Economics, Northwestern University, 2001 Sheridan Road, Evanston, IL 60208. Phone: (847) 491-8218. Support from the National Science Foundation is gratefully acknowledged.

1 Introduction

The identification problem in econometrics is a first step that one must explore and carefully analyze before any meaningful inferences can be reached. Identification clarifies what can be learned about a parameter of interest in a given model. There is a class of models in econometrics, mainly arising in limited dependent variable models, that attain identification by requiring that covariate variables take support in regions with arbitrary small probability mass. These identification strategies sometimes lead to estimators that are weighted by a density or a conditional probability, and these weights take arbitrarily large values on these regions of small mass. For example, an estimator that is weighted by the inverse of the density of a continuous random variable, effectively only uses observations for which that density is small. Taking values on these "thin sets," is essential (often necessary) for point identification in these models. This thin set identification is an essential feature of the models we consider in this paper. We explore the relationship between the inverse weighting method in these models and the rates of convergence of resulting estimators, and show that under general conditions, it is not possible to attain the regular rate (square root of the sample size) with these models. We label these identification strategies as "thin set identified" and argue that they are "irregularly identified", or belong to the class of "identified at infinity" models. Note that it is part of econometrics folklore to equate the parameter being identified at infinity with slow rates of convergence, as was also established in Chamberlain(1986) for some particular models. See also Andrews and Schafgans(1998).

The results in this paper are connected to a set of models that we examine in detail. In these models, the parameter of interest can be written as a weighted expectation. For comparison purposes, we first consider the binary choice model under a median restriction where point identification of the regression parameters was shown in Manski(1975,1985), though it was shown to be irregular in Chamberlain(1986), who established an impossibility theorem for estimating these parameters at the parametric rate. Analogous impossibility theorems hold for estimating parameters in the set of models that we study in detail and are the basis of this paper. The first model is the binary choice model under an exclusion restriction and mean restrictions. This model was introduced to the literature by Lewbel(1997, 1998, 2000). There, Lewbel demonstrates that this binary model is point identified with only a mean restriction on the unobservables, by requiring the presence of a special regressor that is conditionally independent of the error. Under these conditions, Lewbel provides a density weighted estimator for the finite dimensional parameter (including the constant term). We show that the parameters in a simple version of that model are thin set identified. We derive the efficiency bound and show that it is infinite for those parameters, unless special relative tail behavior conditions are imposed. Two such conditions are: a support condition such that the propensity score hits limiting values with positive measure, in which case root n estimation can be attained; or infinite moment restrictions on the regressors. As we show below, the optimal rate of convergence in these cases depend on the tail behavior of the special regressor relative to the error distribution.

We also examine the censored regression model under mean restriction and random independent censoring. An inverse weighting method for identification was proposed in Koul, Van Rysin and Susarla(1981). Again, under general conditions, we show that the regression parameters are identified on thin sets and hence under general conditions on the censoring process, non-regular rates of convergence are obtained. Newey (2003) derived efficiency bounds for a constant censoring version of this model, and concluded the parametric rate was unattainable.

Finally, we consider the treatment effects model under binary treatment and exogenous selection. Hahn(1998) in important work derived the efficiency bound for this model and provided a series based estimator that reaches that bound (See also Hirano, Imbens and Ridder(2000)). In the case where the covariates take infinite support, the propensity score is arbitrarily close to zero or one (as opposed to bounded away from zero or one). We show that the average treatment effect in this case can be irregularly identified (the bound can be infinite) and this identification results in non-regular rates of convergence for the ATE in general: the optimal rates will depend on the relative tail thickness of the error in the treatment equation and the covariates.

The next section reintroduces the maximum score model, which will serve as the base model to introduce the concepts of *irregular identification* and relates it to the models we study in details which is done in sections 3 - 6. Specifically, sections 3 - 5 consider the binary choice model under median and mean restrictions, and censored regression models under mean restrictions, respectively. Section 6 considers the average treatment effect. Section 7 concludes by summarizing and suggesting areas for future research.

2 Irregular Identification, Identification on Thin Sets and Inverse Weighting

The notion of regularity of statistical models is linked to efficient estimation of parameters-see, e.g. Stein(1956). The concept of irregular identification has been related to consistent estimation of the parameters of interest at the parametric rate. An important early reference to irregular identification in the literature appears in Chamberlain(1986) dealing with efficiency bounds for sample selection models. There, Chamberlain showed that even though a slope parameter in the selection equation is identified, it is not possible to estimate it at the parametric root n rate (under the assumptions he maintains in the paper). Chamberlain added that point identification in this case relies "on the behavior of the regression function at infinity." In essence, *since the disturbance terms are allowed to take support on the real line*, one requires that the regression function takes "large" values so that at the limit the propensity score is equal to one (or is arbitrarily close to one). The identified set in the model shrinks to a point when the propensity score hits one (or no selection). Heckman(1990) highlighted the importance of identified at infinity parameters in various selection models and linked this type of identification essentially to the propensity score. Andrews and Schafgans(1998) used an estimator proposed by Heckman(1990) and confirmed slow rates of convergence of the estimator for the constant and also showed that this rate depends on the thickness of the tails of the regression function distribution relative to that of the error term. This rate can be as slow as cube root and can be arbitrarily close to root n.

The class of models we consider in this paper are ones where the parameter of interest can be written as a closed form solution to a weighted moment condition, where the weight functions take arbitrarily large values on sets of observables we call "thin sets." Most importantly, values on these thin sets are necessary conditions for point identification of the parameters. In this class of models where point identification is fragile, the rate of convergence is generally slower than the regular rate. The exact rate is determined by the properties of the weight function on these thin sets. In some models, not only is point identification delicate, but also if the regressors fail to take values on the "thin sets", then the model will not be able to have any information about the true parameter, i.e., the identified set is the whole parameter space. This leads to a non-robust identification in the sense that one either learns the parameter precisely, or, the model provides no information about the parameter of interest. We view this lack of robustness as an added drawback of inference in this class of models.

Heuristically, we consider models where the parameter of interest can be written as a weighted expectation of some observed variables. We illustrate exactly these weight functions in three examples: the binary response model under various point identification assumptions, the censored regression model under mean restrictions, and the average treatment effects model. Generally, we say that a parameter is thin set identified if it is only identified if some observables take values on a set of arbitrary small probability mass and the weight function on this set is unbounded. We first give a definition of irregular identification which we tie in later with our thin identified models we consider below. Following the statistical and econometric literature, we define irregular identification as follows:

Definition 2.1 (Irregular Identification of a parameter vector) *We say that a parameter vector is irregularly point identified if it cannot be estimated regularly*[4] *at the parametric rate.*

As we will show, the models studied in the following sections will all be irregularly identified. We begin with the maximum score model introduced in Manski(1975) so we can compare its properties to models studied more recently in the literature.

3 Binary Choice Model with Median Restriction

Consider the binary choice relation

$$y = 1[\alpha + v - \epsilon \geq 0] \tag{3.1}$$

where 1[.] is a function returning a 1 if . is true and 0 otherwise. We consider a set of assumptions relating the stochastic relationship between ϵ and v. We go through a set of models that make different assumptions on the joint distribution of ϵ and v. We start with the maximum score model.

The first model we consider is Manski's maximum score model in which it is assumed that $med(\epsilon|v) = 0$. Under a set of support conditions on v, Manski showed that the parameter α is point identified . However, with ϵ having support on the real line, it is easy to show that α is point identified even if we do not require v to have large support. In fact, all that is needed for point identification of α is for $\alpha + v$ to be continuous in a neighborhood of zero . Nonetheless, identification is irregular by our definitions since α cannot be estimated at the parametric rate. (see, e.g. Chamberlain(1986)).
In addition, we see that if we remove the subset of regressor values such that $|\alpha + v| < \eta$ for any arbitrarily small positive η, we lose point identification. However, the parameter space can be reduced, i.e., the parameter is set identified[5] using the maximum score procedure even after removing this subset of the regressor space.

[4] see, e.g. Newey(1990) for the conditions for an estimator to be regular.

[5] For example, if for a given v, $P(y = 1|v) > \frac{1}{2}$, then this implies that $\alpha + v \geq 0$ which by itself restricts the identified set to a strict subset of the parameter space.

To compare this with the other models we consider below, we note one can easily write

$$\alpha = -v^* \quad \text{where} \quad P(y = 1|v^*) = \frac{1}{2}$$

Hence, α can be written as

$$\alpha = -E[\frac{\delta(P(y = 1|v) - \frac{1}{2})}{f(v)}\, v] \equiv -E[w(v)\, v]$$

where $\delta()$ is the Dirac delta function and $f()$ denotes the density function of v. The above can be seen as a weighted expectation of v where the weight function, in this case the ratio of the delta function to the density[6], takes arbitrarily large values at v^*. Moreover, if we exclude the value v^* from the support of v, then the model does not point identify[7]α. Given this structure, we can define thin set identification in this particular context. Specifically, we say that α is thin set identified in the maximum score model above since the weight function $w(.)$ takes arbitrarily large values on the set $|\alpha + v| < \eta$ for any arbitarily small positive η. This set, which has arbitrarily small measure, is **necessary** for point identification.

