Entropy-Based Estimation and Inference in Binary Response Models under Endogeneity

Douglas J. Miller, Ron C. Mittelhammer, and George G. Judge
University of California, Washington State University, and Purdue University

Selected Paper prepared for presentation at the American Agricultural Economics Association Annual Meeting, Denver, Colorado, August 1-4, 2004

Copyright 2004 by Douglas J. Miller, Ron C. Mittelhammer, and George G. Judge. All rights reserved. Readers may make verbatim copies of this document for non-commercial purposes by any means, provided that this copyright notice appears on all such copies.

Abstract

This paper considers estimation and inference for the binary response model in the case where endogenous variables are included as arguments of the unknown link function. Semiparametric estimators are proposed that avoid the parametric assumptions underlying the likelihood approach as well as the loss of precision when using nonparametric estimation. Suggestions are made for how the utility maximization decision model can be altered to permit attributes to vary across alternatives.

Keywords: multinomial process, endogeneity, empirical likelihood procedures, semiparametric estimation and inference, quasi-likelihood estimation.

1. Introduction

In this paper, we consider conventional estimators of latent variables models typically are based on strong assumptions involving a *particular* finitely parameterized error distribution specification. Economic theories that motivate these models and estimators rarely, if ever, justify such restrictions on the error specification. This uncertainty regarding the specification of the data sampling process implies that, in reality, a broad range of statistical models and estimators should not logically be ruled out as potential generators of the observed data. Within the context of this challenging model specification scenario, in this paper we consider the case of a multinomial response model involving endogenous covariates as arguments in the unknown link function. To recover the unknown response parameters and marginal probabilities, we demonstrate a semiparametric estimator that avoids many of the assumptions of the likelihood approach and the loss of precision that occurs in fully nonparametric estimation.

1.1 Some Background

In the context of multinomial response models, assume that on trial $i = 1, 2, \ldots, n$, one of $j = 1, 2, \ldots, J$ alternatives is observed to occur among the binary random variables $\{y_{i1}, ..., y_{iJ}\}$ having p_{ij}, , as their respective probabilities of success. Assume further that the p_{ij}'s are related to a set of k covariates through link functions of the form $G_j(\mathbf{x}_i, \boldsymbol{\beta})$, where the vector $\mathbf{n}_i$ contains attributes of the decision maker and/or the

alternatives, $\boldsymbol{\beta}$ is a vector of unknown parameters, and $G_j : \mathbb{R} \rightarrow [0,1]$ may be either known or unknown. The data sampling process is represented as

$$y_{ij} = p_{ij} + \varepsilon_{ij} = G_j(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_{ij} \tag{1.1}$$

where the ε_{ij} are unobservable independent noise components and $E[y_{ij} \mid \mathbf{x}_i] = G_j(\mathbf{x}_i, \boldsymbol{\beta})$.

In those rare instances where the parametric functional form of $G_j(\mathbf{x}_i, \boldsymbol{\beta})$ and the parametric family of probability density functions underlying the decision model are known, one can use the traditional maximum likelihood (ML) approach and the multinomial log-likelihood function

$$L(\boldsymbol{\beta}; \mathbf{y}) = \sum_i \sum_j \left(y_{ij} \ln G_j(\mathbf{x}_i, \boldsymbol{\beta}) \right) \tag{1.2}$$

to obtain estimates of the parameters of the model. Depending on the specific parametric family of distributions assumed for the noise term of latent variables that govern the decision process (discussed in section 2 ahead), logit, probit, or other formulations arise. Whatever the distribution underlying the likelihood specification, if the choice of distribution happens to be correct, then the usual properties of ML estimation hold including consistency, asymptotic normality and efficiency. However, if these conditions do not hold, then standard ML estimating procedures do not attain their usual attractive sampling properties. For detailed discussions concerning these types of models, see Maddala (1983) and McCullough and Nelder (1995).

Several estimating procedures for $\boldsymbol{\beta}$ that do not require a parametric formulation for the G_j's exist. For example, Ichimura (1993) demonstrates a least squares estimator of $\boldsymbol{\beta}$, and Klein and Spady (1993) demonstrate a quasi-maximum likelihood estimator when y_{ij} is binary. These estimates are consistent and asymptotically normal under their prescribed regularity conditions. Unfortunately, they involve nonlinear optimization problems whose solutions are difficult to compute. Using an information theoretic formulation, Golan, Judge, and Perloff (1996) demonstrate a semiparametric estimator for the traditional multinomial response problem that has asymptotic properties in line with parametric counterparts. In terms of multinomial problems with endogenous explanatory variables the formulations of Newey (1986, 1987) and Blundell and Powell (1999) are important examples.

Building on these productive efforts, in this paper we seek a semiparametric basis for recovering $\boldsymbol{\beta}$ in (1.1) when the functional form of the link functions $G_j(\mathbf{x}_i, \boldsymbol{\beta})$ is unknown and the covariates in the untransformed structural model contain endogenous or random components such that $E[\mathbf{x}_i \varepsilon_{ij}] \neq \mathbf{0}$. In this context, one objective is to demonstrate an estimator that avoids many of the assumptions of the likelihood approach and permits us to cope with endogeneity-measurement error problems that often arise in practice.

1.2 The Format

In Section 2, we define a particular multinomial response model that reflects the endogenous nature of the sampling process, formulate a semiparametric estimation procedure in the form of an extremum problem, and provide a solution to the semiparametric estimation problem that has the sampling properties of consistency and asymptotic normality. In Section 3 we discuss alternative multinomial response model formulations and indicate corresponding semiparametric estimation methods. Finally, in Section 4 the estimation and inference implications of our proposed models are summarized.

2. A Multinomial Response Model and a Semiparametric Solution

Assume the multinomial response model

$$\begin{aligned} y_{1ij} &= \prod_{k \neq j} I_{(0,\infty)}\left(y^*_{1ij} - y^*_{1ik}\right) \\ &= 1 \text{ iff } y^*_{1ij} > y^*_{1ik}, \forall k \neq j \end{aligned} \tag{2.1}$$

where the latent variable y^*_{1ij} is assumed to be generated from the linear model

$$y^*_{1ij} = \mathbf{x}_i' \boldsymbol{\beta}_j + u_{ij} \quad , \tag{2.2}$$

$\mathbf{x}_i$ is now a $(k \times 1)$ vector of explanatory covariates over $i = 1, 2, \ldots, n$ observations relating to decision maker attributes, u_{ij} is an unobservable noise component, and $I_{(0,\infty)}(\upsilon)$ is a standard indicator function that takes the value one if $\upsilon \in (0, \infty)$ and equals zero otherwise. This particular multinomial formulation is based explicitly on the

decision maker's attributes represented by $\mathbf{x}_i$, i = 1, ... , n, which clearly do not vary across the J alternatives. The decision maker attributes are translated into a utility index via alternative-specific $\boldsymbol{\beta}_j$'s that indicates how attributes specific to the decision maker affect the rankings for each of the J alternatives. In this formulation, the utility index associated with alternative j, conditional on a decision-maker's attributes, is given by $\mathbf{x}_i'\boldsymbol{\beta}_j$, for each j, apart from random noise in the random utility framework. The formulation suppresses any explicit *alternative-specific attributes*.

