Title

Estimating the Link Function in Multinomial Response Models under Endogeneity and Quadratic Loss

Authors

Judge, George G.
Mittelhammer, Ron C

Publication Date

2004-02-03

Estimating the Link Function in Multinomial Response Models under Endogeneity and Quadratic Loss[*]

George G. Judge[1] and Ron C. Mittelhammer[2]

University of California, Berkeley and Washington State University

Abstract

This paper considers estimation and inference for the multinomial response model in the case where endogenous variables are 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. A data based shrinkage estimator that seeks an optimal combination of estimators and results in superior risk performance under quadratic loss is also developed.

Keywords: Multinomial Process, endogeneity, empirical likelihood procedures, quadratic loss, semiparametric estimation and inference, data dependent shrinkage, asymptotic and finite sample risk.

AMS 1991 Classifications: Primary 62E20

JEL Classifications: C10, C24

---

[*] Contribution to a volume in honor of the eminent economist-econometrician and raconteur Stanley R. Johnson

[1] George G. Judge is Professor in the Graduate School, 207 Giannini Hall, University of California, Berkeley, CA, 94720, e-mail: judge@ARE.Berkeley.edu

[2] Ron C. Mittelhammer is Professor of Statistics and Agricultural and Resource Economics at Washington State University, e-mail: mittelha@wsu.edu

We gratefully acknowledge the helpful comments of Doug Miller and Ken Train.

1. Introduction

Stan Johnson is a person that likes to think about different ways to analyze and solve economic-econometric problems. In this spirit we focus on the following problem. 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 i) a semiparametric estimator that avoids many of the assumptions of the likelihood approach and the loss of precision that occurs in fully nonparametric estimation, and ii) a combining model methodology, in the form of a Stein-like estimator, whose objective is to produce an optimal combination, under quadratic loss, of estimators that are considered feasible candidates for the data sampling process.

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}, $j = 1, ..., J$, 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 $\mathbf{x}_i$ 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 $\mathrm{E}\left[y_{ij} \mid \mathbf{x}_i \right] = G_j\left(\mathbf{x}_i, \boldsymbol{\beta}\right)$.

In those rare instances where the parametric functional form of $G_j\left(\mathbf{x}_i, \boldsymbol{\beta}\right)$ 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 log-likelihood function

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

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. Ahn, et al. (1993) demonstrate a semiparametric estimator applicable to censored selection models, whereas Ahn, et al. (1996) develop a semiparametric estimator for the traditional single index problem. As an extension of Ahn, et al. (1996), Blundell and Powell (1999) demonstrate an estimator for the single index problem that involves endogeneity of the explanatory variables. Building on work by Armstrong (1985) and Carroll, et al. (1995), Spiegelman, Rosner, and Logan (2000)

investigate and propose a semiparametric method useful in logistic regression models that involve covariate misclassification and measurement error. Hong and Tanner (2003) have recently suggested a semiparametric approach for estimating the binary choice model based on median restrictions. Their approach involves extremum estimation based on an estimation objective function that characterizes the median of the noise distribution as zero, conditional on a vector of instruments. The method requires estimating unknown distributional components of the objective function using kernel density estimation techniques.

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 $G_j\left(\mathbf{x}_i, \boldsymbol{\beta}\right)$ is unknown and the covariates in the untransformed structural model contain endogenous or random components such that $\mathrm{E}\left[\mathbf{x}_i \varepsilon_{ij}\right] \neq \mathbf{0}$. 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. A second objective involves demonstrating a risk superior estimator that combines, in a Stein-like way, an estimator that is consistent and asymptotic normal with one that has only the property of superior precision.

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 define a semiparametric estimator that is asymptotically biased and demonstrate a Stein-like estimator that combines estimation problems with different sampling attributes. In section 4 we discuss alternative multinomial response model formulations and indicate corresponding semiparametric estimation methods. Finally, in Section 5 the estimation and inference implications of our proposed models are summarized.