The combination of *both* the loss of identification without the thin set and the fact that the weight function above takes arbitrarily large values there are the two properties that are shared by the identification strategies in the models we consider in the rest of this paper. As we will also show, this will result in impossibility theorems analogous to those attained in Chamberlain(1986).

4 Binary Choice with Mean Restrictions

Now, consider the mean independent binary choice model where (3.1) is combined with the assumption that $E(\epsilon) = 0$. Manski(1988) showed that this model does not identify the parameter α. In fact, he showed that the mean independence assumption is not strong enough in binary response models to even bound α. Hence, we modify this model by adding more assumptions to ensure point identification. The ensuing model is a simplified version of the one introduced by Lewbel(1997, 2000):

$$y = 1[\alpha + v - \epsilon \geq 0] \tag{4.2}$$

[6] By no means are we suggesting that the weight function is unique for identifying α as a weighted expectation. In fact, the maximum score is not a weighted estimator. We are adopting it here to illustrate similarities to identification strategies that have been adopted for the parameters in the models we consider in this paper.

[7] Nonetheless, we can still attain set identification which is in contrast to the subsequent models we study. To distinguish the two cases, we will say that the maximum score model is thin set *robustly* identified.

where v is a scalar random variable that is independent of ϵ and *both* ϵ and v have support on the real line. The object of interest is the parameter α. The location restriction on ϵ point identifies α as follows where we also point out that this point identification here is non-robust in a manner that will made clear below. We start with the following:

$$P(y=1|v) \quad = \quad F_\epsilon(\alpha+v) \tag{4.3}$$

where $F_\epsilon(.)$ is the cdf of ϵ where we assume this cdf is a strictly increasing function. Since v and ϵ have support on the real line, then we see that the cdf of ϵ, shifted by α, is identified on *all* its support. Hence, the density of ϵ, f_ϵ is also identified (up to location):

$$f_\epsilon(\alpha+v) = \partial_v P(y=1|v) \tag{4.4}$$

Exploiting the mean zero condition on ϵ, we have:

$$\begin{aligned} 0 \quad &= \quad E[\epsilon] = \int_{-\infty}^{+\infty} u f_\epsilon(u)du \\ &=_{(i)} \quad \int_{-\infty}^{+\infty} (\alpha+v)f_\epsilon(\alpha+v)dv =_{(ii)} \alpha + \int_{-\infty}^{+\infty} v\partial_v P(y=1|v)dv \end{aligned}$$

where (i) follows from a change of variable and the fact that both v and ϵ have infinite support and (ii) follows from (4.4). Hence this means that

$$\alpha = -\int_{-\infty}^{+\infty} v\partial_v P(y=1|x=v)dv$$

Since $\partial P(y=1|v)$ is identified on the support of v which is the real line, then one can estimate α by simply taking a sample analog of the above. Lewbel(1997) derived the following equivalent relation

$$\alpha = E[(y_i - I[v_i>0])f(v_i)^{-1}] \equiv E[(y_i - I[v_i>0])w(v_i)] \tag{4.5}$$

with the weight function $w(v_i) = f(v_i)^{-1}$ where $f(.)$ here denotes the density of v_i. We note this identification result satisfies the two conditions for "thin set identification" mentioned previously. Identification of α is lost[8] when the support of $\alpha + v_i$ is exceeded by the support of ϵ_i, so in this case (where ϵ_i has unbounded support), $v^* = \pm\infty$. The density of v_i vanishes at these points, implying like in the Manski model, that there is a set of arbitrarily small measure, in this case $|v| > M$, for an arbitrarily large constant M, where the weight function becomes unbounded. This suggests α is irregularly identified, a result we will formally establish by showing an impossibility theorem analogous to Chamberlain(1986).

[8] If one excludes the thin set, then the model will not contain any information about α (trivial bounds). This is a case of **non-robust** identification. To see this, note that if we restrict the support of v to lie on the set $[-K, K]$ *for any* $K > 0$, α will NOT be point identified. To see this, note that in 4.3, we can only

4.1 Infinite Bounds

In this section, we formally show that efficiency bounds are infinite for a variant of model (4.2) above. Now, introducing regressors x_i, we alter the assumption so that $\epsilon|x, v =^d \epsilon|x$ and that $E[\epsilon|x] = 0$ (here, $=^d$ means "has the same distribution as"). We also impose other conditions such that $\epsilon|x$ and $v|x$ have support on the real line for all x. The model we consider now is

$$y = 1[\alpha + x\beta + v - \epsilon \geq 0] \tag{4.6}$$

The following theorem, proven in the appendix, shows the identification is irregular. Specifically, it states that there does not exist a regular root-n estimator for α and β in this model.

Theorem 4.1 *In the binary choice model 4.6 with exclusion restriction and unbounded support on the error terms, and where $\epsilon|x, v =^d \epsilon|x$ and $E[\epsilon|x] = 0$, if the second moments of x, v are finite, then the semiparametric efficiency bound for β is not finite.*

A related result, focusing exclusively on the intercept term is the following:

Theorem 4.2 *In the homoskedastic binary choice model with unbounded support on the error terms and no regressors except for v, if the second moment of v is finite, the semiparametric efficiency bound for the intercept term is not finite.*

The proof of the above result is omitted as it follows from identical arguments used in proving Theorem 4.1.

Remark 4.1 *The proof of the above results are based on the assumption of second moments of the regressors being finite. This type of assumption was also made in Chamberlain(1986) in establishing his impossibility result for the maximum score model.*

learn F_ϵ from $\alpha - K$ to $\alpha + K$. Hence,

$$\alpha = \underbrace{-\int_{-\infty}^{-K} v f_\epsilon(\alpha + v)dv}_{(1)} - \underbrace{\int_{-K}^{K} v\partial_v P(y = 1|v)dv}_{(2)} - \underbrace{\int_{K}^{\infty} v f_\epsilon(\alpha + v)dv}_{(3)}$$

We see that only (2) is identified, while (1) and (3) are not. However, one can bound (1) and (2). So, we see that no matter how large K is, the model is set identified. In addition, we see that one can easily choose the unidentified portion of $f_\epsilon(.)$ (parts (1) and (2) above) in such a way that the model provides NO information about α. This was shown for the Lewbel model in the recent paper by Magnac and Maurin (2007). This is in contrast to the maximum score model where one is able to nontrivially bound α even when the thin set, $|\alpha + v| < \eta$, is excluded.

We see that in general, it is difficult to estimate parameters in semiparametric binary choice models with mean restrictions. Moreover, in mean based models, the point identification is fickle or non-robust. The above impossibility result motivates us to characterize *achievable* rates of convergence for particular cases of model (4.6) above. We do this next.

4.2 Relative Tail Behavior and Rates of Convergence: Density Weighting

The previous section established impossibility theorems for the binary response model under mean restriction and hence showed that those are irregularly identified. Here we derive the rates of convergence and show that these depend on the relative tail behavior. In generic cases, the rate of convergence for the estimator of the intercept term in (4.2) is slower than parametric rate. The result seems to hold for any model where the tail of the special regressor v is as thin or thinner than the tail of the error term. Moreover, there are cases (when the moment conditions in the above theorem are violated) where the rate of convergence reaches the regular parametric rate. As we show below, when v is Cauchy, then the estimator of α converges at root n rate[9].

We start with the following estimator α that was proposed by Lewbel (1997)[10]:

$$\hat{\alpha}_n = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i - I[v_i > 0]}{f(v_i)} I[v_i \leq \gamma_n] \tag{4.7}$$

The estimator includes the additional trimming term $I[v_i \leq \gamma_n]$ where γ_n is a deterministic sequence of numbers that satisfies $\lim_{n\to\infty} \gamma_n = \infty$ and $\lim_{n\to\infty} \gamma_n/n = 0$. Effectively, this extra term helps govern tail behavior. Trimming this way suffices (under the assumptions in Lewbel(1997))to deal with the denominator problem associated with the density function $f(v_i)$ getting small.

Next we define $\bar{\alpha}_n = E[\hat{\alpha_n}]$. In what follows we will establish a rate of convergence and limiting distribution for $\hat{\alpha}_n - \bar{\alpha}_n$. To do so we first define the sequence of constants $v(\gamma_n) = \mathrm{Var}(\frac{y_i - I[v_i>0]}{f(v_i)} I[v_i \leq \gamma_n])$, and in what will follow we let $h_n = v(\gamma_n)^{-1}$. Based on our

[9] The notion of attaining a faster rate of convergence when moments are infinite is not new. For example, it is well known that when regressors have a cauchy distribution in the basic linear model, OLS is super-consistent.

