Simple Estimators for Semiparametric Multinomial Choice Models

James L. Powell and Paul A. Ruud
University of California, Berkeley

March 2008

Preliminary and Incomplete

Comments Welcome

Abstract

This paper considers estimation of the coefficients in a semiparametric multinomial choice model with linear indirect utility functions (with common coefficients but differing regressors) and errors that are assumed to be independent of the regressors. This implies that the conditional mean of the vector of dependent indicator variables is a smooth and invertible function of a corresponding vector of linear indices. The estimation method is an extension of an approach proposed by Ahn, Ichimura, and Powell (2004) for monotone single-index regression models to a multi-index setting, estimating the unknown index coefficients (up to scale) by an eigenvector of a matrix defined in terms of a first-step nonparametric estimator of the conditional choice probabilities. Under suitable conditions, the proposed estimator is root-n-consistent and asymptotically normal.

JEL Classification: C24, C14, C13.

1. Introduction

While a large literature exists for estimation of single-index regression and semiparametric binary response models – examples include Ahn, Ichimura, and Powell (1996), Han (1987), Härdle and Stoker (1989), Härdle and Horowitz (1996), Ichimura (1993), Klein and Spady (1993), Manski (1975, 1985), Newey and Ruud (1991), and Powell, Stock and Stoker (1989), among many others – there are fewer results available on estimation of multiple-index regression models and semiparametric multinomial choice models. Lee (1995) constructs a multinomial analogue to Klein and Spady's (1993) estimator of the semiparametric binary choice model, estimating the index coefficients by minimizing a semiparametric "profile likelihood" constructed using nonparametric estimators of the choice probabilities as functions of the indices. Lee demonstrates semiparametric efficiency of the estimator under the "distributional index" restriction that the conditional distribution of the errors depends on the regressors only through the indices. As Thompson (1993) shows, though, the semiparametric efficiency bound for multinomial choice under the assumption of independence of the errors – which coincides with the bound the weaker distributional index restriction for binary choice – differs when the number of choices exceeds two, indicating possible efficiency improvements from the stronger independence restriction. Ruud (2000) shows that the stronger independence restrictions yield choice probabilities that are invertible functions of the indices and whose derivative matrix (with respect to the indices) is symmetric, neither of which needs hold under only the distributional index restriction.

The object of this paper is to construct a computationally-simple estimator of the index coefficients for the multinomial choice model under independence of the errors and regressors. The estimator is a multinomial analogue of that proposed by Ahn, Ichimura, and Powell (1996) for the binary choice model under that restriction, exploiting the invertibility of the choice probabilites in the index functions under this restriction. The parameter vector is identified as the eigenvector corresponding to the unique zero eigenvalue for a matrix defined as the conditional expectation of a quadratic form in differences in regressors for observations with equal choice probability vectors; the coefficient estimator is defined analogously, as the eigenvector of the smallest (in magnitude) eigenvalue of a sample version of the expected matrix quadratic form. In a sense, the proposed estimator is a semiparametric version of Amemiya's (1976) extension of Berkson's (1955) minimum chi-square logit estimator to the multinomial choice model with known error distribution. As for the minimum chi-square estimation

methods, the choice probability vectors are first nonparametrically estimated – here, using general nonparametric regression methods instead of cell means for regressors with finite support – and then the index coefficients are estimated using a (nearly) closed-form second-stage procedure. And the efficient "weight matrix" for the matrix quadratic form defining the estimator is the same as the efficient weight matrix for the Amemiya's (1976) GLS estimator for parametric multinomial choice with grouped data.

2. The Model and Estimator

For the semiparametric *multinomial choice (MNC)* model considered here, the J-dimensional dependent variable $\mathbf{d}_i$ is a vector of indicator variables denoting which of $J+1$ mutually-exclusive and exhaustive alternatives (numbered from $j=0$ to $j=J$) is chosen. Specifically, for individual i, alternative j is assumed to have an unobservable indirect utility y_{ij}^* for that individual, and the alternative with the highest indirect utility is assumed chosen. Thus an individual component d_{ij} of the vector $\mathbf{d}_i$ has the form

$$d_{ij} = 1\{y_{ij}^* \geq y_{ik}^* \text{ for } k = 0, ..., J\}, \tag{2.1}$$

with the convention that $\mathbf{d}_i = \mathbf{0}$ indicates choice of alternative $j=0$. An assumption of joint continuity of the indirect utilities rules out ties (with probability one); in this model, the indirect utilities are further restricted to have the linear form

$$y_{ij}^* = \mathbf{x}_{ij}'\boldsymbol{\beta}_0 + \varepsilon_{ij} \tag{2.2}$$

for $j = 1, ..., J$, where the vector $\boldsymbol{\varepsilon}_i$ of unobserved error terms is assumed to be jointly continuously distributed and independent of the $J \times r$-dimensional matrix of regressors $\mathbf{X}_i$ (whose j^{th} row is $\mathbf{x}_{ij}'$). For alternative $j=0$, the usual normalization $y_{i0}^* = 0$ is imposed.