2. A Multinomial Response Model and a Semi Parametric Solution

Assume the multinomial response model

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

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

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

$\mathbf{x}_i$ 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*. However this does not necessarily imply that the attributes of the alternatives are unobserved and/or not considered by the decision maker. Rather, consistent with representations of Neoclassical utility functions, the attributes of the alternatives may be "bundled into" the definition of the alternative, and it is assumed that the decision maker processes them accordingly. In effect, the utility function is specified at the level of a reduced form in which the bundled alternative-specific attributes are codified by the name/description of the alternative, and different individuals (as differentiated by individual-specific attributes) can value each bundle of attributes differently. We will consider alternative multinomial response model formulations, and in particular the case where alternative-specific attributes appear explicitly in the formulation of the utility index, in section 4.

To characterize in an expository manner a situation that is consistent with the covariate endogeneity or measurement error problem, assume that $\mathbf{x}_i' = \left[\mathbf{z}_{1i}', \mathrm{y}_{2i}\right]$ 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 $\mathrm{E}\left[y_{2i}u_{ij}\right] \neq 0$. Then as in Blundell and Powell (1999) 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 $\mathrm{E}\left[\mathbf{z}_i v_i\right] = \mathbf{0}$. 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 $p\lim\left(n^{-1}\sum_{i=1}^{n}\mathbf{w}_i e_{ij}\right) = \mathbf{0}$.

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 $(nJ \times 1)$ 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. In this form, 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)

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

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. In the limit as γ ranges from -2 to 1, a family of estimation and inference procedures emerges. Three main variants of $I\left(\mathbf{p}, \mathbf{q}, \gamma\right)$ have received explicit attention in the literature (see Mittelhammer, Judge and Miller,

2000). Assuming that the $q_i's$ represent the reference distribution of the CR statistic and that this reference distribution is specified to be the uniform distribution, i.e., $q_i = J^{-1}, \forall i$, then when $\gamma \to -1$, $I(\mathbf{p},\mathbf{q},\gamma)$ converges in the limit to an estimation criterion equivalent to the negative of Owen's (1988, 1991, 2000) empirical likelihood (EL) metric $J^{-1}\sum_{i=1}^{J}\ln(p_i)$. The second prominent case corresponds to letting $\gamma \to 0$ and leads to the negative of the information theoretic measure of discrepancy $-\sum_{i=1}^{J} p_i \ln(p_i)$ as the estimation criterion, the latter being referred to in the literature as the empirical exponential likelihood-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, $J^{-1}\sum_{i=1}^{J}(J^2 p_i^2 - 1)$. We can then define a generalized extremum 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[3]

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

for a given choice of γ and a uniform reference distribution $\mathbf{q} = J^{-1}\mathbf{1}_{nJ}$ representing the usual case of uninformative priors, where $\mathbf{1}_\ell$ denotes a $(\ell \times 1)$ vector of 1's.

2.3 Problem Formulation and Solution

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

[3] 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

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

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

subject to the information-moment constraint

$$(\mathbf{I}_J \otimes \mathbf{w}')\mathbf{y}_1 = (\mathbf{I}_J \otimes \mathbf{w}')\mathbf{p} \tag{2.15}$$

and the n normalization (adding up) conditions

$$[\mathbf{1}_J' \otimes \mathbf{I}_n]\mathbf{p} = \mathbf{1}_n \tag{2.16}$$

Note that maximization of (2.14) subject to the 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 $(p_{i1}, p_{i2} \ldots, p_{iJ})$, for $i = 1, 2, \ldots, n$ and subject to the same moment constraints. For the case of binary data this leads to searching for the maximum entropy distribution for nonnegative valued data that matches the first and second order statistics of the data (Downs, 2003).

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 $\boldsymbol{\beta}_j$'s through the Lagrange multipliers. In particular, the Lagrangian for the KL-maximum entropy optimization problem is

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

The solution to this optimization problem is

$$\hat{p}_{ij} = \frac{\exp\left(-\mathbf{w}_i'\hat{\boldsymbol{\lambda}}_j\right)}{\Omega_i\left(-\hat{\boldsymbol{\lambda}}\right)} = \frac{\exp\left(\mathbf{w}_i'\hat{\boldsymbol{\beta}}_j\right)}{\Omega_i\left(\hat{\boldsymbol{\beta}}\right)} = \frac{\exp\left(\mathbf{w}_i'\hat{\boldsymbol{\beta}}_j\right)}{1 + \sum_{k=2}^{J} \exp\left(\mathbf{w}_i'\hat{\boldsymbol{\beta}}_k\right)} \tag{2.18}$$

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 $\Omega_i\left(\hat{\boldsymbol{\beta}}\right)$ term is a normalization factor. We 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 and the corresponding first-order

conditions, the Hessian 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

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

The negative definite Hessian matrix ensures a unique global solution for the p_{ij}'s.