[10] Actually, it is an infeasible version since we assume here the density function $f(v_i)$ is known. It is also different in the sense trimming should be based on both tails -i.e. $I[|v_i| \leq \gamma_n]$. Both changes are made only for notational convenience and do not effect the main results of upper bounds for rates of convergence.

previous impossibility result we will generally have $\lim_{n\to\infty} v(\gamma_n) = \infty$, which will result in a slower rate of convergence.

To do so, we will apply the Lindeberg condition to the following triangular array:

$$\sqrt{nh_n}(\hat{\alpha_n} - \bar{\alpha}_n) = \sum_{i=1}^{n} \sqrt{\frac{h_n}{n}} \left(\frac{y_i - I[v_i > 0]}{f(v_i)} I[v_i \leq \gamma_n] - \bar{\alpha_n} \right) \tag{4.8}$$

Before claiming asymptotic normality of the above term we verify that the Lindeberg condition for the array is indeed satisfied. Specifically, we need to establish the limit of

$$h_n E\left[\left(\frac{(y_i - I[v_i > 0])}{f(v_i)} I[v_i \leq \gamma_n] - \bar{\alpha}_n \right)^2 I[\left(\frac{(y_i - I[v_i > 0])}{f(v_i)} - \bar{\alpha}_n \right)^2 > \frac{n}{h_n}\epsilon^2] \right] \tag{4.9}$$

for $\epsilon > 0$. The above limit is indeed zero since $h_n E[\left[\left(\frac{(y_i - I[v_i > 0])}{f(v_i)} I[v_i \leq \gamma_n] - \bar{\alpha}_n \right)^2 \right] = 1$ and $\frac{n}{h_n} \to \infty$. Therefore, we can conclude that

$$\sqrt{nh_n}(\hat{\alpha_n} - \bar{\alpha}_n) \Rightarrow N(0,1) \tag{4.10}$$

So we have established that the rate of convergence of the centered estimator is governed by the limiting behavior of h_n. As we will show below, in most cases, $h_n \to 0$, resulting in a slower rate of convergence, as is to be expected given our previous efficiency bound calculations.

Of course, the above calculations only derives rates for the centered estimator, and not the estimator itself. To complete the distribution theory we need to derive the rate of convergence of the bias term:

$$\sqrt{nh_n}(\bar{\alpha}_n - \alpha_0) \tag{4.11}$$

where from Lewbel(1997) we know that

$$\alpha_0 = E[\frac{y_i - I[v_i > 0]}{f(v_i)}] \tag{4.12}$$

so we have:

$$b_n \equiv \bar{\alpha}_n - \alpha_0 = \int_{\gamma_n}^{\infty} (p(v) - 1) dv \tag{4.13}$$

where $p(v) = P(y_i = 1 | v_i = v)$. Clearly we have $\lim_{n\to\infty} b_n = 0$ if we maintain that $\lim_{n\to\infty} \gamma_n = \infty$.

There we can see that the limiting distribution of the estimator can be characterized by two components, the variance, which we see depends on the limiting ratio of the propensity score divided by the regressor density, and the bias, which depends on the rate at which the propensity score converges to 1. Our results are "high level" as stated since we have not yet stated conditions on γ_n except that it increases to infinity. Nor have we stated what the sequences h_n and b_n look like as functions of γ_n. We show now how they relate to tail behavior assumptions on the regressor distribution and latent error term.

First we calculate the form of the variance term $v(\gamma_n)$ as a function of γ_n. Note we can express $v(\gamma_n)$ as the following integral:

$$v(\gamma_n) = \int_{-\infty}^{\gamma_n} \frac{p(v)(1-p(v)}{f(v)}dv + \mathrm{Var}((E[(\frac{y_i - I[v_i > 0]}{f(v_i)}I[v_i \leq \gamma_n])|v_i] - \bar{\alpha}_n)) \tag{4.14}$$

We will focus on the first term because we will show the second term is negligible when compared to the first. The second term in the variance is of the form:

$$\mathrm{Var}(E[(\frac{y_i - I[v_i > 0]}{f(v_i)}I[v_i \leq \gamma_n])|v_i] - \bar{\alpha}_n) \tag{4.15}$$

The above variance is of the form:

$$E[(\bar{\alpha}_n(v_i) - \bar{\alpha}_n)^2] \tag{4.16}$$

where $\bar{\alpha}_n(v_i)$ denotes the conditional expectation in (4.15). Note that since $\bar{\alpha}_n$ converges to α_0 and $E[\bar{\alpha}_n(v_i)] = \bar{\alpha}_n$ the only term in the above expression that may diverge is

$$E[\bar{\alpha_n}(v_i)^2] \tag{4.17}$$

This term can be expressed as the integral:

$$\int_{-\infty}^{\gamma_n} \frac{(1-p(v))^2}{f(v)}dv \tag{4.18}$$

which will generally be dominated by the first piece of the variance term, which recall was of the form:

$$\int_{-\infty}^{\gamma_n} \frac{p(v)(1-p(v))}{f(v)}dv \tag{4.19}$$

So generally speaking, as far as deriving the optimal rate of convergence for the estimator, we can ignore this component of the variance.

Using similar arguments we can show that the bias term behaves asymptotically like:

$$b_n \approx \gamma_n(1 - p(\gamma_n)) \tag{4.20}$$

With these general results, we now calculate the rate of convergence for some special cases corresponding to various tail behavior conditions on the regressors and the error terms.

4.2.1 Rates of Convergence in Special Cases

Here we derive the rates of convergence under various tail behavior conditions on both the latent error and the regressor.

- [Logit Errors/Logit Regressors] Here we assume the latent error term and the regressors both have a standard logistic distribution. We consider this to be the main example, as the results that arise here (a slower than root-n rate) will generally also arise anytime we have distributions whose tails decline exponentially, such as the normal, logistic and Laplace distributions. From our calculations in the previous section we can see that

$$v(\gamma_n) = \gamma_n \quad \text{and} \quad b_n = \gamma_n \frac{\exp(-\gamma_n)}{1+\exp(-\gamma_n)} \tag{4.21}$$

So to minimize mean squared error we set $\frac{\gamma_n}{n} = \gamma_n^2 \frac{\exp(-2\gamma_n)}{(1+\exp(-\gamma_n))^2}$ to get $\gamma_n = O(\log n^{1/2})$, resulting in the rate of convergence of the estimator:

$$\sqrt{\frac{n}{\log n^{1/2}}}(\hat{\alpha} - \alpha_0) = O_p(1) \tag{4.22}$$

Furthermore, from the results in the previous section, the estimator will be asymptotically normally distributed, and have a limiting bias.

- [Normal Errors/Normal Regressors] Here we assume that the latent error term and the regressors both have a standard normal distribution. To calculate $v(\gamma_n)$ in this case, we can use the asymptotic properties of the Mills Ratio (see, e.g. Gordon(1941, AMS, 12)), $r(\cdot)$:

$$r(t) \sim \frac{1}{t} \text{ as } t \to \infty \tag{4.23}$$

to approximate $v(\gamma_n) \approx \log(\gamma_n)$. For this case we have $b_n \approx \gamma_n \exp(-\gamma_n^2/2)$. So to minimize MSE, we set $\gamma_n = \sqrt{\log n}$ and the estimator is $O_p(\sqrt{\frac{\log(\sqrt{\log n})}{n}})$.

- [Logit Errors/ Normal Regressors] Here we assume the latent error term has a logistic distribution but the regressors have a normal distribution. As we show here the rate of convergence for the estimator will be very slow. The variance is of the form:

$$v(\gamma_n) = \int^{\gamma_n} \exp(v^2/2)dv \tag{4.24}$$

which can be approximated as: $v(\gamma_n) \approx \exp(\gamma_n^2/2)\gamma_n^{-1}$ see, e.g. Weisstein, Eric W. (1999)[11]. The bias of the form:

$$b_n \approx \gamma_n \exp(-\gamma_n) \tag{4.25}$$

So the MSE minimizing sequence is of the form: $\gamma_n = \sqrt{\log n}$ Resulting in the rate of convergence

$$O_p\left(\frac{n^{-1/4}}{\sqrt[4]{\log n}}\right)$$

- [Logit Errors/Cauchy Regressors] Here we assume the latent error term has a standard logistic distribution but the regressor has a standard cauchy distribution. From our calculations in the previous section we can see that

$$v(\gamma_n) = \gamma_n \frac{\exp(-\gamma_n)}{(1+\exp(-\gamma_n))^2}(1+\gamma_n^2) \tag{4.26}$$

and

$$b_n = \gamma_n \frac{\exp(-\gamma_n)}{1+\exp(-\gamma_n)} \tag{4.27}$$

We note in this situation $v(\gamma_n)$ remains bounded as $\gamma_n \to \infty$, so the variance of the estimator is $O(\frac{1}{n})$. Therefore we can let γ_n increase to infinity as quickly as desired to ensure $b_n = o_p(n^{-1/2})$. Therefore in this case we can conclude that the estimator is root-n consistent and asymptotically normal with no asymptotic bias.

