Local NLLS Estimation of Semiparametric Binary Choice Models

Jason R. Blevins
The Ohio State University

Shakeeb Khan
Duke University

September 2012

Abstract

In this paper, nonlinear least squares (NLLS) estimators are proposed for semiparametric binary response models under conditional median restrictions. The estimators can be identical to NLLS procedures for parametric binary response models (e.g., probit), and consequently have the advantage of being easily implementable using standard software packages such as Stata. This is in contrast to existing estimators for the model, such as the maximum score estimator (Manski, 1975, 1985) and the smoothed maximum score (SMS) estimator (Horowitz, 1992). Two simple bias correction methods—a proposed jackknife method and an alternative nonlinear regression function—result in the same rate of convergence as SMS. The results from a Monte Carlo study show that the new estimators perform well in finite samples.

JEL Classification: C13, C14, C25.

Keywords: binary response, median restriction, nonlinear least squares, bias reduction, jackknife.

1 Introduction

The binary response model studied in this paper is of the form

$$y_i = I[x_i'\beta_0 - \epsilon_i \geq 0],$$

where $I[\cdot]$ is the indicator function, y_i is the observed response variable, taking the values 0 or 1, and x_i is an observed vector of covariates which affect the behavior of y_i. Both the disturbance term ϵ_i and the vector β_0 are unobserved, the latter often being the parameter estimated from a random sample $\{y_i, x_i'\}_{i=1}^n$.

The disturbance term ϵ_i is restricted in ways that ensure identification of β_0. Parametric restrictions specify the distribution of ϵ_i up to a finite number of parameters and assume it is distributed independently of the covariates x_i. The resulting models are often considered too restrictive, as standard estimators are usually inconsistent if the distribution of ϵ_i is misspecified or conditionally heteroskedastic.

Semiparametric, or "distribution-free," restrictions have also been imposed in the literature, resulting in a variety of estimation procedures for β_0. For a thorough survey on the various restrictions and proposed estimators, see Powell (1994). In this paper we focus exclusively on the conditional median restriction,

$$\text{med}(\epsilon_i \mid x_i) = 0,$$

which is widely regarded as the weakest restriction imposed in the literature (cf. Powell, 1994).

Several estimators of β_0 have been proposed under this restriction. The first was the maximum score estimator proposed by Manski (1975), which maximized the objective function

$$M_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left\{I[y_i = 1]I[x_i'\beta \geq 0] + I[y_i = 0]I[x_i'\beta < 0]\right\}. \tag{1.1}$$

Since y_i is a binary variable, this is numerically equivalent to minimizing the least absolute deviations (LAD) objective function:

$$M_n'(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - I[x_i'\beta \geq 0]\right|. \tag{1.2}$$

Manski (1975, 1985) established the estimator's consistency and Kim and Pollard (1990) showed that its rate of convergence is $n^{-1/3}$ and established its limiting distribution, which is non-standard and non-Gaussian, making inference based on this distribution infeasible. As an alternative, Delgado, Rodríguez-Poo, and Wolf (2001) established that inference based on subsampling is possible in such models, but Abrevaya and Huang (2005) showed that the bootstrap does not consistently estimate the asymptotic distribution.

In an effort to improve the situation, Horowitz (1992) modified the maximum score procedure by "smoothing" the objective function in (1.1). Specifically, his approach was to maximize

$$S_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left\{I[y_i = 1]K_h(x_i'\beta) + I[y_i = 0](1 - K_h(x_i'\beta))\right\}, \tag{1.3}$$

where $K_h(\cdot) \equiv K(\cdot/h)$ for some smooth kernel function $K(\cdot)$ and h denotes a smoothing parameter which converges to 0 with the sample size. Under stronger smoothness conditions on the distributions of ϵ_i and x_i, Horowitz showed that the estimator converges at the rate[1] $n^{-2/5}$ and that it is asymptotically normally distributed. Although this makes it possible to carry out standard asymptotic inference with the smoothed maximum score (SMS) estimator, Horowitz (2002) showed that the bootstrap provides asymptotic refinements and provided Monte Carlo evidence of improved finite sample performance relative to first-order asymptotic approximations.

Both Manski and Horowitz assumed that at least one component of x_i had full support on the real line to ensure that β_0 is point identified. More recently, Komarova (2008) developed set estimators based on the maximum score objective function for the case when x_i is discrete, and thus β_0 may only be partially identified. Blevins (2010) extended this idea to the case of fixed effects panel data models where x_i may be either discrete or continuous but bounded.

Although both the maximum score and smoothed maximum score estimators have desirable asymptotic properties, they are difficult to implement in practice. The maximum score estimator has a discontinuous objective function, ruling out gradient-based optimization methods. The smoothed maximum score estimator is also difficult to implement, as the objective function can have several local maxima. Horowitz (1992) suggested using the simulated annealing algorithm (Corana, Marchesi, Martini, and Ridella, 1987; Goffe, Ferrier, and Rogers, 1994) to search for a global maximum. Unfortunately, such an algorithm, which requires the selection of several "tuning" parameters by the researcher, is not available in standard econometric software packages.

The difficulty in implementing the maximum score and smoothed maximum score estimators in practice is precisely what motivates the estimators introduced in this paper.[2] Specifically, we propose procedures that are analogous to NLLS estimators of parametric models such as probit, and can thus be easily implemented using standard software packages such as Stata.

The rest of the paper is organized as follows. The following section describes the new procedures in detail and explores their asymptotic properties. Section 3 discusses bias correction procedures for improving the asymptotic properties of the estimators. Section 4 explores the finite sample properties of the estimator by ways of a small scale simulation study and Section 5 concludes by summarizing and discussing areas for future research. The proofs of the asymptotic properties of the estimators are left to the appendix.

---

[1] Horowitz (1993a) showed that this is the fastest possible rate of convergence under these conditions.

[2] As is the case with standard parametric NLLS estimators and the SMS estimator, the local NLLS estimators developed in this paper do not have globally concave objective functions, so there may be multiple local optima. The local NLLS objective function is smooth, as with SMS, so these estimators still have the advantage that standard gradient-based optimization methods can be used to find local optima. Such methods generally converge faster (to local optima) than other methods such as the Nelder-Mead simplex method or stochastic search algorithms. For all of these estimators, multiple starting values should be used to mitigate the problems of local optima.

2 Local NLLS Estimators

The estimators proposed herein combine ideas from the maximum score and smoothed maximum score objective functions in (1.2) and (1.3). First, note that the maximum score objective function in (1.2) is equivalent to the quadratic loss objective function

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - I[x_i'\beta \geq 0]\right)^2,$$

since both y_i and the indicator function are binary. Next, just as the smoothed maximum score estimator employs a kernel function to smooth the indicator in (1.2), we replace the indicator function above with a kernel function. In the case of SMS, the kernel function serves to approximate a cumulative distribution function (cdf). We take the same approach here and use the standard normal distribution[3] with cdf $\Phi(\cdot)$ and probability density function (pdf) $\phi(\cdot)$.

Formally, let h_n be a positive sequence of real numbers which decreases to zero with the sample size. The sequence h_n can be viewed as a bandwidth sequence used in nonparametric kernel estimation. Because β_0 is only identified up to scale, we use the customary scale normalization used in semiparametric models (e.g., Horowitz, 1992), where we normalize the coefficient on the last regressor and consider estimation of θ_0 only, where $\beta_0 = (\theta_0', 1)'$. Our local NLLS estimator is defined as

$$\hat{\beta} = \arg\min_{\beta\in\Theta\times 1}\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \Phi\left(\frac{x_i'\beta}{h_n}\right)\right)^2, \tag{2.1}$$

where $\hat{\beta} = (\hat{\theta}', 1)'$ and Θ is the parameter space. The primary advantage of this estimator is that it is a simple modification of the standard NLLS probit objective function. Aside from imposing the scale normalization on β and rescaling the index $x_i'\beta$, the objective function is identical to that of the NLLS probit estimator which is widely used to estimate parametric binary choice models. As such, the estimator can be readily computed using standard software packages such as Stata.[4]

As with other semiparametric estimators for the binary choice model, and even the parametric probit and logit models, the estimated coefficients must be interpreted in light of the required scale normalization. That is, only relative magnitudes of the coefficients are identified and the sign of the j-th coefficient, β_j, is the same as the sign of the partial effect of x_{ij} on the response probability. Thus, all coefficients should be interpreted relative to the coefficient on a particular chosen component of x_i. Note that for the local NLLS estimator above, these relative magnitudes are unchanged when scaling by h_n^{-1}.

Our first result regarding the asymptotic properties of the estimator is based on the following

[3] Actually, the cdfs of other random variables can be used as well, so for example NLLS logit can also be used as an estimator. We only use the normal cdf since its values can be easily computed using standard software packages.

[4] For example, in Stata, the `nl` command fits an arbitrary nonlinear function by least squares. The probit regression function can be constructed using Stata's `norm` command, which returns cumulative probabilities from the standard normal distribution.

assumptions. First, let $\tilde{x}_i$ denote the first $k-1$ components of x_i, let $z_i = x_i'\beta_0$, and let $f_{Z|\tilde{X}}(\cdot)$ denote the density function of z_i conditional on $\tilde{x}_i$.

A1 The vectors $(x_i', \epsilon_i)'$ are independent and identically distributed (iid).

A2 $\text{med}(\epsilon_i \mid x_i) = 0$ almost surely.

A3 $\theta_0 \in \Theta$, a compact subset of $\mathbb{R}^{k-1}$.

A4 The support of x_i, denoted $\mathcal{X}$, is not contained in any proper linear subspace of $\mathbb{R}^k$.

A5 $f_{Z|\tilde{X}}(\cdot)$ is positive in a neighborhood of 0.

First, the following theorem, whose proof appears in the appendix, establishes the consistency of the estimator.

Theorem 2.1. *Under Assumptions A1–A5, if* $h_n \to 0$, *then* $\hat{\theta} - \theta_0 \xrightarrow{p} 0$.

Next, we consider the rate of convergence and limiting distribution. We strengthen our assumptions to be able to draw comparisons to the smoothed maximum score estimator and impose conditions that are identical to those in Horowitz (1992).

A5' $f_{Z|\tilde{X}}(\cdot)$ is positive and continuously differentiable with bounded derivative.

A6' θ_0 is contained in the interior of Θ.

A7' $0 < P(y_i = 1 \mid x_i) < 1$ almost surely.

A8' Letting $\|\cdot\|$ denote the Euclidean norm, we have $\text{E}[\|\tilde{x}_i\|^4] < \infty$.

A9' The conditional probability of $y_i = 1$, expressed as a function of $\tilde{x}_i$ and $x_i'\beta_0$, denoted $\tilde{P}(\tilde{x}_i, x_i'\beta_0)$, is twice continuously differentiable with respect to $x_i'\beta_0$ with bounded derivatives for $x_i'\beta_0$ in a neighborhood of 0, for all $\tilde{x}_i$.

A10' The matrix $Q = \text{E}[\tilde{P}_2(\tilde{x}_i, 0)\tilde{x}_i\tilde{x}_i' f_{Z|\tilde{X}}(0|\tilde{x}_i)]$ is nonsingular, where $\tilde{P}_2(\cdot,\cdot)$ denotes the partial derivative of $\tilde{P}(\cdot,\cdot)$ with respect to its second argument.

The following theorem characterizes the rate of convergence and limiting distribution of the local NLLS estimator as a function of h_n. The proof of the theorem is reserved for the appendix.

Theorem 2.2. *Suppose that Assumptions A1–A4 and A5'–A10' hold and* $h_n \to 0$.