2.3.1 The Information Matrix

To obtain an expression for the information matrix of the estimator for $\hat{\boldsymbol{\beta}}$, first rearrange the Hessian implied by (2.19)-(2.20) in terms of J^2 blocks of elements, with the $(i,j)^{th}$ block denoting derivatives with respect to the elements of the $(n\times 1)$ vectors $\mathbf{p}_i$ and $\mathbf{p}_j$, i.e., the n probabilities across observations relating to the i^{th} and j^{th} alternatives, respectively. The j^{th} diagonal block of the Hessian matrix can be represented by defining $\mathbf{1}(i)$ to be a $(n\times 1)$ zero vector, except for a one in row *i*, and summing over the *n* sample observations to obtain

$$I\left(\mathbf{p}_j\right)_{me} = \sum_{i=1}^{n} \frac{\Omega_i(\boldsymbol{\beta})}{\exp\left(\mathbf{w}_i'\boldsymbol{\beta}_j\right)} \mathbf{1}(i)\mathbf{1}(i)' = \sum_{i=1}^{n} \frac{1}{p_{ij}} \mathbf{1}(i)\mathbf{1}(i)'. \tag{2.21}$$

Then, transforming from $\mathbf{p}_i$ to $\boldsymbol{\beta}_j$ space (see Lehmann and Casella, 1998, p.115) yields

$$\sum_j \left(\frac{\partial \mathbf{p}_j}{\partial \boldsymbol{\beta}_l}\right) \mathrm{I}\left(\mathbf{p}_j\right)_{\mathrm{ME}} \left(\frac{\partial \mathbf{p}_j}{\partial \boldsymbol{\beta}_m'}\right) = I\left(\boldsymbol{\beta}_\ell, \boldsymbol{\beta}_m\right)_{ME} = \sum_i \left[p_{im}\mathbf{1}(\ell)'\mathbf{1}(m) - \mathrm{p}_{il}\mathrm{p}_{im} \right] \mathbf{w}_i\mathbf{w}_i' = \mathrm{I}\left(\boldsymbol{\beta}_\ell, \boldsymbol{\beta}_m\right)_{\mathrm{ML}} \tag{2.22}$$

where (2.22) 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) $\boldsymbol{\beta}_1 = \mathbf{0}$. The $\left(K(J-1)\times K(J-1)\right)$ matrix having (2.22) for blocks is identical to the ML multinomial

logit information matrix for $\boldsymbol{\beta}$. The asymptotic covariance matrix for $\hat{\boldsymbol{\beta}}$ can be estimated using the inverse of the $\left(K\left(J-1\right)\times K\left(J-1\right)\right)$ matrix having (2.22) for blocks, evaluated at the classical ML-logit estimates.

2.3.2 Asymptotic Properties

The conceptual bases for the traditional ML multinomial logit and the KL extremum formulations are different, because under the KL formulation no particular functional form linking the p_{ij} and the $\mathbf{w}_i'\boldsymbol{\beta}_j$ is specified. However, the resulting ML logit and KL solutions and information matrices are equivalent, and the usual ML asymptotic properties then follow. Relative to the correspondence between the classical KL and ML logit solutions, note that the estimating equations or moment constraints in the KL formulation are equivalent to the ML logit first-order conditions, and the optimal KL solution has the same post data mathematical form as the logistic multinomial probabilities.