- [Probit Errors/Cauchy Regressors] Here we assume the latent error term has a standard normal distribution but the regressor has a standard cauchy distribution. From our calculations in the previous section we can see that

$$v(\gamma_n) \approx \gamma_n \exp(-\gamma_n^2/2)\gamma_n^{-1}(1+\gamma_n^2) \tag{4.28}$$

where the above approximation is based on the asymptotic series expansion of the erf. Furthermore

$$b_n = \gamma_n \exp(-\gamma_n^2/2)\gamma_n^{-1} \tag{4.29}$$

We can see immediately that no matter how quickly $\gamma_n \to \infty$, $v(\gamma_n) = O(1)$, so we can set $\gamma_n = O(n^2)$ to that

$$\sqrt{n}(\hat{\alpha} - \alpha_0) = O_p(1) \tag{4.30}$$

[11] "Erfi." From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/Erfi.html.

and furthermore the estimator will be asymptotically normal and asymptotically unbiased.

- [Differing Support Conditions: Regular Rate] Here we assume the support of the latent error term is strictly smaller than the support of the regression function. For γ_n sufficiently large $p(\gamma_n)$ takes the value 1, so in this case we need not use the above approximation and $\lim_{\gamma_n \to \infty} v(\gamma_n)$ is finite. This implies we can let γ_n increase as quickly as possible to remove the bias, so the estimator is root-n consistent and asymptotically normal with no limiting bias. This is a case where for example $\alpha + v$ in (4.2) has strictly larger support than ϵ's.

So to summarize our results, we say that the optimal rate is directly tied to the relative tail behavior of the special regressor and the error term. Rates will generally be slower than the parametric rate, which can only be attained with infinite moments on the regressor, or strong support conditions. The latter ensure the propensity scores attains its extreme values with positive measure.

In one sense this is not surprising since when imposing such a strong condition, other existing impossibility theorems can be reversed. For example, in the binary choice conditional median model (maximum score), if the propensity score is assumed to be able to takes values in $\{0, 1\}$, then in the appendix (Section A.1) we provide an estimator for the slope in this maximum score setting that converges at the regular rate.

4.3 Rate Adaptive Inference

As the results in the previous section illustrates, the rates of convergence for estimators of thinly set identified models can vary widely with tail behavior conditions on both observed and unobserved random variables, and the optimal rate rarely coincides with the parametric rate. As mentioned, consequently we feel semiparametric efficiency bounds might not be as useful for this class of models. Instead, we propose the use of rate-adaptive inference procedures as was done in Andrews and Schafgans(1998). We will illustrate this procedure, meant to be applied to all the models considered in this paper, by focusing on the binary choice model under a mean restriction. Here we propose a rate adaptive procedure for the intercept term. Let $\hat{\alpha}_n$ denote the trimmed variant of the estimator proposed in Lewbel(1997) for the intercept term:

$$\hat{\alpha}_n = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i - I[v_i > 0]}{f(v_i)} I[v_i \le \gamma_{2n}] \tag{4.31}$$

And let $\hat{S}_n$ denote a trimmed estimator of its asymptotic variance if conditions were such that the asymptotic variance were finite[12]:

$$\hat{S}_n = \frac{1}{n}\sum_{i=1}^{n}\frac{p(v_i)(1-p(v_i)}{f(v_i)^2}I[v_i \leq \gamma_{2n}] \tag{4.32}$$

Our main result in this section is that the "studentized" estimator converges to a standard normal distribution. We consider this result as rate-adaptive in the sense that the limiting distribution holds regardless of the rate of convergence of the un-studentized estimator. We state this formally as a theorem, which first requires the following definitions:

$$v(\gamma_{2n}) = E[\frac{p(v_i)(1-p(v_i))}{f(v_i)^2}I[v_i \leq \gamma_{2n}]] \tag{4.33}$$

$$b(\gamma_{2n}) = \alpha_0 - E[\frac{y_i - I[v_i > 0]}{f(v_i)}I[v_i \leq \gamma_{2n}]] \tag{4.34}$$

$$X_{ni} = \frac{y_i - I[v_i > 0]}{f(v_i)}I[v_i \leq \gamma_{2n}] \tag{4.35}$$

Theorem 4.3 *Suppose (i)*$\gamma_{2n} \to \infty$*, (ii)* $\forall \epsilon > 0$*,*

$$\lim_{n\to\infty}\frac{1}{v(\gamma_{2n})}E[X_{ni}^2 I[|X_{ni}| > \epsilon\sqrt{v(\gamma_{2n})}] = 0 \tag{4.36}$$

and (iii) $\sqrt{nv(\gamma_{2n})}b(\gamma_{2n}) \to 0$*, then*

$$\sqrt{n}\hat{S}_n^{-1/2}(\hat{\alpha} - \alpha_0) \Rightarrow N(0,1) \tag{4.37}$$

This result is analogous to the results in Andrews and Schafagans(1998) who considered asymptotic properties of the identification at infinity estimator proposed in Heckman(1990).

Proof: We begin by deriving a representation for $\hat{S}_n$. As mentioned previously, we have $\hat{S}_n = v(\gamma_{2n}) + o_p(1)$, where the sequence of constants $v(\gamma_{2n})$ will either converge to infinity or a finite constant, depending on relative tail behaviors. We next turn attention to the term

$$\frac{1}{\sqrt{n}v(\gamma_{2n})^{-1/2}}\sum_{i=1}^{n}\left(\frac{y_i - I[v_i > 0]}{f(v_i)}I[v_i \leq \gamma_{2n}] - \mu_n\right) \tag{4.38}$$

[12] Note here that $\hat{S}_n$ is infeasible since $p(v)$ is unknown to the econometrician. This was assumed to simplify the arguments in the proof below, but can be relaxed.

where

$$\mu_n = E[\frac{y_i - I[v_i > 0]}{f(v_i)} I[v_i \leq \gamma_{2n}] \tag{4.39}$$

Under (ii), we know the above term converges to a standard normal distribution by the Lindeberg theorem. We also note that the bias term vanishes under our conditions for a wide ranges of rates on γ_{2n}:

$$\sqrt{nv(\gamma_{2n})}b(\gamma_{2n}) \to 0 \tag{4.40}$$

establishing the desired result. ■.

5 Censored Regression Models

In this section we characterize various identification approaches for regression parameters in censored regression models under various error restrictions. Honore, Khan and Powell(2002) established regular identification for fixed and randomly censored regression models under a median restriction.

Koul, Susarla and Van Ryzin(1981) (KSV) consider identification of randomly (right) censored regression models under a mean restriction, using an "inverse weighting" method. However, as we establish here this approach is based on thin set identification, using our definition, and we will establish here that no estimator converging at the parametric rate exists for this model. This generalizes the impossibility result in Newey(2003) for the censored regression model with fixed censoring under a conditional mean restriction.

Specifically, we will consider the model:

$$y_i = \min(\alpha_0 + \epsilon_i, c_i) \tag{5.1}$$

where the econometrician observes the vector (y_i, c_i) and we assume that ϵ_i and c_i are independent. Note this is a more informative sampling scheme than considered in KSV in the sense that the censoring variable is always observed, and note for notational convenience we have suppressed the presence of regressors to focus on the intercept term. KSV showed that when c_i and $\alpha_0 + \epsilon_i$ have the same support, α_0 can be identified as:

$$\alpha_0 = E[d_i y_i / S_c(y_i)] = E[d_i(\alpha_0 + \epsilon_i)w(\alpha_0 + \epsilon_i)] \tag{5.2}$$

where d_i is a censoring indicator and $S_c(\cdot)$ denotes the survivor function of the censoring variable, and $w(\alpha_0 + \epsilon_i) = S_c(\alpha_0 + \epsilon_i)^{-1}$. Thus the above moment condition will have the two properties to be characterized as thin set identified. We note identification is lost in this model when the support of the censoring variable is exceeded by the support of $\alpha_0 + \epsilon_i$, so in the Tobit case where $\alpha_0 + \epsilon_i$ has support on the real line the point to be removed to lose identification is $c^* = \infty$, so $w(c^*) = \infty$. This suggests that the above identification result is irregular, as result we confirm for the more general model with regressors x_i:

Theorem 5.1 *In the model*[13]

$$y_i = \max(x_i'\beta_0 + \epsilon_i, c_i) \tag{5.3}$$

with (i)$E[\epsilon_i|x_i] = 0$, (ii)$E[\|x_i\|^2] < \infty$, (iii)c_i and $x_i'\beta_0 + \epsilon_i$ are independent of each other conditional on x_i, and each have unbounded support, the information bound for β_0 is 0.

Proof: See appendix. Hence no estimator can converge at the parametric rate, so the identification is irregular.