To characterize in an expository manner a situation that is consistent with the covariate endogeneity or measurement error problem, assume that $\mathbf{x}_i' = \left[\mathbf{z}_{1i}', y_{2i}\right]$ is a row vector of dimension $m_1 + 1 = k$, $\mathbf{z}_{1i}$ contains a fixed set of exogenous covariates, and y_{2i} is an endogenous random variable where $E\left[y_{2i}u_{ij}\right] \neq 0$. We rewrite (2.2) as the structural equation,

$$y_{1ij}^* = \mathbf{z}_{1i}'\boldsymbol{\beta}_{1j} + y_{2i}\beta_{2j} + u_{ij} \tag{2.3}$$

where y_{1ij} and y_{2i} are jointly determined random variables. To close the system, we define

$$y_{2i} = \mathbf{z}_{1i}'\boldsymbol{\pi}_1 + \mathbf{z}_{2i}'\boldsymbol{\pi}_2 + v_i = \mathbf{z}_i'\boldsymbol{\pi} + v_i \tag{2.4}$$

where $\mathbf{z}_i = \left[\mathbf{z}_{1i}', \mathbf{z}_{2i}'\right]'$ is a column vector of dimension $\left(m_1 + m_2 = m\right)$, $m_1 \geq 1$, and $E\left[\mathbf{z}_i v_i\right] = \mathbf{0}$. Rewriting the structural equation (2.2) in reduced form results in

$$y_{1ij}^* = \mathbf{z}_{1i}'\boldsymbol{\beta}_{1j} + \mathbf{z}_i'\boldsymbol{\pi}\beta_{2j} + v_i\beta_{2j} + u_{ij} = \mathbf{z}_{1i}'\boldsymbol{\beta}_{1j} + \mathbf{z}_i'\boldsymbol{\pi}\beta_{2j} + \nu_{ij}^* \tag{2.5}$$

where $\nu_{ij}^* = v_i\beta_{2j} + u_{ij}$ is a reduced form error term, for $j = 1, 2, \ldots, J$. Since $\boldsymbol{\pi}$ is unknown, we replace it by a consistent least squares estimator $\hat{\boldsymbol{\pi}}$, obtaining

$$\begin{aligned} y_{1ij}^* &= \mathbf{z}_{1i}'\boldsymbol{\beta}_{1j} + \mathbf{z}_i'\hat{\boldsymbol{\pi}}\beta_{2j} + \hat{v}_i\beta_{2j} + u_{ij} \\ &= \mathbf{z}_{1i}'\boldsymbol{\beta}_{1j} + \hat{y}_{2i}\beta_{2j} + \hat{v}_i\beta_{2j} + u_{ij} \\ &= \mathbf{w}_i'\boldsymbol{\beta}_j + e_{ij} \end{aligned} \tag{2.6}$$

and

$$y_{1ij} = I_{[0,\infty)}\left(\mathbf{w}_i'\boldsymbol{\beta}_j + e_{ij}\right) \tag{2.7}$$

where $\mathbf{w}_i = \left[\mathbf{z}'_{1i}, \hat{y}_{2i}\right]'$, $e_{ij} = \hat{v}_i\beta_{2j} + u_{ij}$, $\hat{v}_i = y_{2i} - \mathbf{z}'_i\hat{\boldsymbol{\pi}}$, and $\operatorname{plim}\left(n^{-1}\sum_{i=1}^{n}\mathbf{w}_i e_{ij}\right) = \mathbf{0}$.

Given the statistical model (2.6)-(2.7), the problem is to demonstrate a semiparametric estimator that connects the unknown probabilities, p_{ij}, with the unknown link functions, $G_j\left(\mathbf{x}_i, \boldsymbol{\beta}\right)$ for $j = 1,\ldots, J$, and that also has good sampling properties.

2.1 Problem Formulation

Given the development in (2.1)-(2.7), consider

$$y_{1ij} = G_j\left(\mathbf{x}_i, \boldsymbol{\beta}\right) + \varepsilon_{ij} = p_{ij} + \varepsilon_{ij} \tag{2.8}$$

which, for expository purposes, we rewrite in $\left(nJ \times 1\right)$ vector form by vertically stacking sets of n sample observations, for each of the J responses $j = 1, 2, \ldots, J$, as

$$\mathbf{y}_1 = \mathbf{p} + \boldsymbol{\varepsilon}. \tag{2.9}$$

If we let $\mathbf{w} = \left[\mathbf{z}_1, \hat{\mathbf{y}}_2\right]$ be a matrix of dimension $\left(n \times \left(m_1 + 1\right) = n \times k\right)$, one way to represent information contained in (2.9) is in the form of the empirical moment constraint

$$n^{-1}\left(\mathbf{I}_J \otimes \mathbf{w}'\right)\left(\mathbf{y}_1 - \mathbf{p} - \boldsymbol{\varepsilon}\right) = \mathbf{0}\ . \tag{2.10}$$

If the asymptotic orthogonality conditions $n^{-1}\left(\mathbf{I}_J \otimes \mathbf{w}'\right)\boldsymbol{\varepsilon} \xrightarrow{p} \mathbf{0}$ *hold, then*

$$n^{-1}\left(\mathbf{I}_J \otimes \mathbf{w}'\right)\left(\mathbf{y}_1 - \mathbf{p}\right) = \mathbf{0} \tag{2.11}$$

can be used as an asymptotically valid estimating function. Estimating functions provide one effective path to inference without specifying the underlying probability structure. However, in (2.11) there are kJ moment relations and nJ unknown multinomial parameters, with $nJ > kJ$. Consequently, the inverse problem is ill-posed and cannot be solved for a unique solution by direct matrix inversion methods.

