Comput Stat (2018) 33:1429–1455
https://doi.org/10.1007/s00180-017-0779-2

ORIGINAL PAPER

# Semiparametric estimation of the link function in binary-choice single-index models

**Alan P. Ker**[1] · **Abdoul G. Sam**[2]

Received: 10 November 2016 / Accepted: 8 November 2017 / Published online: 18 November 2017


**Abstract** We propose a new, easy to implement, semiparametric estimator for binary-choice single-index models which uses parametric information in the form of a known link (probability) function and nonparametrically corrects it. Asymptotic properties are derived and the finite sample performance of the proposed estimator is compared to those of the parametric probit and semiparametric single-index model estimators of Ichimura (J Econ 58:71–120, 1993) and Klein and Spady (Econometrica 61:387–421, 1993). Results indicate that if the parametric start is correct, the proposed estimator achieves significant bias reduction and efficiency gains compared to Ichimura (1993) and Klein and Spady (1993). Interestingly, the proposed estimator still achieves significant bias reduction and efficiency gains even if the parametric start is not correct.

**Keywords** Bias reduction · Link function · Parametric start

## 1 Introduction

Discrete choice models are commonly used in many fields, including economics, medicine, biology, and psychology, to analyze situations where a decision or choice has to be made. Another common use of these models is when dealing with selection bias. Selection bias arises when the observed outcomes are the result of decision makers' selection and hence ignoring the nonrandom nature of the sample would lead

✉ Alan P. Ker
aker@uoguelph.ca

[1] Institute for the Advanced Study of Food and Agricultural Policy, Department of Food, Agricultural and Resource Economics, University of Guelph, Guelph, Canada

[2] Department of Agricultural, Environmental and Development Economics, The Ohio State University, Columbus, OH, USA

to a bias. Estimating a discrete choice model (to estimate the 'selection' equation) is the first step in the popular 2-step Heckman procedure which corrects for selection bias. Discrete choice models usually take the following general form

$$y_i^* = v_i\beta + u_i \tag{1}$$

where $y_i^*$ is a latent variable which is operationalized by defining $y_i = 1$ if $y_i^* \geqslant 0$ and $y_i = 0$ otherwise, $\beta$ is a $(q+1)$x1 vector of unknowns, $v_i \equiv (1, x_i)$ where $x_i$ is a 1x$q$ vector of explanatory variables, and $u_i$ is the error term. The literature is dominated by parametric estimation where the distribution function (cdf) of $u_i$, say $F(u)$, is assumed to be normal (probit model) or logistic (logit model). Because parametric assumptions that are not consistent with the data could invalidate the results[1], some have considered nonparametric and semiparametric methods. In discrete choice models, the estimated effects of the regressors are of interest as well as the estimated conditional mean. Thus instead of a fully nonparametric approach, semiparametric methods have been the focus to circumvent any distributional assumptions yet recover the desired estimates.[2]

There has been significant research on semiparametric estimation of single-index models that contain parametric discrete choice models as a special case (see Ichimura 1993; Klein and Spady 1993; Horowitz and Härdle 1996; Cosslett 1987; Frölich et al. 2017; Horowitz 1998, chapters 2 and 3; Pagan and Ullah 1999, chapter 7; Birke et al. 2017). A single-index model has the following form

$$E(y|v) = F(v\beta) \tag{2}$$

where $F$ is an unknown (not necessarily a distribution) function, called the *link* function. The term $v\beta$ is the *index*.[3] Note that if $F$ is the normal or logistic distribution function, the function in (2) is the binary probit or logit model and if it is the identity function, equation (2) becomes the usual linear regression model. Among the advantages of single-index models is dimension reduction. The index $v\beta$ is a scalar and thus single-index models do not suffer from the curse of dimensionality; if $\beta$ were known it would be possible to estimate $F$ as the nonparametric mean regression of $y_i$ on $z_i = v_i\beta$ which is a scalar. Therefore in single-index models it is possible to estimate $F$ at the nonparametric rate as if there is a single regressor[4] and the coefficient vector $\beta$ at the parametric rate $O(n^{-1/2})$ (see Horowitz 1998, chapter 2 for further advantages of single-index models).

Along a completely different vein, a paper by Hjort and Glad (1995) proposes a semiparametric method for density estimation which starts with a parametric estimator

[1] For an exception see Ruud (1983).

[2] Furthermore, fully nonparametric methods suffer from the so-called 'curse of dimensionality', i.e., as the number of regressors increases, estimation precision decreases rapidly. Single-index models, which are explained below, reduce this dimensionality problem to a scalar.

[3] This article will be concerned with a linear index as in (2) instead of a general form $m(v;\beta)$ where $m$ is a scalar valued function. Ichimura (1993) has a general analysis of single-index models and Ichimura and Lee (1991) extend that general framework to multiple-index models.

[4] The fact that $\beta$ is unknown and has to replaced with an estimator does not change this result as long as the estimator of $\beta$ is $\sqrt{n}$-consistent (see Horowitz 1998, pp. 21–22).

and multiplies this parametric start with a correction factor (the unknown density divided by the parametric start) which is estimated nonparametrically. The idea is based on bias reduction. If the parametric start captures a sufficient amount of the curvature of the unknown density, the correction factor will be close to a constant and less rough. Thus the bias associated with the nonparametric estimation of this correction factor will be less than that associated with the underlying density. Neither this paper nor a companion piece by Glad (1998) consider single-index models.

Unlike the semiparametric papers in the literature that estimate the link function nonparametrically, this article proposes to estimate the link function semiparametrically by employing the Hjort and Glad (1995) bias reduction idea. Potentially relevant information is introduced in the form of a parametric function which is the parametric guide for the link function. The distinguishing characteristic of the proposed estimator is how the unknown link function $F$ is estimated using prior parametric information about its shape.

Asymptotic properties for the proposed estimator are derived and an extensive simulation analysis is conducted in which the finite sample performance of the popular parametric probit estimator and semiparametric estimator of Ichimura (1993) are compared with the proposed estimator. Note that finite sample simulation analysis is especially important because bias reduction is not always realized in samples of reasonable size (see Jones and Signorini 1997).

The article is organized as follows: In the next section, some of the semiparametric and nonparametric estimators for binary data are reviewed with emphasis on the semiparametric estimators based on the index restriction. The new estimator is presented in Sect. 3. Section 4 contains the simulation results. Finally, concluding thoughts are discussed in Sect. 6. All proofs are collected in the appendix.

## 2 Semi and nonparametric estimators of binary data

In this section we briefly review the semi and nonparametric models for binary data. This review is by no means exhaustive; instead, it reflects their importance and relevance for the proposed estimator in this article. For a more comprehensive review, see, for instance, Horowitz (1993), Pagan and Ullah (1999) and Powell (1994). Common features of semiparametric models are that it is assumed that the distribution of $u$ depends on $x$ only through the index[5] (index restriction) and that $F$ is completely unknown (no centering assumptions will be made thus the intercept term can not be identified).[6]

[5] Single-index models, unlike models which assume independence of $u$ and $x$, allow for limited forms of heteroscedasticity (general but known form and unknown form if it depends only on the index). This limitation can be serious since, for instance, the assumption that $Pr(y = 1|v)$ depends only on the index does not allow a certain form of heteroscedasticity (random coefficients model) which may be important in applications.

[6] The maximum score estimator of Manski (1975) and its smoothed version by Horowitz (1992) make zero conditional median assumption (median$(u|x) = 0$) which identifies the intercept term (zero conditional mean assumption is not sufficient for identification in a binary response model, see Manski (1988, p. 731); Horowitz 1998, section 3.2). These models allow for different forms of heteroscedasticity including random coefficients models although at the cost of a rate of convergence slower than $\sqrt{n}$. In fact under this

### 2.1 Quasi-maximum likelihood estimator

The semiparametric single-index model of Klein and Klein and Spady (1993) can be considered as quasi log-likelihood estimation. Note that for binary data $Pr(y = 1|v) = F(v\beta)$ and if $F$ (the link function) were known the maximum likelihood estimator would maximize the log-likelihood

$$\sum_{i=1}^{n}(y_i \log[F(v_i\beta)] + (1 - y_i)\log[1 - F(v_i\beta)]). \tag{3}$$

The idea is to consider the link function in (3) unknown and replace it with a nonparametric estimator. Since $\beta$ in single-index models is not fully identified, a location and scale normalization is required (see Horowitz 1998, section 2.4). Location normalization is achieved by requiring $v$ to contain no intercept term and scale normalization is achieved by setting the $\beta$ coefficient of a (continuous) regressor equal to one.[7] The $\beta$ vector after scale and location normalizations is denoted by $b$, i.e., $b \equiv (1, \beta_2, \ldots, \beta_q)^T$ assuming the first regressor has a continuous distribution. The following Nadaraya-Watson nonparametric estimator for the link function is used

$$\tilde{F}(x_i b) = \frac{\sum_{j\neq i} y_j K\left(\frac{x_i b - x_j b}{h}\right)}{\sum_{j\neq i} K\left(\frac{x_i b - x_j b}{h}\right)} \tag{4}$$

where $K$ is the Kernel function (usually a symmetric density function) and $h = h(n)$ is the smoothing parameter such that $h \to 0$ as $n \to \infty$.[8] By replacing the unknown link function $F$ in (3) by (4), the quasi log-likelihood function is obtained and by maximizing the quasi log-likelihood with respect to $\tilde{b}$ where $\tilde{b}$ is $b$ without its first component, the semiparametric estimator $\hat{\tilde{b}}_{KS}$ is obtained. Klein and Spady (1993) show that $\hat{\tilde{b}}_{KS}$ satisfies $\sqrt{n}(\hat{\tilde{b}}_{KS} - \tilde{b}_0) \xrightarrow{d} N(0, \Omega_{QL})$ where $\Omega_{QL}$ can be consistently estimated by the Hessian and the outer product of the gradient matrices and it attains the semiparametric efficiency bound of Cosslett (1987) if the errors are independent of the regressors.[9] Note that in (4) the denominator can get arbitrarily close to zero so care

Footnote 6 continued
conditional median independence assumption $\sqrt{n}$ consistency is not possible, see Pagan and Ullah 1999, p. 278 and Horowitz (1993).