To show the correspondence of the two approaches explicitly, the extremum KL approach can be reformulated as an unconstrained problem. Combining the Lagrangian (2.17) and the solution for the p_{ij}'s (2.18), we can rewrite the constrained KL optimization problem in an unconstrained or concentrated form as the minimization, with respect to $\boldsymbol{\lambda}$, of

$$M\left(\boldsymbol{\lambda}\right)=-\mathbf{y}'\left(\mathbf{I}\otimes\mathbf{w}\right)\boldsymbol{\lambda}+\sum_{i=1}^{n}\ln\left[\Omega_i\left(-\boldsymbol{\lambda}\right)\right] \tag{2.23}$$

which is equivalent to maximizing the multinomial log-likelihood function,

$$\begin{aligned}\ln\left(L\left(\mathbf{p};\mathbf{y}\right)\right)&=\sum_i\sum_j y_{1ij}\ln\left(p_{ij}\right)\\&=\sum_i\sum_j y_{1ij}\ln\frac{\exp\left(\mathbf{w}_i'\boldsymbol{\beta}_j\right)}{\Sigma_j\exp\left(\mathbf{w}_i'\boldsymbol{\beta}_j\right)}\\&=\sum_i\sum_j y_{1ij}\left[\mathbf{w}_i'\boldsymbol{\beta}_j\right]-\sum_i\ln\left[\Omega_i\left(\boldsymbol{\beta}\right)\right]\end{aligned} \tag{2.24}$$

where as before $\boldsymbol{\beta} = -\boldsymbol{\lambda}$ and the usual logit asymptotic properties follow. The unconstrained concentrated approach substantially reduces the computational complexity of the optimization problem.

2.3.3 Alternative Estimation Objective Functions

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

$$\hat{p}_{ij} = \left[\mathbf{w}_i'\hat{\boldsymbol{\beta}}_j + \hat{\tau}_i\right]^{-1} \tag{2.23}$$

where $\hat{\tau}_i$ 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, $\hat{\boldsymbol{\tau}}$ is not in closed form 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). A solution could also be obtained based on the log Euclidean Likelihood objective function.

2.4 A Special Case

The formulation and solution in the previous subsections permit the recovery of estimates for the marginal probabilities and the $\boldsymbol{\beta}_j$ response coefficients under endogeneity. An important special case alternative formulation is one that facilitates the direct recovery of the marginal probabilities and the response coefficients between y_{1ij} and the exogenous variables $\mathbf{z}_i = [\mathbf{z}_{i1}, \mathbf{z}_{i2}]$, where $E[z_i u_{ij}] = 0$. This of course is also the case for the traditional multinomial discrete choice problem and thus provides a

semiparametric basis for estimation and inference (Mittelhammer, et al., 2000). In this case we make use of the empirical moment constraint

$$h(\mathbf{y}_1, \mathbf{z}; \mathbf{p}, \mathbf{u}) = n^{-1}(\mathbf{I}_J \otimes \mathbf{z}')(\mathbf{y}_1 - \mathbf{p} - \mathbf{u}) = \mathbf{0} \tag{2.26}$$

which because $\mathrm{E}\left[(\mathbf{I}_j \otimes \mathbf{z}')\mathbf{u}\right] = \mathbf{0}$, yields the unbiased estimating function

$$\mathrm{E}\left[n^{-1}(\mathbf{I}_J \otimes \mathbf{z}')(\mathbf{y}_1 - \mathbf{p})\right] = \mathbf{0} \tag{2.27}$$

with sample analog

$$n^{-1}(\mathbf{I}_J \otimes \mathbf{z}')(\mathbf{y}_1 - \mathbf{p}) = \mathbf{0}\,. \tag{2.28}$$

Given this information base, we have in the context of the previous sections, the following extremum problem:

$$\underset{p_{ij} \in (0,1), \forall i \text{ and } j}{max} \left[-\mathbf{p}' \ln(\mathbf{p}) \; s.t. \; (\mathbf{I}_J \otimes \mathbf{z}')(\mathbf{y}_1 - \mathbf{p}) = \mathbf{0}, \; (\mathbf{1}_J' \otimes \mathbf{I}_n)\mathbf{p} = \mathbf{1}_n\right]. \tag{2.29}$$