2.2 An Estimation Criterion – Distance Measures

One way to solve the ill-posed inverse problem for the unknown parameters, without making a large number of assumptions or introducing additional information, is to formulate it as an extremum problem. In this context, the Cressie-Read statistic (Cressie and Read, 1984; Read and Cressie, 1988; Corcoran, 2000)

$$I(\mathbf{p},\mathbf{q},\gamma) = \frac{1}{\gamma(\gamma+1)}\sum_{j=1}^{J} p_j\left[\left(\frac{p_j}{q_j}\right)^{\gamma} - 1\right] \tag{2.12}$$

where we focus on discrete probability distributions with J nonzero probability elements, represents an estimating criterion that is particularly useful since the unknowns of the problem are contained within the unit simplex. The result is a multinomial allocation that assigns probability p_{ij} to the possible outcomes of y_{1ij}. In the limit as γ ranges from -2 to 1, a family of estimation and inference procedures emerges. Three main variants of $I(\mathbf{p},\mathbf{q},\gamma)$ have received explicit attention in the literature (see Mittelhammer, Judge and Miller, 2000). Assuming that the q_j's represent the reference distribution of the CR statistic and that this reference distribution is specified to be the uniform distribution, i.e., $q_j = J^{-1}, \forall j$, then $I(\mathbf{p},\mathbf{q},\gamma)$ converges to an estimation criterion equivalent to the negative of Owen's (1988, 1991, 2000) empirical likelihood (EL) metric $J^{-1}\sum_{j=1}^{J}\ln(p_j)$, when $\gamma \to -1,$. The second prominent case corresponds to letting $\gamma \to 0$ and leads the estimation criterion $-\sum_{j=1}^{J} p_j \ln(p_j)$, which is the negative of the empirical exponential likelihood (EEL) or Kullback-Leibler (1959) distance. As Csiszar (1998) has noted, the Kullback-Leibler (KL) distance is not a true distance metric, but in many respects, it is an analogue to the squared Euclidean distance measure. Finally $\gamma = 1$ results in an estimation objective that is proportional to the log Euclidian likelihood function, $J^{-1}\sum_{j=1}^{J}(J^2 p_j^2 - 1)$. We can then define a generalized extremum, global optimization with respect to γ, formulation for our problem, with the estimation objective being to maximize the negative of a Cressie-Read statistic that has been extended to represent n multinomial distributions, each with J alternatives, as[1]

---

[1] Letting $\mathbf{p}_i$ denote the $J \times 1$ vector of multinomial probabilities associated with sample observation i, and letting $\mathbf{q}_i$ denoted the associated reference distribution, the extended Cressie-Read statistic is of the form

$$I(\mathbf{p},\mathbf{q},\gamma) = \frac{1}{\gamma(\gamma+1)}\sum_{i=1}^{n}\sum_{j=1}^{J}\mathbf{p}_i[j]\left[\left(\frac{\mathbf{p}_i[j]}{\mathbf{q}_i[j]}\right)^{\gamma} - 1\right].$$

$$\mathrm{l}(\mathbf{p}) = \max_{p_{ij}\in(0,1),\forall i \text{ and } j} \left\{ -\mathrm{I}(\mathbf{p},\mathbf{q},\gamma) \mid \mathrm{n}^{-1}\left(\mathbf{I}_\mathrm{J} \otimes \mathbf{w}'\right)\left(\mathbf{y}-\mathbf{p}\right) = \mathbf{0},\, \left[\mathbf{1}_\mathrm{J}' \otimes \mathbf{I}_\mathrm{n}\right]\mathbf{p} = \mathbf{1}_\mathrm{n} \right\} \tag{2.13}$$

for a given choice of γ and a uniform reference distribution $\mathbf{q} = \mathrm{J}^{-1}\mathbf{1}_{\mathrm{nJ}}$ representing the usual case of uninformative priors, where $\mathbf{1}_\ell$ denotes a $(\ell \times 1)$ vector of 1's. The integer values of γ that are noted above they become special cases.

2.3 Problem Formulation and Solution

Focusing on the case where $\gamma \to 0$, the KL estimation problem is defined by

$$\max_{\mathbf{p}} \mathrm{H}(\mathbf{p}) = -\mathbf{p}' \ln(\mathbf{p}) \tag{2.14}$$

subject to the assumed information-moment constraint

$$\left(\mathbf{I}_\mathrm{J} \otimes \mathbf{w}'\right)\mathbf{y}_1 = \left(\mathbf{I}_\mathrm{J} \otimes \mathbf{w}'\right)\mathbf{p} \tag{2.15}$$

and the n normalization (adding up) conditions

$$\left[\mathbf{1}_\mathrm{J}' \otimes \mathbf{I}_\mathrm{n}\right]\mathbf{p} = \mathbf{1}_\mathrm{n} \tag{2.16}$$

Note that maximization of (2.14) subject to the assumed moment constraints (2.15) and the adding up-normalization conditions (2.16) is equivalent to minimization of the KL cross-entropy distance measure relative to a uniform reference distribution for each vector of probabilities $\left(\mathrm{p}_{\mathrm{i}1}, \mathrm{p}_{\mathrm{i}2} \ldots, \mathrm{p}_{\mathrm{i}\mathrm{J}}\right)$, for $\mathrm{i} = 1, 2, \ldots, \mathrm{n}$ and subject to the same moment constraints. For the case of binary data, Downs (2003) discusses an alternative class of maximum entropy distributions that represent other features of the observed data.

Moving in the direction of a solution, the first-order conditions for the Lagrangian form of the optimization problem (2.14)-(2.16) form a basis for recovering the unknown $\mathbf{p}$ and the $\boldsymbol{\beta}_\mathrm{j}$'s through the Lagrange multipliers. In particular, the Lagrangian for the KL-maximum entropy optimization problem is

$$\mathrm{L}(\mathbf{p};\mathbf{y}) = -\mathbf{p}' \ln(\mathbf{p}) + \boldsymbol{\lambda}'\left[\left(\mathbf{I}_\mathrm{J} \otimes \mathbf{w}'\right)\left(\mathbf{y}_1 - \mathbf{p}\right)\right] + \boldsymbol{\tau}'\left[\mathbf{1}_\mathrm{n} - \left[\mathbf{1}_\mathrm{J}' \otimes \mathbf{I}_\mathrm{n}\right]\mathbf{p}\right]. \tag{2.17}$$

The solution to this optimization problem is

$$\hat{\mathrm{p}}_{\mathrm{ij}} = \frac{\exp\left(-\mathbf{w}_\mathrm{i}'\hat{\boldsymbol{\lambda}}_\mathrm{j}\right)}{\Omega_\mathrm{i}\left(-\hat{\boldsymbol{\lambda}}\right)} = \frac{\exp\left(\mathbf{w}_\mathrm{i}'\hat{\boldsymbol{\beta}}_\mathrm{j}\right)}{\Omega_\mathrm{i}\left(\hat{\boldsymbol{\beta}}\right)} = \frac{\exp\left(\mathbf{w}_\mathrm{i}'\hat{\boldsymbol{\beta}}_\mathrm{j}\right)}{1 + \sum_{\mathrm{k}=2}^{\mathrm{J}} \exp\left(\mathbf{w}_\mathrm{i}'\hat{\boldsymbol{\beta}}_\mathrm{k}\right)} \tag{2.18}$$

where $\hat{\boldsymbol{\lambda}}_j$ refers to the $(k \times 1)$ vector of elements associated with alternative j, $\hat{\boldsymbol{\beta}}_j \equiv -\hat{\boldsymbol{\lambda}}_j$ weights the impact of the explanatory variables on the p_{ij}'s, and the $\Omega_i(\hat{\boldsymbol{\beta}})$ term is a normalization factor. We also assume that the standard identification condition $\hat{\boldsymbol{\beta}}_1 = \mathbf{0}$ is imposed.