[7] An alternative scale normalization would be $||\beta|| = 1$ where $||\cdot||$ is the Euclidean norm.

[8] Since there is no location restriction on $u$ in single-index models and thus the intercept term is not identified, (4) is actually a nonparametric estimator of the distribution of $u + \beta_0$ where $\beta_0$ is the intercept. Also, Klein and Spady (1993) have additive terms in the numerator and denominator of (4) to control the rate at which numerator and denominator tend to zero. Ichimura (1993) utilizes indicator variables to trim those observations which correspond to small density values. A similar trimming function and an indicator variable enter multiplicatively to objective functions in (3) and (5) respectively. In this presentation those terms are ignored for simplicity.

[9] A paper by Chen (2000) builds on Klein and Spady (1993) and shows that the intercept can be consistently estimated and there are possible efficiency gains in the estimation of slope coefficients although at the cost

must be taken. Klein and Spady (1993) use trimming procedures without restricting $x$ to be in a specific set as in Ichimura (1993) (see below).

### 2.2 (Weighted) semiparametric least squares estimator

The single-index model of Ichimura (1993) is, in contrast to Klein and Spady (1993), based on minimizing a nonlinear least squares (NLS) loss function. The NLS estimator, denoted $\hat{\tilde{b}}_I$, of $b$ minimizes

$$\frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{F}(x_i b)]^2 \tag{5}$$

where $\hat{F}$ is the nonparametric estimator for the unknown link function in (4). Ichimura (1993) denotes this model semiparametric least squares (SLS) and shows that $\hat{\tilde{b}}_I$ is consistent and $\sqrt{n}(\hat{\tilde{b}}_I - \tilde{b}_0) \xrightarrow{d} N(0, \Omega_{SLS})$ and gives a consistent estimator of $\Omega_{SLS} = \Gamma^{-1}\Sigma\Gamma^{-1}$. $\Gamma$ and $\Sigma$ can be consistently estimated by

$$\hat{\Gamma} = \frac{1}{n}\sum_{i=1}^{n}\tilde{x}_i\tilde{x}_i^T\hat{F}'(x_i\hat{b}_I)^2,$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\tilde{x}_i\tilde{x}_i^T\hat{F}'(x_i\hat{b}_I)^2[y_i - \hat{F}(x_i\hat{b}_I)]^2$$

where $\tilde{x}_i \equiv (x_{2i}, \ldots, x_{qi})$, $\hat{b}_I \equiv (1, \hat{\tilde{b}}_I^T)^T$, and $\hat{F}'$ is the derivative of $\hat{F}$.

Ichimura (1993) also considers weighted SLS (WSLS) in which he weights the objective function (5) and the summands in $\hat{F}$ by a weight function $W(x_i)$. As in parametric NLS, efficiency considerations play a role: the choice of the weight function does not affect the consistency and rate of convergence of the estimator of $b$ but does affect its efficiency. Optimally weighted (the weight function is a consistent estimator of $Var(y|x)^{-1}$) WSLS achieves the semiparametric efficiency bound.[10] Horowitz (1998, p. 31) explains how a consistent estimator of $Var(y|x)$ can be obtained. Note that the first order conditions from (3) are the same conditions that one can obtain from (5) with the *estimated*[11] weight function $W(x) = \{\hat{F}(x\hat{b}_I)[1 - \hat{F}(x\hat{b}_I)]\}^{-1}$. As in the Klein and Spady (1993) estimator, care must be taken to prevent the denominator of (4) from getting arbitrarily close to zero. Ichimura (1993) restricts the summands in (4) and (5) to those observations for which the density of the index is not too small (see Ichimura 1993 for details).

of stronger assumptions: a location restriction in the form of conditional symmetry, i.e., the density of $u$ conditional on the regressors is symmetric around zero and an index restriction stronger than the one in Klein and Spady (1993), namely, the conditional density depends on $x$ only through the squared index.

[10] For this efficiency result, the weight function should depend on $x$ only through the index as in binary-choice models where $Var(y|x) = Var(y|xb)$.

[11] Ichimura (1993) treats $W(\cdot)$ as a known function.

## 3 Parametrically-guided single-index estimator

Suppose one wishes to estimate the conditional mean function $E(y|x) = m(x)$. Hjort and Glad (1995) and Glad (1998) first start with a parametric estimator $m(x, \hat{\beta})$ which could be, for instance, a simple linear regression or a more complex maximum likelihood estimation, and then multiply it with a correction factor $r(x) = m(x)/m(x, \hat{\beta})$ which is estimated nonparametrically. The idea is based on bias reduction: if the parametric start is close to the truth, the correction factor will be close to a constant and thus smoother and (bias-wise) easier to estimate than $m(x)$ itself. Hence the bias associated with nonparametric estimation of this correction factor would be less than the bias from direct nonparametric estimation of the unknown regression function. Their estimator is

$$\hat{m}(x) = m(x, \hat{\beta})\hat{r}(x).$$

When the correction factor is estimated by the Nadaraya–Watson estimator[12] their parametrically guided estimator is

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} y_i \frac{m(x,\hat{\beta})}{m(x_i,\hat{\beta})} K_h(x_i - x)}{\sum_{i=1}^{n} K_h(x_i - x)}.$$

Glad (1998) shows that this estimator has the same large sample variance as the standard nonparametric estimators (Nadaraya–Watson and local linear) while bias reduction is possible if the parametric start belongs to a neighborhood around the true regression curve. Sam and Jiang (2009) and Sam and Ker (2006) report significant efficiency gains of this approach over the Nadaraya–Watson estimator.

Our proposed semiparametric estimator for the single-index models is based on the above idea, that is introducing (potentially) relevant information in the form of a parametric function in an attempt to reduce bias. The estimator starts with a parametric model for the link function, say $G(xb)$, where $G(\cdot)$ is a known function (for instance normal cdf) and multiplies it with the correction factor $r(xb) = F(xb)/G(xb)$ which is estimated nonparametrically. Note that the information that this estimator starts with is only related to the shape of the link function and not to the coefficient estimates. In this sense the estimator is using a fixed start vis-à-vis Hjort and Glad (1995) and Glad (1998). When the Nadaraya–Watson estimator is used to estimate the correction factor, the proposed estimator is

$$\hat{F}(x_i b) = \frac{\sum_{j \neq i} \left\{ y_j \frac{G(x_i b)}{G(x_j b)} \right\} K\left(\frac{x_i b - x_j b}{h}\right)}{\sum_{j \neq i} K\left(\frac{x_i b - x_j b}{h}\right)}. \tag{6}$$

[12] Glad (1998) generalizes this to local $p$th order polynomial estimator which reduces to the Nadaraya–Watson for $p = 0$.

This estimator will be referred to as parametrically-guided single-index model (PGSIM). The unknown $F$ in (3) is replaced with (6) and the resulting quasi log-likelihood function is maximized with respect to $b$.[13]

Note that the semiparametric estimator of the link function in (6) does not nest the nonparametric estimator in (4) so the Klein and Spady (1993) model is not nested in the proposed model.[14] However, a bias and variance comparison of (4) and (6) is useful to see the bias reduction. Equation (4) is the usual Nadaraya–Watson estimator of the link function whereas (6) corresponds to the fixed start of Hjort and Glad (1995, section 2), Glad (1998, section 2) as mentioned above. In the appendix, we prove (see first part of the proof of *theorem 1*) that while both have the same variance

$$Var(\hat{F}(z)) = (nh)^{-1} f(z)^{-1} \sigma^2(z) R(K) + o_p((nh)^{-1})$$

where $f(z)$ is the density of $z$, $\sigma^2(z) = Var(y|z)$, and $R(K) = \int K^2(s)ds$, the bias of (4) is

$$Bias(\tilde{F}(z)) = \frac{h^2 \mu_2(K)}{2f(z)} (F''(z)f(z) + 2f'(z)F'(z)) + o_p(h^2)$$

where $\mu_2(K) = \int s^2 K(s)ds$, whereas the bias of (6) is

$$Bias(\hat{F}(z)) = \frac{h^2 \mu_2(K)}{2f(z)} (r''(z)G(z)f(z) + 2f'(z)r'(z)G(z)) + o_p(h^2).$$

Thus, for the same $h$ and $K$, bias reduction is possible if the parametric start $G$ can be chosen such that

$$|r''(z)G(z)f(z) + 2f'(z)r'(z)G(z)| < |F''(z)f(z) + 2f'(z)F'(z)|. \tag{7}$$

If the parametric start is proportional to $F$, the correction factor $r$ is going to be a constant and thus $r' = r'' = 0$. If it is sufficiently close to $F$, roughness of $r$ will be

[13] While the primary interest in single-index models is generally about the parameter estimates and marginal effects, there are cases—such as economic discrimination analyses—where the focus is on the estimated probabilities. For example, Blinder–Oaxaca-type decomposition studies interested in differences between groups (race, gender etc.) regarding a binary outcome such as computer ownership, teenage pregnancy, or school attendance, are primarily interested in differences in average estimated probabilities between groups (Fairlie 2005; Seah et al. 2017). Correctly estimated probabilities are critical to the accuracy of the gap being investigated.