1. *If* $nh_n^3 \to \infty$, *then* $h_n^{-1}(\hat{\theta} - \theta_0) \xrightarrow{p} \kappa$ *where* κ *is a* k*-dimensional vector of constants.*
2. *If* $h_n = O(n^{-1/3})$, *then* $n^{1/3}(\hat{\theta} - \theta_0) \xrightarrow{d} B$ *where the random vector* B *has non-standard (i.e., non-Gaussian) distribution.*

Thus, the asymptotic properties of the local NLLS estimator are similar to that of the maximum score estimator of Manski (1975, 1985). In particular, for both estimators the rate of convergence can be as fast as $n^{-1/3}$ and the limiting distribution is non-Gaussian.[5]

Note that although the point estimates will be correct, the standard errors reported by a local nonlinear least squares routine will not be correct. Furthermore, because of the complicated nature of the limiting distribution, inference based directly on Theorem 2.2 appears to be infeasible. Alternative methods are necessary, such as subsampling, which Delgado, Rodríguez-Poo, and Wolf (2001) have proposed to use for inference with the closely-related maximum score estimator.[6]

The rate of convergence of the local NLLS estimator is slow, relative to the smoothed maximum score estimator of Horowitz (1992), due to the fact that the bias of the estimator converges at the rate h_n, in contrast to the rate h_n^2 for SMS. Thus, given the different rates of convergence, the situation is similar to the differing rates for one- and two-sided kernel estimators in nonparametric density and regression estimation.

Fortunately, the rate of convergence of the local NLLS estimator can be improved by correcting the bias. The following section considers two procedures that yield the same rate of convergence as SMS while remaining easily implementable in standard statistical software packages.

3 Bias Correction Procedures

To motivate the bias correction procedures we propose, the following theorem, whose proof is left to the appendix, establishes a linear representation for the local NLLS estimator.

Theorem 3.1. *Suppose Assumptions A1–A4 and A5'–A10' hold, $h_n \to 0$, and $nh_n^3 \to \infty$. Then*

$$\hat{\theta} - \theta_0 = Q^{-1} \frac{1}{nh_n} \sum_{i=1}^{n} (\psi_{ni} - \mathrm{E}[\psi_{ni}]) \; + Q^{-1} \frac{\mathrm{E}[\psi_{ni}]}{h_n} + o_p\left(1/\sqrt{nh_n}\right),$$

where

$$\psi_{ni} = \left(y_i - \Phi\left(\frac{x_i'\beta_0}{h_n}\right)\right) \phi\left(\frac{x_i'\beta_0}{h_n}\right) \tilde{x}_i.$$

It can be shown by a standard change of variables argument that the bias term in the linear representation, $Q^{-1}\frac{\mathrm{E}[\psi_{ni}]}{h_n}$, is only of order h_n. As alluded to in the previous section, this is why the local NLLS estimator can only achieve at most cube-root consistency. We propose simple methods for ensuring that the bias of the estimator is $O(h_n^2)$, which will enable a rate of convergence of $\hat{\theta}$ of $O(n^{-2/5})$, as with SMS, if $h_n = O(n^{-1/5})$.

[5]For the local NLLS estimator, the non-Gaussianity stems from the result that the Hessian term in its linear representation converges to a random matrix, implying the estimator has an asymptotically mixed normal distribution (cf. van der Vaart and Wellner, 1996, Section 9.6).

[6]Note, however, that the bias-corrected estimators introduced in the following section, are all asymptotically normal.

3.1 Jackknifed Local NLLS

The first method we propose for reducing the order of bias for the local NLLS estimator is analogous to the "jackknife" method used in nonparametric estimation (Schucany and Sommers, 1977; Bierens, 1987). Our method involves constructing an estimator for θ_0 by simply taking a weighted average of two local NLLS estimators that involve two distinct constants in the smoothing parameter. We note that this procedure can still be performed using standard software packages such as Stata: it merely requires computing the local NLLS estimator twice. Similar jackknife procedures have also been useful for bias reduction in other semiparametric models (see Aradillas-López, Honoré, and Powell, 2007; Cattaneo, Crump, and Jansson, 2011).

To construct our proposed jackknife estimator, let $h_{1n} = \kappa_1 n^{-1/5}$ and $h_{2n} = \kappa_2 n^{-1/5}$ denote two bandwidth sequences, where κ_1 and κ_2 are positive constants. Let w_1 and w_2 denote the weights that will be assigned to the two estimators obtained by using, respectively, bandwidths h_{1n} and h_{2n}. We impose the following conditions on w_1,w_2,κ_1, and κ_2:[7]

$$\begin{aligned} &w_1 + w_2 = 1, \\ &w_1\kappa_1 + w_2\kappa_2 = 0. \end{aligned}$$

Let $\hat{\theta}_1$ and $\hat{\theta}_2$ denote the local NLLS estimators obtained using h_{1n} and h_{2n} as smoothing parameters, respectively. We define the jackknife NLLS estimator as

$$\hat{\theta}_{\text{jk}} = w_1\hat{\theta}_1 + w_2\hat{\theta}_2.$$

The following theorem, whose proof can be found in the appendix, characterizes the asymptotic properties of the jackknife NLLS estimator.

Theorem 3.2. *Under Assumptions A1–A4 and A5'–A10',*

$$n^{2/5}(\hat{\theta}_{\text{jk}} - \theta_0) \xrightarrow{d} \text{N}(\mathcal{B}_{\text{jk}}, Q^{-1}V_{\text{jk}}Q^{-1})$$

where

$$Q = \text{E}[\tilde{P}_2(\tilde{x}_i, 0)\tilde{x}_i\tilde{x}_i' f_{Z|\tilde{X}}(0|\tilde{x}_i)],$$

$$V_{\text{jk}} = V_1(c_1 w_1^2 \kappa_1^{-1} + c_1 w_2^2 \kappa_2^{-1} + 2w_1 w_2 c_2 \kappa_1^{-1}),$$

and

$$V_1 = \text{E}[\tilde{x}_i\tilde{x}_i' f_{Z|\tilde{X}}(0|\tilde{x}_i)],$$

[7] We note that w_1, w_2, κ_1, κ_2 need not be constants—they can all be functions of x_i and the arguments used in this section still carry through. We only assume they are constants for ease of exposition.

with

$$c_1 = \int \Phi^2(u)\phi^2(u)\,du,$$
$$r_\kappa = \kappa_1/\kappa_2,$$
$$c_2 = \int \phi(u)\phi(u/r_\kappa)\left[0.5(1-\Phi(u)-\Phi(u/r_\kappa))+\Phi(u)\Phi(u/r_\kappa)\right]\,du,$$

and

$$\mathcal{B}_{\text{jk}} = (w_1\kappa_1^2 + w_2\kappa_2^2)\frac{1}{2}\,\mathrm{E}\left[\int\left\{\left(\frac{1}{2}-\Phi(u)\right)f_{Z|\tilde{X}}(0|\tilde{x}_i)+2\tilde{P}_2(\tilde{x}_i,0)f'_{Z|\tilde{X}}(0|\tilde{x}_i)\right.\right.$$
$$\left.\left.+\tilde{P}_{22}(\tilde{x}_i,0)f_{Z|\tilde{X}}(0|\tilde{x}_i)\right\}u^2\phi(u)\,du\,\tilde{x}_i\right],$$

where $\tilde{P}_{22}(\cdot,\cdot)$ denotes the second derivative of $\tilde{P}(\cdot,\cdot)$ with respect to its second argument.

Thus, the jackknife NLLS estimator can achieve the same rate of convergence as SMS and is asymptotically normally distributed. In standard nonparametric estimation, the jackknife is used to achieve bias reduction and attain the optimal rate of convergence for estimating a density or regression function (Bierens, 1987). Here, the motivation of combining NLLS estimators is to achieve bias reduction and attain the same rate of convergence as SMS, which is the optimal rate under the maintained assumptions (Horowitz, 1993a).

Furthermore, the form of the limiting distribution, which depends on the weights and constants, suggests choosing those parameters to minimize the asymptotic mean squared error. The optimal choices are discussed in the appendix, along with a procedure to construct a feasible optimal jackknife NLLS estimator.

3.2 A Different Nonlinear Regression Function

As discussed in the proofs of Theorems 2.2 and 3.1, the bias problem of the local NLLS estimator is associated with the fact that the normal cdf is used. An alternative bias correction procedure would be to use a function $F(\cdot)$ in the local NLLS objective function instead of the normal cdf $\Phi(\cdot)$. The bias term of the local NLLS estimator was of order h_n because $\int \Phi(u)\phi(u)u\,du \neq 0$ and so the function $F(\cdot)$ must be such that the analogous integral $\int F(u)f(u)u\,du$, with $f(\cdot) = F'(\cdot)$, is zero.

Importantly, it is no more difficult to implement the estimator using a general function $F(\cdot)$ than with the normal cdf because NLLS procedures in common statistical packages such as Stata allow the user to provide a generic regression function.

The restrictions preclude $F(\cdot)$ from being a cumulative distribution function, making this approach analogous to the use of higher order kernel functions[8] which are not density functions in nonparametric density/regression estimation. Let $\hat{\theta}_F$ denote the local NLLS estimator with $F(\cdot)$

[8]See, for example, Newey, Hsieh, and Robins (2004) on "twicing kernels", which are higher order.

replacing $\Phi(\cdot)$ in (2.1). The theorem below, whose proof can be found in the appendix, establishes that the following conditions on $F(\cdot)$ are sufficient for $\hat{\theta}_F$ to converge at the rate $n^{-2/5}$ with an asymptotic Gaussian distribution.

F1 $\int(\frac{1}{2} - F(u))f(u)\,du = 0$

F2 $\int f(u)u\,du = 0$

F3 $\int F(u)f(u)u\,du = 0$

F4 $\int \left([\frac{1}{2} - F(u)]f'(u) - f^2(u)\right)\,du = 0$

F5 $0 < \left|\int f'(u)u\,du\right| < \infty$

F6 $\left|\int \left([\frac{1}{2} - F(u)]f'(u) - f^2(u)\right)u\,du\right| < \infty$

Theorem 3.3. *Suppose that Assumptions A1–A4 and A5'–A10' hold, that $F(\cdot)$ satisfies conditions F1–F6, and that $h_n = O(n^{-1/5})$. Then*

$$n^{2/5}(\hat{\theta}_F - \theta_0) \xrightarrow{d} \mathrm{N}(\mathcal{B}_F, Q_F^{-1}V_F Q_F^{-1})$$

where

$$\mathcal{B}_F = \frac{1}{2}\int_{\tilde{\mathcal{X}}}\int\left\{\left(\frac{1}{2} - F(u)\right)f_{Z|\tilde{X}}(0 \mid \tilde{x}_i) + 2\tilde{P}_2(\tilde{x}_i, 0)f'_{Z|\tilde{X}}(0 \mid \tilde{x}_i) \right.$$
$$\left. + \tilde{P}_{22}(\tilde{x}_i, 0)f_{Z|\tilde{X}}(0 \mid \tilde{x}_i)\right\}u^2 f(u)\,du\,\tilde{x}_i\,dP_{\tilde{X}}(\tilde{x}_i),$$

$$Q_F = \mathrm{E}\left[\left(c_{F_2}\tilde{P}_2(\tilde{x}_i, 0)f_{Z|\tilde{X}}(0 \mid \tilde{x}_i) + c_{F_3}f'_{Z|\tilde{X}}(0 \mid \tilde{x}_i)\right)\tilde{x}_i\tilde{x}_i'\right],$$

and $V_F = c_{F_1} \cdot \mathrm{E}[\tilde{x}_i\tilde{x}_i' f_{Z|\tilde{X}}(0 \mid \tilde{x}_i)]$, with $c_{F_1} = \int F^2(u)f^2(u)\,du$, $c_{F_2} = \int f'(u)u\,du$, and $c_{F_3} = \int \left([\frac{1}{2} - F(u)]f'(u) - f^2(u)\right)u\,du$.

Remark 3.1. *When the function $F(\cdot)$ satisfies the following two symmetry properties, then the integral in condition F6 and c_{F_3} is zero:*

F7 $F(-u) = 1 - F(u)$,

F8 $f(u) = f(-u)$.

In this case, Q_F simplifies to $Q_F = \mathrm{E}\left[c_{F_2}\tilde{P}_2(\tilde{x}_i, 0)f_{Z|\tilde{X}}(0 \mid \tilde{x}_i)\tilde{x}_i\tilde{x}_i'\right]$. The particular family of regression functions we propose below satisfies these properties.