The unknown $\boldsymbol{\beta}_j$'s that link the p_{ij}'s to the $\mathbf{w}_i$'s are the negative of the kJ Lagrange multiplier parameters that are chosen so that the optimum solution $\hat{p}_{ij}$ satisfies the constraints (2.15). Given the Lagrangian (2.17) and the corresponding first-order conditions, the Hessian matrix with respect to the choice probabilities is a negative definite diagonal matrix characterized by the elements

$$\frac{\partial^2 L}{\partial p_{ij}^2} = -\frac{\Omega_i(\boldsymbol{\beta})}{\exp\left(\mathbf{w}_i'\boldsymbol{\beta}_j\right)} = -\frac{1}{p_{ij}} \tag{2.19}$$

and

$$\frac{\partial^2 L}{\partial p_{ij}\partial p_{k\ell}} = 0 \text{ when } (i,j) \neq (k,\ell). \tag{2.20}$$

The negative definite Hessian matrix ensures a unique global solution for the p_{ij}'s provided the constraint set includes an interior feasible point. To reduce the computational burden of the estimation problem, we note that the minimum KL approach can be reformulated as an unconstrained problem. By substitution of the solution outcomes (2.18) back into the Lagrangian (2.17), we can rewrite the constrained KL optimization problem in an unconstrained or concentrated form

$$M(\boldsymbol{\lambda}) = \sum_i \sum_{j=2}^{J} y_{1ij}\mathbf{w}_i'\boldsymbol{\lambda}_j + \sum_i \ln\left[\Omega_i(-\boldsymbol{\lambda})\right] \quad . \tag{2.21}$$

By the saddle-point property of the minimum KL problem, $M(\boldsymbol{\lambda})$ is strictly convex in $\boldsymbol{\lambda}$, and the optimal values of the Lagrange multipliers may be computed by minimizing $M(\boldsymbol{\lambda})$ with respect to $\boldsymbol{\lambda}$ (or maximizing $-M(\boldsymbol{\lambda})$ with respect to $\boldsymbol{\lambda}$). We also use $M(\boldsymbol{\lambda})$ to derive the asymptotic properties of the minimum KL estimator.

2.3.1 The Traditional Multinomial Logit Estimator

The maximum likelihood (ML) multinomial logit estimator is a special case of the minimum KL solution stated in (2.18) if the model (2.3) does not include the endogeneity component (i.e., $\beta_{2j} = 0$ for all j). In this case, the minimum KL solution to the restricted version of the problem in (2.14)-(2.16) is

$$\hat{p}_{ij} = \frac{\exp\left(-\mathbf{z}_{1i}'\hat{\boldsymbol{\lambda}}_{1j}\right)}{\Omega_i\left(-\hat{\boldsymbol{\lambda}}_1\right)} = \frac{\exp\left(\mathbf{z}_{1i}'\hat{\boldsymbol{\beta}}_{1j}\right)}{1+\sum_{k=2}^{J}\exp\left(\mathbf{z}_{1i}'\hat{\boldsymbol{\beta}}_{1k}\right)} \tag{2.22}$$

where $\hat{\boldsymbol{\beta}}_{1j} \equiv -\hat{\boldsymbol{\lambda}}_{1j}$ for each j and $\boldsymbol{\beta}_{11} = \mathbf{0}$ is imposed. Both the general choice probability formulation in (2.18) and the traditional multinomial logit model in (2.22) are consistent with utility maximization (see Train, 2003, p. 41). To show the correspondence of the two approaches explicitly, we consider the special case of (2.21) associated with (2.22) (i.e., under the restriction $\beta_{2j} = 0$). The optimal Lagrange multipliers are selected by maximizing $-M(\boldsymbol{\lambda}_1)$

$$-M(\boldsymbol{\lambda}_1) = -\sum_i\sum_{j=2}^{J} y_{1ij}\mathbf{z}_{1i}'\boldsymbol{\lambda}_{1j} - \sum_i \ln\left[\Omega_i\left(-\boldsymbol{\lambda}_1\right)\right] \quad . \tag{2.23}$$

with respect to $\boldsymbol{\lambda}$. This concentrated objective function is equivalent to the multinomial logit log-likelihood function

$$\begin{aligned} \ln\left(L\left(\boldsymbol{\beta}_1;\mathbf{y}\right)\right) &= \sum_i\sum_j y_{1ij}\ln\left[\frac{\exp\left(\mathbf{z}_{1i}'\boldsymbol{\beta}_{1j}\right)}{1+\sum_{k=2}^{J}\exp\left(\mathbf{z}_{1i}'\boldsymbol{\beta}_{1k}\right)}\right] \\ &= \sum_i\sum_{j=2}^{J} y_{1ij}\mathbf{z}_{1i}'\boldsymbol{\beta}_{1j} - \sum_i \ln\left[\Omega_i\left(\boldsymbol{\beta}_1\right)\right] \end{aligned} \tag{2.24}$$

where $\boldsymbol{\beta}_{1j} \equiv -\boldsymbol{\lambda}_{1j}$. Although the conceptual bases for the traditional ML multinomial logit and the minimum KL formulations are different, the ML and minimum KL parameter estimates are identical.

The equivalence of the ML and minimum KL estimators also implies that they share identical finite and large sample properties. If the logistic model specification is correct, we know that the ML and minimum KL estimators are $\sqrt{n}$-consistent such that

$\hat{\boldsymbol{\beta}}_1 \xrightarrow{p} \boldsymbol{\beta}_1^0$ under the standard regularity conditions for ML estimators. The estimators are also asymptotically normal so that $\sqrt{n}\left(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^0\right) \xrightarrow{d} N\left(\mathbf{0}, \boldsymbol{\Delta}_0^{-1}\right)$ where $\boldsymbol{\Delta}_0$ is the limiting information matrix, $\boldsymbol{\Delta}_0 \equiv \lim_{n\to\infty} E\left[-n^{-1} \frac{\partial^2 \ln L(\boldsymbol{\beta}_1; \mathbf{y})}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1'} \Big|_{\boldsymbol{\beta}_1^0} \right]$.