[14] For this to happen, the parametric start $G(\cdot)$ should be a constant function. In density estimation, the Hjort and Glad (1995) estimator nests the usual Kernel density estimator if the parametric start is the uniform density over the space. But here a distribution function which is constant and which satisfies the continuity (of the index) assumption can not be found. One obvious example of such a constant distribution function, which does not satisfy the continuity assumption, is the unit point mass at $a$ when $z = a$ a.s.:

$$G(z) = \begin{cases} 0 & \text{if } z < a \\ 1 & \text{if } z \geqslant a. \end{cases}$$

Here not only is the continuity assumption not satisfied but also $\beta$ is not identified with this $G(\cdot)$.

less than roughness of $F$ and $r$ will have a smaller second derivative. Thus (7) defines a 'neighborhood' of $F$ where bias reduction is possible by choosing a parametric start from this neighborhood.

The above argument indicates that even though the estimator can not asymptotically outperform Klein and Spady (1993) and Ichimura (1993) (when optimally weighted in binary-choice model estimation) estimators with respect to the coefficients as they attain the semiparametric efficiency bound, in finite samples there is potential to significantly increase efficiency by using a parametric guide.

### 3.1 How improvement in the link function estimate affects precision of parameter estimates

The "neighborhood" provided in equation (7) is directly relevant for the finite dimensional parameters and associated marginal effects. This can be seen from the gradient vectors (first-order conditions) for the quasi-maximum likelihood and nonlinear least squares objective functions which are, respectively:

$$\sum_{i=1}^{n} \frac{\partial \tilde{F}(x_i \hat{b}_{KS})}{\partial b} [\tilde{F}(x_i \hat{b}_{KS})]^{-1} [1 - \tilde{F}(x_i \hat{b}_{KS})]^{-1} [y_i - \tilde{F}(x_i \hat{b}_{KS})], \quad \text{and}$$

$$\sum_{i=1}^{n} \frac{\partial \tilde{F}(x_i \hat{b}_{I})}{\partial b} [y_i - \tilde{F}(x_i \hat{b}_{I})].$$

Consistency of the Nadaraya–Watson estimator of the link function $\tilde{F}(.)$ ensures consistency of the finite dimensional parameter vector. However, in finite samples, $\tilde{F}(.)$ and its derivative may substantially deviate from $F(.)$, the true underlying error distribution and its derivative because of the bias (see above for bias expression). It's clear from the gradient functions that such deviation will result in less precise estimation of the parameters. As noted in Pagan and Ullah (1999, p. 275), the gradient vectors above generally do not have an expectation of zero under specification error (bias in our case) and therefore $\hat{b}$ may substantially deviate from the true parameter vector in finite samples. The smaller the finite sample bias of the link function estimate, the more precise the finite dimensional parameters are going to be. This is why Klein and Spady (1993) advocate the use of a higher order Kernel rather than the Nadaraya–Watson to estimate $F(.)$. Higher order Kernels, however, may generate negative probability estimates for any given finite sample (Klein and Spady 1993). The approach we are proposing seeks to remedy the flaws of both the Nadaraya–Watson and higher order Kernel link function estimates under certain circumstances (equation (7)) in order to yield more precise finite dimensional parameters.

Note that the proposed link function estimator (6) collapses to the parametric start [(e.g., Normal (Probit) or Logistic (Logit)] if the parametric start happens to be the correct functional form of the error distribution; that is if $G$ is identifical to $F$. In this case, the correction factor ($r(z) = 1 \; \forall z$) is estimated unbiasedly (slope and curvature are both equal to zero). Hence, while finite-dimensional parameters obtained using the parametrically-guided single index model and those obtained using either Ichimura

or Klein and Spady are all consistent, our parameter estimates will be more efficient since our link function is devoid of functional form misspecification. In general, if the parametric start is suffciently close to $F$, roughness of $r$ will be less than roughness of $F$, implying finite dimensional parameters that are more precisely estimated than those based on the Nadaraya–Watson link function estimator.

### 3.2 Consistency and asymptotic normality

As mentioned above, in the semiparametric estimators of Ichimura (1993) and Klein and Spady (1993) care must be taken as the nonparametric density estimator in the denominator can get arbitrarily small. Also, in Klein and Spady (1993), any estimator of $F$ should be kept in the (0, 1) open interval. They use likelihood trimming to downweight observations for which the corresponding densities are small and probability trimming to control the rate at which numerator and denominator of $\hat{F}$ tend to zero. Ichimura (1993) on the other hand, restricts $x$ to a set by indicator variables on which the above mentioned problems are avoided. This latter approach is easier to deal with in asymptotics.[15] Hence we employ this approach. In our case, these restrictions should include one more thing, namely, that the parametric start $G(\cdot)$ should be nonzero throughout the support of the index. In what follows, our notation for the trimming terms will follow Ichimura (1993, p. 78) and Horowitz (1998, pp. 23–24) closely.

Our semiparametric estimator for binary-choice models maximizes

$$\hat{Q}_n(b) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{[x_i \in A_x]}(y_i \log[\hat{F}(x_i b)] + (1 - y_i)\log[1 - \hat{F}(x_i b)]) \tag{8}$$

where $\mathbf{1}_{[\cdot]}$ is an indicator variable, $A_x \subset \mathbb{R}^q$ is such that $A_x = \{x : p(xb) \geq \eta \quad \forall\, b \in B\}$, $p(\cdot)$ is the density of the index, $\eta$ is a positive constant, and

$$\hat{F}(x_i b) = \frac{\frac{1}{(n-1)h}\sum_{j\neq i} y_j \mathbf{1}_{[x_j \in A_{nx}]}\left\{\frac{G(x_i b)}{G(x_j b)}\right\} K\left(\frac{x_i b - x_j b}{h}\right)}{\frac{1}{(n-1)h}\sum_{j\neq i} \mathbf{1}_{[x_j \in A_{nx}]} K\left(\frac{x_i b - x_j b}{h}\right)} \tag{9}$$

where $K : \mathbb{R} \to \mathbb{R}$ is a density function, $h > 0$ and $h \to 0$ as $n \to \infty$, and $A_{nx}$ is a set such that, as Ichimura (1993, p. 79) explains "…includes $[A_x]$ in such a way that all boundary points in $[A_x]$ are interior to $[A_{nx}]$, in a neighborhood of $x$, with probability approaching 1, there are data in all directions to take a local average". Clearly the purpose is to reduce bias that may otherwise result close to the boundary points. His suggestion is to use $A_{nx} = \{x : ||x - x'|| \leqslant 2h \text{ for some } x' \in A_x\}$. Note

[15] In actual estimations, trimming has very little effect on the performance of the estimators. Klein and Spady paper reports simulation results from untrimmed estimator: "…the estimate obtained without any trimming performed quite similar to that under the trimming that we employed. Accordingly, we report results for the semiparametric estimator obtained without probability or likelihood trimming (Klein and Spady 1993, p. 406)."

that as $n \to \infty$, $A_{nx}$ seems to get smaller but as the sample size increases there will be more and more sample points close to boundary as well and sufficient to take a local average.

The identification of single-index models, in the most general context, has been analyzed by Ichimura (1993). Manski (1975) looks at identification of binary-choice models with linear index under different assumptions including index restriction. Klein and Spady (1993) give conditions for identification of single-index models in binary response models where the index is a general but known function. We refer the reader to original papers for details. Here we make the following assumptions:

| | |
|---|---|
| Assumption I1 | The model in (1) satisfies the index restriction. |
| Assumption I2 | $F$ is continuously differentiable and not a constant function of the index over its support. |
| Assumption I3 | At least one regressor, with nonzero coefficient, has a continuous distribution. Its distribution conditional on the remaining regressors is absolutely continuous. |
| Assumption I4 | Varying the values of discrete regressors must not divide the support of the index into disjoint subsets. |
| Assumption I5 | $Pr(y = 1\|xb_0) = Pr(y = 1\|xb_*) \Rightarrow b_0 = b_*$. |

For I4, see Horowitz (1998, pp. 16–17) for an example how its violation turns the slope coefficient on the discrete regressor into an intercept term which is not identified. Assumption I5 is a necessary restriction for identification of binary-choice maximum-likelihood models in particular. Klein and Spady (1993, pp. 395–397) provide sufficient conditions under this assumption.[16]

The consistency proof boils down to showing uniform convergence, over $x$ and $b$, of $\hat{F}$ to $F$. Once this is accomplished, the estimator that maximizes the quasi-likelihood in (8) asymptotically behaves like the estimator that maximizes the likelihood function for a known $F$ since $\sup_{b\in B} |\hat{Q}_n(b) - Q_n(b)| = o_p(1)$ where $Q_n(b) = n^{-1} \sum_{i=1}^{n} \mathbf{1}_{[x_i \in A_x]}(y_i \log[F(x_i b)] + (1 - y_i) \log[1 - F(x_i b)])$. The estimator which maximizes this likelihood can be analyzed by standard methods for parametric estimators. Lemma 1 shows the uniform convergence of $\hat{F}$ and is proved in the appendix.

**Lemma 1** *Under assumptions 1–7 (in appendix), if* $h \to 0$ *and* $h\sqrt{n} \to \infty$ *as* $n \to \infty$, *then for any* $\epsilon > 0$

$$Pr\left( \sup_{(x,b)\in A_x \times B} |\hat{F}(xb) - F(xb)| > \epsilon \right) \to 0 \quad as \quad n \to \infty.$$

*Proof* See the appendix. □

[16] There are two cases to consider: when the link function $F$ is monotonic in the index and when it is not. If the underlying distribution is heteroscedastic, for instance, $F$ need not be monotonic in the index.

As Bierens (1987b, p. 115 ) notes, the best uniform convergence rate is obtained when $\min(h\sqrt{n}, h^{-2})$ (see the appendix) is maximum so $h \propto n^{-1/6}$ and thus $\min(h\sqrt{n}, h^{-2}) \propto n^{2/6}$.[17] This is not the fastest uniform convergence rate achievable for nonparametric regression (see Schuster and Yakowitz 1979), however, this conservative approach has been chosen for its simplicity.

Theorem 1 states that $\hat{F}$ is asymptotically normal and is proved in the appendix.