As Lee (1995) notes, the MNC model with independent errors restricts the conditional choice probabilities to depend upon the regressors only through the vector $\boldsymbol{\mu}_i \equiv \mathbf{X}_i\beta_0$ of linear indices; that is, it takes the form

$$E[\mathbf{d}_i|\mathbf{X}_i] \equiv \mathbf{p}_i = \mathbf{P}(\mathbf{X}_i\boldsymbol{\beta}_0), \tag{2.3}$$

for some unknown function $\mathbf{P}(\cdot)$, so that

$$\mathbf{d}_i = \mathbf{p}_i + \mathbf{u}_i = \mathbf{P}(\mathbf{X}_i\boldsymbol{\beta}_0) + \mathbf{u}_i, \tag{2.4}$$

where $E[\mathbf{u}_i|\mathbf{X}_i] = \mathbf{0}$ by construction. In addition, the assumption of independence of the latent disturbances $\boldsymbol{\varepsilon}_i$ and the regressors $\mathbf{X}_i$ implies that the function $\mathbf{P}(\boldsymbol{\mu})$ is smooth and invertible its argument $\boldsymbol{\mu}$ if $\boldsymbol{\varepsilon}_i$ has nonnegative density (Newey and Ruud, 2007). A weaker condition yielding (2.3) is an assumption that the conditional distribution of $\boldsymbol{\varepsilon}_i$ given $\mathbf{X}_i$ only depends upon the vector of indices $\mathbf{X}_i\boldsymbol{\beta}_0$, but under this restriction the function $\mathbf{P}$ needs not be invertible in its argument, so invertibility would need to be imposed as an additional restriction for the method proposed here to apply. Identification and estimation of an alternative semiparametric multinomial response model – with common covariates but differing coefficient vectors across alternatives – under this weaker "multi-index" restriction was considered in detail by Lee (1995).

For the present model, the coefficient vector $\boldsymbol{\beta}_0$ is clearly only identified up to scale; given such a normalization, though, the parameter vector $\boldsymbol{\beta}_0$ can be identified from inversion of the relation (2.3), provided the matrix $\mathbf{X}_i$ of regressors is sufficiently variable given the vector $\mathbf{X}_i'\boldsymbol{\beta}_0$ of indices, or, equivalently, given the vector $\mathbf{p}_i$ of conditional expectations of $\mathbf{d}_i$ given $\mathbf{X}_i$. Writing

$$\Pi(\mathbf{p}_i) \equiv \mathbf{P}^{-1}(\mathbf{p}_i) = \mathbf{X}_i\boldsymbol{\beta}_0, \tag{2.5}$$

for the inverse relation between $\mathbf{p}_i$ and the linear index vector $\mathbf{X}_i\boldsymbol{\beta}_0$, identification of $\boldsymbol{\beta}_0$ can be based upon the fact that, for values of $\mathbf{p}_i$ that are nearly equal, the corresponding values of $\mathbf{X}_i\boldsymbol{\beta}_0$ will also be nearly equal. That is, following Ahn, Ichimura, and Powell (2004), the vector $\boldsymbol{\beta}_0$ can be identified by matching observations (numbered i and m) with the same conditional expectation $\mathbf{p}_i = \mathbf{p}_m$ but different matrices of regressors. Conditional on

$$\mathbf{p}_i = \mathbf{p}_m, \tag{2.6}$$

relation (2.5) implies that

$$(\mathbf{X}_i - \mathbf{X}_m)\boldsymbol{\beta}_0 = \mathbf{0}. \tag{2.7}$$

It follows that

$$E[\mathbf{L}_{ij}(\mathbf{X}_i - \mathbf{X}_m)|\mathbf{p}_i = \mathbf{p}_m]\boldsymbol{\beta}_0 = \mathbf{0} \tag{2.8}$$

for any random $r \times J$ matrix $\mathbf{L}_{im}$ for which $E||\mathbf{L}_{im}(\mathbf{X}_i - \mathbf{X}_m)||$ exists. A convenient class of such matrices is

$$\mathbf{L}_{im} = (\mathbf{X}_i - \mathbf{X}_m)'\mathbf{W}_{im}, \tag{2.9}$$

for some suitable $J \times J$, nonnegative-definite "weight/trimming" matrix $\mathbf{W}_{im} \equiv \mathbf{W}(\mathbf{X}_i, \mathbf{X}_m)$; this implies that the (identified) $r \times r$ matrix

$$\begin{aligned} \boldsymbol{\Sigma}_0 &\equiv E[(\mathbf{X}_i - \mathbf{X}_m)\mathbf{W}_{im}(\mathbf{X}_i - \mathbf{X}_m)|\mathbf{p}_i = \mathbf{p}_m] \\ &\equiv \lim_{\varepsilon \to 0} E[(\mathbf{X}_i - \mathbf{X}_m)\mathbf{W}_{im}(\mathbf{X}_i - \mathbf{X}_m)| \, \|\mathbf{p}_i = \mathbf{p}_m\| < \varepsilon] \end{aligned} \tag{2.10}$$

