Distribution Free Estimation of Heteroskedastic Binary Response Models Using Probit/Logit Criterion Functions

Shakeeb Khan*
Duke University

Revised: March 2010

Abstract

In this paper estimators for distribution free heteroskedastic binary response models are proposed. The estimation procedures are based on relationships between distribution free models with a conditional median restriction and parametric models (such as Logit/Probit) exhibiting (multiplicative) heteroskedasticity. The first proposed estimator is based on the observational equivalence between the two models, and is a semiparametric sieve estimator (see, e.g. Gallant and Nychka(1987), Ai and Chen(2003), Chen, Hong and Tamer(2005)) for the regression coefficients, based on maximizing standard Logit/Probit criterion functions, such as NLLS and MLE. This procedure has the advantage that choice probabilities and regression coefficients are estimated simultaneously. The second proposed procedure is based on the equivalence between existing semiparametric estimators for the conditional median model (Manski(1975,1985), Horowitz(1992)) and the standard parametric (Probit/Logit) NLLS estimator. This estimator has the advantage of being implementable with standard software packages such as Stata. Distribution theory is developed for both estimators and a Monte Carlo study indicates they both perform well in finite samples.

JEL Classification: C13,C14,C24

Key Words: binary response, heteroskedasticity, Probit/Logit, sieve estimation.

*Corresponding author. Department of Economics, Duke University, Durham, NC 27708; e-mail:shakeebk@duke.edu. I am grateful to Co-editor T. Amemiya, an Anonymous Associate Editor and 2 Anonymous referees, S. Chen, X. Chen, M. Coppejans, B. Honoré, A. Lewbel, W. Newey, W. Ploberger, J. Powell, and seminar participants at Boston College, Brown, McGill, Harvard/MIT, Rice, and Texas A&M for helpful comments. This research was supported in part by the National Science Foundation through grant SES-0213621.
1 Introduction

The binary response model has received a great deal of attention in both the theoretical and applied econometrics literature, as many economic variables of interest are of a qualitative nature. The model is usually represented by some variation of the following equation:

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

(1.1)

where \( I[\cdot] \) is the usual 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 effect the behavior of \( y_i \). Both the disturbance term \( \epsilon_i \), and the vector \( \beta_0 \) are unobserved, the latter often being the parameter estimated from a random sample of \((y_i, x_i') i = 1, 2, ...n\).

The disturbance term \( \epsilon_i \) is restricted in ways that ensure identification of \( \beta_0 \). Parametric restrictions specify the distribution of \( \epsilon_i \) up to a finite number of parameters and assume it is distributed independently of the covariates \( x_i \). Under such a restriction, \( \beta_0 \) can be estimated (up to scale) using maximum likelihood or nonlinear least squares. However, except in special cases, these estimators are inconsistent if the distribution of \( \epsilon_i \) 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 \( \beta_0 \). The first was the “maximum score” estimator proposed in Manski(1975). Identification of \( \beta_0 \) was based on a conditional median restriction:

\[ \text{med}(\epsilon_i | x_i) = 0 \]  

(1.2)

Manski’s estimator maximized the following objective function

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

(1.3)

Manski(1975,1985) established the estimator’s consistency. Kim and Pollard(1991) established its rate of convergence and limiting distribution, which were \( n^{-1/3} \) and non-Gaussian, respectively.

Horowitz(1992) modified the procedure by “smoothing” the objective function in (1.3). Specifically, his approach was to maximize the following objective function:

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

(1.4)
where $K_h(\cdot) \equiv \frac{1}{h}K(\cdot/h)$ with $K(\cdot)$ denoting a smooth kernel function, and $h$ denoting a smoothing parameter, converging to 0 with the sample size. Under stronger smoothness conditions on the distributions of $\epsilon_i$ and $x_i$, Horowitz showed that the estimator converges at the rate of $n^{-2/5}$ with an asymptotically normal distribution. By strengthening the conditions further, he was able to attain a rate of $n^{-p/(2p+1)}$ where $p \geq 2$ is an integer related to the order of smoothness of the distributions of $\epsilon_i$ and $x_i'\beta_0$ in neighborhoods of 0.

These estimators have two disadvantages which this paper attempts to address. For one, both the maximum score and smooth maximum score estimation procedures only provided an estimator of $\beta_0$. As discussed in Manski(1988), an estimator of $\beta_0$ permits “structural analysis”, which may be of interest for one of two reasons. For one, the researcher may have a scientific interest in learning about the process yielding binary outcomes. The other motive is prediction\(^1\), where structural analysis enables more precise and tractable prediction, as well as extrapolation. However, choice probabilities and marginal effects are also of interest in most practical applications- see Greene(1997) for an explanation. Unfortunately, the maximum score and smooth maximum score procedures do not estimate these variables.

Alternative semiparametric restrictions used in the literature were independence/index restrictions. These restrictions are much stronger than the median restriction mentioned, as they require the error term to be distributed independently of $x_i$, or depend on $x_i$ through the index $x_i'\beta_0$. Estimation procedures under this restriction include those proposed by Cosslett(1983), Powell et al.(1989), Ichimura(1993), Klein and Spady(1993), and Coppejans(2001). An advantage of most of these procedures is that they enable joint estimation of the regression coefficients and choice probabilities. However a drawback of these procedures is the restrictions they are based on are much stronger than the median restriction mentioned - they require the error term to be distributed independently of $x_i$, or depend on $x_i$ through the index $x_i'\beta_0$. They do not permit the general forms of heteroskedasticity that the conditional median restriction allows for.

Therefore, the first procedure proposed in this paper aims to address the drawbacks of the existing estimators mentioned. Specifically, the general heteroskedasticity of the conditional median restriction is maintained, yet the joint estimation of the regression coefficients and the choice probabilities is also permitted. The idea behind this approach is based on the observational equivalence between a distribution free model under a conditional median restriction, and a (multiplicative) heteroskedastic parametric (e.g. probit, logit) model.

---

\(^1\)By prediction, we mean in a somewhat crude sense. That is, one predicts the value of 1 or 0 based on the sign of the estimated index.
This equivalence result motivates an estimator of the heteroskedastic parametric model, and the estimators proposed permit joint estimation of regression coefficients and choice probabilities. The procedures involve maximizing standard parametric criteria functions, such as MLE, and NLLS probit/logit.

A second drawback of maximum and smoothed maximum score estimators is implementation. Specifically, their objective functions are non-standard and thus they cannot be computed using standard software packages. This motivates the second estimator which can compute regression coefficients in the semiparametric binary choice model under median restrictions using the NLLS objective function for a parametric model such as Logit or Probit. Consequently, the regression coefficients can be estimated using standard software packages such as Stata.

The paper is organized as follows. The following section formally establishes an equivalence result which motivates the first estimation procedure. Section 3 proposes the estimation procedure and establishes its asymptotic properties. Section 4 proposes the estimation procedure for the regression coefficients that is very simple to implement on standard software packages. Section 5 explores the finite sample performance of these estimators via a simulation study. Section 6 concludes. Proofs of the asymptotic properties of the proposed estimators are left to the appendix.

2 An Equivalence Result

The equivalence result is based on the following two models:

\[ y_i = I[x_i' \beta_0 - \epsilon_i \geq 0] \tag{2.1} \]

where

Model 1: Conditional Median Restriction

CM1 $x_i \in \mathbb{R}^k$ is assume to have density with respect to Lebesgue measure\(^2\), which is positive on the set $\mathcal{X} \subseteq \mathbb{R}^k$. In what follows, we will let $x_i^{[j,1]}$ denote the $j$-th component of $x_i$,

\[ j = 1, 2, .., k. \]

\(^2\)This assumption is not required but will be maintained throughout the paper for notational convenience. Technically we require only one regressor to be continuously distributed and have positive density on the real line.
CM2 Letting $\wp_0(t, x)$ denote $P(\epsilon_i \leq t|x_i = x)$ we assume

CM2.1 $\wp_0(\cdot, \cdot)$ is continuous on $\mathbb{R} \times \mathcal{X}$.

CM2.2 $\wp_0'(t, x) \equiv \partial \wp_0(t, x)/\partial t$ exists and is continuous and positive on $R$ for all $x \in \mathcal{X}$.

CM2.3 $\wp_0(0, x) = 1/2$ for $x \in \mathcal{X}$.

CM2.4 $\lim_{t \to -\infty} \wp(t, x) = 0 \lim_{t \to +\infty} \wp(t, x) = 1$.

Model 2: Heteroskedastic Probit/Logit Model

HP1 Assumption CM1.

HP2 $\epsilon_i = \sigma_0(x_i) \cdot u_i$ where $\sigma(\cdot)$ is continuous and positive on $\mathcal{X}$ a.s., and $u_i$ is independent of $x_i$ with any known (e.g. logistic, normal) distribution with median 0 and has a density function which is positive and continuous on the real line.

Theorem 2.1 Under Assumptions CM1,CM2,HP1,HP2, Models 1 and 2 are observationally equivalent.

Proof: Note that the assumptions in Model 2 easily imply the assumptions in Model 1 are satisfied. Now assume the assumptions of Model 1 are satisfied. We will show that there exists a scale function $\sigma_0(\cdot)$ which satisfies Assumption HP2 such that the conditional distribution of the observed dependent variable is the same under the two models. Note it will suffice to show that $P(y_i = 1|x_i = x)$ is the same ($x_i$ a.s.) in both models. Let $P_0(x) = \wp_0(x'\beta_0, x)$ denote this probability function for the Model 1. Now define $\sigma_0(x) = x'\beta_0/\Phi^{-1}(P_0(x))I[x'\beta_0 \neq 0]$ where $\Phi(\cdot)$ denotes the known c.d.f. of $u_i$. Note that $\sigma_0(x) > 0$ for all $x$ such that $x'\beta_0 \neq 0$. This is because $x'\beta_0 > 0 \Rightarrow P_0(x) > 1/2 \Rightarrow \Phi^{-1}(P_0(x)) > 0$, and similarly $x'\beta_0 < 0 \Rightarrow \Phi^{-1}(P_0(x)) < 0$. We immediately see that for the heteroskedastic probit model, $P(y_i = 1|x_i = x) = \Phi(x'\beta_0/\sigma_0(x_i)) = \Phi((\Phi^{-1}(P_0(x)))) = P_0(x)$. Since $x'\beta_0 = 0$ with probability 0 under Assumption CM1, establishing the equivalence of the two models.