**Theorem 1** *Under assumptions 1, 5, and 8 to 11 (in appendix),*

$$\sqrt{nh}(\hat{F}(z) - F(z) - Bias(\hat{F}(z))) \to \mathcal{N}(0, f(z)^{-1}\sigma^2(z)R(K)).$$

*Proof* See the appendix. □

With lemma 1, we can show that $\hat{Q}_n(b)$ in (8) converges in probability, uniformly in $b$, to a likelihood function for a known $F$. This limiting likelihood, in turn, converges in probability, uniformly in $b$, to its expectation which is maximized by $b_*$. This $b_*$ satisfies (see Klein and Spady 1993, p. 400) $Pr(y = 1|x) = Pr(y = 1|xb_0) = Pr(y = 1|xb_*)$ where the first equality is the index restriction and the second equality is a necessary condition for $b_*$ to be maximum. Hence, from I5, $b_* = b_0$ is unique maximum. Theorem 2 below gives the result and is proved in the appendix.

**Theorem 2** *Under assumptions 1–7 (in appendix),*

$$\hat{b} \equiv \arg\sup_b \hat{Q}_n(b) \xrightarrow{p} b_0.$$

*Proof* See the appendix. □

## 4 Simulations

In this section, finite sample performance of the probit and Ichimura estimators are compared with the proposed estimator. The Klein and Spady estimator yields almost identical results and are available from the authors.

### 4.1 Design

The model generating the data is

$$y_i^* = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$$

and $y_i$ takes a value of 1 if the latent $y_i^* \geqslant 0$ and a value of 0 otherwise. The values of the true parameters are $\beta_0 = 0$, $\beta_1 = 1$, and $\beta_2 = 1$. The regressors $x_1$ and $x_2$ are independent. The data generating process (DGP) for $x_1$ is chi-square distribution with 3 degrees of freedom truncated at 6. The DGP for $x_2$ is standard normal truncated at

[17] Klein and Spady (1993) obtain a similar uniform convergence rate.

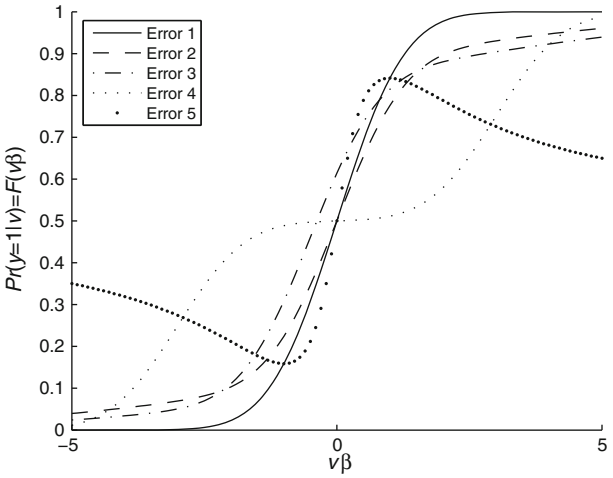


**Fig. 1** Link function for design errors

±2. Both $x_1$ and $x_2$ are first trimmed by 2%, i.e., lower and upper tails of their empirical distributions are trimmed by one percent and then standardized to have zero mean and unit variance. For the DGP of the error term, four homoscedastic normal mixtures and a heteroscedastic distribution are considered. Normal mixtures are (1) standard normal, (2) 0.75·N(0,1) + 0.25·N(0,25), (3) 0.75·N(-0.5,1) + 0.25·N(1.5,25), and (4) 0.5·N(3,1) + 0.5·N(-3,1). The second distribution is leptokurtic, the third distribution is skewed and leptokurtic, and the fourth distribution is bimodal. The second (third) (fourth) distribution has standard error 2.65 (2.78) (3.16), skewness 0 (1.29) (0), and kurtosis 6.61 (6.29) (1.38). The heteroscedastic distribution is normal with zero mean and variance $0.25(1 + (v_i\beta)^2)^2$ where $v_i\beta = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$.[18] Figure 1 gives a graph of the link function for different errors in the simulation design. Note that when the error distribution is heteroscedastic (error 5), the link function is not monotonic in the index.

[18] Even though it is not in the simulation design, at this point, it is instructive to digress to discuss maximum likelihood estimation of misspecified binary choice models. Ruud (1983) showed that when the explanatory variables are multivariate normal, maximum likelihood estimates of slope coefficients can still be estimated consistently up to scale even when the distributional assumption is not correct. More generally the result holds when

$$E(\tilde{x}|xb = t) = c_0 + c_1 t \tag{10}$$

where $c_0$ and $c_1$ are constants, which is satisfied when the explanatory variables are multivariate normal. Note that this consistency result does not hold for the probability estimates.
Another interesting implication of (10) is that when it holds, the semiparametric efficiency bound is the same as the parametric (Cramér–Rao) efficiency bound (see Cosslett 1987).

**Table 1** Simulation results - $\hat{\beta}_2$

| | Probit | | | Ichimura | | | PGSIM | | |
|---|---|---|---|---|---|---|---|---|---|
| N | Bias | RMSE | Cov | Bias | RMSE | Cov | Bias | RMSE | Cov |
| *Error 1* | | | | | | | | | |
| 100 | 0.0082 | 0.1687 | 0.952 | 0.1906 | 0.5163 | 0.968 | 0.0264 | 0.2151 | 0.962 |
| 250 | 0.0110 | 0.1117 | 0.936 | 0.0698 | 0.1611 | 0.944 | 0.0140 | 0.1285 | 0.956 |
| 500 | 0.0040 | 0.0822 | 0.938 | 0.0663 | 0.1278 | 0.922 | 0.0094 | 0.1059 | 0.958 |
| *Error 2* | | | | | | | | | |
| 100 | −0.1840 | 0.2575 | 0.862 | 0.6480 | 1.2734 | 0.858 | −0.0009 | 0.2544 | 0.932 |
| 250 | −0.2060 | 0.2364 | 0.432 | 0.1051 | 0.2951 | 0.922 | −0.0060 | 0.1548 | 0.944 |
| 500 | −0.2805 | 0.2913 | 0.094 | 0.0876 | 0.1640 | 0.956 | −0.0288 | 0.1163 | 0.952 |
| *Error 3* | | | | | | | | | |
| 100 | −0.2064 | 0.2678 | 0.802 | 0.6800 | 1.2959 | 0.828 | −0.0287 | 0.2959 | 0.928 |
| 250 | −0.2434 | 0.2674 | 0.246 | 0.1179 | 0.3085 | 0.918 | 0.0023 | 0.1572 | 0.914 |
| 500 | −0.3419 | 0.3512 | 0.028 | 0.0638 | 0.1597 | 0.912 | −0.0453 | 0.1234 | 0.950 |
| *Error 4* | | | | | | | | | |
| 100 | −0.7222 | 0.7387 | 0.002 | 2.7573 | 2.8366 | 0.010 | −0.3791 | 0.6015 | 0.776 |
| 250 | −0.7049 | 0.7137 | 0.004 | 2.7987 | 2.8599 | 0.178 | −0.3935 | 0.5629 | 0.946 |
| 500 | −0.8762 | 0.8789 | 0.002 | 2.5483 | 2.7039 | 0.250 | −0.3828 | 0.4568 | 0.814 |
| *Error 5* | | | | | | | | | |
| 100 | −0.3826 | 0.4241 | 0.790 | 0.3554 | 0.6532 | 0.964 | 0.1929 | 0.3993 | 0.934 |
| 250 | −0.3637 | 0.3793 | 0.026 | 0.0569 | 0.1405 | 0.856 | 0.0475 | 0.1326 | 0.912 |
| 500 | −0.3664 | 0.3752 | 0.030 | 0.0988 | 0.1367 | 0.734 | 0.0808 | 0.1233 | 0.858 |

### 4.2 Results

Table 1 has the bias, root mean squared error (RMSE) and coverage probability of $\hat{\beta}_2$ and Table 2 has the $L_1$ and $L_2$ errors for $\hat{F}$ over 500 simulations for sample sizes of 100, 250 and 500.[19] In the table, PGSIM is the parametrically-guided single-index model estimator. For identification purposes, in single-index models, $\beta_0$ is not estimated and $\beta_1$ is set equal to 1. A probit cdf is the parametric guide in the PGSIM. For all three semiparametric estimators, a normal density function is used as the Kernel. The smoothing parameter is fixed at $h = \hat{\sigma}(v_i\beta)n^{-1/7.5}$ as in Horowitz and Härdle (1996) where $\hat{\sigma}(v_i\beta)$ is the standard deviation of the index.[20] Note that in practice if, for some $z_j$, $G(z_j)$ is zero or near zero while $G(z_i)$ is not then the ratio $G(z_i)/G(z_j)$ blows up in PGSIM. Following Hjort and Glad (1995) and Glad (1998), trimming is conducted below 0.1 and above 10.

[19] $L_1$ norm is $\int |F - \hat{F}|dx$ while $L_2$ norm is $\int (F - \hat{F})^2 dx$ where $F$ is the true unknown link function and $\hat{F}$ is an estimate of the unknown link function.

[20] We also run the simulations with $h = n^{-1/6.02}$ as in Klein and Spady (1993) and the results changed immaterially. In practical applications, the smoothing parameter should be chosen along with the parameter estimates by mazimizing the quasi likelihood function.