Following the discussion in Golan, Judge, and Perloff, the sample information matrix used to estimate the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}_1$ may be derived from the information about the underlying conditional choice probabilities. First, we rearrange the Hessian matrix composed of (2.19) and (2.20) to form J^2 blocks of elements. The $(j,k)^{th}$ block denotes the second partial derivatives of the Lagrangian with respect to elements of the $(n \times 1)$ vectors $\mathbf{p}_j$ and $\mathbf{p}_k$ (i.e., the n probabilities across observations for the j^{th} and k^{th} alternatives, respectively). The j^{th} diagonal block of the Hessian matrix can be represented by defining $\mathbf{1}(i)$ to be a $(n \times 1)$ null vector except for a one in row i and by summing over the n sample observations to obtain

$$I(\mathbf{p}_j)_{ME} = \sum_{i=1}^{n} \frac{\Omega_i(\boldsymbol{\beta}_1)}{\exp\left(\mathbf{z}_{1i}'\boldsymbol{\beta}_{1j}\right)} \mathbf{1}(i)\mathbf{1}(i)' = \sum_{i=1}^{n} \frac{1}{p_{ij}} \mathbf{1}(i)\mathbf{1}(i)' . \tag{2.25}$$

Then, we transform from $\mathbf{p}_i$ to $\boldsymbol{\beta}_{1j}$ space (see Lehmann and Casella, 1998, p. 115) to derive

$$\begin{aligned} \sum_j \left(\frac{\partial \mathbf{p}_j}{\partial \boldsymbol{\beta}_{1\ell}}\right) I(\mathbf{p}_j)_{ME} \left(\frac{\partial \mathbf{p}_j}{\partial \boldsymbol{\beta}_{1m}'}\right) &= I(\boldsymbol{\beta}_{1\ell}, \boldsymbol{\beta}_{1m})_{ME} \\ &= \sum_{i=1}^{n} \left[p_{im}\mathbf{1}(i)\mathbf{1}(i)' - p_{i\ell}p_{im} \right] \mathbf{z}_{1i}\mathbf{z}_{1i}' \\ &= I(\boldsymbol{\beta}_{1\ell}, \boldsymbol{\beta}_{1m})_{ML} \end{aligned} \tag{2.26}$$

where (2.26) is the $(\ell, m)^{th}$ block of $(J-1)^2$ blocks of dimension $(K \times K)$ referring to all parameter vectors other than the fixed (for identification purposes) $\boldsymbol{\beta}_{11} = \mathbf{0}$. The matrix composed of blocks (2.26) is $\left(K(J-1) \times K(J-1)\right)$ in dimension and is identical to the sample information matrix for the ML multinomial logit estimator. The estimated

asymptotic covariance matrix for $\hat{\boldsymbol{\beta}}_1$, $\hat{\mathbf{cov}}\left(\hat{\boldsymbol{\beta}}_1\right) = n\hat{\boldsymbol{\Delta}}^{-1}$, is the inverse of this sample information matrix evaluated at $\hat{\boldsymbol{\beta}}_1$.

Given that we view the multinomial choice model from the semiparametric perspective, it is important to note that the large sample properties may also hold if the logistic model specification is incorrect. The key regularity condition (in addition to those required for the ML logit model) is the existence of some vector of model parameters $\boldsymbol{\beta}_1^0$ such that $n^{-1}\left(\mathbf{I}_J \otimes \mathbf{z}_1'\right)\left(\mathbf{y}_1 - \mathbf{p}\left(\boldsymbol{\beta}_1^0\right)\right) \xrightarrow{p} \mathbf{0}$ as $n \to \infty$. Under these conditions, the estimators are also consistent such that $\hat{\boldsymbol{\beta}}_1 \xrightarrow{p} \boldsymbol{\beta}_1^0$ and asymptotically normal as $\sqrt{n}\left(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^0\right) \xrightarrow{d} N\left(\mathbf{0}, \boldsymbol{\Delta}_0^{-1}\boldsymbol{\Xi}_0\boldsymbol{\Delta}_0^{-1}\right)$ where $\boldsymbol{\Xi}_0$ is the limiting covariance matrix of the normalized necessary conditions, $\boldsymbol{\Xi}_0 \equiv \lim_{n\to\infty} E\left[n^{-1} \left.\frac{\partial \ln L}{\partial \boldsymbol{\beta}_1}\right|_{\boldsymbol{\beta}_1^0} \left.\frac{\partial \ln L}{\partial \boldsymbol{\beta}_1'}\right|_{\boldsymbol{\beta}_1^0} \right]$. If the model is correctly specified, the limiting covariance matrix reduces to $\boldsymbol{\Delta}_0^{-1}$ under the information matrix equality, $\boldsymbol{\Xi}_0 = -\boldsymbol{\Delta}_0$.

2.3.2. Sampling Properties under Endogeneity

The asymptotic properties of the minimum KL estimator in (2.18) do not carry over under the unrestricted version of the model (2.3) due to the endogeneity of y_{2i}. The key problem is that the asymptotic orthogonality condition $n^{-1}\left(\mathbf{I}_j \otimes \mathbf{w}'\right)\boldsymbol{\varepsilon} \xrightarrow{p} 0$ underlying (2.11) does not hold. Although $\hat{y}_{2_i}$ is uncorrelated with the errors e_{ij} in the latent regression model (2.6), $\hat{y}_{2_i}$ may be correlated with ε_{ij} such that $E\left[\left(\mathbf{I}_J \otimes \mathbf{w}'\right)\boldsymbol{\varepsilon}\right] \neq 0$ because the errors in the observed regression model (2.8) are nonlinear functions of the latent noise components. This point was illustrated with a Monte Carlo simulation example presented by Dagenais (1999)[2].

[2] We note that while his conceptual point remains valid, there is an error in the numerical simulation results reported by Degenais. In particular, he utilized a standard normal distribution when in fact a normal distribution, with variance $\sigma_v^2 = 4$, should have been used in generating the outcomes of the latent variable in his structural equation. The corrected correlation between instruments and the disturbance term

We performed a limited set of Monte Carlo experiments based on the data sampling process characterized by (2.3)-(2.4) in which the key comparisons were the impact of the sample size (n) and the trade-off between the noise components, u_i and v_i. We consider the following specific implementation of (2.3) and (2.4),

$$y_{1i}^{*} = \mathbf{z}_{11i} + 2\,\mathbf{z}_{12i} + y_{2i}\beta_2 + u_i \tag{2.27}$$

$$y_{2i} = \mathbf{z}_{11i} - 2\,\mathbf{z}_{12i} + \mathbf{z}_{21i} - \mathbf{z}_{22i} + v_i, \tag{2.28}$$

where $y_{1i} = I\left(y_{1i}^{*} > 0\right)$. The exogenous (instrumental) variables $\mathbf{z}_{1i}$ and $\mathbf{z}_{2i}$ are generated as pseudo-random Uniform(0,2) outcomes and held fixed in repeated Monte Carlo trials. We also choose $\beta_2 \in \{0,1\}$ to consider the behavior of estimators in models for which there is endogeneity (i.e., $\beta_2 = 1$) and no endogeneity (i.e., $\beta_2 = 0$).