Remark 2.1 Here we note the following implications of the established equivalence result:
The above equivalence result is similar to the Lemma on page 737 in Manski (1988) who established a class of “dual models”. These models had nonlinear regression functions and homoskedastic disturbance terms with known distribution. Here we have a linear regression function and a heteroskedastic normal error term which makes it relatively simple to extract the structural component of the model from the choice probabilities. This is enabled by two properties of the model- 1) the normal distribution has median zero and positive density everywhere 2) the scale function is positive everywhere. These constraints can be easily imposed to simultaneously estimate $\beta_0$ and $\sigma_0(\cdot)$, as will be illustrated later in the paper.

Another useful feature of the equivalence result is that it suggests other methods of estimating the model. The first model is generally estimated using the $L_1$ and smoothed $L_1$ norm estimators proposed in Manski (1975) and Horowitz (1992). This is a natural approach in the sense that models with conditional median restrictions are often estimated minimizing least absolute deviation (LAD) objective functions. In the following section, we propose an estimator based on the observationally equivalent Model 2, and describe its advantages over the aforementioned existing estimators.

3 Estimation Procedure

Results in the previous section suggest that one could estimate a heteroskedastic probit model which is “distribution free”. We note that the result matching choice probabilities to a distribution free model restricted the sign of the scale function to be positive everywhere on the support of $x_i$. This will have to be incorporated into the estimation procedure for consistent, distribution free estimation of $\beta_0$. The proposed estimators will consider joint estimation of the “parameter” $(\beta_0', \sigma_0(\cdot))'$. This is analogous to existing estimators (e.g. Cosslett (1983), Klein and Spady (1993), Coppejans (2001)) of $(\beta_0', F(\cdot))'$ where the function $F(\cdot)$ denotes the c.d.f. of the error term. As mentioned previously, these estimators assume independence between $x_i$ and the error term, ruling out conditional heteroskedasticity. On the surface it appears that the approach adopted here is allowing for heteroskedasticity at the expense of requiring a parametrically specified error distribution, as well as restricting

---

3 In fact there are several other structures that enable the choice probabilities to match up with those attained from Model 1. I am grateful to a referee for pointing this out to me.

4 The multiplicative form of the heteroskedasticity has been imposed elsewhere in the literature- see, e.g, Klein and Vella (2005).
the heteroskedasticity to be multiplicative. However, this is not the case. The nonparametric component\(^5\) \(\sigma_0(\cdot)\) permits both an unknown error distribution and the conditional heteroskedasticity of a conditional median restriction. The normality assumption only serves to impose the 0 conditional median restriction, and any distributional assumption on \(u_i\) that has median 0 can be used for distribution free estimation\(^6\).

Before introducing the estimator, we introduce the notation we will adopt to account for the fact that the regression coefficients are only identified up to scale. Following convention we set the last coefficient value to 1 and estimate the \(k-1\) vector \(\theta_0\), where \((\theta'_0, 1)' = \beta_0\).

The heteroskedastic probit model can be viewed as a likelihood model with infinite dimensional parameter space. This class of models has been studied extensively in the econometric and statistics literature. Work in this area includes Geman and Huang\((1982)\), Gallant and Nychka\((1987)\), Wong and Severini\((1991)\), Shen and Wong\((1994)\), Shen\((1997)\), Chen and Shen\((1998)\), and Coppejans\((2001)\), Ai and Chen\((2003)\), Chen et al.\((2005)\), Chen and Pouzo\((2009)\). Most of these papers focus on the method of sieves, which will be used in the construction of an estimator in this paper.

The estimator introduced here is based on treating the scale function as an infinite dimensional parameter. This motivates constructing an estimator which maximize a probit/logit criterion function which includes this function. Specifically we define the criterion function as\(^7\)

\[
\gamma_n(\theta, \ell) = -\frac{1}{n} \sum_{i=1}^n \left( y_i - \Phi \left( (\tilde{x}'_i \theta + x_i^{[k, 1]}) \exp(\ell(x_i)) \right) \right)^2
\]

for \(\alpha \equiv (\theta, \ell)\) in the (infinite dimensional) parameter space \(\mathcal{A}\), whose properties will be detailed shortly.

---

\(^5\)While the previous theorem illustrated identification of \(\beta_0\) and \(P_0\), \(\sigma_0\) is also identified and easily estimable using the procedure discussed in the following section. It should be emphasized that this parameter by itself is of less interest, as it only provides the functional form of the heteroskedasticity when the errors are indeed normally distributed.

\(^6\)Consequently, the normal c.d.f. used here can be interpreted as a particular “kernel” function, analogous to kernel functions used in smoothed maximum score estimation.

\(^7\)Here we have used a probit function, with \(\Phi(\cdot)\) denoting the normal c.d.f. and have adopted the NLLS objective function, as its boundedness properties facilitate proofs. The MLE objective function could also be used, but as will be argued later on, this results in the same asymptotic variance matrix as NLLS in this context. We also note that this NLLS objective function is similar to the smoothed maximum score estimator when one sets \(\exp(\ell(x_i)) = h_n^{-1}\) where \(h_n \to 0\). The properties of this estimator are discussed in Section 4. Finally, note that the infinite dimensional parameter \(\ell(\cdot)\) is the log of the scale function.
To be able to implement the NLLS procedure, since the parameter space is infinite dimensional, we propose a linear in parameters sieve estimator. Let \( b_0(x_i) \) denote a sequence of known basis functions\(^8\). Denote \( b^{\kappa_n}(x_i) = (b_{01}(x_i), \ldots b_{0\kappa_n}(x_i))^T \) for some integer \( \kappa_n \). An approximator of \( g(x_i) \equiv \exp(\ell(x_i)) \) in the above objective function is \( g_n(x_i) = \exp(b^{\kappa_n}(x_i)\Pi_n) \) where \( \Pi_n \) is a vector of constants, and the exponential function serves to impose the positivity of the scale function needed for identification.\(^9\) Let \( \alpha_n \equiv (\theta, g_n) \in A_n \) where \( A_n \) is the sieve space. We can formally define the estimator as\(^{10}\):\(^\star\)

\[
\hat{\alpha}_n = \min_{\alpha \in A_n} \frac{1}{n} \sum_{i=1}^{n} (y_i - \Phi(x_i'\beta \cdot g_n(x_i)))^2
\]

**Remark 3.1** The idea of minimizing a probit or logit criterion function that includes a growing number of basis functions is not new to the econometrics or statistics literature. The first was in the seminal work of McFadden(1974) who introduced the “Mother Logit” model. Stone(1994) estimated choice probabilities in a binary choice model by replacing the index \( x_i'\beta_0 \) with a linear in parameters series, inside a probit or logit likelihood function. While his approach can estimate the probability function by estimating the function \( \Phi(g(x_i)) \), it cannot estimate the structural parameter\(^{11}\) \( \beta_0 \) as the proposed procedure can.

\(^8\)See, e.g. Chen and Shen(1998) for examples of basis functions. For the problem at hand with a regressor that has unbounded support, certain basis functions (e.g. power series) will not achieve the desired approximation for the asymptotic theory to be valid. Consequently, we restrict ourselves to basis functions suitable for approximating functions of regressors with unbounded support- see. e.g. Chen et al.(2005) who use polynomial splines.

\(^9\)The exponential function is not necessary and only adopted here for convenience. One could simply use the approximator \( b^{\kappa_n}(x_i)^T\Pi_n \) and impose constraints on \( \Pi_n \) to ensure positivity of the scale function. Sieve estimators can easily incorporate such parameter constraints- see e.g. Shen(1997).

\(^{10}\)Effectively, we are simply optimizing the objective function with respect to the parameters \( \beta, \Pi_n \). We note the objective function is smooth in these parameters and standard optimization routines can be used to find local optima. However, the objective function is not concave in the parameters, and a search amongst these local maxima needs to be conducted. A similar problem is encountered with the smoothed maximum score estimator and Horowitz(1992) suggested the use of the generalized simulated annealing algorithm in Bohachevsky et al.(1986). We note it is not difficult to implement a procedure where we impose positivity of the scale function by imposing parameter constraints in the optimization. In fact, since the objective function is smooth in the parameters, CO - an application module written in GAUSS, can be used for the problem at hand.

\(^{11}\)However, this estimator can be used in a first stage to estimate choice probabilities which can then be projected onto \( \Phi(x_i'\beta_0, g_n(x_i)) \) to form an estimator of \( \beta_0 \). Since the first stage involves a concave objective function if MLE is used, this approach may have computational advantages over the approach suggested here.
We now detail the conditions under which the asymptotic properties of this estimator will be derived. The first property we will establish is consistency. We first introduce some notation which will be used in imposing smoothness and compactness conditions. This will require introducing new notation, and the notation adopted here is identical to that used in Ai and Chen(2003), Chen et al.(2005). For any $k \times 1$ vector $\mathbf{v} = (v_1, v_2, ... v_k)'$, let $|\mathbf{v}|$ denote $\sum_{i=1}^{k} v_i$. Let $h(\cdot)$ denote any function on $\mathcal{X}$. We denote the $|\mathbf{v}|$-th derivative of the function $h(\cdot)$ as:

$$\nabla^{|\mathbf{v}|} h(x) = \frac{\partial^{|\mathbf{v}|}}{\partial x_1^{v_1} \cdots \partial x_k^{v_k}} h(x)$$

Also, for $\gamma > 0$ we let $\Lambda^\gamma(\mathcal{X})$ denote the space of functions which have up to $[\gamma]$ (here $[\cdot]$ denotes the integer operator) continuous derivatives with the highest derivatives that are Holder continuous$^{12}$ with exponent $(\gamma - [\gamma])$.

Let $\| \cdot \|_E$ denote the Euclidean norm. For a real valued function $h(\cdot) \in \Lambda^\gamma(\mathcal{X})$ we define its Holder norm as

$$\| h \|_{\Lambda^\gamma} = \sup_{x \in \mathcal{X}} |h(x)| + \max_{|\mathbf{v}| = [\gamma]} \sup_{x \neq \bar{x}} \frac{\| \nabla^{|\mathbf{v}|} h(x) - \nabla^{|\mathbf{v}|} h(\bar{x}) \|_E}{\sqrt{(x - \bar{x})'(x - \bar{x})^{\gamma - [\gamma]}}}$$

Finally we denote a space of functions that will be used in defining the parameter space:

$$\Lambda^\gamma_c(\mathcal{X}, w_1) \equiv \{ h \in \Lambda^\gamma(\mathcal{X}) : \| h(\cdot) (1 + x' x)^{-w_1/2} \|_{\Lambda^\gamma} \leq c < \infty \}$$

where $w_1 > 0$ and $c$ is a known constant.