**Table 2** Simulation results (multiplied by 100) - $\hat{F}$

| | | Probit | | Klein and Spady | | Ichimura | | PGSIM | |
|---|---|---|---|---|---|---|---|---|---|
| | N | $L_1$ | $L_2$ | $L_1$ | $L_2$ | $L_1$ | $L_2$ | $L_1$ | $L_2$ |
| Error 1 | 100 | 0.0000 | 0.0000 | 1.1798 | 1.0862 | 0.5348 | 0.7313 | 0.2385 | 0.4883 |
| | 250 | 0.0000 | 0.0000 | 0.5336 | 0.7305 | 0.2209 | 0.4700 | 0.1482 | 0.3850 |
| | 500 | 0.0000 | 0.0000 | 0.3143 | 0.5606 | 0.1777 | 0.4215 | 0.1082 | 0.3289 |
| Error 2 | 100 | 0.5900 | 0.7681 | 3.2217 | 1.7949 | 1.3821 | 1.1756 | 0.5911 | 0.7689 |
| | 250 | 0.5900 | 0.7681 | 1.8863 | 1.3734 | 0.8747 | 0.9353 | 0.4994 | 0.7067 |
| | 500 | 0.5900 | 0.7681 | 1.2750 | 1.1291 | 0.8159 | 0.9033 | 0.4286 | 0.6547 |
| Error 3 | 100 | 0.7749 | 0.8803 | 3.2705 | 1.8084 | 1.7271 | 1.3142 | 0.7216 | 0.8495 |
| | 250 | 0.7749 | 0.8803 | 2.4436 | 1.5632 | 1.2962 | 1.1385 | 0.7074 | 0.8411 |
| | 500 | 0.7749 | 0.8803 | 1.9443 | 1.3944 | 1.2781 | 1.1307 | 0.6824 | 0.8261 |
| Error 4 | 100 | 2.1945 | 1.4814 | 3.2728 | 1.8091 | 2.9982 | 1.7315 | 1.7678 | 1.3296 |
| | 250 | 2.1945 | 1.4814 | 2.3210 | 1.5235 | 2.0055 | 1.4162 | 1.6064 | 1.2674 |
| | 500 | 2.1945 | 1.4814 | 2.0753 | 1.4406 | 1.7104 | 1.3078 | 1.4807 | 1.2168 |
| Error 5 | 100 | 2.0785 | 1.4417 | 8.0167 | 2.8314 | 2.2951 | 1.5150 | 1.6913 | 1.3005 |
| | 250 | 2.0785 | 1.4417 | 4.2085 | 2.0515 | 2.4252 | 1.5573 | 1.6386 | 1.2801 |
| | 500 | 2.0785 | 1.4417 | 3.3469 | 1.8295 | 2.3198 | 1.5231 | 1.6314 | 1.2772 |

When the error distribution is standard normal (error 1), the parametric guide is the true link function hence probit performs the best. The PGSIM also performs quite well with lower bias and RMSE than Ichimura (and Klein and Spady) estimator across all sample sizes. The efficiency gains of the proposed estimator relative to the other semiparametric estimators are expected in this case since the parametric start is correct. The coverage probability is also as expected, roughly around 95%, equivalent to the confidence interval.

Under errors 2, 3 and 4, the parametric guide for the link function is not correct but the estimator still achieves significant bias reduction and efficiency gain compared to the competing semiparametric estimators at all sample sizes. Not surprisingly, as The PGSIM estimator outperforms the Ichimura estimator quite significantly in small sample sizes but as $n$ increases that efficiency gain is reduced (with the exception of error 4). When the error distribution is heteroscedastic (error 5), unlike linear models and as in most nonlinear models, probit maximum likelihood estimators are inconsistent (Yatchew and Griliches 1985). The proposed estimator under error 5 still dominates the Ichimura estimator.

Overall, the PGSIM achieves efficiency gain with respect to the Ichimura estimator in all the 15 simulations scenarios (5 errors and 3 sample sizes). This dominant performance is often due to the considerably lower bias of the PGSIM. Not surprisingly, the relative efficiency gain of the PGSIM is reduced as the sample size increases.

The performances of single-index models are also compared with respect to their ability to estimate probabilities ($\hat{F}$). Table 2 has the results from this comparison. In the table, $L_1$ ($L_2$) is the difference of $\hat{F}$ from $F$ measured by $L_1$-norm ($L_2$-norm) mul-

tiplied by 100.[21] The proposed estimator outperforms Ichimura and Klein and Spady in all simualtion scenarios according to both $L_1$ and $L_2$ metrics. The performance of the PGSIM is not surprizing since, as mentioned above, all that is required from the parametric start is to smooth the object to be estimated nonparametrically and not that the parametric start be correct all the time. This is the strength of the estimator and shows its usefulness in many situations.

We also conducted simulations (results not shown here) where we added an additional variable, considered a index variable, and changed the scale of the coefficient. Not surprisingly, the results are qualitatively same as the results in Tables 1 and 2.

## 5 Application

To illustrate the applicability of the PGSIM estimator, we conduct an empirical analysis to explore the determinants of technology (inorganic fertilizer) adoption among smallholder farmers in Uganda. The dependent variable is a binary decision to apply inorganic fertilizer or not. Besides the fact that the dependent variable lends itself well to the proposed estimator, the rationale for the empirical application is that the adoption of improved production technologies is seen as crucial to increasing yields and thereby reducing food insecurity. The agricultural sector in many Sub-Saharan Africa (SSA) countries is characterized by persistently low crop yields in part because of little reliance on said technologies such as fertilizer and high-yielding seeds (Diiro and Sam 2015). Several statistics illustrate this fact. For instance, an average SSA farmer applies only about 8 kg per hectare of fertilizer compared to 101 kg per hectare in South Asia and over 145 kg per hectare in the developed world; only 28% of the land area allocated to maize in the region is planted with improved maize varieties (see Diiro and Sam 2015). Economists have put forth several theories about the barriers to technology adoption in SSA including capital constraints and uncertainty of agricultural returns (Mishra et al. 2017).

The application extends the empirical study of fertilizer use in Uganda in Diiro et al. (2015) by adding estimates from the PGSIM. The dataset was collected as part of the 2009–2010 Uganda National Household Survey and consists of 1,357 male famers (84% of fertilizer users). We draw from the extensive economics literature on agricultural technology adoption to come up with a set of control variables in the analysis. These are farm size, distance to the nearest trading center to sell crops or buy inputs, access to extension services, age and education of the household head, nonfarm income as proxy for ability to self-finance, household size as proxy for labor force, and dummy variables for the country's agroclecological zones.

The results are presented in Table 3. As can be seen from the results table the three sets of parameter estimates are generally similar in magnitude, sign and statistical significance. There is no standard error for the household size in the nonparametric (KS, PGSIM) models as the coefficient is necessarily restricted to the Probit coefficient. Also, the intercept is not identified in the nonparametric models. There are two notable differences between the models. First, the coefficient on farm size is positive in all

[21] In general, $L_p$-norm is defined as $(E(|x|^p))^{1/p}$.

**Table 3** Determinants of fertilizer adoption by a sample of Ugandan farmers

| Variables | Probit Model | KS Model | PGSIM model |
|---|---|---|---|
| Farm size (hectares) | 0.0430 | 0.0541 | 0.0627** |
| | (0.0395) | (0.0555) | (0.0302) |
| ln(Distance to trading center (km)) | −0.0786 | −0.0838 | −0.0561 |
| | (0.0768) | (0.0975) | (0.0403) |
| No. of Extension visits | 0.0473*** | 0.0595*** | 0.0451** |
| | (0.0159) | (0.0191) | (0.0187) |
| Age of head of household (years) | −0.0111** | −0.0103*** | 0.0014 |
| | (0.0047) | (0.0045) | (0.0010) |
| Average education of household (years) | 0.0006 | −0.0048 | −0.0023 |
| | (0.0121) | (0.0165) | (0.0036) |
| ln(Nonfarm income (US$)) | 0.0629** | 0.0716*** | 0.0521** |
| | (0.0247) | (0.0298) | (0.0247) |
| Central region | −0.1858 | −0.2240 | −0.2220* |
| | (0.1506) | (0.2244) | (0.1220) |
| Western region | −0.2910* | −0.3622 | −0.2596* |
| | (0.1636) | (0.2342) | (0.1485) |
| Northern region | −0.9420*** | −1.0397*** | −0.9352** |
| | (0.2168) | (0.2091) | (0.3489) |
| Household size (No. of Persons) | 0.0231 | 0.0231 | 0.0231 |
| | (0.0191) | N/A | N/A |
| Intercept | −1.2386*** | N/A | N/A |
| | (0.2859) | N/A | N/A |
| Log likelihood | −262.2334 | −281.30 | −276.58 |

* denotes statistical significance at the 10% level; ** denotes statistical significance at the 5% level; *** denotes statistical significance at the 1% level

three models but only statistically significant in PGSIM estimator. This suggests that farmers with bigger farms are more likely to rely on inorganic fertilizer. Second, the age of the household head is statistically significant in the Probit and Klein and Spady models but not in for our proposed model. A priori, it is not clear why the age of the household head should cause an increase or decrease in fertilizer use. On one hand older farmers are more risk-averse and therefore may be less inclined to invest in inorganic fertilizer due to uncertainty of returns. On the other hand, age of the farmer may serve as a proxy for experience which is likely to spur fertilizer use.

## 6 Conclusions

In this article, a new semiparametric estimator for binary-choice single-index models is proposed which uses parametric information in the form of a known (parametric) link function and nonparametrically corrects it. A significant advantage is its ease of

use. Asymptotic properties are derived and an extensive simulation study is conducted to compare the finite sample performances of the new estimator, the semiparametric estimators of Klein and Spady (1993) and Ichimura (1993), and the parametric probit. Results indicate that if the parametric start is correct, the proposed estimator achieves significant bias reduction and efficiency gains. The proposed estimator still achieves significant bias reduction and efficiency gains even if the parametric start is not correct assuming the errors are not heteroskedastic. The performance of the proposed estimator is robust to the misspecification of the parametric guide because all that is required from the guide is to smooth the function to be estimated nonparametrically so as to achieve bias reduction, not that it be correct.