Functions satisfying the required conditions can be constructed using the normal cdf. For example, functions of the form

$$F(u) = (1/2 - \alpha_F - \beta_F) + 2\alpha_F\Phi(u) + 2\beta_F\Phi(\sqrt{2}u), \tag{3.1}$$

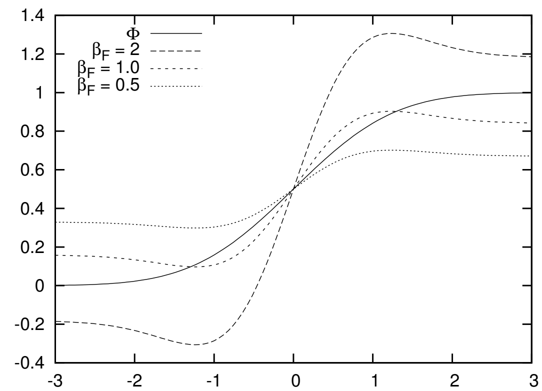

Figure 1: Nonlinear regression functions F compared to the normal cdf Φ

with $\alpha_F = -\frac{1}{2}\left(1-\sqrt{2}+\sqrt{3}\right)\beta_F$ and $\beta_F \neq 0$ satisfy the main conditions F1–F6 as well as the symmetry properties F7 and F8.

This family of functions is plotted in Figure 1 for several values of β_F alongside the standard normal cdf $\Phi(\cdot)$. Note that $\Phi(\cdot)$ has the form above, but with $\alpha_F = 1/2$ and $\beta_F = 0$, so it is not a member of the same family of functions because the coefficient β_F is zero, violating the conditions required for unbiasedness, which cdfs do not satisfy. This family of alternative regression functions is examined in the following section, which discusses a series of Monte Carlo experiments designed to shed light on the small-sample properties of the estimator.

With the more standard rate and limiting distribution of this NLLS estimator, a natural extension to consider is then a weighted NLLS procedure. It is well known that in parametric settings, the efficiency of the NLLS probit estimator can be improved by weighting the observations, and one can make the NLLS estimator as efficient as the MLE using optimal weights. For the problem at hand, the weight function can be chosen to minimize the asymptotic mean squared error of the estimator. The appendix provides the form of this optimal weighting matrix and further discusses implementation of a feasible weighted NLLS approach.

4 Monte Carlo Results

In this section, we investigate the small-sample performance of the estimators introduced in this paper by ways of a small-scale Monte Carlo study. The model used in this simulation study is

$$y_i = I[\alpha_0 + x_{1i}\beta_0 + x_{2i} + \epsilon_i > 0]$$

where x_{1i} has a chi-square distribution with 1 degree of freedom (minus its mean of 1), x_{2i} has a standard normal distribution, α_0 was set at -0.5 and β_0 at -1. Three different distributions for ϵ_i were simulated: standard normal, chi-square with 1 degree of freedom minus its median, and Cauchy. Both homoskedastic and heteroskedastic designs were simulated. The heteroskedastic designs involved a multiplicative scale factor of the form $\exp(x_{1i} \cdot |x_{2i}|)$.

Table I: Homoskedastic Normal

	α				β			
Estimator	Mean Bias	Median Bias	RMSE	MAD	Mean Bias	Median Bias	RMSE	MAD
100 obs.								
NLLS	-0.066	-0.024	0.295	0.157	-0.102	-0.040	0.404	0.216
JKNLLS-1	-0.071	0.007	0.572	0.253	-0.104	0.056	0.749	0.336
JKNLLS-2	-0.014	0.061	0.463	0.204	0.018	0.146	0.622	0.271
NLLSF	-0.066	-0.001	0.389	0.172	-0.101	0.008	0.537	0.238
SMS-1	-0.412	-0.210	1.098	0.278	-0.757	-0.415	1.956	0.394
SMS-2	-0.157	-0.059	0.584	0.242	-0.264	-0.089	0.785	0.329
SMS-3	-0.145	-0.043	0.789	0.241	-0.267	-0.068	1.576	0.324
200 obs.								
NLLS	-0.029	-0.009	0.177	0.107	-0.048	-0.022	0.243	0.142
JKNLLS-1	-0.043	0.012	0.406	0.202	-0.057	0.050	0.554	0.262
JKNLLS-2	0.012	0.055	0.314	0.161	0.052	0.134	0.428	0.215
NLLSF	-0.022	0.009	0.235	0.127	-0.032	0.021	0.319	0.172
SMS-1	-0.227	-0.149	0.505	0.184	-0.438	-0.299	1.807	0.263
SMS-2	-0.092	-0.037	0.355	0.184	-0.152	-0.056	0.486	0.246
SMS-3	-0.077	-0.029	0.346	0.183	-0.134	-0.050	0.471	0.249
400 obs.								
NLLS	-0.014	-0.009	0.115	0.075	-0.024	-0.014	0.161	0.102
JKNLLS-1	-0.025	0.017	0.304	0.160	-0.022	0.052	0.404	0.210
JKNLLS-2	0.025	0.050	0.229	0.129	0.067	0.118	0.317	0.172
NLLSF	-0.001	0.013	0.158	0.094	0.005	0.037	0.222	0.127
SMS-1	-0.139	-0.111	0.258	0.127	-0.266	-0.224	0.407	0.183
SMS-2	-0.058	-0.028	0.243	0.140	-0.095	-0.045	0.337	0.196
SMS-3	-0.049	-0.025	0.241	0.147	-0.084	-0.033	0.332	0.196

Tables I–VI report results for comparing the performance of the estimators discussed in this paper: the local NLLS (NLLS), jackknifed NLLS (JKNLLS), local NLLS with an alternative regression function (NLLSF), and smoothed maximum score (SMS) estimators. For NLLSF, we use the regression function in (3.1) with $\beta_F = 1$. Reported are the mean bias, median bias, root mean square error (RMSE), and median absolute deviation (MAD) for sample sizes $n = 100$, 200, and 400, with 4001 replications each.

For each estimator, we selected the bandwidth for each sample as follows. For NLLS, we chose h_n using cross-validation to minimize the leave-one-out sum of squared residuals. For JKNLLS, the weights and bandwidth constants were chosen for each sample according to the procedures outlined in the appendix. JKNLLS-1 indicates the first method, which chooses w_1, w_2, κ_1, and κ_2 to minimize the constant portion of the asymptotic mean square error. For JKNLLS-2, we chose these constants to minimize an estimate of the asymptotic mean square error using a finite sample estimate of the asymptotic variance matrix. For NLLSF, we used the optimal bandwidth selection procedure suggested by Horowitz (1992) for SMS, since both estimators have a similar asymptotically linear structure which yields asymptotic normality with bias on the order of h_n^2 in

Table II: Heteroskedastic Normal

	α				β			
Estimator	Mean Bias	Median Bias	RMSE	MAD	Mean Bias	Median Bias	RMSE	MAD
100 obs.								
NLLS	0.088	0.137	0.326	0.181	0.200	0.258	0.476	0.273
JKNLLS-1	0.015	0.105	0.517	0.272	0.081	0.205	0.690	0.391
JKNLLS-2	0.059	0.150	0.474	0.229	0.173	0.301	0.654	0.332
NLLSF	0.012	0.077	0.372	0.215	0.084	0.164	0.507	0.317
SMS-1	-0.301	-0.211	0.638	0.309	-0.434	-0.319	0.850	0.409
SMS-2	-0.066	0.012	0.467	0.273	-0.049	0.037	0.598	0.374
SMS-3	-0.036	0.037	0.466	0.273	-0.010	0.082	0.607	0.380
200 obs.								
NLLS	0.111	0.140	0.242	0.135	0.229	0.260	0.378	0.203
JKNLLS-1	0.024	0.092	0.431	0.238	0.073	0.166	0.581	0.336
JKNLLS-2	0.072	0.135	0.365	0.205	0.172	0.253	0.516	0.289
NLLSF	0.037	0.077	0.272	0.174	0.116	0.156	0.386	0.242
SMS-1	-0.214	-0.173	0.416	0.225	-0.297	-0.245	0.550	0.293
SMS-2	-0.040	0.009	0.351	0.224	-0.019	0.028	0.455	0.296
SMS-3	-0.020	0.026	0.352	0.224	0.005	0.052	0.460	0.302
400 obs.								
NLLS	0.124	0.140	0.197	0.099	0.247	0.262	0.329	0.144
JKNLLS-1	0.023	0.079	0.353	0.201	0.060	0.125	0.476	0.277
JKNLLS-2	0.066	0.121	0.304	0.175	0.149	0.204	0.433	0.249
NLLSF	0.044	0.066	0.212	0.136	0.126	0.145	0.308	0.190
SMS-1	-0.154	-0.136	0.289	0.160	-0.201	-0.180	0.378	0.213
SMS-2	-0.032	-0.008	0.275	0.182	-0.014	0.012	0.356	0.241
SMS-3	-0.011	0.016	0.278	0.187	0.011	0.029	0.360	0.243

both cases. For SMS, a normal kernel function was used and we compared three bandwidth selection procedures. For SMS-1, we again used the bandwidth selection procedure suggested by Horowitz (1992).[9] For SMS-2, we selected the bandwidth using Silverman's rule of thumb, $h_n = 1.06 \cdot \hat{\sigma} \cdot n^{-1/5}$, where $\hat{\sigma}$ is the sample standard deviation of y_i. Finally, for SMS-3 we chose the bandwidth using leave-one-out least absolute deviations cross-validation.

All the estimators were computed[10] using the simplex algorithm of Nelder and Mead (1965) with multiple starting values, including the OLS, LAD, and probit estimates, their average, the zero vector, and seven random starting values.

As the results indicate, the finite sample performance is mostly, but not entirely in accordance with the asymptotic theory. The biggest surprise is that in terms of RMSE, for some designs, the

[9]For both NLLSF and SMS-1 we iterated this procedure, as suggested by Horowitz (1992). For each sample, we first obtained an estimate of $\hat{\theta}^{(0)}$ using the bandwidth $h_n^{(0)} = n^{-1/5}$. We then estimated the optimal bandwidth $h_n^{(1)}$, which we used to obtain a second estimate $\hat{\theta}^{(1)}$. Then, using the second estimate, we obtained a second estimate of the optimal bandwidth, $h_n^{(2)}$. Finally, we reported the estimate $\hat{\theta}^{(2)}$ obtained using the bandwidth $h_n^{(2)}$.

[10]The simulation study was performed in Fortran, despite the fact that the new estimators were motivated by the fact that they could be computed with Stata. Fortran was used so all estimators could be computed using a common random number generator, as SMS is difficult to compute with Stata.