The weighting function of the regressors $(1 + x' x)^{-w_1/2}$ goes to 0 as $\| x \|_E$ goes to infinity and permits $h(\cdot)$ and its derivatives to be unbounded.$^{13}$

With our weighting function we can introduce the weighted sup norm defined as:

$$\| h(x) \|_{\infty, w_1} = \sup_{x \in \mathcal{X}} \| h(x) (1 + x' x)^{-w_1/2} \|_E$$

Our assumptions for consistency are:

**RC1** (Parameter Space) Recall our notation that $\beta = (\theta', 1)'$. Let $\mathcal{B} = \Theta \times 1$. The parameter space $\mathcal{A}$ consists of all pairs $\beta, \ell(\cdot)$ such that

---

$^{12}$See Ai and Chen(2003) for a formal definition of Holder continuity and a more detailed discussion on Holder Spaces.

$^{13}$This is the weighting function used in Chen et al.(2005). Examples of other weighting functions, such as $\exp(-x' x_i)$, can be found in Gallant and Nychka(1987).
i \( \beta \in B \), a compact subset of \( \mathbb{R}^k \).

ii \( \ell(x) \in A^p_{\ell}(X, w_1) \), where \( p > 0 \).

**RC2** (Regressor Distribution) Recall that \( \mathcal{X} \) denotes the support of the regressors. For simplicity, we assume the regressor vector is continuously distributed and denote its joint density function as \( f_{X}(\cdot) \).

i The \( k^{th} \) regressor, conditional on the other regressors, has density function with respect to Lebesgue measure that is positive on \( \mathbb{R} \). The first \( k-1 \) components of \( x_i \), denoted by \( \tilde{x}_i \), are assumed to have bounded support.

ii The support of the distribution of \( x_i \) is not contained in any proper linear subspace of \( \mathbb{R}^k \).

iii \( \int (1 + \|x\|^2) w_2 f_{X}(x) dx < \infty \) where \( w_2 > w_1 \).

**RC3** \( E[b^n(x_i)b^n(x_i)'] \) is nonsingular for all \( n \).

**RC4** The vector \( (y_i, x_i')' \) is i.i.d. and satisfies

\[
P(y_i = 1|x_i) = \Phi(x_i'\beta_0g_0(x_i)) \equiv \Phi(x_i'\beta_0\exp(\ell_0(x_i))]
\]

**Remark 3.2** Before establishing consistency, we comment on some of the regularity conditions imposed:

- **RC1**ii is a type of compactness condition on the functional space, and often imposed in the sieve literature. See, e.g. Chen et al. (2005). With our definition of the sieve space, we will have a sieve approximation error which converges to 0 with respect to a weighted sup norm.

- Assumption RC2i imposes regressor support conditions. The condition on the \( k^{th} \) regressor is used for identification. The bounded support condition on \( \tilde{x}_i \) is only made to simplify arguments in the proofs and be relaxed to this subvector having finite fourth moments.

- Assumption RC3 is mainly useful to ensure point identification of the sieve coefficients.\(^{14}\)

\(^{14}\)It may not be necessary to consistently estimate the regression coefficients and choice probability considered here. I am grateful to a referee for pointing this out.
The above conditions are sufficient to establish consistency of the estimator of the regression coefficients. The proof is omitted as it follows from virtually identical arguments as in Chen et al. (2003).

**Theorem 3.1** Under assumptions RC1-RC4, if \( \kappa_n \to \infty \) and \( \kappa_n/n \to 0 \), we have

\[
\hat{\beta} - \beta_0 = o_p(1) \tag{3.3}
\]

While the above result is an important first step, as mentioned in the introduction, there are several estimators for the model considered here for which the regression coefficients can be estimated consistently. The motivation for SNLLS estimator proposed here was also to consistently estimate the choice probability function, which we now turn attention to. Before doing this, we will introduce the Fisher norm. This norm will prove useful on many fronts. For one, a convergence result for this norm will directly tied to a convergence rate for the probability function, as this function is a functional satisfying a Lipschitz condition. Second, the asymptotic distribution of the regression coefficient estimator, which we will also establish shortly, will be related to features of this norm.

We define the Fisher norm as follows: on \( \mathcal{A} \), and denote it by \( \| \cdot \|_F \). For \( \alpha_1 = (\theta_1, \ell_1) \) and \( \alpha_2 = (\theta_2, \ell_2) \) we define

\[
\| \alpha_1 - \alpha_2 \|_F^2 \equiv E[\phi_0^2 \varphi_0^2 (\tilde{x}'_i (\theta_1 - \theta_2) - (x'_i \beta_0) (\ell_1 - \ell_2))^\prime (\tilde{x}'_i (\theta_1 - \theta_2) - (x'_i \beta_0) (\ell_1 - \ell_2))]
\]

Our rate result\(^{15}\) is stated in the following theorem, whose proof is omitted as it can be shown using similar arguments to those used in Ai and Chen (2003) and Chen et al. (2005). Before stating the theorem, we will impose the following locally quadratic condition on the objective function we adopted:

**RC5** There exist positive constants \( c_1, c_2 \) such that

\[
c_1 E[(y_i - \Phi(x'_i \beta \exp(\ell(x_i))))^2] \leq \| \alpha - \alpha_0 \|_F^2 \leq c_2 E[(y_i - \Phi(x'_i \beta \exp(\ell(x_i))))^2] \tag{3.4}
\]

\(^{15}\)This particular rate result is with respect to the Fisher norm, which, as we will see shortly, will provide us rates for the choice probability functional as well. We then will derive a rate result and distribution theory for the estimator of \( \beta_0 \).
Theorem 3.2 Suppose assumptions RC1–RC5 hold, but with the added conditions \( p > k/2 \) and \( w_2 > p \). Then

\[
\|\hat{\alpha} - \alpha_0\|_F = O_p\left(\sqrt{\frac{\kappa_n}{n}} + \kappa_n^{-p/k}\right) \tag{3.5}
\]

Remark 3.3 Assumption RC5 imposes that the population criterion function can be approximated locally by a quadratic function, effectively assuming that the remainder term in a mean value expansion get small as the parameter \( \alpha \) approaches \( \alpha_0 \). This condition will also be used when deriving the limiting distribution theory for \( \hat{\theta} \).

Remark 3.4 We note the above rate coincides with the attained in Newey(1997) for estimating a regression function using series estimation. From our conditions it will also imply the same rate of convergence for choice probability function estimator\(^{16}\):

\[
E[(\Phi(x'_i\beta_0 \exp(\ell(x_i))) - \Phi(x'_i\hat{\beta} \exp(\ell(x_i))))^2] \tag{3.6}
\]

We next turn attention to the limiting distribution theory of the estimator \( \hat{\theta} \). For this we require the additional assumptions:

AD1 \( \beta_0 \in \text{int} \mathcal{B} \).

AD2 Reparameterizing the function \( g_0(x_i) \equiv \sigma_0(x_i)^{-1} \) as \( \tilde{g}_0(z_{0i}, \tilde{x}_i) \) with \( z_{0i} \equiv x'_i\beta_0 \), the matrix

\[
Q = E[\phi(0)^2\tilde{x}_i\tilde{x}'_i\tilde{g}_0(0, \tilde{x}_i)^2f_{Z|\tilde{X}}(0|\tilde{x})]
\]

is non-singular where \( \phi(\cdot) \) is the standard normal density function and \( f_{Z|\tilde{X}}(\cdot|\cdot) \) denotes the conditional density of \( z_{0i} \equiv x'_i\beta_0 \) given \( \tilde{x}_i \).

AD3 \( f_{Z|\tilde{X}}(z_0|\tilde{x}) \) is continuously differentiable \( z_0 \) in a neighborhood of 0 and all \( \tilde{x} \).

The main theorem establishes a linear representation for the sieve estimator. The linear representation exposes the bias and variance of the estimator as a function of the number

\(^{16}\)As discussed in Ai and Chen(2003), attaining \( L_2 \) rates generally requires stronger conditions than those needed for rates with respect to the Fisher norm. In the current setting, Assumption RC5 is what enables us to get the same rate under both norms.
of basis functions in the sieve, $\kappa_n$, and can be used to derive the rate at which $\kappa_n \to \infty$ in order for $\hat{\beta}$ to converge to $\beta_0$ at the fastest rate in terms of MSE. The linear representation below requires the introduction of some new notation. Let $\phi_0, g_0$ denote $\phi(x_i^t\beta_0, g_0(x_i))$ and $g_0(x_i)$ respectively. Let $t$ be a $(k - 1) \times 1$ non-zero vector and let $w_{ni}^*$ be a $(k - 1) \times 1$ vector, which satisfies

$$w_{ni}^* = \arg \inf_{w_{ni}:\theta_{ni} \in A_n, \theta \in \Theta} E[\phi_{0i}^2 g_{0i}^2 (\tilde{x}_i + (x_i^t \beta_0) w_{ni})' (\tilde{x}_i + (x_i^t \beta_0) w_{ni})]$$

and let $\tilde{x}_n$ denote $x_i^t \beta_0 w_{ni}^*$.

Let $\ell_n(x_i)$ satisfy $(\theta_0, \ell_n(x_i)) \in A_n$ and minimize:

$$\|\alpha_0 - \alpha_{0n}\|_F$$

where $\alpha_0 = (\beta_0, \ell_0(x_i))$ and $\alpha_{0n} = (\beta_0, \ell_n(x_i))$.

**Theorem 3.3** Under assumptions RC1-RC5, AD1-AD3, if $\kappa_n^{-1/k} \to 0$ and $n \kappa_n^{-1/k} \to \infty$, then

$$\hat{\theta} - \theta_0 = \hat{c}(p)^{-1} Q^{-1} \kappa_n^{1/k} \frac{1}{n} \sum_{i=1}^n \phi(x_i^t \beta_n \tilde{g}_n(x_i)) \tilde{g}_n(x_i) (\tilde{x}_i - \tilde{x}_n) (y_i - \Phi(x_i^t \beta_n \tilde{g}_n(x_i)))$$

$$+ o_p \left( \sqrt{n^{-1} \kappa_n^{1/k}} \right) + o_p(n^{-p/k})$$

(3.7)

where $\beta_n$ denotes a sequence of values in between $\hat{\beta}$ and $\beta_0$, $\tilde{g}_n(x_i)$ denotes analogous intermediate values for the scale function, and where $\hat{c}(p)$ is a constant depending on the assumed order of smoothness $p$, and whose expression can be found in (A.8).