## Appendix

We make the following assumptions:

Assumption 1 Observed sample $(x_i, y_i)$, $i = 1, \ldots, n$ is $i.i.d$.
Assumption 2 $B \subset \mathbb{R}^q$ is compact and the true parameter vector $b_0$ is in the interior of $B$.
Assumption 3 $A_x$ is compact.
Assumption 4 $K(s)$ is a density. Furthermore $\int sK(s)ds = 0$, $K(s) = 0$ for $s < -1$ and $s > 1$, and its second derivative satisfies a Lipschitz condition.
Assumption 5 Parametric start $G$ is uniformly bounded over $x$, $b$ and $G(xb) \neq 0 \; \forall x, b \in A_x \times B$.
Assumption 6 $\int |\phi(t)|dt < \infty$ where $\phi(t)$ is the characteristic function of $K$.
Assumption 7 There exist $\underline{F}$ and $\overline{F}$ that do not depend on $x$ such that $0 < \underline{F} \leqslant F(xb) \leqslant \overline{F} < 1 \quad \forall b \in B$.

*Proof of lemma 1* Our proof of lemma 1 follows closely Bierens (1987a, b) and Pagan and Ullah (1999, pp. 36–39). Note that $\hat{F}$ in (9) can be written as $\hat{F}_1/\hat{F}_2$ where

$$\hat{F}_1 = \frac{1}{(n-1)h} \sum_{j \neq i} y_j \mathbf{1}_{[x_j \in A_{nx}]} \left\{ \frac{G(x_i b)}{G(x_j b)} \right\} K\left( \frac{x_i b - x_j b}{h} \right)$$

$$\hat{F}_2 = \frac{1}{(n-1)h} \sum_{j \neq i} \mathbf{1}_{[x_j \in A_{nx}]} K\left( \frac{x_i b - x_j b}{h} \right).$$

Since $\hat{F}_2$ is $\hat{F}_1$ with $y_j = 1$ and $G(\cdot)$ a constant function, we will only show uniform convergence of $\hat{F}_1$.[22] Now observe that

[22] The derivation below assumes that $(y_i, x_i)$ is absolutely continuous. Obviously in binary-choice models this is not true as $y_i$ is a Bernoulli random variable. We will keep the absolute continuity interpretation as it is more general and give the necessary changes here for the binary-response case. Using a notation similar to Klein and Spady (1993), let $g_x$ be the unconditional density for $x$ and $g_{x|y}$ be the density for $x$ conditional on $y$ for $y = 0, 1$. We have the following series of equalities

$$E(y|x) = F(x) = Pr(y = 1|x) = \frac{Pr(y = 1)g_{x|1}}{g_x} = \frac{g_{1x}}{g_x} = G(x)\frac{g_{1x}/G(x)}{g_x} = G(x)g(x)$$

$$E(y|x) = F(x) = \frac{G(x)\int y\frac{1}{G(x)}f(y,x)dy}{\int f(y,x)dy}.$$

Let $g(x) = \int y\frac{1}{G(x)}f(y,x)dy/\int f(y,x)dy$ and $h(x) = \int f(y,x)dy$. So $F(x) = G(x)g(x)$ and $G(x)g(x)h(x) = G(x)\int y\frac{1}{G(x)}f(y,x)dy$. Thus $\hat{F}_1 = G(x)\hat{g}(x)\hat{h}(x)$. Notice that by assumption 5, $G$ is uniformly bounded and we have $\sup_x |G(x)| = O(1)$. In fact a plausible start is a distribution function in which case $\sup_x |G(x)| = 1$. Hence it suffices to show $\sup_x |\hat{g}(x)\hat{h}(x) - g(x)h(x)| \to 0$. So we have

$$\begin{aligned}\sup_x |\hat{g}(x)\hat{h}(x) - g(x)h(x)| &\leqslant \sup_x |\hat{g}(x)\hat{h}(x) - E\hat{g}(x)\hat{h}(x)| \\ &+ \sup_x |E\hat{g}(x)\hat{h}(x) - g(x)h(x)|. \end{aligned} \tag{11}$$

Like Ichimura (1993), we will refer to the second term of the right-hand side as bias term and show that it converges to 0 at the rate $h^2$. But first notice that

$$\hat{g}(x)\hat{h}(x) = \frac{1}{nh}\sum_{j=1}^{n}\frac{y_j}{G(x_j)}K\left(\frac{x-x_j}{h}\right). \tag{12}$$

From the inversion formula (see Fristedt et al. 1997, p. 231) and by assumption 6 we have

$$K(a) = \frac{1}{2\pi}\int \exp(-ita)\phi(t)dt \tag{13}$$

where $\phi(t)$ is the characteristic function of $K$ and $i^2 = -1$. Using (12) and (13) and letting $s = t/h$ we get

$$\hat{g}(x)\hat{h}(x) = \frac{1}{2\pi}\int\left\{\frac{1}{n}\sum_{j=1}^{n}\frac{y_j}{G(x_j)}\exp(itx_j)\right\}\exp(-itx)\phi(ht)dt. \tag{14}$$

From (14) we get

$$E\hat{g}(x)\hat{h}(x) = \frac{1}{2\pi}\int E\left[\frac{y_j}{G(x_j)}\exp(itx_j)\right]\exp(-itx)\phi(ht)dt. \tag{15}$$

Footnote 22 continued
where $g(x) = (g_{1x}/G(x))/g_x$. Thus $\hat{F}_1 = G(x)\hat{g}(x)\hat{g}_x$ where $\hat{g}(x)\hat{g}_x = ((n-1)h)^{-1}\sum_{j\neq i}\frac{y_j}{G(x_j)}K((x-x_j)/h)$. So there is no change from (11) to (18). In equation (18), we can take an iterated expectation to get $E_X\left[1/G(x_j)K((z-x_j)/h)Pr(y_j = 1|x)\right] = E_X\left[K((z-x_j)/h)g(x)\right] = \int K((z-x)/h)g(x)g_x dx$. And now $1/h$ times this last term would replace equation (19) and we can apply Taylor expansion to $\psi(x) = g(x)g_x$.

From (14) and (15) and noting that $|\exp(-itx)| = 1$

$$|\hat{g}(x)\hat{h}(x) - E\hat{g}(x)\hat{h}(x)| \leqslant \frac{1}{2\pi}\int \left|\frac{1}{n}\sum_{j=1}^{n}\left\{\frac{y_j}{G(x_j)}\exp(itx_j) - E\left[\frac{y_j}{G(x_j)}\exp(itx_j)\right]\right\}\right| |\phi(ht)|dt.$$

So

$$\sup_x |\hat{g}(x)\hat{h}(x) - E\hat{g}(x)\hat{h}(x)| \leqslant \frac{1}{2\pi}\int \left|\frac{1}{n}\sum_{j=1}^{n}\left\{\frac{y_j}{G(x_j)}\exp(itx_j) - E\left[\frac{y_j}{G(x_j)}\exp(itx_j)\right]\right\}\right| |\phi(ht)|dt.$$

Using $\exp(itx_j) = \cos(tx_j) + i\sin(txj)$ we can write

$$\begin{aligned} & E\left|\frac{1}{n}\sum_{j=1}^{n}\left\{\frac{y_j}{G(x_j)}\exp(itx_j) - E\left[\frac{y_j}{G(x_j)}\exp(itx_j)\right]\right\}\right| \\ & = E\Bigg|\underbrace{\frac{1}{n}\sum_{j=1}^{n}\left\{\frac{y_j}{G(x_j)}\cos tx_j - E\left[\frac{y_j}{G(x_j)}\cos tx_j\right]\right\}}_{A} \\ & \quad + \underbrace{i\frac{1}{n}\sum_{j=1}^{n}\left\{\frac{y_j}{G(x_j)}\sin tx_j - E\left[\frac{y_j}{G(x_j)}\sin tx_j\right]\right\}}_{B}\Bigg| \end{aligned} \tag{16}$$

Note that we can write $|A+iB| = (A^2+B^2)^{1/2}$ and so $E|A+iB| = E(A^2+B^2)^{1/2} \leqslant (EA^2+EB^2)^{1/2} = (Var(A)+Var(B))^{1/2}$ where the inequality comes from Jensen's inequality and by construction $EA = 0$ and $EB = 0$. So (16) is

$$\leqslant \left\{Var\left[\frac{1}{n}\sum_{j=1}^{n}\left\{\frac{y_j}{G(x_j)}\cos tx_j - E\left[\frac{y_j}{G(x_j)}\cos tx_j\right]\right\}\right] + Var\left[\frac{1}{n}\sum_{j=1}^{n}\left\{\frac{y_j}{G(x_j)}\sin tx_j - E\left[\frac{y_j}{G(x_j)}\sin tx_j\right]\right\}\right]\right\}^{1/2}$$

$$= \left\{ Var\left[ \frac{1}{n}\sum_{j=1}^{n}\left\{ \frac{y_j}{G(x_j)}\cos tx_j \right\} \right] + Var\left[ \frac{1}{n}\sum_{j=1}^{n}\left\{ \frac{y_j}{G(x_j)}\sin tx_j \right\} \right] \right\}^{1/2}$$
$$= \left\{ \frac{1}{n}\left( Var\left[ \frac{y_j}{G(x_j)}\cos tx_j \right] + Var\left[ \frac{y_j}{G(x_j)}\sin tx_j \right] \right) \right\}^{1/2}.$$

Note that $VarX \leqslant EX^2$ so

$$\leqslant \left\{ \frac{1}{n}\left( E\left[ \left( \frac{y_j}{G(x_j)} \right)^2 \cos^2 tx_j \right] + E\left[ \left( \frac{y_j}{G(x_j)} \right)^2 \sin^2 tx_j \right] \right) \right\}^{1/2}$$
$$= \frac{1}{\sqrt{n}} \left\{ E\left[ \left( \frac{y_j}{G(x_j)} \right)^2 \right] \right\}^{1/2}$$

noting that $\cos^2 tx_j + \sin^2 tx_j = 1$. So

$$E \sup_x |\hat{g}(x)\hat{h}(x) - E\hat{g}(x)\hat{h}(x)| \leqslant \frac{1}{2\pi}\frac{1}{\sqrt{n}} \left\{ E\left[ \left( \frac{y_j}{G(x_j)} \right)^2 \right] \right\}^{1/2} \int |\phi(ht)|dt$$
$$= \frac{1}{2\pi}\frac{1}{h\sqrt{n}} \left\{ E\left[ \left( \frac{y_j}{G(x_j)} \right)^2 \right] \right\}^{1/2} \int |\phi(s)|ds \tag{17}$$

after a change of variables ($s = ht$) and the last term goes to zero as $h\sqrt{n} \to \infty$. Finally using Markov's inequality with (17) we get

$$Pr\left( \sup_x |\hat{g}(x)\hat{h}(x) - E\hat{g}(x)\hat{h}(x)| > \epsilon \right) \to 0 \quad \text{as} \quad n \to \infty.$$