Table III: Homoskedastic Chi-Square

Estimator	α Mean Bias	α Median Bias	α RMSE	α MAD	β Mean Bias	β Median Bias	β RMSE	β MAD
100 obs.								
NLLS	0.188	0.197	0.304	0.154	-0.019	0.015	0.327	0.198
JKNLLS-1	0.025	0.038	0.390	0.229	0.008	0.089	0.539	0.287
JKNLLS-2	0.065	0.082	0.328	0.183	0.088	0.151	0.464	0.247
NLLSF	0.112	0.123	0.287	0.167	-0.046	-0.003	0.407	0.221
SMS-1	0.006	0.013	1.179	0.219	-0.750	-0.375	4.223	0.294
SMS-2	0.017	0.022	0.351	0.202	-0.161	-0.059	0.753	0.272
SMS-3	0.030	0.038	0.541	0.206	-0.160	-0.046	1.891	0.283
200 obs.								
NLLS	0.198	0.202	0.255	0.106	0.009	0.021	0.212	0.138
JKNLLS-1	0.010	0.018	0.304	0.177	0.021	0.082	0.417	0.238
JKNLLS-2	0.040	0.049	0.240	0.148	0.089	0.123	0.335	0.197
NLLSF	0.103	0.109	0.210	0.118	-0.007	0.002	0.247	0.157
SMS-1	0.012	0.026	0.235	0.140	-0.325	-0.284	0.478	0.194
SMS-2	0.016	0.010	0.230	0.147	-0.073	-0.040	0.334	0.194
SMS-3	0.026	0.019	0.237	0.149	-0.062	-0.036	0.340	0.207
400 obs.								
NLLS	0.207	0.209	0.234	0.074	0.024	0.030	0.146	0.097
JKNLLS-1	0.002	0.003	0.232	0.142	0.033	0.069	0.320	0.195
JKNLLS-2	0.021	0.023	0.181	0.115	0.085	0.099	0.258	0.151
NLLSF	0.089	0.089	0.162	0.090	0.018	0.022	0.172	0.111
SMS-1	0.028	0.031	0.147	0.092	-0.223	-0.209	0.302	0.128
SMS-2	0.008	-0.001	0.169	0.109	-0.044	-0.033	0.241	0.149
SMS-3	0.014	0.005	0.181	0.116	-0.033	-0.024	0.249	0.160

standard NLLS performs as well as, if not better than the other estimators despite its slower rate of convergence. The jackknife bias correction procedure (JKNLLS) generally results in a lower bias than NLLS, but it appears this sometimes comes at the expense of a larger variance, leading to a higher finite-sample RMSE in some designs. Furthermore, the alternative regression function (NLLSF) achieves a relatively low RMSE uniformly across the experiments, having the lowest or second-lowest RMSE among all estimators considered. In all of the normal and Cauchy designs, and in all but the largest sample size chi-square specifications, it is second only to the baseline NLLS estimator. For the chi-square designs with our largest sample size, SMS-1 has the lowest RMSE for α.

The NLLS estimators appear to perform better in the homoskedastic designs. The situation here is similar to that of parametric NLLS estimators, for which weighting can improve efficiency under heteroskedasticity. For example, in the probit model, an optimally weighted NLLS estimator is asymptotically equivalent to MLE. We discuss a weighted NLLS extension in the appendix.

Interestingly, both NLLS and NLLSF outperform the SMS in many designs, especially for the smaller sample sizes, though it may well be the case that relative performance depends on the

Table IV: Heteroskedastic Chi-Square

	α				β			
Estimator	Mean Bias	Median Bias	RMSE	MAD	Mean Bias	Median Bias	RMSE	MAD
100 obs.								
NLLS	0.247	0.275	0.382	0.197	0.205	0.239	0.444	0.274
JKNLLS-1	0.085	0.118	0.438	0.268	0.120	0.198	0.585	0.354
JKNLLS-2	0.114	0.163	0.410	0.238	0.160	0.244	0.557	0.333
NLLSF	0.154	0.174	0.355	0.223	0.082	0.109	0.441	0.308
SMS-1	-0.026	-0.028	0.797	0.265	-0.424	-0.341	2.019	0.337
SMS-2	0.055	0.059	0.370	0.255	-0.003	0.025	0.505	0.337
SMS-3	0.089	0.094	0.496	0.257	0.025	0.071	1.005	0.339
200 obs.								
NLLS	0.242	0.261	0.329	0.155	0.207	0.217	0.361	0.208
JKNLLS-1	0.054	0.082	0.368	0.233	0.092	0.155	0.492	0.304
JKNLLS-2	0.084	0.122	0.343	0.208	0.135	0.195	0.469	0.284
NLLSF	0.134	0.143	0.278	0.175	0.072	0.081	0.335	0.232
SMS-1	-0.020	-0.019	0.277	0.185	-0.287	-0.281	0.456	0.232
SMS-2	0.038	0.030	0.280	0.200	0.004	0.009	0.369	0.256
SMS-3	0.059	0.052	0.289	0.203	0.034	0.040	0.383	0.270
400 obs.								
NLLS	0.244	0.253	0.296	0.116	0.211	0.215	0.302	0.150
JKNLLS-1	0.030	0.051	0.311	0.198	0.073	0.119	0.417	0.250
JKNLLS-2	0.063	0.088	0.279	0.177	0.117	0.156	0.385	0.229
NLLSF	0.122	0.124	0.223	0.130	0.075	0.069	0.255	0.167
SMS-1	-0.007	-0.007	0.191	0.128	-0.221	-0.220	0.334	0.167
SMS-2	0.016	0.006	0.220	0.156	-0.005	-0.013	0.292	0.199
SMS-3	0.034	0.024	0.232	0.160	0.019	0.013	0.305	0.212

bandwidth choice. Again, it is quite surprising that the standard NLLS sometimes outperforms SMS, as the latter converges at a faster rate.

Overall, these results indicate that the NLLS estimators introduced in this paper are a viable alternative to SMS in empirical applications, since it appears that their ease in implementation does not come at the expense of finite sample performance.

5 Conclusions

In this paper, new estimation procedures for binary response models under conditional median restrictions were proposed. The estimators were based on applying NLLS procedures for parametric models to this semiparametric model. Their primary advantage is their relative computational simplicity, as they can be applied using standard software packages such as Stata. A simulation study indicates these estimators perform adequately well in finite samples.

The work here suggests areas for future research. First we note that variations of the (smoothed) maximum score estimator have been developed for the analysis of binary choice in panel data (Man-

Table V: Homoskedastic Cauchy

Estimator	α Mean Bias	Median Bias	RMSE	MAD	β Mean Bias	Median Bias	RMSE	MAD
100 obs.								
NLLS	-0.070	-0.008	0.441	0.215	-0.122	0.001	0.666	0.285
JKNLLS-1	-0.069	0.038	0.739	0.305	-0.113	0.102	1.004	0.394
JKNLLS-2	-0.024	0.084	0.660	0.257	-0.003	0.181	0.891	0.340
NLLSF	-0.064	0.029	0.560	0.218	-0.088	0.068	0.746	0.300
SMS-1	-0.676	-0.259	2.852	0.405	-1.757	-0.520	6.303	0.577
SMS-2	-0.184	-0.040	0.815	0.296	-0.349	-0.062	1.494	0.408
SMS-3	-0.231	-0.018	2.816	0.298	-0.587	-0.037	6.039	0.394
200 obs.								
NLLS	-0.015	0.012	0.242	0.148	-0.037	0.014	0.335	0.201
JKNLLS-1	-0.039	0.030	0.507	0.242	-0.071	0.063	0.668	0.323
JKNLLS-2	0.006	0.065	0.432	0.205	0.028	0.141	0.569	0.270
NLLSF	0.000	0.042	0.313	0.161	0.004	0.077	0.416	0.221
SMS-1	-0.384	-0.186	1.200	0.258	-0.705	-0.395	1.897	0.368
SMS-2	-0.116	-0.035	0.490	0.228	-0.210	-0.070	0.676	0.319
SMS-3	-0.105	-0.020	0.543	0.229	-0.177	-0.054	1.427	0.316
400 obs.								
NLLS	0.006	0.017	0.159	0.101	0.000	0.022	0.217	0.137
JKNLLS-1	-0.020	0.024	0.367	0.194	-0.037	0.064	0.501	0.251
JKNLLS-2	0.020	0.059	0.298	0.159	0.051	0.123	0.400	0.211
NLLSF	0.013	0.042	0.217	0.117	0.032	0.081	0.300	0.163
SMS-1	-0.199	-0.132	0.418	0.173	-0.374	-0.271	0.624	0.242
SMS-2	-0.067	-0.026	0.318	0.169	-0.121	-0.047	0.435	0.230
SMS-3	-0.057	-0.018	0.317	0.172	-0.109	-0.043	0.432	0.240

ski, 1987; Charlier, Melenberg, and van Soest, 1995) and choice-based sampling model (Manski, 1986; Horowitz, 1993b, 2009), so the local NLLS approach proposed in this paper can be extended to those settings as well. Thus, future work can derive the asymptotic properties of these estimators.

Second, the relative efficiency of the procedures introduced here needs to be explored, specifically in comparison to the SMS and its more efficient variant in Kotlyarova and Zinde-Walsh (2004). Related to this, efficiency gains of the NLLS estimator, either by optimally selecting the weights in the proposed jackknife or by a weighted local nonlinear least squares (WNLLS) estimator, needs to be studied.

Furthermore, it would be useful to explore whether rates of convergence arbitrarily close to $n^{-1/2}$ can be attained by the proposed estimators under stronger smoothness conditions,[11] as is the case with the smoothed maximum score estimator.

[11] Specifically, the rate of convergence of the smoothed maximum score estimator is $n^{-h/(2h+1)}$ if a smoothing kernel function of order h is used, if $f^{(i)}_{Z|\tilde{X}}(\cdot)$ is continuous and bounded for $i = 1, \ldots, h-1$, and $\tilde{P}^{(i)}_2(\tilde{x}_i, x_i'\beta_0)$ is continuous and bounded in a neighborhood of 0 for almost every $\tilde{x}_i$ for $i = 1, \ldots, h$. As $h \to \infty$, the rate of convergence can be made to approach $n^{-1/2}$.

Table VI: Heteroskedastic Cauchy

	α				β			
Estimator	Mean Bias	Median Bias	RMSE	MAD	Mean Bias	Median Bias	RMSE	MAD
100 obs.								
NLLS	0.172	0.225	0.398	0.197	0.327	0.411	0.586	0.289
JKNLLS-1	0.050	0.151	0.642	0.300	0.131	0.294	0.843	0.425
JKNLLS-2	0.060	0.176	0.596	0.258	0.171	0.368	0.807	0.377
NLLSF	0.042	0.135	0.474	0.234	0.135	0.256	0.655	0.354
SMS-1	-0.419	-0.213	1.227	0.399	-0.735	-0.367	6.133	0.537
SMS-2	-0.043	0.060	0.562	0.294	-0.028	0.106	0.744	0.427
SMS-3	-0.016	0.095	0.641	0.297	0.003	0.163	0.891	0.429
200 obs.								
NLLS	0.192	0.228	0.317	0.145	0.354	0.404	0.499	0.221
JKNLLS-1	0.043	0.130	0.508	0.256	0.089	0.215	0.670	0.365
JKNLLS-2	0.075	0.168	0.451	0.226	0.171	0.307	0.628	0.332
NLLSF	0.076	0.130	0.331	0.189	0.175	0.243	0.472	0.276
SMS-1	-0.275	-0.192	0.595	0.290	-0.399	-0.302	0.780	0.386
SMS-2	-0.037	0.034	0.438	0.251	-0.019	0.064	0.566	0.351
SMS-3	-0.012	0.050	0.439	0.253	0.011	0.087	0.567	0.355
400 obs.								
NLLS	0.218	0.240	0.281	0.108	0.394	0.423	0.468	0.160
JKNLLS-1	0.027	0.098	0.419	0.232	0.063	0.163	0.566	0.325
JKNLLS-2	0.075	0.149	0.369	0.195	0.159	0.257	0.530	0.292
NLLSF	0.076	0.114	0.262	0.154	0.178	0.217	0.386	0.222
SMS-1	-0.196	-0.158	0.388	0.215	-0.273	-0.216	0.514	0.273
SMS-2	-0.022	0.019	0.329	0.206	-0.001	0.046	0.431	0.279
SMS-3	-0.005	0.033	0.326	0.206	0.022	0.067	0.429	0.278

References

Abrevaya, J. and J. Huang (2005). On the bootstrap of the maximum score estimator. *Econometrica 73*, 1175–1204.

Amemiya, T. (1985). *Advanced Econometrics.* Cambridge, Massachusetts: Harvard University Press.

Aradillas-López, A., B. E. Honoré, and J. L. Powell (2007). Pairwise difference estimation with nonparametric control variables. *International Economic Review 48*, 1119–1158.

Bierens, H. J. (1987). Kernel estimators of regression functions. In T. F. Bewley (Ed.), *Advances in Econometrics: Fifth World Congress*, Volume I, New York, pp. 99–144. Cambridge University Press.

Blevins, J. R. (2010). Partial identification and inference in binary choice and duration panel data models. Working Paper, The Ohio State University.