**Remark 3.5** From the linear representation in the theorem, we can see that the bias is of order $\kappa_n^{-p/k}$, the rate at which the functions $\Phi(x_i^t \beta_0 g_0(x_i))$ can be approximated well (with respect to our weighted norm) by our basis function approximation (see, e.g. Chen et al.(2005)). The variance is of order $\kappa_n^{1/k}/n$, the rate we are dividing the summation by. Equating the two to derive the optimal rate at which the sequence $\kappa_n$ increases, we get

---

\[\text{Details on how these rates are derived can be found in the derivation of (A.12)}\]
\[ \kappa_n = O(n^{k/(2p+1)}) \]. This implies the rate of convergence of the MSE of \( \hat{\theta} \) is \( O(n^{-p/(2p+1)}) \) which is slower than the parametric (root-\( n \)) rate. Chen and Khan (2003) show that the parametric rate is not achievable for a similar model, and it is conjectured that the MSE rate attained here is the optimal rate of convergence for the model under Assumptions RC1-RC5, AD1-AD3. A formal proof of this is left for future work.

An immediate corollary to the above theorem is the limiting distribution theory for the sieve NLLS estimator:

**Corollary 3.1** Consider the sequence \( \kappa_n = O(n^{k+\epsilon_3/2p+k}) \) where \( \epsilon_3 > 0 \) is an arbitrarily small constant, and \( p > k/2 \). It follows that:

\[
\sqrt{n\kappa_n^{-1/k}}(\hat{\theta} - \theta_0) \Rightarrow N(0, \frac{1}{4} \tilde{c}(p)^{-1}Q^{-1})
\] (3.8)

We conclude this section with some comments the form of the limiting distribution.

**Remark 3.6** Recall the estimator was motivated by the fact that the heteroskedastic probit model probabilities could be equated to the probabilities in a distribution free model by setting

\[
\tilde{g}_0(z_{0i}, \tilde{x}_i) = \frac{\Phi^{-1}(\tilde{P}_0(z_{0i}, \tilde{x}_i))}{z_{0i}}
\]

where recall \( z_{0i} = x_i' \beta_0 \) and \( \tilde{P}_0(\cdot, \cdot) \) denotes the probability function reparameterized as a function of two arguments. Taking limits as \( z_{0i} \to 0 \) (keeping \( \tilde{x}_i \) fixed) we get

\[
\tilde{g}_0(0, \tilde{x}_i) = \frac{1}{\phi(0)} \tilde{P}_1(0, \tilde{x}_i)
\]

where here \( \tilde{P}_1(\cdot, \cdot) \) denotes the partial derivative of \( \tilde{P}_0(\cdot, \cdot) \) with respect to its first argument. This we see

\[
Q = E[\tilde{x}_i \tilde{x}'_i \tilde{P}_1(0, \tilde{x}_i)^2 f_Z(z)(0|\tilde{x})]
\] (3.9)

and we note the form of \( Q \) is independent of the fact that the normal c.d.f. was used in the objective function.
Remark 3.7 We note the variance covariance matrix is not of a sandwich form. While this feature usually occurs for MLE estimators it is a feature of the sieve NLLS estimator because all the information for $\beta_0$ is at $x_i'\beta_0 = 0$. This causes the usual sandwich form found in NLLS estimators to “collapse”, since theouterscore term, which has the term $\text{Var}(y_i|x_i)$, is now equal to the constant $\frac{1}{4}$. This makes the outerscore term proportional to the hessian term, causing the “collapse”.

We conclude this section by illustrating a further advantage of the proposed estimation procedure. In addition to estimating the structural parameters $\beta_0$, the sieve approach also permits estimation of other functionals of the probability function. One relevant functional is the (weighted) average marginal effect, which we define here as:

$$ W = \int w_W(x)\partial P_0(x)/\partial x dx $$

(3.10)

where recall $P_0(\cdot)$ denotes the choice probability function and $w_W(\cdot)$ denotes a weighting function (assumed here to have compact support) satisfying $w_W(x) \geq 0$ and

$$ \int w_W(x)dx = 1 $$

(3.11)

Letting $\hat{W}$ denote the estimator obtained by replacing $P_0$ with our proposed sieve estimator of the choice probability in (3.10). The following theorem establishes the limiting distribution theory of this estimator. Its proof is omitted as it follows virtually identical arguments as used in the proving the previous theorems.

Theorem 3.4 Under the conditions imposed in Theorem 3.2, if $\sqrt{n}\kappa_n^{-p/k} \to 0$, then

$$ \sqrt{n}(\hat{W} - W) \Rightarrow N(0, V_W) $$

(3.12)

where

$$ V_W = E_X[v_W(x_i)v_W(x_i)'P_0(x_i)(1 - P_0(x_i))] $$

(3.13)

with

$$ v_W(x_i) = -f_X(x_i)^{-1}\partial w_W(x_i)/\partial x_i $$

(3.14)
Remark 3.8 We note that this limiting distribution corresponds to that obtained in Theorem 3 in Newey(1997), who estimated the probability function by a series regression and did not attain an estimator of $\beta_0$. This result agrees with the general conclusion in Shen(1997) which is that the two main conditions affecting the limiting distribution of a smooth functional are the rate of convergence of the sieve estimator and the smoothness of the functional. Since the rate of convergence attained in Theorem 3.2 aligns with Theorem 1 in Newey(1997), one would then expect the limiting distributions of the same smooth functional to coincide.

4 Local NLLS Estimators

This section proposes a procedure which again relates median based semiparametric estimators for binary choice models to standard estimation procedures for parametric binary choice models. Like the previous proposed estimator, the estimator optimizes a NLLS parametric objective function. It differs in the sense that it does not estimate choice probabilities like the previous procedure, but it has the advantage of being implementable in standard software packages such as Stata.

The estimators we propose involve combining the maximum score and smoothed maximum score objective functions in (1.3) and (1.4) respectively. First we note that the objective function of the maximum score estimator:

$$\frac{1}{n} \sum_{i=1}^{n} |y_i - I[x_i'\beta \geq 0]|$$

is identical to the squared loss objective function

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

since both $y_i$ and $I[\cdot]$ are 0-1 variables. Next we smooth this objective function as was done in (1.4), by replacing the indicator function with a kernel function. For the smoothed maximum score estimator, the kernel function serves to approximate a c.d.f. We do the same here, using the c.d.f. of the standard normal distribution\(^{18}\) which as before we denote by $\Phi(\cdot)$, and whose p.d.f we denote by $\phi(\cdot)$.

\(^{18}\)Actually, the c.d.f. 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 c.d.f. since its values can be easily computed using standard software packages.
To formally define the estimator, we let $h_n$ denote a sequence of positive numbers, decreasing to 0 with the sample size. ($h_n$ can be viewed as a bandwidth sequence found in nonparametric kernel estimation). We adopt the usual scale normalization in semiparametric models (e.g. Horowitz(1992)), where we set the coefficient on the $k^{th}$ regressor to be 1, and consider estimation of $\theta_0$, where $\beta_0 = (\theta_0', 1)'$. Our NLLS estimator $\hat{\beta} = (\hat{\theta}', 1)'$ is defined as

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

(4.3)

The main advantage of this procedure is that it involves the standard NLLS objective function. In fact, it is the standard NLLS Probit estimator used to estimate parametric binary choice models. Thus standard software packages, such as Stata, can be used to compute\footnote{For example, in Stata, the \texttt{nl} command fits an arbitrary nonlinear function by least squares. The Probit regression function can be constructed using Stata’s \texttt{norm()} command, which returns cumulative probabilities from the standard normal distribution.} the estimator of $\theta_0$.

Regarding asymptotic properties of this estimator, we impose conditions that are identical to those in Horowitz(1992).

A1' $\theta_0$ is in the interior of a compact set $\Theta$.

A2' The vector $\tilde{x}_i$ has bounded support.

A3' The density function of $x'_i \beta_0$ conditional on $\tilde{x}_i$, denoted by $f_Z|\tilde{X}(\cdot)$ is positive and continuously differentiable with bounded derivative.

A4' The conditional probability function of $y_i$, expressed as a function of $\tilde{x}_i$ and $x'_i \beta_0$, 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 $\tilde{x}_i$.

A5' The matrix

$$
Q_H = E[\tilde{P}_t(0, \tilde{x}_i)\tilde{x}_i\tilde{x}_i f_Z|\tilde{X}(0|\tilde{x}_i)]
$$

(4.4)

is nonsingular, where $\tilde{P}(x'_i \beta_0, \tilde{x}_i)$ denotes the conditional probability of $y_i = 1$ given $x_i$, which we reparamaterized as a function of $\tilde{x}_i, x'_i \beta_0$, and $\tilde{P}(\cdot, \cdot)$ denotes the partial derivative of $\tilde{P}(\cdot, \cdot)$ with respect to its first argument.
The following theorem characterizes the estimators rate of convergence and limiting distribution as a function of \( h_n \). The proof of the theorem is omitted as it follows from arguments that are similar to those used in Horowitz(1992).

**Theorem 4.1** Assume CM1,CM2, A1'-A5' hold and \( h_n \to 0 \), then,

1. if \( n h_n^3 \to \infty \) then \( h_n^{-1}(\hat{\theta} - \theta_0) \xrightarrow{p} \kappa \) where \( \kappa \) is a \( k \) dimensional vector of constants.

2. At the rate \( h_n = O(n^{-1/3}) \) then \( n^{1/3}(\hat{\theta} - \theta_0) \xrightarrow{d} B \) where the random vector \( B \) has non-standard (i.e. non-Gaussian) distribution.

As the above theorem indicates, the local NLLS estimator has asymptotic properties that are similar to the maximum score estimator proposed in Manski(1975,1985). Specifically, its rate of convergence can be as fast as \( O(n^{-1/3}) \), the same rate of the maximum score estimator, and it has a non-Gaussian limiting distribution\(^{20}\).

However, the slow rate of convergence (relative to the smoothed maximum score estimator in Horowitz(1992)) is due to a bias condition, where the bias of the estimator converges at the rate of \( h_n \), which is in contrast to the rate of \( h_n^2 \) for the smoothed maximum score estimator. Thus the different rates of convergence for the two estimators (NLLS and SMS) is loosely analogous to differing rates of convergence for one-sided and two-sided kernel estimators in nonparametric density and regression estimation. Fortunately, this bias condition in NLLS is easily correctible. For example, an alternative kernel function to the normal c.d.f. could be used to reduce the order of the bias, or other bias reducing mechanisms, such as jackknifing could be implemented, to achieve the same rate as SMS, as well as an asymptotic normal distribution. The asymptotic properties of such approaches is left for future work.