The extremum problem can be cast in Lagrangian form as

$$L = -\mathbf{p}' \ln(\mathbf{p}) + \boldsymbol{\delta}'\left[(\mathbf{I}_J \otimes \mathbf{z}')(\mathbf{y}_1 - \mathbf{p})\right] + \boldsymbol{\gamma}'\left[\mathbf{1}_n - (\mathbf{1}_J' \otimes \mathbf{I}_n)\mathbf{p}\right] \tag{2.30}$$

with solution

$$\hat{p}_{ij} = \frac{\exp\left(-\mathbf{z}_i'\boldsymbol{\delta}_j\right)}{\Omega_i(-\boldsymbol{\delta})} = \frac{\exp\left(\mathbf{z}_i'\boldsymbol{\alpha}_j\right)}{\Omega_i(\boldsymbol{\alpha})} = \frac{\exp\left(\mathbf{z}_i'\boldsymbol{\alpha}_j\right)}{1 + \sum_{m=2}^{J} \exp\left(\mathbf{z}_i'\boldsymbol{\alpha}_m\right)} \tag{2.31}$$

where $\boldsymbol{\delta}_j$ refers to the $(k \times 1)$ vector of elements associated with alternative j,

$$\Omega_i(\boldsymbol{\alpha}) \equiv 1 + \sum_{m=2}^{J} \exp\left(\mathbf{z}_i'\boldsymbol{\alpha}_m\right) \tag{2.32}$$

and $\boldsymbol{\alpha}_j \equiv -\boldsymbol{\delta}_j$ measures the impact of the explanatory exogenous variables on the p_{ij}'s, where the standard identification condition $\boldsymbol{\alpha}_1 = \mathbf{0}$ is imposed. The term $\Omega_i(\boldsymbol{\alpha})$ is a normalization factor. The unknown $\boldsymbol{\alpha}_j$ links the p_{ij} to the $\mathbf{z}_i$. The development of the Hessian and the asymptotic covariance matrix proceeds as in Section 2.3.1.

3. A Competing Estimator

The semiparametric estimator demonstrated in Section 3 has the nice first-order properties of consistency and asymptotic normality. In cases where endogeneity is present in the underlying data sampling process, other estimators exist that do not have

the property of first-order consistency. However, these estimators may possess good precision characteristics. This raises questions concerning the sampling characteristics that would emerge if two estimators with different sampling attributes were combined.

One variant of this type of estimator is produced if, in the context of Section 2, we proceed as if y_{2i} is *uncorrelated* with the noise component. In this case, if we incorrectly assume $\mathrm{E}\left[\left(\mathbf{I}_J \otimes \mathbf{x}'\right)\mathbf{u}\right] = \mathbf{0}$, and now let $\boldsymbol{\pi}$ denote the vertically concatenated vector of choice probabilities, the empirical moment constraint

$$n^{-1}\left(\mathbf{I}_J \otimes \mathbf{x}'\right)\left(\mathbf{y}_1 - \boldsymbol{\pi} - \mathbf{u}\right) = \mathbf{0} \tag{3.1}$$

yields the biased estimating function

$$\mathrm{E}\left[n^{-1}\left(\mathbf{I}_J \otimes \mathbf{x}'\right)\left(\mathbf{y}_1 - \boldsymbol{\pi}\right)\right] \neq \mathbf{0}. \tag{3.2}$$

with sample analog

$$\left[n^{-1}\left(\mathbf{I}_J \otimes \mathbf{x}'\right)\left(\mathbf{y}_1 - \boldsymbol{\pi}\right)\right] = \mathbf{0}, \tag{3.3}$$

and we are lead to the following extremum problem:

$$\max_{\pi_{ij} \in (0,1), \forall i \text{ and } j} \left[-\boldsymbol{\pi}' \ln\left(\boldsymbol{\pi}\right) \mid \left(\mathbf{I}_J \otimes \mathbf{x}'\right)\left(\mathbf{y}_1 - \boldsymbol{\pi}\right) = \mathbf{0},\ \left(\mathbf{1}_J' \otimes \mathbf{I}_n\right)\boldsymbol{\pi} = \mathbf{1}_n\right]. \tag{3.4}$$