Cattaneo, M. D., R. K. Crump, and M. Jansson (2011). Generalized jackknife estimators of weighted average derivatives. CREATES Research Paper 2011-12, Aarhus University.

Charlier, E., B. Melenberg, and A. H. O. van Soest (1995). A smoothed maximum score estimator for the binary choice panel data model with an application to labour force participation. *Statistica Neerlandica 49*, 324–342.

Corana, A., M. Marchesi, C. Martini, and S. Ridella (1987). Minimizing multimodal functions of continuous variables with the simulated annealing algorithm. *ACM Transactions on Mathematical Software 31*, 262–280.

de Jong, R. M. and T. M. Woutersen (2011). Dynamic time series binary choice. *Econometric Theory 27*, 673–702.

Delgado, M. A., J. M. Rodríguez-Poo, and M. Wolf (2001). Subsampling inference in cube root asymptotics with an application to Manski's maximum score estimator. *Economics Letters 73*, 241–250.

Goffe, W. L., G. D. Ferrier, and J. Rogers (1994). Global optimization of statistical functions with simulated annealing. *Journal of Econometrics 60*, 65–99.

Horowitz, J. L. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica 60*, 505–531.

Horowitz, J. L. (1993a). Optimal rates of convergence of parameter estimators in the binary response model with weak distributional assumptions. *Econometric Theory 9*, 1–18.

Horowitz, J. L. (1993b). Semiparametric and nonparametric estimation of quantal response models. In G. S. Maddala, C. R. Rao, and H. D. Vinod (Eds.), *Handbook of Statistics*, Volume 11, pp. 45–72. Amsterdam: North Holland.

Horowitz, J. L. (2002). Bootstrap critical values for tests based on the smoothed maximum score estimator. *Journal of Econometrics 111*, 141–167.

Horowitz, J. L. (2009). *Semiparametric and Nonparametric Methods in Econometrics*. New York: Springer.

Kim, J. and D. Pollard (1990). Cube root asymptotics. *Annals of Statistics 18*, 191–219.

Komarova, T. (2008). Binary choice models with discrete regressors: Identification and misspecification. Working Paper, London School of Economics.

Kotlyarova, Y. and V. Zinde-Walsh (2004). Improving the efficiency and robustness of the smoothed maximum score estimator. Working Paper, McGill University.

Manski, C. F. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics 3*, 205–228.

Manski, C. F. (1985). Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of Econometrics 27*, 313–333.

Manski, C. F. (1986). Semiparametric analysis of binary response from response-based samples. *Journal of Econometrics 31*, 31–40.

Manski, C. F. (1987). Semiparametric analysis of random effects linear models from binary panel data. *Econometrica 55*, 357–362.

Nelder, J. A. and R. Mead (1965). A simplex method for function minimization. *Computer Journal 7*, 308–313.

Newey, W. K., F. Hsieh, and J. M. Robins (2004). Twicing kernels and a small bias property of semiparametric estimators. *Econometrica 72*, 947–962.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. In R. F. Engle and D. McFadden (Eds.), *Handbook of Econometrics*, Volume 4, Amsterdam. North Holland.

Powell, J. L. (1994). Estimation of semiparametric models. In R. F. Engle and D. L. McFadden (Eds.), *Handbook of Econometrics*, Volume 4, Amsterdam, pp. 2443–2521. North Holland.

Schucany, W. R. and J. P. Sommers (1977). Improvement of kernel type density estimators. *Journal of the American Statistical Association 72*, 420–423.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes*. New York: Springer-Verlag.

A Appendix

A.1 Proof of Theorem 2.1

We first establish consistency using the standard consistency theorem of Newey and McFadden (1994, Theorem 2.1). The proof is similar to those found in Manski (1985) and Horowitz (1992).

Since the observations are iid (Assumption A1), Θ is compact (Assumption A3), and the objective function is a sample average of bounded functions that are continuous in the parameters, uniform convergence follows from the uniform law of large numbers of Amemiya (1985, Theorem 4.2.1). Continuity of the limiting objective function follows from Assumption A5. We note by the assumption that $h_n \to 0$, that the component of the limiting objective function that depends on β is $\mathrm{E}[(1-2\tilde{P}_i)(I[x_i'\beta > 0] - I[x_i'\beta_0 > 0])]$, where $\tilde{P}_i$ denotes $\tilde{P}(\tilde{x}_i, x_i'\beta_0) \equiv P(y_i = 1 \mid x_i)$. The above expectation is clearly 0 for $\beta = \beta_0$. By the strict monotonicity of $\Phi(\cdot)$, which is greater than (less than) 1/2 if its argument is greater than (less than) zero, and Assumptions A2, A4, and A5, it follows that this component of the objective function is strictly positive if $\beta \neq \beta_0$. Therefore, it is uniquely minimized at β_0. This establishes consistency. ■

A.2 Proof of Theorems 2.2 and 3.1

This section derives the asymptotic theory for the local NLLS estimator. The strategy adopted is to expand the first order condition, as is typically done in standard parametric distributional theory, and separately derive the asymptotic properties of the Hessian and score terms. This approach permits the proof of both theorems in a common setting.

Before proceeding with the proofs, we briefly discuss some additional conditions imposed as well as notation used throughout. Throughout this section, $\|\cdot\|$ will denote the Euclidean norm. Ranges of integration are denoted by subscripts, or otherwise taken to be the real line. Also, here we assume $\tilde{x}_i$, whose distribution function will be denoted by $P_{\tilde{X}}(\cdot)$, has bounded support, denoted by $\tilde{\mathcal{X}}$. This can be relaxed at the expense of longer proofs, either by decomposing the support of $\tilde{x}_i$ and using Assumption A8', or along the lines of the proofs in de Jong and Woutersen (2011).

Furthermore, the following notation will also be adopted: let

$$\Phi_{ni}, \phi_{ni}, \phi'_{ni}, \hat{\Phi}_{ni}, \hat{\phi}_{ni}, \hat{\phi}'_{ni}, \Phi^*_{ni}, \phi^*_{ni}, \phi'^*_{ni}$$

denote

$$\begin{aligned}&\Phi(x_i'\beta_0/h_n), \phi(x_i'\beta_0/h_n), \phi'(x_i'\beta_0/h_n),\\&\Phi(x_i'\hat{\beta}/h_n), \phi(x_i'\hat{\beta}/h_n), \phi'(x_i'\hat{\beta}/h_n),\\&\Phi(x_i'\beta^*/h_n), \phi(x_i'\beta^*/h_n), \phi'(x_i'\beta^*/h_n),\end{aligned}$$

respectively, where β^* denotes a point on the line segment between the zero vector and $\hat{\beta}$.

Throughout the proofs, we will use the following properties of the standard normal distribution (all integrals below are from $-\infty$ to $+\infty$ and $\phi'(\cdot)$ denotes the derivative of the standard normal density function):[12]

- $\int \phi'(u)\, du = 0$
- $\int u\phi'(u)\, du = -1$
- $\int (\Phi(u)\phi'(u) + \phi^2(u))\, du = 0$
- $\int \Phi(u)\phi(u)\, du = \frac{1}{2}$
- $\int \left[(\frac{1}{2} - \Phi(u))\phi'(u) - \phi^2(u)\right]\, du = 0$
- $\int \left[(\frac{1}{2} - \Phi(u))\phi'(u) - \phi^2(u)\right] u\, du = 0$
- $\int \phi(u)^2 \left[\frac{1}{2} - \Phi(u)\right]\, du = 0$

Proceeding with the proof, we will let $\varepsilon > 0$ denote an arbitrarily small constant such that for $|x_i'\beta_0| < \varepsilon$, the smoothness conditions in Assumptions A5' and A9' hold.

[12]Note the list of properties is not minimal in the sense that some on the list follow from others. They are listed in this fashion with the hope of making arguments in the proofs easier to follow.

The first order condition can now be expressed as

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\Phi}_{ni})\hat{\phi}_{ni}\tilde{x}_i = 0.$$

The usual mean value expansion yields

$$\hat{\theta} - \theta_0 = \left(\frac{1}{nh_n^2}\sum_{i=1}^{n}\left((y_i - \Phi_{ni}^*)\phi_{ni}^{\prime *} - \phi_{ni}^{2*}\right)\tilde{x}_i\tilde{x}_i'\right)^{-1}\frac{1}{nh_n}\sum_{i=1}^{n}(y_i - \Phi_{ni})\phi_{ni}\tilde{x}_i. \tag{A.1}$$

Before proceeding with the remainder of the proofs, we first establish the following preliminary results. The first result will be used to establish limits of several integrals we encounter in the main proof. The second result is similar to (A16) in Horowitz (1992).

Lemma A.1. *Let the functions $g_u : \mathbb{R} \to \mathbb{R}$, $g_x : \tilde{\mathcal{X}} \to \mathbb{R}$, and $g_{zx} : \mathbb{R} \times \tilde{\mathcal{X}} \to \mathbb{R}$ and the vector $\delta \in \mathbb{R}^{k-1}$ be given. Suppose that Assumption A5' holds, that $\tilde{x}_i$ has bounded support, and that there is a constant $M < \infty$ such that $\sup|g_u| < M$, $\sup|g_x| < M$, and $\sup|g_{zx}| < M$. Let $\phi_{ni\delta}$ denote $\phi(\frac{z_i}{h_n} + \tilde{x}_i'\delta)$. Then,*

$$h_n^{-1}\int_{\tilde{\mathcal{X}}}\int_{|z_i|>\varepsilon} g_{zx}(z_i, \tilde{x}_i)\phi_{ni\delta} f_{Z|\tilde{X}}(z_i \mid \tilde{x}_i)\,dz_i\, g_x(\tilde{x}_i)\,dP_{\tilde{X}}(\tilde{x}_i) = o(h_n) \tag{A.2}$$

and

$$\int_{\tilde{\mathcal{X}}}\int_{|u-\tilde{x}_i'\delta|\leq\varepsilon/h_n} g_u(u)\phi(u)\,du\, g_x(\tilde{x}_i)\,dP_{\tilde{X}}(\tilde{x}_i) = \mathrm{E}[Cg_x(\tilde{x}_i)] + o(h_n), \tag{A.3}$$

where $C = \int_{-\infty}^{\infty} g_u(u)\phi(u)\,du$.

Proof. First we establish (A.2). Because $f_{Z|\tilde{X}}(\cdot \mid \cdot)$ is bounded for $|z_i| > \varepsilon$ by Assumption A5', the Euclidean norm of (A.2) is bounded above by a constant times $h_n^{-1}\int_{\tilde{\mathcal{X}}}\int_{|z_i|>\varepsilon}\phi_{ni\delta}\,dz_i\,g_x(\tilde{x}_i)\,dP_{\tilde{X}}(\tilde{x}_i)$. With the change of variables $u = \frac{z_i}{h_n} + \tilde{x}_i'\delta$, noting that $\tilde{x}_i$ is bounded, this term is bounded by a constant times $\int_{|u|>\varepsilon/h_n}\phi(u)\,du$, which is $o(h_n)$ by the tail behavior properties of the normal pdf.

Next we consider (A.3), letting I_n denote the double integral on the left hand side. We note that if the range of this integral were $u \in (-\infty, +\infty)$, then it would evaluate to $\mathrm{E}[Cg_x(\tilde{x}_i)]$. Intuitively, the range of integration approaches the real line as $n \to \infty$, however, we need to formally establish that the difference is $o(h_n)$.