## 5 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. We begin by considering the designs used in Horowitz(1992). These are based on the model:

\[
y_i = I[x_{1i} + \beta_0 x_{2i} - u_i \geq 0]
\]

\(^{20}\)For the 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. See, for example Section 9.6 in van der Vaart(1998).
\[ \beta_0 = 1, \quad x_{1i} \sim N(0,1) \text{ and } x_{2i} \sim N(1,1). \]

There are 4 designs corresponding to 4 different distributions of \( u_i \). They are:

1. \( u_i \sim \text{logistic, median 0, variance 1.} \)
2. \( u_i \sim \text{uniform, median 0, variance 1.} \)
3. \( u_i \sim \text{Student’s t with 3 degrees of freedom, normalized to have variance 1.} \)
4. \( u_i = 0.25 \times (1 + 2z_{0i}^2 + z_{0i}^4)\nu_i \) where \( z_{0i} = x_{1i} + x_{2i} \) and \( \nu_i \sim \text{logistic with median 0 and variance 1.} \)

The estimators studied in the study are the sieve NLLS (SNLLS), the sieve MLE (SMLE), maximum score (MS), the smoothed maximum score (SMS), the proposed local NLLS estimator (LNLLS) and its jackknifed version (JKNLLS). To implement the estimators for SMS the feasible optimal bandwidth sequence introduced in Horowitz(1992) was used. For the sieve estimators a series was used in the expansion of the log scale function with a polynomial of degree 1 for \( n = 250 \) and 2 otherwise.\(^{21}\) For the LNLLS a bandwidth sequence of \( n^{-1/3} \) was used. For JKNLLS, the weights used were 4/3 and -1/3, and the bandwidths were \( c_1 n^{-1/5}, c_2 n^{-1/3} \) with constants 1/4 and 1.

Tables I-IV report the mean bias and MSE for each of the estimators for \( n = 250, 500, 1000 \) with 1000 replications. The MS and SMS results reported are those found in Horowitz(1992). The sieve estimators were computed using the Nelder-Mead simplex algorithm\(^{22}\), with 15 randomly generated starting values\(^{23}\). The sieve estimators generally perform better than SMS across designs with the exception of Design 3 where results are very similar. The SMLE and SNLLS perform quite similarly, also in accordance with the theory, as the MSE for the SMLE is not smaller than the SNLLS. The local NLLS estimators also perform quite well. One surprise in the simulation results is that in terms of RMSE, for some designs, the standard NLLS performs as well as, if not better than the other estimators despite its slower rate

\(^{21}\)Precisely, to estimate with a polynomial of order 2 the scale function was approximated with \( \exp(\Pi_0 + \Pi_1 x_{1i} + \Pi_2 x_{2i} + \Pi_3 x_{1i}^2 + \Pi_4 x_{2i}^2 + \Pi_5 x_{1i} x_{2i}). \) Results for similar orders were experimented with but did not change results much, and are not reported.

\(^{22}\)As mentioned previously, since the objective function is smooth in the parameters, more standard, gradient based algorithms may be used. They were not adopted here to avoid potential instability problems associated with near singularity of hessian matrices, and also because the relatively low dimensionality of designs permit the Nelder-Mead algorithm to be computationally feasible.

\(^{23}\)The simulation was performed in GAUSS, on a Pentium 4, 2.80 GhZ PC. Computation time was roughly 16 seconds per replication for the 1000 observation designs.
of convergence. The jackknife procedure generally results in a lower bias than the LNLLS, but it appears this sometimes comes at the expense of a larger variance.

As mentioned in the paper an advantage of the sieve NLLS and the sieve MLE is that they simultaneously estimate choice probabilities as well as regression coefficients. Figures I-IV plot the mean value of the estimated choice probabilities using SNLLS on a grid of 2500 regressor values for each of the 4 designs, for sample sizes of \( n = 250, 500, 1000 \), again using 1000 replications. Also reported in parentheses are the values of the average mean square errors (AMSE) which averages MSE across the points on the grid. As the results indicate the SNLLS does and adequate job of estimating choice probabilities, and the values of the AMSE go down with the sample size. The estimator performs the worst in the heteroskedastic design, both in terms of the level of the AMSE, and the rate at which it decreases with the sample size.

As a final component of our simulation study, we explore how each of the estimators perform in a higher dimensional, more complicated design. Specifically, we allow for more covariates and a form of heteroskedasticity that is not a function of the index \( x_i' \beta_0 \), but a more general form of the covariates. The following model was simulated:

\[
y_i = I[x_{1i} + \beta_0^{(1)} x_{2i} + \beta_0^{(2)} x_{3i} + \beta_0^{(3)} x_{4i} - u_i \geq 0]
\]

where here \( \beta_0^{(1)} = \beta_0^{(2)} = \beta_0^{(3)} = 1 \), \( x_{1i} \sim N(0,1), x_{2i} \sim N(1,1), x_{3i} \sim \chi^2_1, x_{4i} \sim N(0,1) \). The heteroskedastic error term was distributed logistically, with scale function \( exp(|x_{2i}| \ast x_{3i}) \).

Table V reports results for the estimator of \( \beta_0^{(1)} \) for the same 4 estimators for the same sample sizes and number of replications. For implementation, for the sieve estimator we increased the order of polynomial by 1 for each sample to account for the fact there are more regressors. To implement the SMS, we used the fourth-order kernel function described in Horowitz(1992), and at first used the “plug-in method” described in Horowitz(1992) to select the smoothing parameter. However this resulted in unstable results for \( n = 250 \), so we implemented an extra “iteration” in the plug in strategy. That is, we implemented the plug-in method to get an initial estimator of the regression coefficients which we used to estimate the constant in the smoothing parameter. This led to improved results for \( n = 250 \).

As the results in the table indicate, all estimators perform reasonable well. The SMLE is smaller than the SNLLS for \( n = 250, 500 \) but the reverse is true for \( n = 1000 \), providing further evidence that neither estimator is more efficient. The SMS exhibits large values of MSE for \( n = 250 \) but stabilizes afterwards. Nonetheless, even for \( n = 1000 \) it has a larger MSE than either sieve estimator. MS exhibits the largest bias and MSE except for \( n = 250 \),
when its MSE is smaller than SMS, though larger than the sieve estimators. The NLLS and JKNLLS estimators perform well in this design as well, but not as well as the sieve estimators for large sample sizes. The results for this design are encouraging for the sieve estimators, demonstrating they do not suffer any more in higher dimensional designs than existing estimators.

6 Conclusions

In this paper new estimation procedures for a distribution free heteroskedastic binary response model were proposed. The sieve estimators enable joint estimation of the regression coefficients, choice probabilities. The proposes local NLLS estimators estimated only regression coefficients but had the advantage of being very simple to implement with standard software packages. A simulation study indicates these estimators perform adequately well in finite samples.

The work here suggests areas for future research. Limiting distribution theory for the choice probability, and marginal effects estimators, as well as smooth functionals thereof, needs to be derived. Also it would also be useful to explore if further restrictions on the model, by constraining the behavior of $\sigma_0(x_i)$, would enable improving upon the optimal rates attained here and in Horowitz(1993). Such further restrictions would be relatively easy to impose using the sieve estimation approach adopted here.

References


A Appendix

A.1 Proof of Theorem 3.3

Before we derive the linear representation for the estimator \( \hat{\theta} \), recall we defined the Fisher norm, denoted by \( \| \cdot \|_F \), as

\[
\| \alpha_1 - \alpha_2 \|^2_F \equiv E[\phi_0^2(x'_i g_0(x_i)) (\theta_1 - \theta_2) - (x'_i \beta_0) (\ell_1 - \ell_2)]^2
\]

(A.1)

where \( \phi_0, g_0, \ell_1, \ell_2 \) denote

\( \phi(x'_i \beta_0 g_0(x_i)), g_0(x_i), \ell_1(x_i), \ell_2(x_i) \) respectively. As we will see, deriving the form of the linear representation will rely heavily on convergence of certain terms with respect to this norm. We note that similar arguments as used in, e.g. Ai and Chen(2002), can be used to conclude that

\[
\| \hat{\alpha} - \alpha_0 \|_F = O_p \left( \sqrt{\frac{r}{n} + \kappa_n^{-p/k}} \right)
\]

(A.2)

To establish the limiting distribution theory of the estimator we note there are many results in the literature for the asymptotic theory for smooth functionals—see, e.g. Shen(1997), Chen and
Shen(1998), Ai and Chen(2003) and Chen et al.(2005). However, these results apply only to the root-$n$ case, which is not possible here.\textsuperscript{24}

Our proof strategy is to follow the arguments used in Ai and Chen(2003) Chen et al.(2005), but make the necessary modifications to account for the fact that the estimator does not converge at the parametric (root-$n$) rate. In the rest of this section we will “scalarize” the problem by deriving the linear representation for \( t'(\hat{\theta} - \theta_0) \) where \( t \) is a \((k-1) \times 1\) non zero vector.

Following Ai and Chen(2003) we wish to find the \((k-1) \times 1\) vector \( w^*_{ni} \) which minimizes:

\[
\inf_{w_{ni}} t' E [\phi_0^2 g_0^2 (\tilde{x}_i + z_{0i} w_{ni}^*)(\tilde{x}_i + z_{0i} w_{ni}^*)'] t
\]

and satisfies \((\theta, w^*_{ni}) \in A_n\) for each \( \theta \in \Theta \). Clearly, the above expectation can be set to 0 by setting \( w^*_{ni} = I[z_{0i} \neq 0](-\tilde{x}_i/z_{0i}) \), as \( z_{0i} \) is continuously distributed around 0. The fact that we can make this expectation as small as possible relates to the impossibility of attaining the root-$n$ rate for \( \hat{\theta} \). What will determine the rate of convergence of the estimator is the rate of convergence of the above expectation to 0 when we replace \( w^*_{ni} \) with \( w_{ni} \) where \((\theta, w_{ni}) \in A_n\). So we will aim to find

\[
\inf_{w_{ni}} t' E [\phi_0^2 g_0^2 (\tilde{x}_i + z_{0i} w_{ni})(\tilde{x}_i + z_{0i} w_{ni})'] t
\]

\[(A.3)\]

as a function of \( \kappa_n \).