has

$$\boldsymbol{\beta}_0' \boldsymbol{\Sigma}_0 \boldsymbol{\beta}_0 = \mathbf{0}, \tag{2.11}$$

that is, $\boldsymbol{\Sigma}_0$ has a zero eigenvalue with corresponding eigenvector equal to the true parameter $\boldsymbol{\beta}_0$. If the matrix $\mathbf{W}_{im}$ is chosen so that the zero eigenvalue of $\boldsymbol{\Sigma}_0$ is unique – which requires a sufficiently rich support of the conditional distribution of $\mathbf{X}_i$ given $\mathbf{X}_i'\boldsymbol{\beta}_0$ – this suffices to identify $\boldsymbol{\beta}_0$ up to scale as the unique solution to (2.11). The weight/trimming matrix might be chosen for technical convenience and/or to improve the asymptotic efficiency of the corresponding estimator of $\boldsymbol{\beta}_0$.

Given a random sample of size n from this model, the preceding identification result can be transformed into an estimation method for $\boldsymbol{\beta}_0$ by first estimating the unobservable conditional expectation terms $\mathbf{p}_i \equiv E[\mathbf{d}_i|\mathbf{X}_i]$ by some nonparametric method and then, as in Ahn, Ichimura, and Powell (2004), estimating a sample analogue to the matrix $\boldsymbol{\Sigma}_0$ using pairs of observations with estimated values $\hat{\mathbf{p}}_i$ of $\mathbf{p}_i$ that were approximately equal. For example, the first-step nonparametric estimator of $\mathbf{p}_i$ may be the familiar kernel regression estimator, which takes the form of a weighted average of the dependent variable,

$$\hat{\mathbf{p}}_i \equiv \frac{\sum_{m=1}^n k_{im} \cdot \mathbf{d}_m}{\sum_{i=1}^n k_{im}}, \tag{2.12}$$

with weights K_{ij} given by

$$k_{im} \equiv k\left(\frac{\mathbf{x}_i - \mathbf{x}_m}{h_1}\right), \tag{2.13}$$

for $\mathbf{x}_i \equiv vec(\mathbf{X}_i')$, $k(\cdot)$ a "kernel" function which tends to zero as the magnitude of its argument increases, and $h_1 \equiv h_{1n}$ a first-step "bandwidth" which is chosen to tend to zero as the sample size n increases. Given this estimator $\hat{\mathbf{p}}_i$ of the conditional mean variable $\mathbf{p}$ – or a nonparametric estimator of $\mathbf{p}_i$ with comparable properties – a second-step estimator of a matrix analogous to $\boldsymbol{\Sigma}_0$ can be

$$\hat{\mathbf{S}} \equiv \left(\begin{array}{c} n \\ 2 \end{array}\right)^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{1}{h_2^J} \, K\left(\frac{\hat{\mathbf{p}}_i - \hat{\mathbf{p}}_j}{h_2}\right) \cdot (\mathbf{X}_i - \mathbf{X}_j)' \mathbf{A}_{ij} (\mathbf{X}_i - \mathbf{X}_j), \tag{2.14}$$

where $K(\cdot)$ is a univariate kernel analogous to k above, $h_2 \equiv h_{2n}$ is a bandwidth sequence for the second-step estimator $\hat{\mathbf{S}}$, and $\mathbf{A}_{ij} = \mathbf{A}(\mathbf{X}_i, \mathbf{X}_j)$ is a $J \times J$, nonnegative-definite "weight/trimming" matrix which is constructed to equal zero for observations where $\hat{\mathbf{p}}_i$ or $\hat{\mathbf{p}}_j$ is imprecise (i.e., where $\mathbf{X}_i$ or $\mathbf{X}_j$ is outside some compact subset of its support). The term $K\left((\hat{\mathbf{p}}_i - \hat{\mathbf{p}}_j)/h_2\right)$ declines to zero as $\hat{\mathbf{p}}_i - \hat{\mathbf{p}}_j$ increases relative to the bandwidth h_2; thus, the conditioning event "$\mathbf{p}_i = \mathbf{p}_j$" in the definition of $\mathbf{\Sigma}_0$ is ultimately imposed as this bandwidth shrinks with the sample size (and the nonparametric estimator of $\mathbf{p}_i$ converges to its true value in probability).

Given the estimator $\hat{\mathbf{S}}$ of $\mathbf{\Sigma}_0$ – which corresponds to a particular structure for the population weight matrix $\mathbf{V}_{im}$ in the definition of $\mathbf{\Sigma}_0$, as discussed below – construction of an estimator of $\boldsymbol{\beta}_0$ follows exactly the same form as in Ahn, Ichimura, and Powell (2004) and Blundell and Powell (2004), exploiting a sample analogue of relation (2.11) based on the eigenvalues and eigenvectors of $\hat{\mathbf{S}}$. Since the estimator $\hat{\mathbf{S}}$ may not be in finite samples if the kernel function $K(\cdot)$ is not constrained to be nonnegative, the estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}_0$ is defined here as the eigenvector for the eigenvalue of $\hat{\mathbf{S}}$ that is closest to zero in magnitude. That is, defining $(\hat{\nu}_1, ..., \hat{\nu}_p)$ to be the p solutions to the determinantal equation