Note that we can write $I_n = I_1 + I_{2n}$ where

$$I_1 = \int_{\tilde{\mathcal{X}}}\int_{-\infty}^{\infty} g_u(u)\phi(u)\,du\, g_x(\tilde{x}_i)\,dP_{\tilde{X}}(\tilde{x}_i) = \mathrm{E}[Cg_x(\tilde{x}_i)]$$

and

$$I_{2n} = \int_{\tilde{\mathcal{X}}}\int_{|u-\tilde{x}_i'\delta|>\varepsilon/h_n} g_u(u)\phi(u)\,du\, g_x(\tilde{x}_i)\,dP_{\tilde{X}}(\tilde{x}_i).$$

Note that since $\|\tilde{x}_i\|$ is bounded, and consequently $|h_n \tilde{x}_i' \delta|$ can be made arbitrarily small, we have

$$h_n^{-1} \int_{|u - \tilde{x}_i'\delta| > \varepsilon/h_n} g_u(u)\phi(u)\, du \leq h_n^{-1} M \int_{|u| > \varepsilon/h_n} \phi(u)\, du.$$

The right hand side converges to 0 as $n \to \infty$ by the tail behavior properties of the normal pdf. Thus $h_n^{-1} I_{2n} = o(1)$ by the dominated convergence theorem, which permits us to exchange limits and integrals, as g_x is bounded over the support of $\tilde{x}_i$. ■

Lemma A.2. *Under Assumptions A1–A4 and A5'–A10', if $h_n \to 0$ and $nh_n \to \infty$, then $\hat{\theta} - \theta_0 = O_p(h_n)$.*

Proof. Let $z_i = x_i'\beta_0$, let δ be any $(k-1) \times 1$ vector, let $\Phi_{ni\delta}$ and $\phi_{ni\delta}$ denote $\Phi(\frac{z_i}{h_n} + \tilde{x}_i'\delta)$ and $\phi(\frac{z_i}{h_n} + \tilde{x}_i'\delta)$, respectively, and define the random process

$$T_n(\delta) = \frac{1}{nh_n^2} \sum_{i=1}^{n} (y_i - \Phi_{ni\delta})\phi_{ni\delta}\tilde{x}_i.$$

We will first show that

$$\| \mathrm{E}[T_n(\delta)] - Q\delta \| = O(1) + O(h_n\|\delta\|) + O(h_n\|\delta\|^2). \tag{A.4}$$

Before proving (A.4), we explain why it is being shown. If we let $\delta = h_n^{-1}(\hat{\theta} - \theta_0)$, then by the first order condition, $T_n(\delta) = o_p(1)$ and the conclusion of Lemma A.2 will follow from the assumption that Q is full rank (Assumption A10').

To show (A.4), we first note that

$$\begin{aligned} \mathrm{E}[T_n(\delta)] &= h_n^{-2} \int_{\tilde{\mathcal{X}}} \int_{|z_i| \leq \varepsilon} \left(\tilde{P}(\tilde{x}_i, z_i) - \Phi_{ni\delta} \right) \phi_{ni\delta} f_{Z|\tilde{X}}(z_i \mid \tilde{x}_i)\, dz_i\, \tilde{x}_i\, dP_{\tilde{X}}(\tilde{x}_i) & \text{(A.5)} \\ &+ h_n^{-2} \int_{\tilde{\mathcal{X}}} \int_{|z_i| > \varepsilon} \left(\tilde{P}(\tilde{x}_i, z_i) - \Phi_{ni\delta} \right) \phi_{ni\delta} f_{Z|\tilde{X}}(z_i \mid \tilde{x}_i)\, dz_i\, \tilde{x}_i\, dP_{\tilde{X}}(\tilde{x}_i). & \text{(A.6)} \end{aligned}$$

The integral in (A.6) converges to zero by Lemma A.1.

Turing attention to (A.5), since the integral is over z_i in a neighborhood of 0, we take a second order expansion of $\tilde{P}(\tilde{x}_i, z_i)$ and $f_{Z|\tilde{X}}(z_i \mid \tilde{x}_i)$ around $z_i = 0$. Note this is permitted by Assumptions A5' and A9'. This gives us the sum of three terms and a remainder term:

$$h_n^{-2} \int_{\tilde{\mathcal{X}}} \int_{|z_i| \leq \varepsilon} \left(\frac{1}{2} - \Phi_{ni\delta} \right) \phi_{ni\delta} f_{Z|\tilde{X}}(0 \mid \tilde{x}_i)\, dz_i\, \tilde{x}_i\, dP_{\tilde{X}}(\tilde{x}_i), \tag{A.7}$$

$$h_n^{-2} \int_{\tilde{\mathcal{X}}} \int_{|z_i| \leq \varepsilon} \tilde{P}_2(\tilde{x}_i, 0) \phi_{ni\delta} f_{Z|\tilde{X}}(0 \mid \tilde{x}_i) z_i\, dz_i\, \tilde{x}_i\, dP_{\tilde{X}}(\tilde{x}_i), \tag{A.8}$$

and

$$h_n^{-2}\int_{\tilde{\mathcal{X}}}\int_{|z_i|\le\varepsilon}\left(\frac{1}{2}-\Phi_{ni\delta}\right)\phi_{ni\delta}f'_{Z|\tilde{X}}(0\mid\tilde{x}_i)z_i\,dz_i\,\tilde{x}_i\,dP_{\tilde{X}}(\tilde{x}_i). \tag{A.9}$$

In (A.8), $\tilde{P}_2$ denotes the partial derivative of $\tilde{P}$ with respect to the second argument, z_i. The remainder term involves all second order derivatives and will be dealt with after deriving the properties of each of the above three terms.

We first show (A.7) is $o(1)$. We use the change of variables $u=\frac{z_i}{h_n}+\tilde{x}_i'\delta$, and obtain

$$h_n^{-1}\int_{\tilde{\mathcal{X}}}\int_{|u-\tilde{x}_i'\delta|\le\varepsilon/h_n}\left(\frac{1}{2}-\Phi(u)\right)\phi(u)f_{Z|\tilde{X}}(0\mid\tilde{x}_i)\,du\,\tilde{x}_i\,dP_{\tilde{X}}(\tilde{x}_i).$$

Then, we obtain the result by applying Lemma A.1 with $g_u(u)=\frac{1}{2}-\Phi(u)$, $g_x(\tilde{x}_i)=f_{Z|\tilde{X}}(0\mid\tilde{x}_i)\tilde{x}_i$, and $\int g_u(u)\phi(u)\,du=0$, noting that the conditional density of z_i is bounded near 0 by Assumption A5', as is $\tilde{x}_i$ over its support.

Turning attention to (A.8), the same change of variables as before yields

$$\int_{\tilde{\mathcal{X}}}\int_{|u-\tilde{x}_i'\delta|\le\varepsilon/h_n}\tilde{P}_2(\tilde{x}_i,0)\phi(u)f_{Z|\tilde{X}}(0\mid\tilde{x}_i)(u-\tilde{x}_i'\delta)\,du\,\tilde{x}_i\,dP_{\tilde{X}}(\tilde{x}_i).$$

We apply Lemma A.1 separately to the integrals involving u and $\tilde{x}_i'\delta$ respectively, with $g_u(u)=u$, $\int g_u(u)\phi(u)\,du=0$, $g_x(\tilde{x}_i)=\tilde{P}_2(\tilde{x}_i,0)f_{Z|\tilde{X}}(0\mid\tilde{x}_i)\tilde{x}_i$ for the first integral and $g_u(u)=1$, $\int g_u(u)\phi(u)\,du=1$, $g_x(\tilde{x}_i)=\tilde{P}_2(\tilde{x}_i,0)f_{Z|\tilde{X}}(0\mid\tilde{x}_i)\tilde{x}_i\tilde{x}_i'\delta$ for the second. Combining our results, we may conclude that (A.8) is $Q\delta+o(1)$.

We now derive the limit of (A.9). With the same change of variables we get

$$\int_{\tilde{\mathcal{X}}}\int_{|u-\tilde{x}_i'\delta|\le\varepsilon/h_n}\left(\frac{1}{2}-\Phi(u)\right)\phi(u)f'_{Z|\tilde{X}}(0\mid\tilde{x}_i)(u-\tilde{x}_i'\delta)\,du\,\tilde{x}_i\,dP_{\tilde{X}}(\tilde{x}_i).$$

As before, we focus on the integrals involving u and $\tilde{x}_i'\delta$ separately. We apply Lemma A.1 twice, with $g_u(u)=(1/2-\Phi(u))u$, $g_x(\tilde{x}_i)=f_{Z|\tilde{X}}(0\mid\tilde{x}_i)\tilde{x}_i$, and $\int g_u(u)\phi(u)\,du=c_\phi\approx 0.28$ for the first integral and $g_u(u)=(1/2-\Phi(u))$, $g_x(\tilde{x}_i)=f_{Z|\tilde{X}}(0\mid\tilde{x}_i)\tilde{x}_i\tilde{x}_i'\delta$, and $\int g_u(u)\phi(u)\,du=0$ for the second. Combining our results, we find that (A.9) is $\mathrm{E}[c_\phi f'_{Z|\tilde{X}}(0\mid\tilde{x}_i)\tilde{x}_i]+o(1)$. This establishes the asymptotic properties of the three terms (A.7), (A.8), and (A.9). Combined, their sum is $Q\delta+\mathrm{E}[c_\phi f'_{Z|\tilde{X}}(0\mid\tilde{x}_i)\tilde{x}_i]+o(1)$.

Finally, we deal with the remainder term, which involves the integral evaluated at second order derivatives times z_i^2. Using the same change of variables and limit arguments used to establish the limits of (A.7), (A.8), and (A.9), it follows that the Euclidean norm of the remainder term is $o(1)+O(h_n\|\delta\|)+O(h_n\|\delta\|^2)$. Collecting all results establishes (A.4).

Therefore the conclusion of the lemma follows since by setting $\delta=h_n^{-1}(\hat{\theta}-\theta_0)$ as in this case $T_n(\delta)=o_p(1)$ by the first order condition and the established consistency of the estimator. ■

Hessian Term As mentioned at the beginning of this section, both Theorems 2.2 and 3.1 will be proven by expanding the first order condition of the local NLLS estimator. We first derive the probability limit of the Hessian term in (A.1), which we denote by $\hat{H}$:

$$\hat{H} = \frac{1}{nh_n^2} \sum_{i=1}^{n} \left((y_i - \Phi_{ni}^*)\phi_{ni}^{\prime *} - \phi_{ni}^{2*} \right) \tilde{x}_i \tilde{x}_i'. \tag{A.10}$$

To do so, we first evaluate

$$\mathrm{E}\left[\left((y_i - \Phi_{ni})\phi_{ni}' - \phi_{ni}^2 \right) \tilde{x}_i \tilde{x}_i' \right] / h_n^2. \tag{A.11}$$

As before, decompose the support of z_i into the regions $|z_i| \leq \varepsilon$ and $|z_i| > \varepsilon$, where $\varepsilon > 0$ is small enough so the smoothness assumptions in A5' and A9' can be applied for $|z_i| \leq \varepsilon$. When $|z_i| > \varepsilon$, the integral is negligible (i.e., $o(1)$) as before, by Lemma A.1. When $|z_i| \leq \varepsilon$, by the change of variables $u = z_i/h_n$, and a first order expansion around $h_n = 0$ (permitted by Assumptions A5' and A9'), it follows that (using Assumptions A2 and A8' and the properties of the normal integrals stated previously) this term can be expressed as

$$\mathrm{E}\left[\tilde{P}_2(\tilde{x}_i, 0) f_{Z|\tilde{X}}(0 \mid \tilde{x}_i) \tilde{x}_i \tilde{x}_i' \right] + O(h_n).$$

Note that the derivation of the above expectation uses the property that $\int u^2 \phi(u)\, du = 1$ and that the term in the expansion involving $f_{Z|\tilde{X}}'(0 \mid \tilde{x}_i)$ vanished because of the last property of the normal distribution stated previously. Now consider the expectation in (A.11) evaluated at $\beta = \beta_n$ where $\beta_n - \beta_0 = O(h_n)$. We do this because the Hessian term is evaluated not at β_0, but at the value β^* on the line segment between $\hat{\beta}$ and β_0, and we have established that $\hat{\beta} - \beta_0 = O(h_n)$.