Now for $|E\hat{g}(x)\hat{h}(x) - g(x)h(x)|$ note that

$$E\hat{g}(z)\hat{h}(z) = E\left[ \frac{1}{nh}\sum_{j=1}^{n} y_j \frac{1}{G(x_j)} K\left( \frac{z - x_j}{h} \right) \right]$$
$$= \frac{1}{h} E\left[ y_j \frac{1}{G(x_j)} K\left( \frac{z - x_j}{h} \right) \right] \tag{18}$$
$$= \frac{1}{h} \int K\left( \frac{z - x}{h} \right) \underbrace{\int y \frac{1}{G(x)} f(y, x)dy}_{g(x)h(x)} dx. \tag{19}$$

Now let $\psi(x) = g(x)h(x)$ and $s = (z - x)/h$ for the Taylor expansion

$$\psi(x) = \psi(z - sh) = \psi(z) - hs\psi'(z) + \frac{1}{2}h^2 s^2 \psi''(z) + o(h^2).$$

So

$$\frac{1}{h}\int \left(\psi(z) - hs\psi'(z) + \frac{1}{2}h^2 s^2 \psi''(z)\right) K(s)hds = \psi(z) + \frac{1}{2}h^2 \psi''(z) \int s^2 K(s)ds.$$

Thus we can write

$$\begin{aligned}\sup_x \left|E\hat{\psi}(x) - \psi(x)\right| &= \sup_x \left|\psi(x) + \frac{1}{2}h^2\psi''(x)\int s^2 K(s)ds - \psi(x)\right| \\ &\leqslant \frac{1}{2}h^2 \sup_x |\psi''(x)| \int |s^2 K(s)|ds\end{aligned}$$

so the last term goes to zero at the rate $h^2$. This completes the proof of lemma 1.

*Proof of theorem 1* Let $z$ represent a point on the support of $z_i$. Given the binary nature of the dependent variable $y_i$ and the fact that $P(y_i = 1) = F(z_i)$, we have $E(y_i|z_i) = F(z_i)$. Hence: $y_i = F(z_i) + \epsilon_i$ where $\epsilon$ is an i.i.d error term such that $E(\epsilon_i|z_i = 0)$ and $Var(\epsilon_i|z_i) = F(z_i)(1 - F(z_i)) \ \ \forall i$. In deriving the asymptotic properties of the proposed link function estimator (PGSIM), we require, in addition to Assumptions 1 and 5, the following assumptions:

Assumption 8 The density function of the index $f(z)$ with bounded support $Z$, and the unknown link function $F(z) \in \mathcal{C}^2(\Theta)$ with finite second derivatives and $f(z) \neq 0$ in $\Theta$, the neighborhood of point $z$.

Assumption 9 The Kernel function $K(s)$ is bounded, real-valued, with the following characteristics: (i) $\int K(s)ds = 1$, (ii) $K(s)$ is symmetric about 0, (iii) $\int s^2 K(s)ds < \infty$, (iv) $|s|K(|s|) \to 0$ as $|s| \to \infty$, (v) $\int K^2(s)ds \leq \infty$.

Assumption 10 $h \to 0$ and $nh \to \infty$.

Assumption 11 $E\left|\frac{G(z)}{G(z_i)}\right|^{2+\delta}$, $E|\epsilon_i|^{2+\delta}$, and $\int |K(\omega)|^{2+\delta}$ are finite for some $\delta > 0$.

The nonparametric Kernel estimator of the link function (Ichimura 1993; Klein and Spady 1993) is:

$$\tilde{F}(z) = \frac{\sum_i y_i K_h(z_i - z)}{\sum_i K_h(z_i - z)} \tag{20}$$

Denoting $\mu_2 = \int s^2 K(s)ds$ and $\mathrm{R}(K) = \int K^2(s)dz$, standard properties of $\tilde{F}(z)$ are:

$$E(\tilde{F}(x) - F(z)) = \frac{1}{2}\mu_2 h^2 \left( F''(z) + 2F'(z)\frac{f'(z)}{f(z)} \right) + o_p(h^2), \tag{21}$$

$$\mathrm{Var}(\tilde{F}(z)) = \frac{\sigma^2(z)R(K)}{(nh)f(z)} + O_p(h/n) \tag{22}$$

where $\sigma^2(z) = F(z)(1 - F(z))$.

Now consider our parametrically guided nonparametric estimator of the link function:

$$\hat{F}(z) = \frac{\sum_{i=1}^{n} y_i \left[ \frac{G(z)}{G(z_i)} \right] K_h(z_i - z)}{\sum_{i=1}^{n} K_h(z_i - z)} \tag{23}$$

$$= \frac{n^{-1} \sum_{i=1}^{n} y_i \left[ \frac{G(z)}{G(z_i)} \right] K_h(z_i - z)}{\hat{f}(z)} \tag{24}$$

First, we derive the bias and variance of the proposed estimator. We have

$$\begin{aligned}
\left( \hat{F(z)} - F(z) \right) \hat{f}(z) &= n^{-1} \sum_i K_h(z_i - z) \left[ \frac{G(z)}{G(z_i)} \right] y_i - n^{-1} \sum_i K_h(z_i - z) F(z) \\
&= n^{-1} \sum_i K_h(z_i - z) \left( \left[ \frac{G(z)}{G(z_i)} \right] y_i - F(z) \right) \\
&= n^{-1} \sum_i K_h(z_i - z) \left( G(z) \frac{F(z_i) + \epsilon_i}{G(z_i)} - r(z)G(z) \right) \\
&= n^{-1} G(z) \sum_i K_h(z_i - z) \left( r(z_i) - r(z) \right) \qquad (\ C_n) \\
&+ n^{-1} G(z) \sum_i K_h(z_i - z) \frac{\epsilon_i}{G(z_i)} \qquad (\ D_n)
\end{aligned}$$

Per assumption 1, we have,

$$\begin{aligned}
E(C_n) &= G(z) \int K_h(z_1 - z) \left( r(z_1) - r(z) \right) f(z_1) dz_1 \\
&= G(z) \int K(s) \left( r(z + hs) - r(z) \right) f(z + hs) ds \\
&\quad \text{after a change of variable} \\
&= \frac{h^2}{2} \left( G(z) f(z) r''(z) + 2G(z) f\prime(z) r'(z) \right) \mu_2(K) + o_p(h^2).
\end{aligned}$$

It can be easily seen that $E(D_n) = 0$. Using the fact that $\hat{f}(z) = f(z) + o_p(1)$, we obtain

$$E(\hat{F}(z) - F(z)) = \frac{1}{2}\mu_2 h^2 \left( r''(z) + 2r'(z)\frac{f'(z)}{f(z)} \right) G(z) + o_p(h^2) \qquad (25)$$

Turning to the variance of the estimator, we have

$$\begin{aligned}
Var(C_n) &= Var\left[ n^{-1} G(z) \sum_i^n K_h(z_i - z)(r(z_i) - r(z)) \right] \\
&= (n)^{-2} G^2(z) \left( E\left[ \sum_i^n K_h^2(z_i - z)(r(z_i - r(z))^2 \right] \right. \\
&\quad \left. - \left[ E \sum_i^n K_h(z_i - z)(r(z_i) - r(z)) \right]^2 \right) \\
&= (nh)^{-1} G^2(z) \int K^2(s)(r(z + hs) - r(z))^2 f(z + hs) ds \\
&\quad + \frac{n(n-1)}{n^2} G^2(z) \left[ \int K(s)(r(z + hs) - r(z)) f(z + hs) ds \right]^2 \\
&\quad - G^2(z) \left[ \int K(s)(r(z + hs) - r(z)) f(z + hs) ds \right]^2 \\
&= (nh)^{-1} \int K^2(s)[hs\, r'(z)]^2 f(z) ds + O(n^{-1}) + o_p((nh)^{-1})
\end{aligned}$$

Hence $Var(C_n) = o_p((nh)^{-1})$.

$$\begin{aligned}
Var(D_n) &= E(Var_{z_i}(D_n)) \ \text{ since} E_{z_i}(D_n) = 0, \text{where} Var_{z_i} \\
&\quad \text{denotes the conditional variance of} D_n \\
&= E\left[ n^{-2} G(z)^2 \sum_i K_h^2(z_i - z) G^{-2}(z_i) \sigma^2(z_i) \right] \\
&= (nh)^{-1} \sigma^2(z) G(z)^2 \int K^2(s) ds [G(z)^{-2}(z) f(z)] + o((nh)^{-1}) \\
&= (nh)^{-1} \sigma^2(z) R(K) f(z) + o_p((nh)^{-1})
\end{aligned}$$

Finally,

$$\begin{aligned}
Cov(C_n, D_n) &= E(C_n D_n) \quad \text{since } E(D_n) = 0 \\
&= (n)^{-2} G^2(z) E\left[ \sum_i K_h^2(z_i - z)(r(z_i) - r(z))^2 \epsilon(z_i)^2 \right] \\
&\quad + \frac{n(n-1)}{(n)^2} G^2(z) E\left[ K_h(z_1 - z)(r(z_1 - r(z)) \right] E\left[ K_h(z_1 - z)\epsilon(z_1) \right]
\end{aligned}$$

$$= (nh)^{-1} G^2(z)\sigma^2(z)h^2 r'(z) f(z) \int s^2 K^2(s) ds$$
$$= o_p((nh)^{-1})$$

Piecing the results together, we have

$$Var(\hat{F}(z)) = \frac{\sigma^2(z)R(K)}{(nh)f(z)} + o_p(nh)^{-1}.$$

Note that the variance function is the same as the variance of the Klein and Spady or Ichimura estimator.