$$|\hat{\mathbf{S}} - \nu \mathbf{I}| = 0, \tag{2.15}$$

the estimator $\hat{\boldsymbol{\beta}}$ is defined as an appropriately-normalized solution to

$$(\hat{\mathbf{S}} - \hat{\nu}\mathbf{I})\hat{\boldsymbol{\beta}} = \mathbf{0}, \tag{2.16}$$

where

$$\hat{\nu} \equiv \arg\min_j \{|\hat{\nu}_j|\}. \tag{2.17}$$

Normalizing the first component of $\boldsymbol{\beta}_0$ to zero, with remaining coefficients defined as -$\boldsymbol{\theta}_0$, i.e.,

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} 1 \\ -\hat{\boldsymbol{\theta}} \end{pmatrix}, \qquad \boldsymbol{\beta}_0 = \begin{pmatrix} 1 \\ -\boldsymbol{\theta}_0 \end{pmatrix}, \tag{2.18}$$

and partitioning $\hat{\mathbf{S}}$ conformably as

$$\hat{\mathbf{S}} \equiv \begin{bmatrix} \hat{\mathbf{S}}_{11} & \hat{\mathbf{S}}_{12} \\ \hat{\mathbf{S}}_{21} & \hat{\mathbf{S}}_{22} \end{bmatrix}, \tag{2.19}$$

the solution to (2.16) takes the form

$$\hat{\boldsymbol{\theta}} \equiv [\hat{\mathbf{S}}_{22} - \hat{\nu}\mathbf{I}]^{-1} \cdot \hat{\mathbf{S}}_{21} \tag{2.20}$$

for this normalization. An alternative "closed form" estimator $\tilde{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_0$ can be defined as

$$\tilde{\theta} \equiv [\hat{\mathbf{S}}_{22}]^{-1} \cdot \hat{\mathbf{S}}_{21}; \tag{2.21}$$

this estimator exploits the fact (to be verified below) that $\hat{\nu}$ tends to zero in probability, since the smallest eigenvalue of the probability limit $\boldsymbol{\Sigma}_0$ of $\hat{\mathbf{S}}$ is zero.

As discussed by Ahn, Ichimura, and Powell (2004), the relation of $\hat{\boldsymbol{\theta}}$ to $\tilde{\boldsymbol{\theta}}$ here is analogous to the relationship between the two classical single-equation estimators for simultaneous equations systems, namely, limited-information maximum likelihood, which has an alternative derivation as a least-variance ratio (LVR) estimator, and two-stage least squares (2SLS), which can be viewed as a modification of LVR which replaces an estimated eigenvalue by its known (zero) probability limit. The analogy to these classical estimators extends to the asymptotic distribution theory for $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$, which, under the conditions imposed below, will be asymptotically equivalent, like LVR and 2SLS. Their relative advantages and disadvantages are also analogous – e.g., $\tilde{\boldsymbol{\theta}}$ is slightly easier to compute, while $\hat{\boldsymbol{\theta}}$ will be equivariant with respect to choice of which (nonzero) component of $\boldsymbol{\beta}_0$ to normalize to unity. As noted in the introduction, both estimators are semiparametric analogues to the minimum chi-square estimators of Berkson (1955) and Amemiya (1976), using a particular semilinear regression estimator applied to (2.5), which treats the $\boldsymbol{\Pi}(\cdot)$ function as the unknown nonparametric component and the first component of $\mathbf{X}_i$ (with coefficient normalized to unity) as the dependent variable. The proposed estimator will be easier to calculate than Lee's (1995) estimator for the multinomial response model under index restrictions, which requires solution of a r-dimensional minimization problem with a criterion involving simultaneous estimation of J nonparametric regressions (with the J index functions as arguments) and minimization over the parameter vector $\boldsymbol{\theta}$.. Unlike the estimator proposed here, however, Lee's estimator does not impose invertibility of the vector of choice probabilities in the vector of indices.

3. Large Sample Properties of the Estimator

Since the definition of the estimator $\hat{\boldsymbol{\beta}} = (1, -\hat{\boldsymbol{\theta}}')'$ is based on the same form of a "pairwise difference" matrix estimator $\hat{\mathbf{S}}$ analyzed in Ahn, Ichimura, and Powell (2004) and Blundell and Powell (2004), the regularity conditions imposed here will be quite similar to those imposed in these earlier papers. Rather than restrict the first-stage nonparametric estimator of the choice probabilities $\mathbf{p}(\mathbf{X}_i)$ to have a

particular form (e.g., the kernel estimator in (2.12)„ it is assumed to satisfy some higher-level restrictions on its rate of convergence and Bahadur representation which would need to be verified for the particular nonparametric estimation method utilized. Specifically, given a random sample of size n for $\{\mathbf{d}_i, \mathbf{X}_i\}$, it is assumed that the first-stage estimator $\hat{\mathbf{p}}(\mathbf{X}_i)$ has a relatively high convergence rate and the same asymptotic linear representation as for a kernel estimator. That is, defining the "trimming" indicator