Although the scale parameter for u_i is not identified for estimation purposes, we alter the value of this parameter within the experimental design to control the relative noise composition of y_{1i}^{*}. We draw pseudo-random outcomes from the bivariate normal distribution

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{pmatrix}\right] . \tag{2.29}$$

To vary the relative importance of the noise components, we set $\sigma_v^2 = 1$ and $\sigma_u^2 = 1$, and the correlation is $\rho \in \{-0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75\}$. We also set the number of observations as $n \in \{100, 250, 500, 1000\}$ to represent relatively small to large sample sizes. Under these model variations, the experimental design included a total of 28 sampling combinations. The simulations results are presented in Tables 1 and 2 for both the KL and Logit estimators based on one thousand simulated sample repetitions and with and without endogeneity, respectively.

The results suggest that in the exogenous regressors case, the KL method is very competitive in MSE with the Logit estimator across all sampling conditions, and as correlation and sample sizes increase, the relative superiority of the KL estimator is very

---

of the censoring equation in this case is -.095, based on one million repetitions, as opposed to the reported value of -.46 by Degenais. However, in any case, the correlation is nonzero, illustrating his conceptual point.

substantial. Empirical evidence of the consistency of the KL estimator is evident in Table 1, whereas the inconsistency of the Logit estimator is also evident particularly for highly correlated situations with large sample sizes. In the endogenous regressors case, the Logit estimator is more often the MSE superior estimator, although the KL estimator maintains superiority when the sample size is small and the correlation is positive and large. Empirical evidence of inconsistency is apparent in both estimators especially in cases of higher correlation.

2.3.3 Alternative Estimation Objective Functions

Finally we note that in (2.13) as γ approaches -1, maximization of the limit of $-\mathrm{I}(\mathbf{p},\mathbf{q},\gamma)$ for $\mathbf{q} = \mathrm{J}^{-1}\mathbf{1}_{\mathrm{nJ}}$ is equivalent to maximization of the empirical likelihood (EL) criterion, namely $\mathrm{H}(\mathbf{p}) = \mathrm{J}^{-1}\mathbf{1}'_{\mathrm{nJ}}\ln(\mathbf{p})$. Replacing the objective $-\mathrm{I}(\mathbf{p},\mathbf{q},\gamma)$ in (2.13) with $\mathrm{H}(\mathbf{p})$ leads to a constrained optimization problem that can be solved by the method of Lagrange multipliers to yield, for each i,j, the following optimal probabilities,

$$\hat{\mathrm{p}}_{\mathrm{ij}} = \left[\mathbf{w}_{\mathrm{i}}'\hat{\boldsymbol{\beta}}_{\mathrm{j}} + \hat{\tau}_{\mathrm{i}}\right]^{-1} \tag{2.27}$$

where $\hat{\tau}_{\mathrm{i}}$ is the Lagrange multiplier associated with the i^{th} probability additivity constraint on $\mathbf{p}$, and $\hat{\boldsymbol{\beta}}$ weights the impact of the explanatory variables on the unknown probabilities, where again $\hat{\boldsymbol{\beta}}_1 = \mathbf{0}$. As before, the probabilities are implicitly defined through the Lagrange multipliers $\hat{\boldsymbol{\tau}}$ and do not have a closed form solution, which prevents direct evaluation of the functional form to ascertain the estimator's finite sample properties. For finite sample and limiting sampling properties of this and the KL formulation, see Mittelhammer, Judge, and Schoenberg (2003). An alternative semiparametric model of the choice probabilities could also be derived under the log Euclidean Likelihood objective function.

3. Alternative Multinomial Choice Models

The multinomial formulation that was presented in section 2 is based exclusively on decision maker's attributes represented by $\mathbf{x}_{\mathrm{i}}$, i = 1, ... , n, which clearly do not vary

across the J alternatives. We now consider alternative multinomial response models, and suggest how semiparametric estimates of these models might be defined based on the KL information theoretic framework.

3.1 Alternative-Specific Attributes

The utility maximization-decision model underlying the multinomial choice problem can be altered in a number of ways. One prominent model variation is the case where alternative-specific attributes are accounted for explicitly, allowing for estimates of the impacts on decision making of marginal changes in the levels of attributes contained in the J alternatives. Suppressing decision maker-specific attributes, in this formulation there is a *common (across alternatives)* parameter vector $\boldsymbol{\beta}$ representing marginal utilities of attributes associated with each of the alternatives. The overall utility of each alternative is derived by accumulating the utility of the bundle of attributes associated with the alternative as $\mathbf{x}_j'\boldsymbol{\beta}$, for j = 1, …, J, and then the alternative with the highest realization of the accumulated utility, also accounting for random noise in the random utility formulation, is the alternative chosen.

The preceding model variant can be accommodated within the KL-problem context with minor changes to the formulation of section 2. First of all, we alter the representation in (2.8) to the following:

$$y_{1ij} = G_j\left(\mathbf{z}_{ij}, \boldsymbol{\beta}\right) + \varepsilon_{ij} = p_{ij} + \varepsilon_{ij} \tag{3.1}$$

where $\mathbf{z}_{ij}$ now refers to a vector of observed attribute levels corresponding to alternative j and observation i. Note the formulation in (3.1) is consistent with utility maximization, as noted and motivated in Train (2003, p. 41). For expository purposes, we rewrite the information in (3.1) in $\left(nJ \times 1\right)$ vector form by vertically stacking sets of *n* sample observations, for each of the *J* responses $j = 1, 2, \ldots, J$, as

$$\mathbf{y}_1 = \mathbf{p} + \boldsymbol{\varepsilon}. \tag{3.2}$$

Then we can utilize the information contained in (3.2) in the form of the empirical moment constraint

$$\left(nJ\right)^{-1} \mathbf{z}'\left(\mathbf{y}_1 - \mathbf{p} - \boldsymbol{\varepsilon}\right) = \mathbf{0}\ . \tag{3.3}$$

If the asymptotic orthogonality conditions $(nJ)^{-1}\mathbf{z}'\boldsymbol{\varepsilon} \xrightarrow{p} \mathbf{0}$ hold, then

$$(nJ)^{-1}\mathbf{z}'(\mathbf{y}_1 - \mathbf{p}) = \mathbf{0} \tag{3.4}$$

can be used as an asymptotically valid estimating function. In this form, there are k moment relations and nJ unknown multinomial probability parameters, with $nJ > k$. Consequently, the inverse problem is ill-posed as before and cannot be solved for a unique solution by direct matrix inversion methods.