Remark 5.1 *The results here are completely analogous to the binary choice model discussed previously. There, the regular rate could only be attained if extreme support or tail conditions were satisfied. The same is true here where now one needs strong support or tail thickness of the censoring variable, relative to the error term.*

5.1 Estimation and Rates of Convergence

Here we follow the arguments used when considering the binary choice model to derive the optimal rates of convergence for the inverse weight estimator proposed in Koul et al.(1981). To clarify our arguments, we will focus on estimating the intercept term α_0 in the model

$$y_i = \min(\alpha_0 + \epsilon_i, c_i) \tag{5.4}$$

and let d_i denote an indicator variable taking one if the observation is uncensored. As before we assume the censoring point c_i is observed, and the latent error term ϵ_i has mean 0. From Koul et al.(1981) we have the moment condition:

$$\alpha_0 = E[d_i y_i / S_c(y_i)] \tag{5.5}$$

[13] Note our impossibility result is expressed for the left censoring model so that the result can be compared to Newey(2003). Such a result will also hold for the right censored model.

where $S_c(\cdot)$ denotes the survivor function of the censoring variable. This suggests the analog estimator:

$$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} \frac{d_i y_i}{S_c(y_i)} \tag{5.6}$$

And as before there will be many situations when the variance of the term inside the summation will be infinite. Consequently, we will trim the variable y_i at the sequence of constants γ_n, and denote the variance by $v(\gamma_n)$. Here we have:

$$v(\gamma_n) = E[I[\alpha_0 + \epsilon_i \leq \gamma_n]\frac{\alpha_0^2 + \epsilon_i^2 + 2\alpha_0\epsilon_i}{S_c(\alpha_0 + \epsilon_i)}] \tag{5.7}$$

whose lead term as ϵ_i approaches infinity behaves like

$$v(\gamma_n) \approx \int_{-\infty}^{\gamma_n - \alpha_0} \frac{\epsilon_i^2 f(\epsilon_i)}{S_c(\alpha_0 + \epsilon_i)} d\epsilon_i \tag{5.8}$$

and the bias induced by trimming will be of the form

$$b(\gamma_n) = \alpha_0 - \int_{-\infty}^{\infty} I[\alpha_0 + \epsilon_i \leq \gamma_n](\alpha_0 + \epsilon_i) f(\epsilon_i) d\epsilon_i \tag{5.9}$$

which can be decomposed as the sum of

$$-\alpha_0 \int_{\gamma_n - \alpha_0}^{\infty} f(\epsilon_i) d\epsilon_i + \int_{-\infty}^{\gamma_n - \alpha_0} \epsilon_i f(\epsilon_i) d\epsilon_i \tag{5.10}$$

whose terms behave asymptotically like

$$\gamma_n^{-1} f(\gamma_n) + f(\gamma_n) \approx f(\gamma_n) \tag{5.11}$$

So as before we can attain optimal rates of γ_n. We now consider some special cases of error term and censoring variable tail behavior:

- [Logit Errors/Logit Censoring] Here we assume the latent error term has a standard logistic distribution as does the censoring variable. From our calculations in the previous section we can see that

  $$v(\gamma_n) \approx O(\gamma_n^3) \tag{5.12}$$

  and

  $$b_n \approx \exp(-\gamma_n) \tag{5.13}$$

So equating the variance with the square of the bias we get

$$\gamma_n = \frac{1}{2}\log n \tag{5.14}$$

So

$$\hat{\alpha} - \alpha_0 = O_p(\sqrt{(\log n)^3/n}) \tag{5.15}$$

Differing Support Conditions: $\sqrt{n}$ rate Here we assume the support of the latent dependent variable is strictly smaller than the support of the censoring variable. Using analogous arguments we can show the estimator is root-n consistent and asymptotically normal with no limiting bias. This result is analogous to what we attained previously since in this setting the propensity score, in this case the probability of being uncensored becomes one with positive measure.

6 Treatment Effects Model under Exogenous Selection

This section shows that another parameter of interest- the Average Treatment Effect is only point identified under support conditions and thus can only be estimated at the parametric rate under strong tail behavior restrictions.

A central problem in evaluation studies is that potential outcomes that program participants would have received in the absence of the program is not observed. Letting d_i denote a binary variable taking the value 1 if treatment was given to agent i, and 0 otherwise, and letting y_{0i},y_{1i} denote potential outcome variables, we refer to $y_{1i} - y_{0i}$ as the *treatment effect* for the i'th individual. A parameter of interest for identification and estimation is the *average treatment effect*, defined as:

$$\beta = E[y_{1i} - y_{0i}] \tag{6.1}$$

One identification strategy for β was proposed in Rosenbaum and Rubin(1983), who imposed the following:

(i) There exists an observed variable x_i s.t. d_i and (y_{0i}, y_{1i}) are independent of each other given x_i.

(ii) $0 < P(d_i = 1|x_i) < 1 \;\; \forall x_i$

from which we can identify β as

$$\beta = E_X[E[y_i|d_i = 1, x_i] - E[y_i|d_i = 0, x_i]] \tag{6.2}$$

or

$$\beta = E_X[E[y_i|d_i = 1, p(x_i)] - E[y_i|d_i = 0, p(x_i)]] \tag{6.3}$$

where $p(x_i) = P(d_i = 1|x_i)$. Hirano et al.(1999) showed the following inverse weighting identification result:

$$\beta = E[\frac{y_i(d_i - p(x_i))}{p(x_i)(1 - p(x_i))}] \tag{6.4}$$

Clearly identification is lost when we remove any region in the support of x_i. Note however, that regions in the tails of x_i where the weight function $p(x_i)^{-1}(1 - p(x_i))^{-1}$ becomes unbounded, putting the above identification result into our class of thin set identification, suggesting the irregularity of the identification without further restrictions than those stated above.

Whether or not this identification result is regular is a delicate equation, and depends on the support conditions on the propensity score. It is *regular* if we strengthen the above condition on the propensity scores to be such that they are *bounded away* from 0 and 1:

(ii') There exists and $\epsilon > 0$ s.t. $\epsilon < P(d_i = 1|x_i) < 1 - \epsilon \ \ \forall x_i$

The efficiency bound derived in Hahn(1998) holds under (i) and (ii). If we assume (ii') we can efficiently estimate β at the parametric rate as Hahn(1998)'s bound is finite. If on the other hand (ii') does not hold (but (ii) does), there will be *many* situations where the bound is infinite.

For example, in the homoskedastic setting with one continuous regressor, anytime the distribution of the regressor is the same as that of the error term in the treatment equation, so (ii) is satisfied but not (ii'),the variance in Hahn(1998) is infinity.

So if we assume (ii) but not (ii'), the efficiency bound for β can be infinite if we rule out models which $E[p(x_i)^{-1}]$ is finite, such as the class of models considered before. Thus we see that, such as was the case in the binary choice and censored regression models, rates of convergence will depend on relative tail behaviors of regressors and error terms. Consequently, we do not feel it is useful to compute efficiency bounds in this setting under Assumptions (i) and (ii).

An impossibility theorem for this model is stated in the following theorem, whose proof is omitted as it follows from similar arguments to prove the other theorems.

Theorem 6.1 *Assume:*

(ii") *The support of* $p(x_i)$ *is* $(0,1)$, $E_X[p_0(x_i)^{-1}] = \infty$

Then Under Assumptions (i),(ii"), the variance bound for β *is infinite.*

6.1 Average Treatment Effect Estimation

Here we conduct the same rate of convergence exercises for estimating the average treatment effect. We will explore the asymptotic properties of the following estimator:

$$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{d_i y_i}{p(x_i)} - \frac{(1-d_i)y_i}{1-p(x_i)}\right) I[x_i \leq \gamma_n] \tag{6.5}$$

We note that analogous to before the estimator considered is infeasible, this time because we are assuming the propensity score is known. Like in the binary choice model this will not effect our main conclusions. Also, like before we will assuming it suffices to trim on regressor values, as in this case it is large regressor values which cause the denominator problem. Furthermore, to clarify our arguments we will assume here that the counterfactual outcomes are homoskedastic with a variance of 1.

Like before we will define the following terms:

$$\bar{\alpha}_n = E[\hat{\alpha}] \tag{6.6}$$

and

$$v(\gamma_n) = E[(\frac{1}{p(x_i)} + \frac{1}{1-p(x_i)})I[x_i \leq \gamma_n]] = E[\frac{1}{p(x_i)(1-p(x_i)}I[x_i \leq \gamma_n]] \tag{6.7}$$

and

$$b_n = \bar{\alpha_n} - \alpha_0 \tag{6.8}$$

where here

$$\alpha_0 = E[\left(\frac{d_i y_i}{p(x_i)} - \frac{(1-d_i)y_i}{1-p(x_i)}\right)] \tag{6.9}$$

As before, we will define $h_n = v(\gamma_n)^{-1}$.

Carrying the same arguments as before we explore the asymptotic properties of $v(\gamma_n)$ and b_n as $\gamma_n \to \infty$. Generally speaking, anytime $v(\gamma_n) \to \infty$, the fastest rate for the estimator will be slower than root-n.