$$t_i \equiv 1\{||\mathbf{A}(\mathbf{X}_i, \cdot)|| \neq \mathbf{0}\}, \tag{3.1}$$

the condition

$$\max_i t_i ||\hat{\mathbf{p}}(\mathbf{X}_i) - \mathbf{p}(\mathbf{X}_i)|| = o_p(n^{-3/8}) \tag{3.2}$$

is imposed. This is a restriction on both the nonparametric estimation and the construction of the weight-matrix function $\mathbf{A}(\mathbf{X}_i, \mathbf{X}_m)$, which will generally requiring "trimming" of observations outside a bounded set of $\mathbf{X}_i$ values to ensure that (3.2) is satisfied. Another restriction on the model is that the first column $\mathbf{x}_{i1}$ of the matrix of the regressors $\mathbf{X}_i$ is continuously distributed, with positive density, conditionally on the remaining components, and that the corresponding coefficient $\boldsymbol{\beta}_{0,1}$ is nonzero (and normalized to unity); as discussed by Lee (1995), this restriction helps ensure that the parameters are identified by ensuring that $\mathbf{X}_i$ is sufficiently variable conditional upon a given value of the choice probability vector $\mathbf{p}_i = \mathbf{P}(\mathbf{X}_i\boldsymbol{\beta}_0)$. Other conditions are imposed on the error distribution, kernel function K, and bandwidth h_2; a list of regularity conditions, which are similar to the conditions imposed in Ahn, Ichimura, and Powell (2004) and single-index regression papers, are given in the appendix below.

Under the assumptions in the appendix, consistency of the estimator $\hat{\mathbf{S}}$ for a particular matrix $\boldsymbol{\Sigma}_0$ can be established. That is,

$$\hat{\mathbf{S}} \xrightarrow{p} \boldsymbol{\Sigma}_0, \tag{3.3}$$

where $\boldsymbol{\Sigma}_0$ is of the form given in (2.10) with

$$\mathbf{W}_{im} = \mathbf{A}_{im}\sqrt{\phi(\mathbf{p}_i)\phi(\mathbf{p}_m)}, \tag{3.4}$$

for $\phi(\mathbf{p})$ the density function of the choice probability vector $\mathbf{p}_i = \mathbf{P}(\mathbf{X}_i\boldsymbol{\beta}_0)$. This, with the identification restriction that the true coefficient vector $\boldsymbol{\beta}_0$ is the unique solution of (2.11), implies consistency of the corresponding estimator $\hat{\boldsymbol{\beta}}$ up to scale. The regularity conditions also yield the following asymptotically-linear representation for $\hat{\mathbf{S}}\boldsymbol{\beta}_0$:

$$\sqrt{n}\hat{\mathbf{S}}\boldsymbol{\beta}_0 = \frac{1}{\sqrt{n}}\sum_{\iota=1}^{N}\phi(\mathbf{p}_i)[\mathbf{T}_i\mathbf{X}_i - \mathbf{M}_i]'\left[\frac{\partial\mathbf{P}(\mathbf{X}_i\boldsymbol{\beta}_0)}{\partial\boldsymbol{\mu}'}\right]^{-1}(\mathbf{d}_i - \mathbf{P}(\mathbf{X}_i\boldsymbol{\beta}_0)) + o_p(1), \tag{3.5}$$

for

$$\mathbf{T}_i \equiv E\left[\mathbf{A}_{im}|\mathbf{X}_i, \mathbf{p}_m = \mathbf{p}_i\right], \tag{3.6}$$

and

$$\mathbf{M}_i \equiv E\left[\mathbf{A}_{im}\mathbf{X}_m|\mathbf{X}_i, \mathbf{p}_m = \mathbf{p}_i\right]. \tag{3.7}$$

The terms in the normalized average in expression (3.5) have zero mean and finite variance, so the Lindeberg-Levy central limit theorem implies that $\hat{\mathbf{S}}\,\boldsymbol{\beta}_0$ is asymptotically normal; however, this asymptotic distribution will be singular, since

$$[\mathbf{T}_i\mathbf{X}_i - \mathbf{M}_i]\boldsymbol{\beta}_0 = E[\mathbf{A}_{im}|\mathbf{X}_i][\mathbf{X}_i\boldsymbol{\beta}_0 - E\left(\mathbf{X}_m\boldsymbol{\beta}_0|\mathbf{X}_i, \mathbf{X}_m\boldsymbol{\beta}_0 = \mathbf{X}_i\boldsymbol{\beta}_0\right)] = \mathbf{0}. \tag{3.8}$$

It follows that

$$\sqrt{n}\boldsymbol{\beta}_0'\hat{\mathbf{S}}\,\boldsymbol{\beta}_0 = o_p(1), \tag{3.9}$$

which further implies that the smallest (in magnitude) eigenvalue $\hat{\nu}$ converges in probability to zero faster than the square root of the sample size, because

$$\sqrt{n}\,|\hat{\nu}| = \sqrt{n}\,\min_{\alpha\neq 0}|\boldsymbol{\alpha}'\hat{\mathbf{S}}\,\boldsymbol{\alpha}|/\|\boldsymbol{\alpha}\|^2 \leq \sqrt{n}\,|\boldsymbol{\beta}_0'\hat{\mathbf{S}}\,\boldsymbol{\beta}_0|/\|\boldsymbol{\beta}_0\|^2 = o_p(1). \tag{3.10}$$