Since \( P(z_{0i} = 0) = 0 \), we can reparameterize the above expected value that is to be minimized as:

\[
t' E [I[z_{0i} \neq 0] z_{0i}^2 \phi_0^2 g_0^2 (\tilde{x}_i/z_{0i} + w_{ni})(\tilde{x}_i/z_{0i} + w_{ni})'] t
\]

\[(A.4)\]

One difficulty in deriving the rate at which the above expectation goes to 0 when \( w_{ni} \) is based on an approximation using basis functions is that the function to be approximated, \( \tilde{x}_i/z_{0i} \) is unbounded when \( z_{0i} \) approaches 0. To derive this rate we add and subtract a polynomial spline function, denoted by \( s_{ni} \) which satisfies 3 properties 1) it is exactly equal to \( \tilde{x}_i/z_{0i} \) for \( |z_{0i}| > h_n \) 2) on the interval \( |z_{0i}| \leq h_n \) is a polynomial of degree \( p \) whose \( p-1 \) derivatives are equal to those of \( \tilde{x}_i/z_{0i} \) at the points \( z_{0i} = \pm h_n \) 3) it satisfies the smoothness properties to be approximated by standard basis functions. For a simple illustrative example where we only impose continuity, define \( s_{ni} \) as

\[
s_{ni} = \tilde{x}_i/z_{0i} \quad \text{if} \quad |z_{0i}| > h_n
\]

\[
= \tilde{x}_i(a + bz_{0i}) \quad \text{if} \quad |z_{0i}| \leq h_n
\]

where the constants \( a, b \) solve \( s_{ni} = \tilde{x}_i/z_{0i} \) at \( z_{0i} = h_n, z_{0i} = -h_n \). We see that this corresponds to \( a = 0, b = h_n^{-2} \). The figure below illustrates what this spline function looks like:

\textsuperscript{24}See Chen and Khan(2003) for a related impossibility result. A result on upper bounds on achievable rates is available from the author.
Note we can easily derive the rate of

\[ t' E[I[z_{0i} \neq 0] z_{0i}^2 \phi_{0i}^2 g_{0i}^2 (\tilde{x}_i / z_{0i} - s_{ni})(\tilde{x}_i / z_{0i} - s_{ni})'] t \]  

(A.6)
as a function of \( h_n \). We note the above expectation is

\[ t' \left[ \int_{-h_n}^{h_n} I[z_{0i} \neq 0] z_{0i}^2 \phi_{0i}^2 g_{0i}^2 (\tilde{x}_i / z_{0i} - s_{ni})(\tilde{x}_i / z_{0i} - s_{ni})' f_{Z|X}(z_{0i}|\tilde{x}_i) dz_{0i} \right] dF_{\tilde{x}_i} t \]  

(A.7)

where \( \tilde{\phi}_{0i} \) denotes \( \phi(\tilde{z}_{0i} \tilde{g}(\tilde{z}_{0i}, \tilde{x}_i)) \) and \( \tilde{g}_{0i} \) denote \( \tilde{g}(\tilde{z}_{0i}, \tilde{x}_i) \). Working with the inside integral, we make the change of variables \( u = z_{0i}/h_n \), and let \( \tilde{\phi}_{ni}, \tilde{g}_{ni} \) denote the values obtained by replacing \( z_{0i} \) with \( uh_n \) in \( \tilde{\phi}_{0i}, \tilde{g}_{0i} \). We expand

\[ \tilde{\phi}_{ni}^2 \tilde{g}_{ni}^2 f_{Z|X}(uh_n|\tilde{x}_i) \]  

around its value at \( uh_n = 0 \) (this is permitted by the assumed smoothness of the conditional density and probability functions in a neighborhood of 0). Let \( \tilde{c}(p) \) denote a constant which depends on the coefficients in the polynomial for \( s_{ni} \) defined on \( z_{0i} \in [-h_n, h_n] \). (For the above simple example \( \tilde{c}(p) = 16/15. \) For the general case the expression is of the form:

\[ \tilde{c}(p) = \int_{-1}^{1} (1 - us_p(u))^2 du \]  

(A.8)

where \( s_p(u) = \sum_{j=0}^{p} a_j u^j \) denotes a polynomial whose coefficients \( \{a_j\} \) satisfy that each of the \( p - 1 \) derivatives of \( s_p(u) \) equate with each of the \( p - 1 \) derivatives of the function \( u^{-1} \) at the values \( u = 1 \) and \( u = -1 \).

We get the above integral is of the form

\[ t' \tilde{c}(p) h_n E[\phi(0)^2 \tilde{g}_0(0, \tilde{x}_i)^2 \tilde{x}_i f_{Z|X}(0|\tilde{x}_i)] t + o(h_n) \]  

(A.9)

We next derive the rate of

\[ t' E[I[|z_{0i}| > h_n] z_{0i}^2 \phi_{0i}^2 g_{0i}^2 (w_{ni} - s_{ni})(w_{ni} - s_{ni})'] t \]  

(A.10)
as a function of \( \kappa_n \) and \( h_n \). We note here we may apply standard theorems for approximating functions with basis functions. For example, with polynomials, by Jackson’s theorem (see, e.g.
where \( \pi \) the previous rate to solve for combining our results we may conclude that and we note that rate does not depend on \( \pi \) (though \( p \) does effect the constant term). Furthermore, by combining our results we may conclude that

\[
\kappa_n^{-1/k} t' E[\phi(y, g_0)(\tilde{x}_i + w_{ni})'] t \to \tilde{c}(p) t' E[\phi(0)^2 g_0(0, \tilde{x}_i)^2 \tilde{x}_i' f Z| (0|\tilde{x}_i)] t \quad (A.12)
\]

Note we can use the same arguments (i.e. adding and subtracting a polynomial spline) to derive the rate for

\[
E[\phi(y, g_0)(\tilde{x}_i + w_{ni})' (\Phi(z_0, g_0 - \Phi(x'_i, \beta_n, g_n(x_i)))]
\]

This is how one attains the rate for the bias in (3.7).

Continuing with deriving the asymptotic distribution theory of \( \hat{\theta} \), we continue to follow the strategies in Ai and Chen(2003) and Chen et al.(2005) and make the necessary modifications to account for the slower rate of convergence.

Our proof will be based on three lemmas, which will serve in taking an expansion around the first order condition our estimator solves. Before describing them, we introduce some new notation, chosen deliberately to match the notation used in Ai and Chen(2003). Recall \( \| \cdot \|_F \) denotes the Fisher norm and \( \| \cdot \|_E \) denote the Euclidean norm. Let

\[
N_{0n} = \{ \alpha : \alpha \in A_n, \| \alpha - \alpha_0 \|_F = O(\| \alpha - \alpha_0 \|_F) \}
\]

Let \( Q_n \) denote the \((k-1) \times (k-1)\) matrix

\[
E[\phi(y, g_0)(\tilde{x}_i + z_0 w_{ni})' (\tilde{x}_i + z_0 w_{ni})']
\]

and define \( v_{ni}^* = (v_{ni0}, v_{niq}) \) where

\[
\begin{align*}
v_{ni0}^* &= Q_n^{-1} t \\
v_{niq}^* &= w_{ni}^* v_{ni0}^*
\end{align*}
\]

and we note that \( \| v_{ni}^* \|_F^2 = t' Q_n^{-1} t \). Let \( \delta_{1n} = \| v_{ni}^* \|_F^{-2} = O(\kappa_n^{-1/k}) \) and let \( \delta_{2n} = \frac{1}{\sqrt{\pi n \delta_{1n}}}. \)

Let \( m(y_i, x_i, \alpha) = y_i - \Phi(x_i' \beta \exp(\ell(x_i))) \) where \( \alpha = (\beta, \ell(x_i)) \). Let \( m_n(y_i, x_i, \pi_n \alpha) \) where \( \pi_n \alpha \) denotes a projection of \( \alpha \) onto \( A_n \).
Finally, \( \langle \cdot, \cdot \rangle \) denotes the inner product with respect to the norm \( \| \cdot \|_F \). That is, for
\[
\alpha_1 = (\beta_1, \ell_1(x_i)), \quad \alpha_2 = (\beta_2, \ell_2(x_i)),
\]
\[
\langle \alpha_1, \alpha_2 \rangle = E[\phi_0^2 \phi_0^2 (\tilde{x}'_i \theta_1 + z_{0i} \ell_{1i}) \cdot (\tilde{x}'_i \theta_2 + z_{0i} \ell_{2i})]
\]  
(A.16)

where \( z_{0i} = x'_i \beta_0 \); note \( \langle \alpha_1, \alpha_1 \rangle = \| \alpha_1 \|_F^2 \).

Lastly, we will introduce our notation for pathwise directional derivatives. The derivative of
\( m(y_i, x_i, \alpha) \) evaluated at \( \alpha_0 \) in the direction \( \alpha - \alpha_0 \) is defined as:
\[
m_{\alpha_0}(y_i, x_i, \alpha) [\alpha - \alpha_0] = \lim_{\tau \to 0} \left\{ \frac{m(y_i, x_i, \alpha(\tau)) - m(y_i, x_i, \alpha_0)}{\tau} \right\}
\]  
(A.17)

Having introduced the required notation, we can turn attention to the proof of deriving the
linear representation. Our strategy will be to expand around the first order condition that the
sieve estimator solves, analogous to the steps involved for deriving the linear representation for an
estimator of a parametric model.

Thus we will establish and prove 3 lemmas, which effectively serve to deal with equicontinuity
conditions when expanding the first order condition.

**Lemma A.1** Uniformly over \( \tilde{\alpha} \in \mathcal{N}_{0n} \),
\[
\frac{1}{n} \sum_{i=1}^{n} \frac{dm_n(y_i, x_i, \tilde{\alpha})}{d\alpha} [v_{ni}^*] m_n(y_i, x_i, \tilde{\alpha}) = \frac{1}{n} \sum_{i=1}^{n} \frac{dm(y_i, x_i, \alpha_0)}{d\alpha} [v_{ni}^*] m_n(y_i, x_i, \tilde{\alpha}) + o_p(\delta_{2n})
\]  
(A.18)

**Proof:** We note that
\[
\frac{1}{n} \sum_{i=1}^{n} \left( \frac{dm_n(y_i, x_i, \tilde{\alpha})}{d\alpha} [v_{ni}^*] - \frac{dm(y_i, x_i, \alpha_0)}{d\alpha} [v_{ni}^*] \right) m_n(y_i, x_i, \tilde{\alpha}) =
\]
\[
\frac{1}{n} \sum_{i=1}^{n} \left( \frac{dm_n(y_i, x_i, \tilde{\alpha})}{d\alpha} [v_{ni}^*] - \frac{dm(y_i, x_i, \tilde{\alpha})}{d\alpha} [v_{ni}^*] \right) m_n(y_i, x_i, \tilde{\alpha})
\]  
(A.19)

\[
+ \frac{1}{n} \sum_{i=1}^{n} \left( \frac{dm(y_i, x_i, \alpha_0)}{d\alpha} [v_{ni}^*] - \frac{dm(y_i, x_i, \alpha_0)}{d\alpha} [v_{ni}^*] \right) m_n(y_i, x_i, \tilde{\alpha})
\]  
(A.20)

We shall show each of these two terms is asymptotically negligible- i.e. \( o_p(\delta_{2n}) \).