Let z_{ni} denote $x_i'\beta_n$, let $\Phi_{\beta ni}$ denote $\Phi(z_{ni}/h_n)$, and define $\phi_{\beta ni}$ and $\phi_{\beta ni}'$ analogously. We will evaluate

$$\mathrm{E}\left[\left((y_i - \Phi_{\beta ni})\phi_{\beta ni}' - \phi_{\beta ni}^2 \right) \tilde{x}_i \tilde{x}_i' \right] / h_n^2.$$

To do so we add and subtract $\tilde{P}(\tilde{x}_i, z_{ni}) \equiv P(\epsilon_i \leq z_{ni} \mid x_i)$, which we denote by $\tilde{P}_{\beta i}$. So we will evaluate

$$\mathrm{E}\left[\left((y_i - \tilde{P}_{\beta i})\phi_{\beta ni}' \right) \tilde{x}_i \tilde{x}_i' \right] / h_n^2 \tag{A.12}$$

and

$$\mathrm{E}\left[\left((\tilde{P}_{\beta i} - \Phi_{\beta ni})\phi_{\beta ni}' - \phi_{\beta ni}^2 \right) \tilde{x}_i \tilde{x}_i' \right] / h_n^2. \tag{A.13}$$

Turning attention to (A.12), we express it as the integral

$$h_n^{-2} \int_{\tilde{\mathcal{X}}} \int (\tilde{P}_i - \tilde{P}_{\beta ni})\phi_{\beta ni}' f_{Z|\tilde{X}}(z_{ni} \mid \tilde{x}_i) \tilde{x}_i \tilde{x}_i' \, dz_{ni} \, dP_{\tilde{\mathcal{X}}}(\tilde{x}_i),$$

where recall $\tilde{P}_i$ denotes $\tilde{P}(\tilde{x}_i, z_i)$. Now decompose the integral into the regions $|z_{ni}| \leq \varepsilon$ and $|z_{ni}| > \varepsilon$. The integral in the latter region is negligible by Lemma A.1. In the former region, take a first order expansion of $\tilde{P}_i$ around $\tilde{P}_{\beta i}$, which yields

$$h_n^{-2} \int_{\tilde{\mathcal{X}}} \int_{|z_{ni}| \leq \varepsilon} f_{\epsilon|X}(x_i'\tilde{\beta}_n) \phi'_{\beta ni} f_{Z_n|\tilde{X}}(z_{ni} \mid \tilde{x}_i) \tilde{x}_i \tilde{x}_i' \, dz_{ni} \, dP_{\tilde{X}}(\tilde{x}_i)(\beta - \beta_0), \tag{A.14}$$

where $f_{\epsilon|X}$ denotes the density of ϵ_i conditional on x_i evaluated at $x_i'\tilde{\beta}_n$, with $\tilde{\beta}_n$ being a value on the line segment between β_n and β_0, and where $f_{Z_n|\tilde{X}}(\cdot)$ denotes the conditional density function of z_{ni}. We note that since z_{ni} is in a neighborhood of 0, and $\tilde{\beta}_n$ is in a neighborhood of β_0, with $\tilde{x}_i$ bounded and the compactness of Θ it follows that $x_i'\tilde{\beta}_n$ is in a neighborhood of 0 as well, where the density of ϵ_i is bounded by Assumption A9'. Therefore, since $(\beta_n - \beta_0)/h_n = O(1)$ from Lemma A.2, the above integral in (A.14) will converge to 0 if we can show the integral

$$h_n^{-1} \int_{\tilde{\mathcal{X}}} \int_{|z_{ni}| \leq \varepsilon} \phi'_{\beta ni} f_{Z_n|\tilde{X}}(z_{ni} \mid \tilde{x}_i)(\tilde{x}_i \tilde{x}_i')\tilde{x}_i' \, dz_{ni} \, dP_{\tilde{X}}(\tilde{x}_i)$$

converges to 0.

Next, doing the change of variables $u_i = x_i'\beta_n/h_n$, we express this integral as

$$\int_{\tilde{\mathcal{X}}} \int_{|u| \leq \varepsilon/h_n} \phi'(u) f_{Z_n|\tilde{X}}(uh_n \mid \tilde{x}_i)(\tilde{x}_i \tilde{x}_i')\tilde{x}_i' \, du \, dP_{\tilde{X}}(\tilde{x}_i),$$

take a first order expansion around $h_n = 0$, and, noting that $\int \phi'(u)\, du = 0$, we find that the above integral converges to 0, again using the dominated convergence theorem.

Next, we turn to the term (A.13), which again we write as an integral decomposed over the regions $|z_n| \leq \varepsilon$ and $|z_n| > \varepsilon$. We can show the integral over the latter region is asymptotically negligible using Lemma A.1. In the former region, we make the same change of variables, yielding the integral

$$h_n^{-1} \int_{\tilde{\mathcal{X}}} \int_{|u| \leq \varepsilon/h_n} \left[(\tilde{P}(\tilde{x}_i, uh_n) - \Phi(u))\phi'(u) - \phi^2(u) \right] f_{Z_n|\tilde{X}}(uh_n \mid \tilde{x}_i) \tilde{x}_i \tilde{x}_i' \, du \, dP_{\tilde{X}}(\tilde{x}_i).$$

Taking an expansion around $h_n = 0$, the lead term is

$$h_n^{-1} \int_{\tilde{\mathcal{X}}} \int_{|u| \leq \varepsilon/h_n} \left[(1/2 - \Phi(u))\phi'(u) - \phi^2(u) \right] f_{Z_n|\tilde{X}}(0 \mid \tilde{x}_i) \, du \, \tilde{x}_i \tilde{x}_i' \, dP_{\tilde{X}}(\tilde{x}_i),$$

which converges to 0 as $n \to \infty$, since the integral over $|u| \leq \varepsilon/h_n$ converges to 0 faster than h_n.

The first derivative term in the expansion, which involves the term uh_n, is

$$\int_{\tilde{\mathcal{X}}} \int_{|u|\leq\varepsilon/h_n} \Big\{ \left[(1/2 - \Phi(u))\phi'(u) - \phi^2(u)\right] u f'_{Z_n|\tilde{X}}(0 \mid \tilde{x}_i) \\ - \tilde{P}_2(\tilde{x}_i, 0) f_{Z_n|\tilde{X}}(0 \mid \tilde{x}_i)\phi'(u)u \Big\} \, du \, \tilde{x}_i \tilde{x}_i' \, dP_{\tilde{X}}(\tilde{x}_i)$$

and (by the stated normal integral properties and the tail behavior of the normal distribution) can be written as $\mathrm{E}\left[\tilde{P}_2(\tilde{x}_i, 0) f_{Z|\tilde{X}}(0 \mid \tilde{x}_i)\tilde{x}_i\tilde{x}_i'\right] + o(1)$. The second derivative term in the expansion is $O(h_n)$ by similar arguments.

Therefore, collecting all our results we have the expectation in (A.11), evaluated at $\beta = \beta^*$ instead of $\beta = \beta_0$, is $\mathrm{E}\left[\tilde{P}_2(\tilde{x}_i, 0) f_{Z|\tilde{X}}(0 \mid \tilde{x}_i)\tilde{x}_i\tilde{x}_i'\right] + o(1)$.

As a last step we deal with the average in (A.10) minus its expectation. Here we adopt the notation that $\tilde{\psi}_{ni}(\theta)$ denotes the term in the summation in (A.10) when the parameter is θ. We will derive the asymptotic properties of

$$\frac{1}{nh_n^2} \sum_{i=1}^{n} \tilde{\psi}_{ni}(\theta) - \mathrm{E}[\tilde{\psi}_{ni}(\theta)] \tag{A.15}$$

at $\theta = \theta_0$. Since the above term is mean 0, we evaluate the variance,

$$\frac{1}{nh_n^4} \mathrm{E}\left[\left(\tilde{\psi}_{ni}(\theta) - \mathrm{E}[\tilde{\psi}_{ni}(\theta)]\right)^2\right].$$

The lead term involves $\frac{1}{nh_n^4} \mathrm{E}\left[\tilde{\psi}_{ni}(\theta)^2\right]$, for which, after a change of variables (and decomposing the support of z_i into $|z_i| \leq \varepsilon$ and $|z_i| > \varepsilon$ as before), the first term is the constant $\frac{1}{nh_n^3}$ times the integral

$$\int \left\{ \left[1/2 + \Phi(u)^2 - \Phi(u)\right] \phi'(u)^2 + \phi^4(u) - 2(1/2 - \Phi(u))\phi'(u)\phi^2(u) \right\} du.$$

This integral is nonzero. So, the variance of the demeaned sum is $O(1/nh_n^3)$. Therefore, for (A.15) to converge in probability to zero, we need $nh_n^3 \to \infty$. If $nh_n^3 \to c \neq 0$, the demeaned sum converges to a non-degenerate random variable.

Therefore, we will proceed as if $nh_n^3 \to \infty$. The last step is to account for the fact that the demeaned sum is evaluated not at θ_0 but at θ^*, an intermediate value, so we want to evaluate

$$\frac{1}{nh_n^2} \sum_{i=1}^{n} \tilde{\psi}_{ni}(\theta^*) - \mathrm{E}[\tilde{\psi}_{ni}(\theta^*)].$$

Here we will again use the established result that $\hat{\theta} - \theta_0 = O_p(h_n)$. Subtract from the above term $\tilde{\psi}_{ni}(\theta_0) - \mathrm{E}[\tilde{\psi}_{ni}(\theta_0)]$. The resulting term is still mean zero, so as before we only need evaluate the variance. But by a mean value expansion of $\tilde{\psi}_{ni}(\theta_n)$ around $\tilde{\psi}_{ni}(\theta_0)$ for any θ_n such that

$\theta_n - \theta_0 = O(h_n)$, we get terms involving $(\theta_n - \theta_0)/h_n$ which are $O(1)$, implying that as before, the variance is $O(1/nh_n^3)$, which is $o(1)$ under the assumption that $nh_n^3 \to \infty$. Therefore, it sufficed to work with the asymptotic properties of

$$\frac{1}{nh_n^2}\sum_{i=1}^{n}\tilde{\psi}_{ni}(\theta_0) - \mathrm{E}[\tilde{\psi}_{ni}(\theta_0)].$$

This concludes the asymptotic theory for the Hessian term. To summarize what we have shown, if $nh_n^3 \to \infty$, then $\hat{H} = Q + o_p(1)$, so by the invertibility of Q and Slutsky's theorem, we have $\hat{H}^{-1} = Q^{-1} + o_p(1)$. Also, if $nh_n^3 \to c < \infty$, $\hat{H}$ converges to a nondegenerate random variable.

Score Term We next turn attention to the score term,

$$\frac{1}{nh_n}\sum_{i=1}^{n}(y_i - \Phi_{ni})\phi_{ni}\tilde{x}_i.$$

We add and subtract the term $\mathrm{E}\left[(y_i - \Phi_{ni})\phi_{ni}\tilde{x}_i\right]/h_n$. We note that this expected value is $O(h_n)$ by the same change of variables argument (after decomposing the support of z_i as before). Specifically, by a second order expansion of $\tilde{P}(\tilde{x}_i, z_i)$ around $z_i = 0$, permitted by Assumption A9', it is of the form

$$-\left\{\mathrm{E}[f'_{Z|\tilde{X}}(0 \mid \tilde{x}_i)\tilde{x}_i]\int \Phi(u)\phi(u)u\,du\right\}h_n + O(h_n^2).$$

Finally, we note that by the Lindeberg Theorem, when $nh_n \to \infty$,

$$\frac{1}{\sqrt{nh_n}}\sum_{i=1}^{n}\left\{(y_i - \Phi_{ni})\phi_{ni}\tilde{x}_i - \mathrm{E}\left[(y_i - \Phi_{ni})\phi_{ni}\tilde{x}_i\right]\right\} \xrightarrow{d} \mathrm{N}\left(0, c_1 \cdot \mathrm{E}\left[f_{Z|\tilde{X}}(0 \mid \tilde{x}_i)\tilde{x}_i\tilde{x}_i'\right]\right),$$

where, recall, $c_1 = \int \Phi^2(u)\phi^2(u)\,du$.