To derive the asymptotic distribution of the proposed estimator $\hat{\boldsymbol{\theta}}$ of (2.20), the normalized difference of $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$ can be decomposed as

$$\begin{aligned}\sqrt{n}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) &= [\hat{\mathbf{S}}_{22} - \hat{\nu}\mathbf{I}]^{-1}\sqrt{n}\,[\hat{\mathbf{S}}_{12} - (\hat{\mathbf{S}}_{22} - \hat{\nu}\mathbf{I})\,\boldsymbol{\theta}_0] \\ &= [\hat{\mathbf{S}}_{22} - \hat{\nu}\mathbf{I}]^{-1}\sqrt{n}\,\hat{\mathbf{s}} - \sqrt{n}\,\hat{\nu}\,[\hat{\mathbf{S}}_{22} - \hat{\nu}\mathbf{I}]^{-1}\boldsymbol{\theta}_0,\end{aligned} \tag{3.11}$$

where

$$\hat{\mathbf{s}} \equiv \hat{\mathbf{S}}_{12} - \hat{\mathbf{S}}_{22}\boldsymbol{\theta}_0 \equiv [\hat{\mathbf{S}}\,\boldsymbol{\beta}_0]_2, \tag{3.12}$$

i.e., $\hat{\mathbf{s}}$ is the subvector of $\hat{\mathbf{S}}\boldsymbol{\beta}_0$ corresponding to the free coefficients $\boldsymbol{\theta}_0$. Using conditions (3.10), the same arguments as in Ahn and Powell (1993) yield

$$\hat{\mathbf{S}}_{22} - \hat{\nu}\,\mathbf{I} \rightarrow^p \boldsymbol{\Sigma}_{22}, \tag{3.13}$$

where $\mathbf{\Sigma}_{22}$ is the lower $(r-1)\times(r-1)$ diagonal submatrix of $\mathbf{\Sigma}_0$, and also

$$\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)=[\mathbf{\Sigma}_{22}]^{-1}\sqrt{n}\,\hat{\mathbf{s}}+o(1), \tag{3.14}$$

from which the consistency and asymptotic normality of $\hat{\boldsymbol{\theta}}$ follow from the asymptotic normality of $\hat{\mathbf{S}}\,\boldsymbol{\beta}_0$. Specifically,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\rightarrow^{d}\mathcal{N}(\mathbf{0},\mathbf{\Sigma}_{22}^{-1}\mathbf{\Omega}_{22}\mathbf{\Sigma}_{22}^{-1}),$$

where $\mathbf{\Omega}_{22}$ is the lower $(r-1)\times(r-1)$ diagonal submatrix of

$$\mathbf{\Omega}\equiv E[\boldsymbol{\psi}_i\boldsymbol{\psi}_i'], \tag{3.15}$$

where

$$\boldsymbol{\psi}_i\equiv\phi(\mathbf{p}_i)[\mathbf{T}_i\mathbf{X}_i-\mathbf{M}_i]'\left[\frac{\partial\mathbf{P}(\mathbf{X}_i\boldsymbol{\beta}_0)}{\partial\boldsymbol{\mu}'}\right]^{-1}(\mathbf{d}_i-\mathbf{P}(\mathbf{X}_i\boldsymbol{\beta}_0))\,, \tag{3.16}$$

A similar argument yields the asymptotic equivalence of the "closed form" estimator $\tilde{\boldsymbol{\theta}}$ of (2.21) and the estimator $\hat{\boldsymbol{\theta}}$, since

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}}) &= \sqrt{n}([\hat{\mathbf{S}}_{22}-\hat{\nu}\,\mathbf{I}]^{-1}-[\hat{\mathbf{S}}_{22}]^{-1})\hat{\mathbf{S}}_{12}\\ &= -\sqrt{n}\,\tilde{\nu}\,[\hat{\mathbf{S}}_{22}-\hat{\nu}\,\mathbf{I}]^{-1}\hat{\theta}\\ &= o_p(1)\end{aligned} \tag{3.17}$$

for $\tilde{\nu}$ an intermediate value between $\hat{\nu}$ and zero.