The KL estimation problem can now be defined as

$$\max_{\mathbf{p}} H(\mathbf{p}) = -\mathbf{p}' \ln(\mathbf{p}) \tag{3.5}$$

subject to the information-moment constraint

$$\mathbf{z}'\mathbf{y}_1 = \mathbf{z}'\mathbf{p} \tag{3.6}$$

and the n normalization (adding up) conditions

$$[\mathbf{1}_J' \otimes \mathbf{I}_n]\mathbf{p} = \mathbf{1}_n . \tag{3.7}$$

Note that maximization of (3.5) subject to the moment constraints (3.6) and the adding up-normalization conditions (3.7) is equivalent to minimization of the KL cross-entropy distance measure relative to a uniform reference distribution for each vector of choice probabilities $(p_{i1}, p_{i2} \ldots, p_{iJ})$, for $i = 1, 2, \ldots, n$ and subject to the same moment constraints.

The first-order conditions for the Lagrangian form of the optimization problem (3.5)-(3.7) form a basis for recovering the unknown $\mathbf{p}$ and $\boldsymbol{\beta}$ through the Lagrange multipliers. In particular, the Lagrangian for the maximum entropy optimization problem is now

$$L = -\mathbf{p}' \ln(\mathbf{p}) + \boldsymbol{\lambda}'[\mathbf{z}'(\mathbf{y}_1 - \mathbf{p})] + \boldsymbol{\tau}'[\mathbf{1}_n - [\mathbf{1}_J' \otimes \mathbf{I}_n]\mathbf{p}] . \tag{3.8}$$

The solution to this optimization problem is

$$\hat{p}_{ij} = \frac{\exp(-\mathbf{z}_{ij}'\hat{\boldsymbol{\lambda}})}{\Omega_i(-\hat{\boldsymbol{\lambda}})} = \frac{\exp(\mathbf{z}_{ij}'\hat{\boldsymbol{\beta}})}{\Omega_i(\hat{\boldsymbol{\beta}})} = \frac{\exp(\mathbf{z}_{ij}'\hat{\boldsymbol{\beta}})}{\sum_{k=1}^{J}\exp(\mathbf{z}_{ik}'\hat{\boldsymbol{\beta}})} \tag{3.9}$$

where $\hat{\boldsymbol{\lambda}}$ refers to the $(k \times 1)$ vector of Lagrange multiplier elements and $\hat{\boldsymbol{\beta}} \equiv -\hat{\boldsymbol{\lambda}}$ measures the impact of the explanatory variables on the p_{ij}'s, with $\Omega_i(\hat{\boldsymbol{\beta}})$ being a normalization factor. The unknown $\boldsymbol{\beta}$ that links the p_{ij} to the $\mathbf{z}_{ij}$ is the negative of the Lagrange multiplier vector that is chosen so that the optimum solution $\hat{p}_{ij}$ satisfies the constraints (3.6). The formulation in (3.9) is identical to the standard result for the maximum-utility motivated multinomial (conditional) logit model in the case of alternative-specific attributes (McFadden, 1974; also see Train, 2003, chapter 3).

Following a derivation analogous to the approach underlying (2.25)-(2.26), the information matrix of the current formulation can be derived where

$$I(\mathbf{p}_j)_{me} = \sum_{i=1}^{n} \frac{\Omega_i(\boldsymbol{\beta})}{\exp(\mathbf{z}_{ij}'\boldsymbol{\beta})} \mathbf{1}(i)\mathbf{1}(i)' = \sum_{i=1}^{n} \frac{1}{p_{ij}} \mathbf{1}(i)\mathbf{1}(i)' \tag{3.10}$$

and

$$\sum_{j=1}^{J} \left(\frac{\partial \mathbf{p}_j}{\partial \boldsymbol{\beta}}\right) I(\mathbf{p}_j)_{ME} \left(\frac{\partial \mathbf{p}_j}{\partial \boldsymbol{\beta}'}\right) = I(\boldsymbol{\beta})_{ME} = \sum_{i=1}^{n} \sum_{j=1}^{J} p_{ij} (\mathbf{z}_{ij} - \bar{\mathbf{z}}_i)(\mathbf{z}_{ij} - \bar{\mathbf{z}}_i)' = I(\boldsymbol{\beta})_{ML} \tag{3.11}$$

where $\bar{\mathbf{z}}_i = \sum_{j=1}^{J} p_{ij} \mathbf{w}_{ij}$. The inverse of the latter matrix represents the $(K \times K)$ information matrix for the estimator $\hat{\boldsymbol{\beta}}$, and the result in (3.11) demonstrates that the information matrix of the KL-maximum entropy approach and of the multinomial logit approach are again identical. Following our discussion in Section 2, the asymptotic properties of the minimum KL estimator may be derived analogous to the ML estimator properties.

3.2 Other Model Variants

There are research contexts in which one might want to investigate the impacts of changing attribute levels of alternatives, changing attributes levels of individual decision makers, or *both*. The two formulations in the preceding sections can be extended or combined to accommodate the case where the impacts of both types of attributes are being investigated. The KL-problem framework can accommodate this final model variant by including variables that refer to both types of attributes, and the algebra of the

optimization problem again leads to the multinomial logit result. In fact, the model formulation can be altered from the very beginning by reinterpreting the $\mathbf{x}_i$ vectors as incorporated variables that refer to both types of attributes, with the decision maker-specific observations blocked appropriately to interact with parameters unique to the j^{th} alternative, with an initial block reserved for attribute specific characteristics that interact with common parameters across alternatives. That is, redefine the $\mathbf{x}_i$ vectors to be $\mathbf{x}_i = \left[\mathbf{r}_i'\left[\mathbf{0}\ \mathbf{0}\ ...\mathbf{d}_{ij}'\ \mathbf{0}...\mathbf{0}\right]\right]'$, where $\mathbf{r}_i'$ is a row vector of decision maker-specific attributes for the i^{th} observation, $\mathbf{d}_{ij}'$ is a vector of alternative-specific attributes that are intended to be interacted with the parameters associated with the j^{th} alternative, and $\mathbf{0}$ is a row vector of zeros in placed where blocks of variables interact with parameters that refer to parameters associated with alternatives other than the j^{th}. Then defining the parameter vector to be $\beta = [\delta', \beta_1', \beta_2', ..., \beta_J']'$, it is apparent that a model containing alternative-specific and decision-maker attributes is represented by $\mathbf{x}_i'\beta$.

4. Summary and Implications

Endogeneity is an important and common problem in a range of linear and nonlinear econometric models. Recognizing this, we have focused on semiparametric multinomial choice models and how one may handle the estimation and inference problem under endogeneity. The proposed estimators are semiparametric in the sense that the joint distribution of the data is unspecified apart from a finite number of moment conditions and the conditional mean assumption on the error process. A solution basis is demonstrated that permits the recovery of the unknown response coefficients and the corresponding marginal probabilities and asymptotic sampling characteristics of the estimators are developed. The next steps are i) to develop a consistent non-linear moment based semi parametric estimator under endogeneity, ii) to develop the statistical implications of estimators when there is uncertainty regarding the existence and extent of endogeneity, and iii) to demonstrate how to choose our optimum estimator from the Cressie-Read family.