We can express $v(\gamma_n)$ as the following integral:

$$v(\gamma_n) = \int_{-\infty}^{\gamma_n} \frac{f(x)}{p(x)(1-p(x))} dx \tag{6.10}$$

which using the same expansion as before, will behave asymptotically like

$$v(\gamma_n) \approx \gamma_n \frac{f(\gamma_n)}{p(\gamma_n)(1-p(\gamma_n))} \tag{6.11}$$

For the problem at hand we can write b_n as the following integral:

$$b_n = \gamma_n \int_{\gamma_n}^{\infty} m(x) f(x) dx \tag{6.12}$$

where $m(x)$ is the conditional ATE. We now consider some particular examples corresponding to tail behavior on the error term in the treatment equation and the regressor.

Logit Errors/Regressors/Bounded CATE Here we assume the latent error term and the regressors both have a standard logistic distribution, and to simplify calculations we will assume the conditional average treatment effect is bounded on the regressor support. We consider this to be a main example, as the results that arise here (a slower than root-n rate) will generally also arise anytime we have distributions whose tails decline exponentially, such as the normal, logistic and Laplace distributions. From our calculations in the previous section we can see that

$$v(\gamma_n) = \gamma_n \tag{6.13}$$

and

$$b_n \approx \gamma_n \frac{\exp(-\gamma_n)}{(1+\exp(-\gamma_n))^2} \tag{6.14}$$

Clearly $v(\gamma_n) \to \infty$ resulting in a slower rate of convergence. To get the optimal rate we solve for γ_n that set $v(\gamma_n)/n = b_n^2$. We see this matches up when $\gamma_n = \frac{1}{2}\log n$, resulting in the rate of convergence of the estimator:

$$\sqrt{\frac{n}{\log n^{1/2}}}(\hat{\alpha} - \alpha_0) = O_p(1) \tag{6.15}$$

Furthermore, from the results in the previous section, the estimator will be asymptotically normally distributed, and have a limiting bias.

We notice this the exact same result attained for the binary choice model considered previously. As mentioned, similar slower than parametric rates will be attained when the error distribution and regressor have similar tail behavior.

Normal Errors/Logistic Regressors/Bounded CATE Here we assume the latent error term has a standard normal distribution and the regressors both have a standard logistic distribution, and to simplify calculations we will assume the conditional average treatment effect is bounded on the regressor support.

In this situation we have:

$$v(\gamma_n) \approx \int^{\gamma_n} \frac{\exp(-x)}{1-\Phi(x)} dx \tag{6.16}$$

which by multiplying and dividing the above fraction inside the integral by $\phi(x)$, the standard normal p.d.f, and using the asymptotic approximation to the inverse Mills ratio, we get

$$v(\gamma_n) \approx \int_n^{\gamma} \exp(x^2/2) x dx = \exp(\gamma_n) \tag{6.17}$$

In this setting the bias is of the form

$$b_n = \int_{\gamma_n}^{\infty} f(x) dx \approx \exp(-\gamma_n) \tag{6.18}$$

so the MSE minimizing value of γ_n here is

$$\gamma_n = \log n^{1/3} \tag{6.19}$$

so the estimator is $O_p(n^{-1/3})$.

Differing Support Conditions/Bounded CATE Here we assume the support of the latent error term in the treatment equation is strictly larger than the support of the regressor and the support of the regression function in the treatment equation. Here for γ_n sufficiently large $p(\gamma_n)$ remains bounded away from 0 and 1 so $\lim_{\gamma_n \to \infty} v(\gamma_n)$ is finite. Here we can let γ_n increase as quickly as possible to remove the bias, resulting in a root-n consistent estimator.that is asymptotically normal, completely analogous to the result attained before under differing supports in the binary choice model.

Cauchy Errors/Logistic Regressors/Bounded CATE Here we impose heavy tails on treatment effect error term, assuming it has a Cauchy distribution, but we impose exponential tails on the regressor distribution, assuming here that it has an logistic distribution, though similar results would hold if we assumed normality.

$$v(\gamma_n) \approx \gamma_n \frac{\exp(-\gamma_n)}{\frac{1}{4} - \arctan(\gamma_n)^2} \tag{6.20}$$

where after using L'Hospitale's rule we can conclude that:

$$v(\gamma_n) \approx \gamma_n \exp(-\gamma_n)(1 + \gamma_n^2) \tag{6.21}$$

which remains bounded as $\gamma_n \to \infty$. Therefore we can let γ_n increase arbitrarily quickly in n, say $\gamma_n = O(n^2)$ in which case the estimator will be root-n consistent and asymptotically normal with no asymptotic bias.

7 Conclusion

This paper related the notion of irregular identification to the concept of thin set identification. This was shown to be related to set identification and semiparametric efficiency bounds in a wide class of nonlinear models, which included binary choice, fixed and randomly censored regression models, and treatment effect models. In the process of applying the definition to these models we established new impossibility theorems. The work here questions the usefulness of semiparametric efficiency bounds for this class of models, and instead proposes rate adaptive procedures for conducting inference.

For future research, it would appear useful to study our definition in the context of some nonparametric models. For example, the nonparametric estimators of a censored and truncated regression model proposed in Lewbel and Linton(2003) are based on similar support conditions as those imposed in some of the examples considered in this paper, so a natural question to pose would be how the rates of convergence are affected in those models which also achieve identification at the limit support points.

References

[1] Andrews, D.W.K. and M.M.A. Schafgans (1998), "Semiparametric Estimation of the Intercept of a Sample Selection Model", *Review of Economic Studies*, 65, 497-517. pages

[2] Chamberlain, G. (1986), "Asymptotic Efficiency in Semiparametric Models with Censoring", *Journal of Econometrics*, 32, 189-218 pages

[3] Cosslett, S.R. (1987), "Efficiency Bounds for the Distribution-Free Estimators of the Binary Choice and Censored Regression Models", *Econometrica*, 55, 559-587. pages

[4] Han, A.K. (1987), "Nonparametric Analysis of a Generalized Regression Model: The Maximum Rank Correlation Estimator", *Journal of Econometrics*, 35, 303-316. pages

[5] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.M., Stahel, W.A. (1986), *Robust Statistics – The Approach Based on Influence Functions*, New York: Wiley Interscience. pages

[6] Heckman, J.J. (1990) "Varieties of Selection Bias", *American Economic Review*, 80, 313-18. pages

[7] Horowitz, J.L. (1992), "A Smoothed Maximum Score Estimator for the Binary Response Model", *Econometrica*, 60, 505-531. pages

[8] Horowitz, J.L. (1993a), "Optimal Rates of Convergence of Parameter Estimators in the Binary Response Model with Weak Distributional Assumption" *Econometric Theory*, 9, 1-18. pages

[9] Ichimura, H. (1993) "Semiparametric Least Squares and Weighted SLS Estimation of Single -Index Models", *Journal of Econometrics*, 58, 71-120 pages

[10] Koul H. and V. Susarla and J. Van Rysin (1981) "Regression Analysis with Randomly Right Censored Data", *Annals of Statistics*, 9, pp 1276 - 1288. pages

[11] Kim J., and D. Pollard (1990), "Cube Root Asymptotics", *Annals of Statistics*, 18, 191-219. pages

[12] Klein, R.W. and R.H. Spady (1993), "An Efficient Semiparametric Estimator for Discrete Choice Models", *Econometrica*, 61, 387-421. pages

[13] Koul, H. (1985), "Minimum Distance Estimation in Multiple Linear Regression," *Sankya, Ser. A*, 47, 57-74. pages

[14] Lewbel, A. (1998), "Semiparametric Latent Variable Model Estimation With Endogenous or Mismeasured Regressors," Econometrica, 66, 105–121. pages

[15] Lewbel, A. (2000), "Semiparametric Qualitative Response Model Estimation With Unknown Heteroscedasticity or Instrumental Variables," Journal of Econometrics, 97, 145-177. pages

[16] Lewbel, A., O. Linton (2002), "Nonparametric Censored and Truncated Regression Binomial Data," manuscript. pages

[17] Magnac, T. and E. Maurin (2003), "Identification and Information in Monotone Binary Models," manuscript. pages

[18] Manski, C.F. (1975), "Maximum Score Estimation of the Stochastic Utility Model of Choice", *Journal of Econometrics*, 3, 205-228 pages

[19] Manski, C.F. (1985), "Semiparametric Analysis of Discrete Response: Asymptotic Properties of Maximum Score Estimation", *Journal of Econometrics*, 27, 313-334 pages

[20] Manski, C.F. (1988), "Identification of Binary Response Models", *Journal of the American Statistical Association*, 83, 729-738 pages

[21] Newey, W. K. (2001), "Conditional Moment Restrictions in Censored and Truncated Regression Models", *Econometric Theory*, 17, 863-888. pages

[22] Newey, W.K. and D. McFadden (1994) "Estimation and Hypothesis Testing in Large Samples", in Engle, R.F. and D. McFadden (eds.) , *Handbook of Econometrics, Vol. 4*, Amsterdam: North-Holland. pages