A final requirement for conducting the usual large-sample normal inference procedures is a consistent estimator of the asymptotic covariance matrix $\mathbf{\Sigma}_{22}^{-1}\mathbf{\Omega}\mathbf{\Sigma}_{22}^{-1}$ of $\hat{\theta}$. Estimation of $\mathbf{\Sigma}_{22}^{-1}$ is straightforward; by the results given above, either $[\hat{\mathbf{S}}_{22}-\hat{\nu}\mathbf{I}]^{-1}$ or $[\hat{\mathbf{S}}_{22}^{-1}]^{-1}$ will be consistent, with the former being more natural for $\hat{\theta}$ and the latter for $\tilde{\theta}$. Consistent estimation of the matrix $\mathbf{\Omega}$ is less straightforward; given a suitably-consistent estimator $\hat{\boldsymbol{\psi}}_i$ of the influence function term $\boldsymbol{\psi}_i$ of (3.16), a corresponding estimator of $\mathbf{\Omega}$ could be constructed as

$$\hat{\mathbf{\Omega}}\equiv\frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\psi}}_i\hat{\boldsymbol{\psi}}_i'. \tag{3.18}$$

Based on a similar Taylor's series argument as in Ahn and Powell (1993), a candidate for such an influence function estimator would be

$$\hat{\boldsymbol{\psi}}_i\equiv\frac{2}{n-1}\sum_{i=1}^{n}\frac{1}{h_2^{J+1}}\;D\left(\frac{\hat{\mathbf{p}}_i-\hat{\mathbf{p}}_j}{h_2}\right)(\mathbf{d}_i-\hat{\mathbf{p}}_j)\cdot(\mathbf{X}_i-\mathbf{X}_j)'\mathbf{A}_{ij}(\mathbf{X}_i-\mathbf{X}_j)\hat{\boldsymbol{\beta}}, \tag{3.19}$$

with $D(\mathbf{u}) \equiv \partial K(\mathbf{u})/\partial \mathbf{u}'$. Verification of the consistency of $\hat{\mathbf{\Omega}}$ under the conditions imposed below is a topic of ongoing research.

4. The Ideal Weight Matrix

The results of the preceding section raise the question of the optimal choice of weight matrix $\mathbf{A}_{ij}$ to be used in the construction of $\hat{\mathbf{S}}$ in (2.14) above. The usual efficiency arguments suggest that the optimal choice would yield equality of the matrices $\mathbf{\Sigma}_{22}$ and $\mathbf{\Omega}_{22}$ characterizing the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$. This equality could be obtained using the following infeasible weight matrix:

$$\mathbf{A}^*_{ij} = [\mathbf{H}^*_i]' \, \mathbf{H}^*_j, \tag{4.1}$$

where

$$\mathbf{H}^*_i \equiv \frac{1}{\sqrt{\phi(\mathbf{p}_i)}} \mathbf{V}[\mathbf{d}_i|\mathbf{p}_i]^{-1/2} \left[\frac{\partial \mathbf{P}(\mathbf{X}_i \boldsymbol{\beta}_0)}{\partial \boldsymbol{\mu}'} \right]. \tag{4.2}$$

The corresponding matrix $\mathbf{W}^*_{ij}$ in the definition (2.10) of $\mathbf{\Sigma}_0$ would have

$$\mathbf{W}^*_{ij} = \left[\frac{\partial \mathbf{P}(\mathbf{X}_i \boldsymbol{\beta}_0)}{\partial \boldsymbol{\mu}'} \right]' \mathbf{V}[\mathbf{d}_i|\mathbf{p}_i]^{-1} \left[\frac{\partial \mathbf{P}(\mathbf{X}_i \boldsymbol{\beta}_0)}{\partial \boldsymbol{\mu}'} \right] = \mathbf{W}^*_{ii} \tag{4.3}$$

when $\mathbf{p}_i = \mathbf{p}_j$, and the asymptotic distribution of the estimator $\hat{\boldsymbol{\theta}}^*$ (which uses the infeasible $\mathbf{A}^*_{ij}$ as weight matrix) would be

$$\sqrt{n}(\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_0) \rightarrow^d \mathcal{N}(\mathbf{0}, [\mathbf{\Omega}^*_{22}]^{-1}), \tag{4.4}$$

with

$$\mathbf{\Omega}^* \equiv \mathbf{Var} \left[(\mathbf{X}_i - E[\mathbf{X}_i|\mathbf{p}_i])' \left[\frac{\partial \mathbf{P}(\mathbf{X}_i \boldsymbol{\beta}_0)}{\partial \boldsymbol{\mu}'} \right]' (\mathbf{d}_i - \mathbf{p}_i) \right]. \tag{4.5}$$

Lee (1995) shows that $\mathbf{\Omega}^*_{22}$ is the semiparametric analogue of the information matrix for estimation of $\boldsymbol{\theta}_0$ when only the index restrictions (2.3) are imposed, which implies that $\hat{\boldsymbol{\theta}}^*$ achieves the semiparametric efficiency bound for those restrictions. However, the argument for consistency of $\hat{\boldsymbol{\theta}}^*$ exploits the additional restriction of invertibility of the conditional choice probabilites $\mathbf{P}(\mathbf{X}_i \boldsymbol{\beta}_0)$ in the vector of indices $\mathbf{X}_i \boldsymbol{\beta}_0$, which is not required for consistency of Lee's (1995) semiparametric maximum likelihood estimator.