Next, we show that the proposed semiparametric link function estimator $\hat{F}(z)$ has a limiting normal distribution

$$\sqrt{nh}(\hat{F}(z) - F(z)) \to \mathcal{N}(B(h), \Sigma) \tag{26}$$

where $B(h) = \frac{1}{2}\mu_2 h^2 \left( r''(z) + 2r'(z)\frac{f'(z)}{f(z)} \right) G(z)$ and $\Sigma = \frac{\sigma^2(z)}{f(z)} R(K)$.

From the results above, it can be seen that

$$C_n = f(z)B(h) + o_p(h^2);$$

We have also shown that $E(D_n) = 0$ and $Var(D_n) = (nh)^{-1} \left( \sigma^2(z)R(K)f(z) + o(1) \right)$. $D_n$ is a triangular array of i.i.d. random variables; thus, under assumption A6,

$$(nh)^{-\delta/2} E \left| \frac{G(z)}{G(z_i)} \right|^{2+\delta} E\,|\epsilon_i|^{2+\delta}\, E\,|K_h(z_i - z)|^{2+\delta}\, h^{-1} \to 0 \quad \text{as} \quad n \to \infty.$$

Hence we can apply Liapounov's central limit theorem to obtain $\sqrt{nh}(D_n) \to \mathcal{N}(0, f^2(x)\Sigma)$. Since $plim\, \hat{f}(z) = f(z)$, it also follows that

$$\sqrt{nh}(\hat{F}(z) - F(z) - B(h)) = \sqrt{nh}\frac{D_n}{f(x)} + o_p(1) \to \mathcal{N}(0, \Sigma). \tag{27}$$

*Proof of theorem 2* Note that

$$\sup_b |\hat{Q}_n(b) - Q_0(b)| \leqslant \sup_b |\hat{Q}_n(b) - Q_n(b)| + \sup_b |Q_n(b) - Q_0(b)|$$

where

$$\hat{Q}_n(b) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{[x_i \in A_x]}(y_i \log[\hat{F}(x_i b)] + (1 - y_i)\log[1 - \hat{F}(x_i b)]),$$

$$Q_n(b) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{[x_i \in A_x]}(y_i \log[F(x_i b)] + (1 - y_i)\log[1 - F(x_i b)]),$$

$$Q_0(b) = \frac{1}{n}\sum_{i=1}^{n} E\left[\mathbf{1}_{[x_i \in A_x]}(y_i \log[F(x_i b)] + (1 - y_i)\log[1 - F(x_i b)])\right].$$

Let $\hat{Q}_{1n} = n^{-1}\sum_{i=1}^{n} \mathbf{1}_{[x_i \in A_x]} y_i \log[\hat{F}(x_i b)]$ and similarly for $Q_{1n}$. Let $\hat{F}_i \equiv \hat{F}(x_i b)$ and similarly for $F_i$. $\hat{Q}_{1n}$ can be viewed as a function of $\hat{F}_i$, so from a Taylor expansion of $\hat{Q}_{1n}$ about $F_i$ we get

$$\begin{aligned} |\hat{Q}_{1n} - Q_{1n}| &= \left| \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{[x_i \in A_x]} y_i \log[F_i] \right. \\ &\quad \left. + \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{[x_i \in A_x]} y_i \frac{1}{\tilde{F}_i}(\hat{F}_i - F_i) - \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{[x_i \in A_x]} y_i \log[F_i] \right| \\ &\leqslant \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{[x_i \in A_x]} y_i \frac{1}{\tilde{F}_i}|\hat{F}_i - F_i| \end{aligned}$$

where $\tilde{F}_i$ is between $\hat{F}_i$ and $F_i$. So we have

$$\sup_b |\hat{Q}_{1n} - Q_{1n}| \leqslant \sup_{i,b} |q_i| \frac{1}{n}\sum_{i=1}^{n} \sup_b |\hat{F}_i - F_i|$$

where $q_i = \mathbf{1}_{[x_i \in A_x]} y_i / \tilde{F}_i$. Note that $\sup_{i,b} |q_i| = O(1)$ and from lemma 1 $\sup_b |\hat{F}_i - F_i| \to 0$ in probability so $\sup_b |\hat{Q}_{1n} - Q_{1n}| = o_p(1)$. A similar result can be obtained for $y_i = 0$ part of the likelihood. Thus we have $\sup_b |\hat{Q}_n(b) - Q_n(b)| = o_p(1)$.

Now let $q_i(x_i, y_i, b) = \mathbf{1}_{[x_i \in A_x]}(y_i \log[F(x_i b)] + (1 - y_i)\log[1 - F(x_i b)])$. So

$$\sup_b |Q_n(b) - Q_0(b)| = \sup_b \left| \frac{1}{n}\sum_{i=1}^{n} (q_i(x_i, y_i, b) - E[q_i(x_i, y_i, b)]) \right|. \tag{28}$$

As Ichimura (1993, p. 91), we can use the uniform law of large numbers by Andrews (1987) and so (28) goes to 0 in probability. This completes the proof of theorem 2.

## References


Andrews DWK (1987) Consistency in nonlinear econometric models: a generic uniform law of large numbers. Econometrica 55:1465–1471

Bierens HJ (1987a) Uniform consistency of Kernel estimators of a regression function under generalized conditions. J Am Stat Assoc 78:699–707

Bierens HJ (1987b) Kernel estimators of regression functions. In: Bewley TF (ed) Advances in econometrics, vol 1. Cambridge University Press, Cambridge

Chen S (2000) Efficient estimation of binary choice models under symmetry. J Econ 96:183–199

Cosslett SR (1987) Efficiency bounds for distribution-free estimators of the binary choice and censored regression models. Econometrica 55:559–586

Diiro G, Sam AG (2015) Agricultural technology adoption and nonfarm earnings in Uganda: a semiparametric analysis. J Dev Areas 49(2):145–62

Diiro G, Ker AP, Sam AG (2015) The Role of gender in fertiliser adoption in Uganda. Afr J Agric Resour Econ 10(2):117–30

Fairlie RW (2005) An extension of the Blinder–Oaxaca decomposition technique to logit and probit models. J Econ Soc Meas 30(4):305–316

Fristedt B, Gray L (1997) A modern approach to probability theory. Birkhäuser, Basel

Frölich M, Huber M, Wiesenfarth M (2017) The finite sample performance of semi-and nonparametric estimators for treatment effects and policy evaluation. Comput Stat Data Anal 115:91–102

Glad IK (1998) Parametrically guided nonparametric regression. Scand J Stat 25:649–668

Hjort NL, Glad IK (1995) Nonparametric density estimation with a parametric start. Ann Stat 23:882–904

Horowitz JL (1992) A smoothed maximum score estimator for the binary response model. Econometrica 60:505–531

Horowitz JL (1993) Semiparametric and nonparametric estimation of quantal response models. In: Maddala GS, Rao CR, Vinod HD (eds) Handbook of statistics, vol 11. North-Holland, Amsterdam

Horowitz JL (1998) Semiparametric methods in econometrics. Springer, Berlin

Horowitz JL, Härdle W (1996) Direct semiparametric estimation of single-index models with discrete covariates. J Am Stat Assoc 91(436):1632–1640

Ichimura H (1993) Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. J Econ 58:71–120

Ichimura H, Lee LF (1991) Semiparametric least squares estimation of multiple index models: single equation estimation. In: Barnett WA, Powell J, Tauchen G (eds) Nonparametric and semiparametric methods in econometrics and statistics. Cambridge University Press, Cambridge

Jones MC, Signorini DF (1997) A comparison of higher-order bias kernel density estimators. J Am Stat Assoc 92:1063–1073

Klein RW, Spady RH (1993) An efficient semiparametric estimator for binary response models. Econometrica 61:387–421

Manski CF (1975) The maximum score estimation of the stochastic utility model of choice. J Econ 3:205–228

Manski CF (1988) Identification of binary response models. J Am Stat Assoc 83:729–738

Mishra K, Sam AG, Miranda MJ (2017) You are approved! Insured loans improve credit access and technology adoption of ghanaian farmers. Working paper, The Ohio State University

Pagan A, Ullah A (1999) Nonparametric econometrics. Cambridge University Press, Cambridge

Powell JL (1994) Estimation of semiparametric models. In: Engle RF, McFadden DL (eds) Handbook of econometrics, vol 4. North-Holland, Amsterdam

Ruud PA (1983) Sufficient conditions for the consistency of maximum likelihood estimation despite misspecification of distribution in multinomial discrete choice models. Econometrica 51:225–228

Sam AG, Jiang GJ (2009) Nonparametric estimation of the short rate diffusion process from a panel of yields. J Financ Quant Anal 44:1197–1230

Sam AG, Ker AP (2006) Nonparametric regression under alternative data environments. Stat Probab Lett 76(10):1037–1046

Schuster E, Yakowitz S (1979) Contributions to the theory of nonparametric regression with application to system identification. Ann Stat 7:139–149

Seah KY, Fesselmeyer E, Le K (2017) Estimating and decomposing changes in the whiteblack homeownership gap from 2005 to 2011. Urban Stud 54(1):119–36

Van Birke MS, Bellegem Van, Keilegom I (2017) Semi-parametric estimation in a single-index model with endogenous variables. Scand J Stat 44(1):168–91
Yatchew A, Griliches Z (1985) Specification error in probit models. Rev Econ Stat 67:134–139