[23] Powell, J.L., J.H. Stock, and T.M. Stoker (1989) "Semiparametric Estimation of Index Coefficients", *Econometrica*, 57, 1404-1430. pages

[24] Sherman, R.P. (1994) "U-Processes in the Analysis of a Generalized Semiparametric Regression Estimator", *Econometric Theory*, 10, 372-395 pages

Appendix

A Proof of Theorem 4.1

To prove this theorem, we follow the approach taken in Chamberlain(1986). Specifically, look for a subpath around which the variance of the parametric submodel is unbounded. We first define the function:

$$g_0(t, x) = P(\epsilon_i \leq t | x_i = x)$$

Note it does not depend on the v because of our assumption of conditional independence. The likelihood function we will be working with is based on the following density function:

$$f(y, x, z, \beta, g) = g(x\beta + v, x)^y (1 - g(x\beta + v, x))^{1-y}$$

We first define the family of conditional distributions, Γ:

Definition A.1 Γ *consists of all functions* $g : R^k \rightarrow R$ *such that for all* $(t, x) \in R \times R^{k-1}$ *we have*

1. g *is continuous.*
2. $g'(t, x)$*, the partial derivative of* $g(t, x)$ *with respect to its first argument, is continuous and positive.*
3. $\lim_{s \rightarrow -\infty} g(s, x) = 0, \quad \lim_{s \rightarrow +\infty} g(s, x) = 1.$
4. $\int s g'(s, x) ds = 0$

We next define the set of sub-paths, Λ, we will work with:

Definition A.2 Λ *consists of the paths*[14]*:*

$$\lambda(\delta) = g_0(1 + (\delta - \delta_0)h)$$

where g_0 *is the "true" distribution function, assumed to be an element of* Γ*, and* $h : R^k \rightarrow R$ *is a continuously differentiable function that is 0 outside a compact set, and satisfies* $\int s h'(s, x) ds = 0 \ \ \forall x$

[14] This is just one set of paths that can be used to establish the desired result. For example, one could also work with the paths $g_0 + (\delta - \delta_0)$

We note that the scores of the root likelihood function are:

$$\psi_j(y,x,v) = \frac{1}{2}\left\{ y g_0^{-1/2}(x\beta_0+v,x) - (1-y)(1-g_0(x\beta_0+v,x))^{-1/2}\right\} g_0'(x\beta_0+v,x)x_{(j)}$$

where $x_{(j)}$ the j^{th} component of x.

$$\psi_\lambda(y,x,v) = \frac{1}{2}\left\{ y g_0^{-1/2}(x\beta_0+v,x) - (1-y)(1-g_0(x\beta_0+v,x))^{-1/2}\right\} g_0(x\beta_0+v,x)h(x\beta_0+v,x)$$

We will show that:

Theorem A.1 *Let $I_{\lambda,j}$ denote the partial information for the j^{th} component of β_0 as defined in Chamberlain(1986).*

If $P(x_i\beta_0 + v_i = 0) = 0$ then if the second moment of the vector (x_i, v_i) is finite,

$$\inf_{\lambda\in\Lambda} I_{\lambda,j} = 0$$

.

Note heuristically we will get the desired result if we define $h(t,x) = a(x,v)c(t)$ where $a(x,v)$ is an arbitrarily close to $f_{\epsilon|X}(x\beta_0+v|x)/g_0(x\beta_0+v)\cdot x$, and $c(t)$ is the function that takes the value 1 on its support.

To fill in these details, we define Q_λ and π as:

$$\pi(A) = \int_A g_0(x\beta_0+v,x)(1-g_0(x\beta_0+v))^{-1} f_V(v)dF_X(x)$$

$$Q_\lambda = \int \left[b(x\beta_0+v,x)x - h(x\beta_0+v,x)\right]^2 d\pi(v,x)$$

where $b(\cdot,\cdot) = g_0'(\cdot,\cdot)/g_0(\cdot,\cdot)$. We can then add and subtract $a(x,v)$ to the above integrand, inside the square.

Note we can make the term

$$\int \left[b(x\beta_0+v,x)x - a(x,v)\right]^2 d\pi(v,x)$$

arbitrarily small by the denseness of the space of continuously differentiable functions.

This result, that the above integral can be made arbitrarily small, will follow from Lemma A.2 in Chamberlain(1986) if we can show that $b(x\beta_0 + v) \in \mathcal{L}_2(\pi)$, where

$$d\pi(v, x) = g_0(x\beta_0 + v, x)(1 - g_0(x\beta_0 + v, x))^{-1} f_X(x) f_V(v) dx dv$$

So in other words all we need to show is the finiteness of the following integral:

$$\int \frac{g_0'(x\beta_0 + v, x)^2}{g_0(x\beta_0 + v, x)(1 - g_0(x\beta_0 + v, x))} x_{(j)}^2 f_X(x) f_V(v) dx dv \tag{A.1}$$

Finiteness will follow for all distributions satisfying the assumptions in the definition of Γ, which will imply the uniform (in v, x boundedness of the term

$$\frac{g_0'(x\beta_0 + v, x)^2}{g_0(x\beta_0 + v, x)(1 - g_0(x\beta_0 + v, x))}$$

This is true because under our assumptions, for any finite t, $\frac{g_0'(t,x)}{g_0(t,x)(1-g_0(t,x))}$ is finite, so the only possibility of the fraction becoming unbounded is as $t \to \pm\infty$. However, by considering $t \to \infty$ and applying Hospitale's rule, this would be equivalent to the unboundedness of $\lim_{t\to+\infty} g_0''(t, x)$, (with $g_0''(t, x)$ the second partial derivative of $g_0(t, x)$). But that would contradict $g_0'(t, x)$ being a density function limiting to 0 as $t \to \pm\infty$.

The boundedness of the above ratio will imply finiteness of the integral in (A.1) by the finite second moment assumption of x and the fact that $f_V(v)$ is a density function and integrates to 1.

Finally, the term:

$$\int [a(x, v) - a(x, v)c(t)]^2 \, d\pi(v, x)$$

can be made arbitrarily small by setting $c(t)$ to 1 in most of its (compact) support. Furthermore, with this same definition of $c(t)$ we can satisfy $\int c'(t)tdt = 0$. Notice we can make $c(t)$ "smoother" in the way it drops to 0. For example, we could make the support of $c(t)$ go from -2ρ to +2ρ, where ρ is sufficiently large. Then make $c(t) = 1$ if $|t| < \rho$, and $c(t)$ declines linearly to 0 when t is in $(-2\rho, \rho)$ and $(\rho, 2\rho)$.

A.1 Regular rates with extreme values of the propensity score:

It is interesting to note that we require that the unobserved term has support on the real plane. This is crucial since with bounded support, the propensity score hits zero or one with

high probability (for example if the index has larger support) and it is possible in this case to obtain regular consistent estimators for β (up to scale). For example, consider the binary choice model of section 3 above. Suppose that x_1 is such that $x_1|x_{-1}$ has support on the real line for all x_{-1} a.s., while $x'_{-1}\beta_{(-1)} + \epsilon$ has support in $[K_1, K_2]$ where K_1 and K_2 are finite. Then, it is easy to see that for values of $|x_1|$ that are large, the propensity score is equal to one or zero. In this case, one is able to point identify and estimate the parameter at the parametric rate if there is sufficient "variation" on the set where the propensity score is zero or one. To see that, we first describe heuristically an estimation procedure and in the appendix we show that this estimator is converges at the regular rate.
One can first collect the obervations for which the propensity score is 1 or 0 and then focus on these observations. We normalize $\beta_1 = 1$ and then for these observations (for which propensity score is one or zero), we have (notice here we only condition on x_{-1})

$$\begin{aligned} Pr(y=1|x_{-1}) &= \int F_{\epsilon|x}(x'\beta)1\big[x_1 \in P_1(x_{-1}) \cup P_2(x_{-1})\big]f(x_1|x_{-1})dx_1 && \text{(A.2)} \\ &= \int 1\big[x_1 \in P_1(x_{-1})\big]f(x_1|x_{-1})dx_1 && \text{(A.3)} \\ &= \int 1\big[x_1 \leq x_{-1}\beta_{-1} + K_1(x_{-1})\big]f(x_1|x_{-1})dx_1 && \text{(A.4)} \\ &= F_{x_1|x_{-1}}\big(x_{-1}\beta_{-1} + K_1(x)\big) && \text{(A.5)} \end{aligned}$$

where for a given X_{-1}, P_1 and P_2 are regions for x_1 where the propensity score is equal to 1 and zero respectively, and $K_1(x)$ ($K_2(x)$) is the smallest (largest) value for x_1 for which the propensity is one (zero). These K's are observed in the data. Finally, $F_{x_1|x_{-1}}$ is the distribution function of x_1 conditional on x_{-1} and this too is observed in the data. Hence since K_1 is observed, one can easily solve for β given the usual regularity conditions. Now, we provide an estimator in this case when the propensity score is zero or one and show that this estimator converges at the root n rate.