References

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

Ahn, H., H. Ichimura, and J.L. Powell. (1996). Simple estimators for monotone index models. *Working Paper*.

Armstrong, B.G. (1985). The generalized linear model. *Communications in Statistics* **14**(B):529-544.

Blundell, R. and J.L. Powell. (1999). Endogeneity in single index models. *Working Paper, Department of Economics, UC Berkeley*.

Corcoran, S.A. (2000). Empirical Exponential Family Likelihood using Several Moment Conditions. *Statistic Sinica* **10**:45-557.

Carroll, R.J., D. Ruppert, and L.A. Stefanski. (1995), Measurement Error in Nonlinear Models. London; Chapman and Hall.

Cressie, N. and T. Read. (1984). Multinomial Goodness of Fit Tests. *Journal of the Royal Statistical Society, Series B* **46**:440-464.

Csiszar, I. (1998). Information theoretic methods in probability and statistics. *IEEE Information Theory Society Newsletter* **48**:21-30.

Dagenais, M.G. (1999). Inconsistency of a proposed non-linear instrumental variables estimator for probit and logit models with endogenous regressors. *Economic Letters* **63**:19-21.

Downs, O.B. (2003). Discussion of Slice Sampling. *Annals of Statistics* **31**:743-748.

Golan, A., G.G. Judge, and J. Perloff. (1996). A Maximum Entropy Approach to Recovering Information from Multinomial Response Data. *Journal of the American Statistical Association* **91**:841-853.

Hong, H. and E. Tanner. (2003). Endogenous Binary Choice Model with Median Restrictions. *Economic Letters* **80**:219-225.

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

Judge, G. and M.E. Bock. (1978). The Statistical Implications of Pre-Test and Stein-Rule Estimators. New York: North Holland Publishing.

Judge, G. and R. Mittelhammer. (2003). A Semiparametric Basis for Combining Estimation Problems under Quadratic Loss. *Journal of American Statistical Association*, in press.

Klein, R. and R.H. Spady. (1993). An efficient semiparametric estimator for binary response models. *Econometrica* **61**:387-421.

Kullback, S. and R.A. Leibler. (1951). On information and sufficiency. *The Annals of Mathematical Statistics* 22:79-86.

Lehmann, E.L. and G. Casella. (1998). Theory of Point Estimation. New York: Springer-Verlag.

Maddala, G.S. (1983). Limited Dependent and Qualitative Variables in Econometrics, In: Econometric Society Monograph No. 3. Cambridge University Press, Cambridge.

McCullough, P. and J.A. Nelder. (1995). Generalized Linear Models. New York: Chapman and Hall.

McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior, in P. Zarembka, ed., Frontiers of Econometrics, Academic Press, New York, pp. 105-142.

Mittelhammer, R., G. Judge, and D. Miller. (2000). Econometric Foundations, New York:Cambridge University Press.

Mittelhammer, R. and G. Judge. (2003). Combining Estimators to Improve Structural Model Estimators to Improve Structural Model Estimation and Inference under Quadratic Loss. *Journal of Econometrics* in Press.

Mittelhammer, R., G. Judge, and R. Schoenberg. (2003). Empirical Evidence Concerning the Finite Sample Performance of EL-Type Structural Equation Estimation and Inference Methods. Festschrift in Honor of Thomas Rothenberg, Cambridge University Press, *in press*.

Newey, W. (1987). Efficient Estimation of Limited Dependent Variable Models with Endogenous Explanatory Variables. *Journal of Econometrics* **36**:231-50.

Newey, W. (1986). Linear Instrumental Variable Estimation of Limited Dependent Variable Models with Endogenous Explanatory Variables. *Journal of Econometrics* **32**:127-41.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**:237-249.

Owen, A. (1991). Empirical likelihood for linear models. *The Annals of Statistics* **19**(4):1725-1747.

Owen, A. (2000). Empirical Likelihood. New York: Chapman and Hall.

Read, T.R. and N.A. Cressie. (1988). Goodness of Fit Statistics for Discrete Multivariate Data. New York: Springer Verlag.

Spiegelman, D., B. Rosner, and R. Logan. (2000). Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/ validation study designs, *Journal of the American Statistical Association* **95**:51-61.

Train, K. (2003). Discrete Choice Methods with Simulation. New York: Cambridge University Press.

van Akkeren, M. and G.G. Judge. (1999). Extended empirical likelihood estimation and inference. Working paper, University of California, Berkeley, pp 1-49.

White, H. (1994). Estimation, Inference, and Specification Analysis. New York: Cambridge University Press.

Table 1. MSE Results With No Endogeneity, $\beta_2 = 0$				
Logit Estimator	**MSE**			
Correlation	**n=100**	**n=250**	**n=500**	**n=1000**
-0.75	0.438	0.293	0.213	0.191
-0.5	0.403	0.21	0.13	0.108
-0.25	0.379	0.16	0.079	0.052
0	0.431	0.158	0.07	0.034
0.25	0.469	0.212	0.106	0.07
0.5	0.594	0.338	0.223	0.191
0.75	0.837	0.616	0.467	0.466
KL Estimator	**MSE**			
Correlation	**n=100**	**n=250**	**n=500**	**n=1000**
-0.75	0.415	0.162	0.071	0.036
-0.5	0.419	0.159	0.068	0.036
-0.25	0.406	0.159	0.068	0.035
0	0.435	0.162	0.072	0.035
0.25	0.412	0.167	0.069	0.036
0.5	0.406	0.16	0.069	0.035
0.75	0.418	0.16	0.069	0.034

Table 2. MSE Results With Endogeneity, $\beta_2 = 1$				
Logit Estimator	**MSE**			
Correlation	**n=100**	**n=250**	**n=500**	**n=1000**
-0.75	0.45	0.274	0.185	0.183
-0.5	0.448	0.214	0.134	0.114
-0.25	0.478	0.187	0.093	0.061
0	0.522	0.193	0.085	0.044
0.25	0.686	0.279	0.14	0.092
0.5	0.971	0.498	0.297	0.254
0.75	1.547	0.93	0.642	0.605
KL Estimator	**MSE**			
Correlation	**n=100**	**n=250**	**n=500**	**n=1000**
-0.75	3.003	1.74	1.457	1.31
-0.5	0.907	0.389	0.235	0.18
-0.25	0.445	0.176	0.079	0.045
0	0.391	0.236	0.173	0.151
0.25	0.478	0.372	0.334	0.322
0.5	0.606	0.538	0.514	0.506
0.75	0.721	0.707	0.688	0.688