For (A.19) we add and subtract the expectation of the term inside the summation. We note
the class of functions
\[
\{ \tilde{\alpha} : \left( \frac{dm_n(y_i, x_i, \tilde{\alpha})}{d\alpha} [v_{ni}^*] - \frac{dm(y_i, x_i, \tilde{\alpha})}{d\alpha} [v_{ni}^*] \right) m_n(y_i, x_i, \tilde{\alpha}) \}
\]  
(A.21)
is a Donsker class by the arguments in the proof of Corollary A.1(i) in Ai and Chen (2003) and
Theorem 2.5.6 in van der Vaart and Wellner (1996). Applying Lemma 1 in Chen et al. (2003)
modified to account for the fact that the functions in the class depend on the sample size \( n \),
we have (A.19) minus its expected value is \( o_p(\delta^2_n) \) uniformly in \( \alpha \in \mathcal{N}_0n \). We now turn to the
expectation term:

\[
E \left[ \left( \frac{dm_n(y_i, x_i, \tilde{\alpha})}{d\alpha} [v^*_n] - \frac{dm(y_i, x_i, \tilde{\alpha})}{d\alpha} [v^*_n] \right) m_n(y_i, x_i, \tilde{\alpha}) \right] \tag{A.22}
\]

Let \( \tilde{m}_n(x_i, \alpha) = E[m_n(y_i, x_i, \tilde{\alpha})|x_i] \). To evaluate the rate at which this expectation converges to 0,
by Cauchy Schwartz it will suffice to evaluate the rates for:

\[
E \left[ \left\| \left( \frac{dm_n(y_i, x_i, \tilde{\alpha})}{d\alpha} [v^*_n] - \frac{dm(y_i, x_i, \alpha_0)}{d\alpha} [v^*_n] \right) \right\|^2 \right] \tag{A.23}
\]

(where recall \( \| \cdot \|_E \) denotes the Euclidean norm), and

\[
E[||\tilde{m}(x_i, \alpha_0)||^2] \tag{A.24}
\]

By the smoothness conditions on \( m(\cdot) \) and the rate of convergence of \( \| \tilde{\alpha} - \alpha_0 \|_F \), these expectations
are \( o(n^{-1/2} \delta_1^{-1}) \) and \( o(n^{-1/2}) \) respectively, so the rate of convergence of the expectation in (A.22) is
\( o(n^{-1/2} \delta_1^{-1}) = o(\delta^2_n) \). This establishes the asymptotic negligibility of (A.19). Similar arguments
can be used for (A.20).

**Lemma A.2** Uniformly over \( \tilde{\alpha} \in \mathcal{N}_0n \),

\[
\frac{1}{n} \sum_{i=1}^{n} \frac{dm(y_i, x_i, \alpha_0)}{d\alpha} [v^*_n] (m_n(y_i, x_i, \tilde{\alpha}) - m_n(y_i, x_i, \alpha_0)) = < v^*_n, \tilde{\alpha} - \alpha_0 > + o_p(\delta^2_n) \tag{A.25}
\]

**Proof:** Add and subtract \( E[\frac{dm(x_i, \alpha_0)}{d\alpha} [v^*_n] (m_n(y_i, x_i, \tilde{\alpha}) - m_n(y_i, x_i, \alpha_0))] \) from the lefthand side of (A.25) We note that class of functions:

\[
\left\{ \frac{dm(y_i, x_i, \alpha_0)}{d\alpha} [v^*_n] (m_n(y_i, x_i, \alpha)) : \alpha \in \mathcal{N}_0n \right\} \tag{A.26}
\]

forms a Donsker class by Theorem 2.5.6 in van der Vaart and Wellner (1996). Hence by locally
uniform central limit theorems, (see, Lemma 1 Chen et al. (2003), and Sherman (1994) (for the case
when each function in the class depends on \( n \), as is the case here with the presence of \( v^*_n \))), we have

\[
\sup_{\tilde{\alpha} \in \mathcal{N}_0n} \frac{1}{n} \sum_{i=1}^{n} \frac{dm(y_i, x_i, \alpha_0)}{d\alpha} [v^*_n] (m_n(y_i, x_i, \tilde{\alpha}) - m_n(y_i, x_i, \alpha_0))
\]
\[-E\left[ \frac{dm (y_i, x_i, \alpha_0)}{d\alpha} [v_{ni}^*] (m_n (y_i, x_i, \hat{\alpha}) - m_n (y_i, x_i, \alpha_0))] \right] = o_p (\delta_{2n}) \quad \text{(A.27)}\]

Also, note by the definition of $< \cdot , \cdot >$ and $\| \cdot \|_F$, we have

\[E \left[ \frac{dm (y_i, x_i, \alpha_0)}{d\alpha} [v_{ni}^*] (m (y_i, x_i, \hat{\alpha}) - m (y_i, x_i, \alpha_0))] \right] = < v_{ni}^*, \hat{\alpha} - \alpha_0 > + o_p (\delta_{2n}) \quad \text{(A.28)}\]

which yields the desired result. \hfill \Box

**Lemma A.3** Under the previous assumptions

\[\frac{1}{n} \sum_{i=1}^{n} dm (y_i, x_i, \hat{\alpha}) [u_{ni}^*] m (y_i, x_i, \hat{\alpha}) = o_p (\delta_{2n}) \quad \text{(A.29)}\]

where $u_{ni}^* = \pm v_{ni}^*$. \hfill \Box

**Proof:** We follow the arguments used in Ai and Chen(2003). Let $0 < \epsilon_{2n} = o (\delta_{2n} \cdot \delta_{1n})$, let $u_{ni}^* = \pm v_{ni}^*$. Define $\alpha (t) = \hat{\alpha} + t \times \epsilon_{2n} u_{ni}^*$. Note under our smoothness assumptions $L_n (\alpha (t))$ is twice continuously differentiable with respect to $t$, where recall that

\[L_n (\alpha) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \Phi (x_i' \beta \exp (\ell (x_i))))^2 \quad \text{(A.30)}\]

By a Taylor expansion around $t = 0$ up to second order, we have, noting that $\hat{\alpha}$ minimizes the objective function,

\[0 \leq \frac{1}{n} \sum_{i=1}^{n} \frac{dm (y_i, x_i, \hat{\alpha})}{d\alpha} [\epsilon_{2n} u_{ni}^*] m_n (y_i, x_i, \hat{\alpha}) \]

\[+ \frac{1}{2n} \sum_{i=1}^{n} \frac{d^2 m_n (y_i, x_i, \alpha (s))}{d\alpha^2} [\epsilon_{2n} u_{ni}^*] m_n (y_i, x_i, \alpha (s)) \]

\[+ \frac{1}{2n} \sum_{i=1}^{n} \frac{dm (y_i, x_i, \alpha (s))}{d\alpha} \frac{dm_n (y_i, x_i, \alpha (s))}{d\alpha} [\epsilon_{2n} u_{ni}^*] = o_p (\delta_{1n}^{-1}) \quad \text{(A.31)}\]

where $s \in [0, 1]$. \hfill \Box

We will show that, uniformly over $\alpha (s) \in \mathcal{N}_{0n}$

\[\frac{1}{2n} \sum_{i=1}^{n} \frac{d^2 m_n (y_i, x_i, \hat{\alpha})}{d\alpha^2} [u_{ni}^*, u_{ni}^*] [m_n (y_i, x_i, \alpha (s))] = o_p (\delta_{1n}^{-1}) \quad \text{(A.32)}\]
First, for (A.32) as before, add and subtract the expectation of the term in the summation and apply Lemma 1 in Chen et al. (2003) to infer that the mean 0 summation is $o_p(\delta_{2n}^{-1})$. For the expectation, with the purpose of applying Cauchy Schwartz, we will derive the rates for:

$$\sqrt{E\left[\left\|d^2 d_m(y_i, x_i, \alpha(s)) \frac{d u_n^*}{d\alpha} d\alpha\right\|^2\right]}$$

(A.34)

and

$$\sqrt{E[|\tilde{m}(x_i, \tilde{\alpha})|^2]}$$

(A.35)

The smoothness conditions imposed on $m_n(y_i, x_i, \tilde{\alpha})$ for $\tilde{\alpha} \in \mathcal{N}_0$ imply that the first term is $O_p(\delta_{1n}^{-1})$, and the fact that $\tilde{m}_n(x_i, \alpha_0) = 0$ implies the second term is $O(\|\alpha - \alpha_0\|_F)$ which is negligible. (A.33) follows on the conditions imposed on $m_n(\cdot)$ and the fact that $\tilde{\alpha}$ can be replaced with $\alpha_0$ and the remainder is $o_p(\delta_{2n})$ uniformly in $\tilde{\alpha} \in \mathcal{N}_0$ by the previous arguments used modifying Lemma 1 in Chen et al. (2003).

Combining results we have that

$$0 \leq \epsilon_{2n} - \frac{1}{n} \sum_{i=1}^{n} \frac{d m_n(y_i, x_i, \hat{\alpha})}{d\alpha}[u_{ni}^*] m_n(y_i, x_i, \hat{\alpha}) + O_p(\epsilon_{2n}^2 \delta_{1n}^{-1})$$

(A.36)

Since $u_{ni}^* = \pm v_{ni}^*$ and $\epsilon_{2n} > 0$ we have

$$\frac{1}{n} \sum_{i=1}^{n} \frac{d m_n(y_i, x_i, \hat{\alpha})}{d\alpha}[u_{ni}^*] m_n(y_i, x_i, \hat{\alpha}) = O_p(\epsilon_{2n} \delta_{1n}^{-1}) = o_p(\delta_{2n})$$

(A.37)

which is the desired result.