There is a close connection between the estimation approach proposed here and the "minimum chi-squared logit" estimator proposed for the binary logit by Berkson (1955), and extended to general

parametric multinomial choice models by Amemiya (1976). For that latter estimator, the matrix of regressors $\mathbf{X}_i$ is constant within each of a fixed number of groups, and a nonparametric estimator of the vector of choice probabilities $\mathbf{p}_i$ for each group is constructed from the observed choice frequencies for each group. Rewriting the relation (2.5) as

$$\Pi(\hat{\mathbf{p}}_i) \equiv \mathbf{P}^{-1}(\hat{\mathbf{p}}_i) = \mathbf{X}_i\boldsymbol{\beta}_0 + \mathbf{v}_i, \tag{4.6}$$

with

$$\mathbf{v}_i \equiv \mathbf{P}^{-1}(\hat{\mathbf{p}}_i) - \mathbf{P}^{-1}(\mathbf{p}_i),$$

Amemiya (1976) shows that an efficient estimator of $\boldsymbol{\beta}_0$ for this setup is the coefficient vector of the generalized least squares regression of $\boldsymbol{\Pi}(\hat{\mathbf{p}}_i)$ on $\mathbf{X}_i$, using the matrix $\mathbf{W}_{ii}$ (or a feasible version $\hat{\mathbf{W}}_{ii}$, replacing the unknown probabilities $\mathbf{p}_i$ by their consistent estimators $\hat{\mathbf{p}}_i$) as the weighting matrix. The estimator $\hat{\boldsymbol{\theta}}^*$ is a semiparametric analogue of this minimum chi-squared estimator.

The weight matrix $\mathbf{A}^*_{ij}$ is infeasible in two respects – it does not satisfy the trimming requirement to achieve the uniform nonparametric rate of convergence of $\hat{\mathbf{p}}_i$ to $\mathbf{p}_i$ specified in (3.2), and it involves unknown nuisance parameter functions (the density of the choice probabilities and their derivatives with respect to the indices and the conditional covariance matrix of the choice indicators $\mathbf{d}_i$ given $\mathbf{p}_i$). Construction of a feasible version $\hat{\mathbf{A}}_{ij}$ of $\mathbf{A}^*_{ij}$ that achieves the same asymptotic distribution is another topic of ongoing research.

5. APPENDIX:

[To be completed.]

6. References

Ahn, H., H. Ichimura, and J.L. Powell, 1996, "Simple Estimators for Monotone Index Models," manuscript, Department of Economics, University of California at Berkeley.

Ahn, H. and J.L. Powell, 1993, "Semiparametric estimation of censored selection models with a nonparametric selection mechanism," *Journal of Econometrics*, 58, 3-29.

Amemiya, T., 1976, "The maximum likelihood, the minimum chi-square, and the nonlinear weighted least-squares estimator in the general qualitative response model," *Journal of the American Statistical Association,* 71, 347-351.

Berkson, J., 1955, "Maximum likelihood and minimum χ^2 estimates of the logistic function," *Journal of the American Statistical Association,* 50, 132-162.

Han, A.K., 1987, "Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator," *Journal of Econometrics* 35, 303-316.

Härdle, W. and T.M. Stoker, 1989, "Investigating smooth multiple regression by the method of average derivatives," *Journal of the American Statistical Association*, 84, 986-995.

Härdle, W. and J. Horowitz, 1996, "Direct Semiparametric Estimation of Single-Index Models With Discrete Covariates," *Journal of the American Statistical Association*, 91, 1632-1640.

Ichimura, H., 1993, "Semiparametric least squares (SLS) and weighted SLS estimation of single-index models," *Journal of Econometrics*, 58, 71-120.

Ichimura, H. and L-F. Lee, 1991, "Semiparametric least squares estimation of multiple index models: single equation estimation," in: W.A. Barnett, J.L. Powell, and G. Tauchen, eds., *Nonparametric and semiparametric methods in econometrics and statistics*, Cambridge: Cambridge University Press.

Klein, R.W. and R.S. Spady, 1993, "An efficient semiparametric estimator of the binary response model," *Econometrica*, 61, 387-422.

Lee, L-F., 1995, "Semiparametric maximum likelihood estimation of polychotomous and sequential choice models", *Journal of Econometrics,* 65, 385-428.

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

Manski, C.F., 1985, "Semiparametric Analysis of discrete response, asymptotic properties of the maximum score estimator," *Journal of Econometrics*, 27, 205-228.

Newey, W.K. and P.A. Ruud, 1991, "Density weighted least squares estimation," manuscript, Department of Economics, U.C. Berkeley.

Newey, W.K. and T.M. Stoker, 1993 "Efficiency of weighted average derivative estimators and index models," *Econometrica,* 61, 1199-1223.

Powell, J.L., J.H. Stock and T.M. Stoker, 1989, "Semiparametric estimation of weighted average derivatives," *Econometrica* 57, 1403-1430.

Ruud, P.A., 1986, "Consistent estimation of limited dependent variable models despite misspecification of distribution," *Journal of Econometrics*, 32, 157-187.

Ruud, P.A., 2000, "Semiparametric estimation of discrete choice models," manuscript, Department of Economics, University of California at Berkeley.

Thompson, T.S. , 1993, "Some efficiency bounds for semiparametric discrete choice models," *Journal of Econometrics,* 58, 257–274