Our main result here, of establishing root-n consistency applies to any binary choice model where the propensity scores can take the values 1 and 0. This will include the median model in Manski(1975,1985), the Lewbel model, and a binary choice model under a mean restriction on the error term without the exclusion restriction imposed in Lewbel.

Recall that in the regions where the propensity scores were 0 or 1, we had the following equality:

$$P(y_i = 1|x_{-1}) = F_{x_1|x_{-1}}(x'_{-1}\theta_0 + K_1(x_{-1})) \tag{A.6}$$

where recall $K_1(x_{-1})$ corresponds to the smallest (largest) value of x_1 for which the propensity score is one (zero), and whose value varies with the values of x_{-1}. Here $\theta_0 = \beta_{-1}$. In what follows we will refer to the term on the left hand side of the above equation as the *partial propensity score*, as it is a probability conditional on a subset of the regressors.

Here we propose a method to translate the above equality into an estimation procedure for θ_0. As mentioned previously,the functions $P(y_i = 1|x_{-1} = \cdot)), F_{x_1|x_{-1}}(\cdot), K_1(\cdot)$ are observable from the data. Denote the kernel estimators of the first two by $\hat{p}_{-1}(\cdot)$, and $\hat{F}(\cdot)$. The estimator of $K(\cdot)$ can be obtained by the smallest or largest order statistic, and we denote its estimator by $\hat{K}(\cdot)$. Recall, were are assuming that the special regressor is continuously distributed with positive density on the real line, regardless of the value of the other regressors. This motivates inverting the kernel estimator of the conditional c.d.f of x_1, resulting in a second stage least squares estimator of θ_0

$$\hat{\theta} = \arg\min_{\theta} \sum_i (\hat{F}^{-1}(\hat{p}_{-1}(x_{-1i})) - x'_{-1i}\theta + \hat{K}_1(x_{-1i}))^2 \tag{A.7}$$

where the above summation is over the subset of the n observations in the data set where the propensity scores take the values 0 or 1. We can express this estimator in a closed form. Let $\hat{y}_i$ denote $\hat{F}^{-1}(\hat{p}_{-1}(x_{-1i})) + \hat{K}_1(x_{-1i})$ and let $\tilde{x}_{-1i}$ denote $x_{-1i}I[p_i = 0 \text{ or } p_i = 1]$ with p_i denoting the propensity score. Then the estimator is of the form:

$$\hat{\theta} = \left(\frac{1}{n} \sum_{i=1}^{n} \tilde{x}_{-1i}\tilde{x}'_{-1i} \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} \tilde{x}_{-1i}\hat{y}_i \tag{A.8}$$

We list a set of sufficient conditions for the above estimator to be root-n consistent and asymptotically normal.

A1 The matrix $\Sigma_{xx} \equiv E[\tilde{x}_{-1i}\tilde{x}'_{-1i}]$ is invertible.

A2 The functions $p_{-1}(\cdot)$, $F_{x_1|x_{-1}}(\cdot)$ are both assumed to be s times continuously differentiable with bounded derivatives on their support, with the integer s satisfying $s > d/2$, d being the dimension of the continuous components of x_i.

A3 The bandwidth sequence h_{1n} used in the kernel estimation of $p_{-1}(\cdot)$ and $F_{x_{-1}}(\cdot)$ satisfies $\sqrt{n}h_{1n}^s \to 0$ and $nh_{1n}^{d-1} \to \infty$. The bandwidth sequence h_{2n} used in the kernel estimation of $F_{x_1,x_{-1}}(\cdot)$ satisfies $\sqrt{n}h_{2n}^s \to 0$ and $nh_{2n}^d \to \infty$.

The asymptotic distribution is characterized in the following theorem:

Theorem A.2 *Under Assumptions A1-A3, we have*

$$\sqrt{n}(\hat{\theta} - \theta_0) \Rightarrow N(0, \Sigma_{xx}^{-1}\Omega\Sigma_{xx}^{-1}) \tag{A.9}$$

where

$$\Omega = E[\psi_i \psi_i'] \tag{A.10}$$

with

$$\psi_i = \tilde{x}_{-1i} \frac{1}{f_{x_1|x_{-1}}(F_{x_1 x_{-1}}^{-1}(p_{-1i}))}(d_i - p_{-1i}) \tag{A.11}$$

where p_{-1i} denotes $p_{-1}(x_{-1i})$.

Proof: Note we have the following relationship:

$$\hat{\theta} - \theta_0 = \left(\frac{1}{n}\sum_{i=1}^{n} \tilde{x}_{1i}\tilde{x}_{1i}'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} \tilde{x}_{1i}(\hat{F}^{-1}(\hat{p}_{-1i}) - F^{-1}(p_{-1i})) + o_p(n^{-1/2}) \tag{A.12}$$

where the remainder term of $o_p(n^{-1/2})$ results from the minimum order statistic used to estimate $K_1(\cdot)$ converges to the true value at a faster than parametric rate, given our positive density assumption on x_{1i}.

We note by Assumption A1, the denominator term converges in probability to Σ_{xx}. We can decompose the numerator term as:

$$\frac{1}{n}\sum_{i=1}^{n} \tilde{x}_{1i}(F^{-1}(\hat{p}_{-1i}) - F^{-1}(p_{-1i})) + \tag{A.13}$$

$$\frac{1}{n}\sum_{i=1}^{n} \tilde{x}_{1i}(\hat{F}^{-1}(\hat{p}_{-1i}) - F^{-1}(\hat{p}_{-1i})) \tag{A.14}$$

Under Assumptions A2-A3, we can use the same arguments as in Khan and Powell(2001) to conclude that (A.13) has the representation:

$$\frac{1}{n}\sum_{i=1}^{n} \psi_i + o_p(n^{-1/2}) \tag{A.15}$$

and we can also conclude that (A.14) is $o_p(n^{-1/2})$. The conclusion of the theorem follows from the CLT and Slutsky's theorem. ■.

B Proof of Theorem 5.1

Here we can use arguments completely analogous to those used in Theorem 4.1.

First note that we can write the loglikelihood function for the (left) censored as:

$$L = d_i \log(f(y_i - x_i'\beta)) + (1 - d_i)\log(F(c_i - x_i'\beta))$$

where d_i denotes a censoring indicator, taking the value 1 for uncensored observations. We again define the function:

$$g_0(t, x) = P(\epsilon_i \leq t | x_i = x)$$

We again define the family of conditional distributions, Γ:

Definition B.1 *Γ consists of all functions $g : R^k \to R$ such that for all $(t, x) \in R \times R^{k-1}$ we have*

1. *g is continuous.*
2. *$g'(t, x)$, the partial derivative of $g(t, x)$ with respect to its first argument, is continuous and positive.*
3. $\lim_{s \to -\infty} g(s, x) = 0, \quad \lim_{s \to +\infty} g(s, x) = 1.$
4. $\int s g'(s, x) ds = 0$

We next define the set of sub-paths, Λ; as before we will work with:

Definition B.2 *Λ consists of the paths:*

$$\lambda(\delta) = g_0(1 + (\delta - \delta_0)h)$$

where g_0 is the "true" distribution function, assumed to be an element of Γ, and $h : R^k \to R$ is a continuously differentiable function that is 0 outside a compact set, and satisfies $\int s h'(s, x) ds = 0 \ \ \forall x$

We note that for the censored model the scores of the root likelihood function are:

$$\begin{aligned} \psi_j(y, x, c) &= \frac{1}{2} d g_0'^{-1/2}(y - x\beta_0, x) g_0''(y - x\beta_0, x) x & \text{(B.1)} \\ &+ \frac{1}{2}(1 - d)(g_0(c - x'\beta_0, x))^{-1/2} g_0'(c - x'\beta_0, x) x & \text{(B.2)} \end{aligned}$$

and

$$\begin{aligned}\psi_\lambda(y,x,c) &= \frac{1}{2}dg_0^{\prime -1/2}(y-x\beta_0,x)g_0'(y-x\beta_0,x)h(y-x\beta_0,x)+g_0(y-x\beta_0,x)h'(y-x\beta_0,x) && \text{(B.3)}\\ &+ \frac{1}{2}(1-d)(g_0(c-x'\beta_0,x))^{-1/2}g_0(c-x'\beta_0,x)h(c-x'\beta_0,x) && \text{(B.4)}\end{aligned}$$

This immediately suggests that zero information can be attained by setting $h(t,x)$ to be arbitrarily close in an L_2 sense to $\frac{g_0'(t,x)x}{g_0(t,x)}$. To ensure it will satisfy conditions in the definition of Λ, we can follow the arguments in the proof of Theorem 4.1.

Note that our impossibility result breaks down when $g_0(t,x)$ takes the value 0. This can happen in the random censoring case when c_i takes arbitrarily small values so no censoring occurs. However, under our assumptions on the support of c_i, ϵ_i this cannot happen with positive probability.