Now that we have proved the 3 lemmas, we can easily prove the main theorem, which is the linear representation for $\hat{\theta}$. We have

$$o_p(\delta_{2n}) = \frac{1}{n} \sum_{i=1}^{n} \frac{d m_n(y_i, x_i, \hat{\alpha})}{d\alpha}[v_{ni}^*] m_n(y_i, x_i, \hat{\alpha})$$

(A.38)

$$= \frac{1}{n} \sum_{i=1}^{n} \frac{d m(y_i, x_i, \alpha_0)}{d\alpha}[v_{ni}^*] m_n(y_i, x_i, \hat{\alpha}) + o_p(\delta_{2n})$$

(A.39)

$$= \frac{1}{n} \sum_{i=1}^{n} \frac{d m(y_i, x_i, \alpha_0)}{d\alpha}[v_{ni}^*] (m_n(y_i, x_i, \hat{\alpha}) - m_n(y_i, x_i, \alpha_0))$$

(A.40)
\[ + \frac{1}{n} \sum_{i=1}^{n} \frac{dm(y_i, x_i, \alpha_0)}{d\alpha} [v_{ni}^*] m_n(y_i, x_i, \alpha_0) + o_p(\delta_{2n}) \quad (A.41) \]

\[ = < v_{ni}^*, \hat{\alpha} - \alpha_0 > + \frac{1}{n} \sum_{i=1}^{n} \frac{dm(y_i, x_i, \alpha_0)}{d\alpha} [v_{ni}^*] m_n(y_i, x_i, \alpha_0) + o_p(\delta_{2n}) \quad (A.42) \]

Therefore we have (since \( t'(\hat{\theta} - \theta_0) = < v_{ni}^*, \hat{\alpha} - \alpha_0 > + o_p(\delta_{2n}) \)), :

\[ t'(\hat{\theta} - \theta_0) = \frac{1}{n} \sum_{i=1}^{n} \frac{dm(y_i, x_i, \alpha_0)}{d\alpha} [v_{ni}^*] m_n(y_i, x_i, \alpha_0) + o_p(\delta_{2n}) \quad (A.43) \]

which is the desired result since \( \kappa_n^{1/k} Q_n = \tilde{c}(p)Q + o(1) \).

\[ A.2 \ \text{Proof of Corollary 3.1} \]

Note we can write:

\[ \sqrt{n\kappa_n^{-1/k}} (\hat{\theta} - \theta_0) = \tilde{c}(p)^{-1}Q^{-1} \sqrt{n\kappa_n^{-1/k}} \sum_{i=1}^{n} \phi_{ni} g_{ni}(\tilde{x}_i + z_{0i}w_{ni}^*)(\tilde{x}_i - \Phi_{ni} + o_p(1)) \quad (A.44) \]

Written the above way we can see the right hand side of the above question is loosely analogous to the form of a standard kernel regression estimator (see Bierens(1987)) where here the “bandwidth sequence” is \( \kappa_n^{-1/k} \) which under our assumptions will satisfy \( \kappa_n^{-1/k} \to 0 \) and \( n\kappa_n^{-1/k} \to \infty \). Thus using the same arguments as in Bierens(1987) we can show that the Lindeberg condition is satisfied.

We evaluate the limit of the variance of the above summation. Note this variance is of the form

\[ \tilde{c}(p)^{-1}Q^{-1} \kappa_n^{1/k} E[\phi_{0i}^2 g_{0i}(\tilde{x}_i + z_{0i}w_{ni}^*)^2(\tilde{x}_i + z_{0i}w_{ni}^*)' \Phi_{0i} (1 - \Phi_{0i})] \tilde{c}(p)^{-1}Q^{-1} \]

Using the same arguments to characterize \( w_{ni}^* \) we get

\[ E[\phi_{0i}^2 g_{0i}(\tilde{x}_i + z_{0i}w_{ni}^*)^2(\tilde{x}_i + z_{0i}w_{ni}^*)' \Phi_{0i} (1 - \Phi_{0i})] \to \tilde{c}(p) \frac{1}{4} Q \quad (A.45) \]

therefore the variance converges to \( \frac{1}{4} \tilde{c}(p)^{-1}Q^{-1} \) so the conclusion of the corollary follows from the Lindeberg theorem.

\[ \blacksquare \]
### TABLE I

**Design 1**

<table>
<thead>
<tr>
<th></th>
<th>SMLE</th>
<th>SNLLS</th>
<th>LNLLS</th>
<th>JKNLLS</th>
<th>SMS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>250 obs.</td>
<td>Mean Bias</td>
<td>-0.0141</td>
<td>0.0505</td>
<td>-0.0306</td>
<td>0.0292</td>
<td>0.1092</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.0239</td>
<td>0.0321</td>
<td>0.0524</td>
<td>0.0619</td>
<td>0.0538</td>
</tr>
<tr>
<td>500 obs.</td>
<td>Mean Bias</td>
<td>0.0166</td>
<td>0.0430</td>
<td>-0.0377</td>
<td>0.0119</td>
<td>0.0826</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.0126</td>
<td>0.0115</td>
<td>0.0329</td>
<td>0.0410</td>
<td>0.0260</td>
</tr>
<tr>
<td>1000 obs.</td>
<td>Mean Bias</td>
<td>0.0145</td>
<td>0.0323</td>
<td>-0.0345</td>
<td>-0.0019</td>
<td>0.0620</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.0059</td>
<td>0.0029</td>
<td>0.0178</td>
<td>0.0257</td>
<td>0.0126</td>
</tr>
</tbody>
</table>

### TABLE II

**Design 2**

<table>
<thead>
<tr>
<th></th>
<th>SMLE</th>
<th>SNLLS</th>
<th>LNLLS</th>
<th>JKNLLS</th>
<th>SMS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>250 obs.</td>
<td>Mean Bias</td>
<td>-0.0193</td>
<td>0.0037</td>
<td>-0.0483</td>
<td>0.0616</td>
<td>0.1329</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.0371</td>
<td>0.0383</td>
<td>0.1038</td>
<td>0.1093</td>
<td>0.1380</td>
</tr>
<tr>
<td>500 obs.</td>
<td>Mean Bias</td>
<td>-0.0363</td>
<td>-0.0086</td>
<td>-0.0488</td>
<td>0.0346</td>
<td>0.0760</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.0172</td>
<td>0.0214</td>
<td>0.0619</td>
<td>0.0678</td>
<td>0.0496</td>
</tr>
<tr>
<td>1000 obs.</td>
<td>Mean Bias</td>
<td>-0.0421</td>
<td>-0.0183</td>
<td>-0.0617</td>
<td>0.0023</td>
<td>0.0465</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.0083</td>
<td>0.0094</td>
<td>0.0351</td>
<td>0.0432</td>
<td>0.0202</td>
</tr>
</tbody>
</table>

### TABLE III

**Design 3**

<table>
<thead>
<tr>
<th></th>
<th>SMLE</th>
<th>SNLLS</th>
<th>LNLLS</th>
<th>JKNLLS</th>
<th>SMS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>250 obs.</td>
<td>Mean Bias</td>
<td>0.0863</td>
<td>0.0371</td>
<td>-0.0210</td>
<td>-0.0100</td>
<td>0.0797</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.0475</td>
<td>0.0110</td>
<td>0.0355</td>
<td>0.0414</td>
<td>0.0548</td>
</tr>
<tr>
<td>500 obs.</td>
<td>Mean Bias</td>
<td>0.0674</td>
<td>0.0299</td>
<td>-0.0307</td>
<td>0.0070</td>
<td>0.0652</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.0202</td>
<td>0.0052</td>
<td>0.0167</td>
<td>0.0274</td>
<td>0.0469</td>
</tr>
<tr>
<td>1000 obs.</td>
<td>Mean Bias</td>
<td>0.0704</td>
<td>0.0269</td>
<td>-0.0241</td>
<td>-0.0089</td>
<td>0.0377</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.0131</td>
<td>0.0025</td>
<td>0.0103</td>
<td>0.0147</td>
<td>0.0170</td>
</tr>
</tbody>
</table>
### TABLE IV
Design 4

<table>
<thead>
<tr>
<th></th>
<th>SMLE</th>
<th>SNLLS</th>
<th>LNLLS</th>
<th>JKNLLS</th>
<th>SMS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>250 obs.</td>
<td>Mean Bias</td>
<td>0.3086</td>
<td>0.0353</td>
<td>-0.0033</td>
<td>0.0051</td>
<td>0.0203</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.4313</td>
<td>0.0113</td>
<td>0.0067</td>
<td>0.0110</td>
<td>0.0077</td>
</tr>
<tr>
<td>500 obs.</td>
<td>Mean Bias</td>
<td>0.1259</td>
<td>0.0340</td>
<td>-0.0072</td>
<td>-0.0039</td>
<td>0.0142</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.0504</td>
<td>0.0064</td>
<td>0.0031</td>
<td>0.0055</td>
<td>0.0039</td>
</tr>
<tr>
<td>1000 obs.</td>
<td>Mean Bias</td>
<td>0.1094</td>
<td>0.0337</td>
<td>-0.0076</td>
<td>-0.0038</td>
<td>0.0130</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.0258</td>
<td>0.0033</td>
<td>0.0021</td>
<td>0.0031</td>
<td>0.0019</td>
</tr>
</tbody>
</table>

### TABLE V
High Dimensional Design

<table>
<thead>
<tr>
<th></th>
<th>SMLE</th>
<th>SNLLS</th>
<th>LNLLS</th>
<th>JKNLLS</th>
<th>SMS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>250 obs.</td>
<td>Mean Bias</td>
<td>0.0821</td>
<td>0.1254</td>
<td>0.0301</td>
<td>0.0903</td>
<td>0.0856</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.1733</td>
<td>0.2249</td>
<td>0.2106</td>
<td>0.1421</td>
<td>1.1329</td>
</tr>
<tr>
<td>500 obs.</td>
<td>Mean Bias</td>
<td>0.0801</td>
<td>0.0654</td>
<td>-0.0186</td>
<td>0.0917</td>
<td>0.1022</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.0463</td>
<td>0.0562</td>
<td>0.0841</td>
<td>0.1088</td>
<td>0.1821</td>
</tr>
<tr>
<td>1000 obs.</td>
<td>Mean Bias</td>
<td>0.1094</td>
<td>0.0789</td>
<td>-0.0432</td>
<td>0.0748</td>
<td>0.0598</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.0256</td>
<td>0.0241</td>
<td>0.0511</td>
<td>0.0740</td>
<td>0.0442</td>
</tr>
</tbody>
</table>