However, since the bias is $O(h_n)$ and the variance of the score term is $O(\frac{1}{nh_n})$ (using arguments identical to evaluating the order of the variance in nonparametric density estimation) the optimal rate of convergence (of the score term) in a mean squared error sense is $h_n = O(n^{-1/3})$. However, under this rate $nh_n^3 \to c < \infty$. Thus, the Hessian term does not converge to a degenerate distribution, and the local NLLS estimator is not asymptotically Gaussian.

So by combining our results with the Hessian term, $\hat{H}$, we have the following representation (if $nh_n^3 \to \infty$) in the conclusion in Theorem 3.1:

$$\hat{\theta} - \theta_0 = (Q + o_p(1))^{-1}\left[\left(\frac{1}{nh_n}\sum_{i=1}^{n}(y_i - \Phi_{ni})\phi_{ni}\tilde{x}_i - \mathrm{E}[(y_i - \Phi_{ni})\phi_{ni}\tilde{x}_i]\right)\right.$$
$$\left. - \left\{Q^{-1}\,\mathrm{E}[f'_{Z|\tilde{X}}(0 \mid \tilde{x}_i)\tilde{x}_i]\int \Phi(u)\phi(u)u\,du\right\}h_n + O(h_n^2)\right]. \quad \text{(A.16)}$$

Furthermore, collecting all derived results regarding rates of convergence for the Hessian and score terms as a function of h_n, the conclusions of Theorem 2.2 follow. Specifically, if in the score term, we equate the standard deviation which is $O(1/\sqrt{nh_n})$ with the bias which is $O(h_n)$, we get $h_n = O(n^{-1/3})$, but this violates our assumption that $nh_n^3 \to \infty$ that was needed in the Hessian term. The Hessian condition is also violated if $h_n = o(n^{-1/3})$, which would also result in a slower rate of convergence. If $nh_n^3 \to \infty$, then the Hessian term converges in probability to Q, but the bias in the score term dominates the variance, and we have

$$h_n^{-1}(\hat{\theta} - \theta_0) \xrightarrow{p} -Q^{-1} \left\{ \int_{\tilde{\mathcal{X}}} f_{Z|\tilde{X}}(0 \mid \tilde{x}_i)\tilde{x}_i \, dP_{\tilde{X}}(\tilde{x}_i) \int \Phi(u)\phi(u)u \, du \right\}.$$

■

Finally, as will be formally discussed in the next section, if the bias term in the linear component of the representation is $O(h_n^2)$ by some modified procedure, the optimal sequence is $h_n = O(n^{-1/5})$, in which case the Hessian term converges to a constant matrix. In this case we may apply Slutsky's theorem to conclude that the bias-corrected estimators are asymptotically normal and converge at the rate of $O_p(n^{-2/5})$.

A.3 Proof of Theorem 3.2

The theorem follows almost directly from the results in Theorem 3.1, and follows from establishing that the bias of the jackknifed estimator is $O(h_n^2)$.

For the jackknifed estimator the bias term is of the form

$$-\left\{ Q^{-1} \, \mathrm{E}[f_{Z|\tilde{X}}(0 \mid \tilde{x}_i)\tilde{x}_i] \int \Phi(u)\phi(u)u \, du \right\} (w_1\kappa_1 + w_2\kappa_2)h_n + \mathcal{B}_{\mathrm{jk}} h_n^2 = O(h_n^2),$$

where the equality follows from the second condition imposed on w_1, w_2, κ_1, and κ_2, and

$$\mathcal{B}_{\mathrm{jk}} = (w_1\kappa_1^2 + w_2\kappa_2^2)\frac{1}{2} \int_{\tilde{\mathcal{X}}} \int \left\{ \left(\frac{1}{2} - \Phi(u) \right) f_{Z|\tilde{X}}(0 \mid \tilde{x}_i) + 2\tilde{P}_2(\tilde{x}_i, 0) f'_{Z|\tilde{X}}(0 \mid \tilde{x}_i) + \tilde{P}_{22}(\tilde{x}_i, 0) f_{Z|\tilde{X}}(0 \mid \tilde{x}_i) \right\} u^2 \phi(u) \, du \, \tilde{x}_i \, dP_{\tilde{X}}(\tilde{x}_i).$$

Therefore, we have $n^{2/5}(\hat{\theta}_{\mathrm{jk}} - \theta_0) \xrightarrow{d} \mathrm{N}(\mathcal{B}_{\mathrm{jk}}, Q^{-1}V_{\mathrm{jk}}Q^{-1})$. ■

A.4 Proof of Theorem 3.3

As alluded to in Section 3.2, the function $F(\cdot)$ in the objective function

$$\frac{1}{n} \sum_{i=1}^{n} \left(y_i - F\left(\frac{x_i'\beta}{h_n} \right) \right)^2$$

cannot be a cumulative distribution function if bias reduction is to be achieved. Using the same arguments as in deriving the linear representation, we impose conditions F1–F6 on $F(\cdot)$ and its first and second derivatives, denoted by $f(\cdot)$ and $f'(\cdot)$ respectively. Under these conditions, following the arguments used in the linear representation derivation, the local NLLS estimator using the function $F(\cdot)$ will converge at the rate $O_p(n^{-2/5})$, which is the optimal rate as established in Horowitz (1993a). It will have an asymptotic Gaussian distribution with asymptotic bias $\mathcal{B}_F$ and an asymptotic variance of the form $Q_F^{-1}V_FQ_F^{-1}$. ■

A.5 Jackknife Weights

The form of the liming distribution of the jackknifed estimator is useful for providing guidance for the form of the weights w_1 and w_2. For example, with the analytic form of the asymptotic bias and asymptotic variance, one could attempt to select the two weights that minimize the asymptotic mean squared error subject to the constraints $w_1 + w_2 = 1$ and $w_1\kappa_1 + w_2\kappa_2 = 0$, where, recall, κ_1 and κ_2 denote the two constants for the two bandwidth sequences. Even treating these two constants as given, the optimal values of w_1 and w_2 would depend on the unknown density and distribution function (as well as their derivatives) appearing in the asymptotic bias and variance. These would have to be estimated first, making implementation difficult and requiring the selection of additional bandwidths.

An easier to implement approach would be to only minimize with respect to the constant terms in the asymptotic mean squared error. That is, to minimize the function

$$\frac{1}{4}(w_1\kappa_1^2 + w_2\kappa_2^2)^2 + c_1w_1^2\kappa_1^{-1} + c_1w_2^2\kappa_2^{-1} + 2w_1w_2c_2\kappa_1^{-1}$$

with respect to κ_1 and κ_2 subject to the constraints, which can be solved for κ_1 and κ_2 and substituted in as $w_1 = \kappa_2/(\kappa_2 - \kappa_1)$ and $w_2 = 1 - w_1 = -\kappa_1/(\kappa_2 - \kappa_1)$. The values which minimize this function are approximately $\kappa_1 = 0.56334$, $\kappa_2 = 1.0180$, $w_1 = 2.2389$, and $w_2 = -1.2389$.

A second approach uses a simple "rule of thumb" estimate for the matrix $Q^{-1}V_1Q^{-1}$ appearing in the asymptotic variance given in Theorem 3.2. Let $\hat{E}[\tilde{x}_i\tilde{x}_i']$ denote the sample analog estimator of the expectation and let $\hat{E}[\tilde{x}_i\tilde{x}_i']^{-1}$ denote its inverse. Then we could alter the above objective function to

$$\frac{1}{4}(w_1\kappa_1^2 + w_2\kappa_2^2)^2 + \hat{v}_{\text{ROT}}\left(c_1w_1^2\kappa_1^{-1} + c_1w_2^2\kappa_2^{-1} + 2w_1w_2c_2\kappa_1^{-1}\right)$$

where $\hat{v}_{\text{ROT}} = 0.4^{-3} \cdot \left\|\hat{E}[\tilde{x}_i\tilde{x}_i']^{-1}\right\|_2$ is a rule of thumb approximation of the norm of $Q^{-1}V_1Q^{-1}$, under simplifying normality and independence assumptions. Here 0.4 is the value of the standard normal pdf evaluated at zero and $\|\cdot\|_2$ denotes the Frobenius norm, the square root of the sum of the squared elements of the matrix.

Both of these approaches are evaluated in the simulation studies, labeled JKNLLS-1 and JKNLLS-2 respectively.

A.6 Weighted NLLS

With the limiting Gaussian distribution in hand, a natural extension of the NLLS estimator would be to consider weighting observations to improve efficiency, analogous to generalized least squares. In the parametric probit model, it is well known that NLLS is not as efficient as MLE, but an optimally weighted NLLS is asymptotically equivalent to MLE. For the NLLS estimator, a weighted version would aim to minimize the asymptotic mean squared error (AMSE).

The weighted estimator, referred to here as WNLLS, would minimize the objective function

$$\frac{1}{n}\sum_{i=1}^{n} w(x_i)\left(y_i - F\left(\frac{x_i'\beta}{h_n}\right)\right)^2,$$

where $w(\cdot)$ denotes the weight function. The limiting distribution follows immediately from our linear representation. Let $\tilde{w}(\tilde{x}_i, x_i'\beta_0)$ denote the reparametrized weight function, expressed as a function of the subset of regressors $\tilde{x}_i$ and the index $x_i'\beta_0$. Now the asymptotic bias is of the form

$$\mathcal{B}_F^w = \frac{1}{2}\int_{\tilde{\mathcal{X}}}\int\left\{\left(\frac{1}{2} - F(u)\right) f_{Z|\tilde{X}}(0 \mid \tilde{x}) + 2\tilde{P}_2(\tilde{x}_i, 0) f'_{Z|\tilde{X}}(0 \mid \tilde{x}) \right.$$
$$\left. + \tilde{P}_{22}(\tilde{x}_i, 0) f_{Z|\tilde{X}}(0 \mid \tilde{x})\right\} u^2 f(u)\, du\, \tilde{w}(\tilde{x}_i, 0)\tilde{x}_i\, dP_{\tilde{X}}(\tilde{x}_i)$$

and the components of the asymptotic variance matrix are

$$V_F^w = c_{F_1}\cdot \mathrm{E}[\tilde{w}^2(\tilde{x}_i, 0)\tilde{x}_i\tilde{x}_i' f_{Z|\tilde{X}}(0 \mid \tilde{x}_i)]$$

and

$$Q_F^w = \mathrm{E}\left[\left(c_{F_2}\tilde{P}_2(\tilde{x}_i, 0) f_{Z|\tilde{X}}(0 \mid \tilde{x}_i) + c_{F_3} f'_{Z|\tilde{X}}(0 \mid \tilde{x}_i)\right)\tilde{w}(\tilde{x}_i, 0)\tilde{x}_i\tilde{x}_i'\right].$$

This immediately suggests an infeasible weighting function. If we condition on a particular value of $\tilde{x}_i$, $\tilde{x}$, we can treat all the functions inside the expectations $\mathcal{B}_F^w$,V_F^w, and Q_F^w as given and minimize the conditional mean squared error with respect to $w(\tilde{x}, 0)$, which we refer to here as $w^*(\tilde{x}, 0)$. This minimized valued will obviously depend on the values of the other functions evaluated at $\tilde{x}$. Then our infeasible estimator minimizes the objective function

$$\frac{1}{n}\sum_{i=1}^{n} \tilde{w}^*(\tilde{x}_i, 0)\left(y_i - F\left(\frac{x_i'\beta}{h_n}\right)\right)^2.$$

Of course what makes this approach infeasible is that the optimal function $w^*(\tilde{x}, 0)$ depends on the other functions, such as $f_{Z|\tilde{X}}(0 \mid \tilde{x})$ and $\tilde{P}_2(\tilde{x}_i, 0)$, which are unknown. But analogous to feasible GLS for the linear model, one can first estimate β_0 using a suboptimal weighting function, say, $w(x_i) = 1$, and use that to nonparametrically estimate the unknown nuisance functions $f_{Z|\tilde{X}}(0 \mid \tilde{x})$ and $\tilde{P}_2(\tilde{x}_i, 0)$ that can then be used to obtain a feasible estimator of $\tilde{w}^*(\tilde{x}_i, 0)$.