The Lagrangian form of the extremum problem is given by

$$L = -\boldsymbol{\pi}' \ln\left(\boldsymbol{\pi}\right) + \boldsymbol{\delta}'\left[\left(\mathbf{I}_j \otimes \mathbf{x}'\right)\left(\mathbf{y}_1 - \boldsymbol{\pi}\right)\right] + \boldsymbol{\eta}'\left[\mathbf{1}_n - \left(\mathbf{1}_J' \otimes \mathbf{I}_n\right)\boldsymbol{\pi}\right] \tag{3.5}$$

with solution,

$$\tilde{\pi}_{ij} = \frac{\exp\left(-\mathbf{x}_i'\tilde{\boldsymbol{\delta}}_j\right)}{\Omega_i\left(-\boldsymbol{\delta}\right)} = \frac{\exp\left(\mathbf{x}_i'\tilde{\boldsymbol{\beta}}_j\right)}{\Omega_i\left(\tilde{\boldsymbol{\beta}}\right)} = \frac{\exp\left(\mathbf{x}_i'\tilde{\boldsymbol{\beta}}_j\right)}{1 + \sum_{m=2}^{J}\exp\left(\mathbf{x}_i'\tilde{\boldsymbol{\beta}}_m\right)} \tag{3.6}$$

where $\boldsymbol{\delta}_j$ refers to the $(k \times 1)$ vector of Lagrange moment constraint multipliers associated with alternative j, $\Omega_i\left(\tilde{\boldsymbol{\beta}}\right) \equiv 1 + \sum_{m=2}^{J}\exp\left(\mathbf{x}_i'\tilde{\boldsymbol{\beta}}_m\right)$, and $\tilde{\boldsymbol{\beta}}_j \equiv -\tilde{\boldsymbol{\delta}}_j$ measures the impact of the explanatory exogenous variables on the π_{ij}'s, where the standard identification condition $\tilde{\boldsymbol{\beta}}_1 = \mathbf{0}$ is imposed. The unknown $\tilde{\boldsymbol{\beta}}_j$ links the π_{ij} to the $\mathbf{z}_i$. The development of the Hessian and the asymptotic covariance matrix proceeds as in Section 2.3.1.

Given that the estimating equations on which the solution (3.6) is biased, even in the limit, it is to be expected that the estimates derived from them do not possess the property of consistency. However, analogous to the case of comparing OLS to instrumental variable-based estimators of parameters in linear and nonlinear models, it is possible that the estimator based on the misspecified moment constraints has lower variation than the estimator that replaces $\mathbf{y}_2$ with a projection of itself through an instrument space. We consider next a method that attempts to exploit this potential characteristic of the alternative estimator through combinations of estimators.

3.1. Combined estimators formulation

The semiparametric estimator demonstrated in Section 2 has the attractive first-order asymptotic properties of consistency and asymptotic normality. A variant of this estimator when y_{2i} replaces $\hat{y}_{2i}$ in the structural moment condition is not consistent but its variance and/or quadratic risk performance may be superior to that of its competitors. Since each of these estimators can have superior sampling characteristics in some respects, this leads us to consider, in the spirit of Judge and Mittelhammer (2003) and Mittelhammer and Judge (2003), an estimator that is a weighted combination of the two. In this context, we consider the estimator that results from the following linear combination

$$\overline{\boldsymbol{\beta}}(\alpha) = \alpha\hat{\boldsymbol{\beta}} + (1-\alpha)\tilde{\boldsymbol{\beta}} \tag{3.7}$$

and ask whether, under quadratic loss $\| \overline{\boldsymbol{\beta}} - \boldsymbol{\beta} \|^2$, a combination of the estimators can be devised that has asymptotic risk that is superior to $\hat{\boldsymbol{\beta}}$ and that also performs well in finite samples. The asymptotic risk of $\overline{\boldsymbol{\beta}}(\alpha)$ is

$$\rho(\overline{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \alpha^2 tr\left(\Sigma_{\hat{\boldsymbol{\beta}}}\right) + (1-\alpha)^2 \left(tr\left(\Sigma_{\tilde{\boldsymbol{\beta}}}\right) + \boldsymbol{\mu}'\boldsymbol{\mu}\right) + 2\alpha(1-\alpha) tr\left(\Sigma_{\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{\beta}}}\right) \tag{3.8}$$

where $\boldsymbol{\mu}$ is the asymptotic bias of $\tilde{\boldsymbol{\beta}}$, $\Sigma_{\hat{\boldsymbol{\beta}}}$ *and* $\Sigma_{\tilde{\boldsymbol{\beta}}}$ are the asymptotic covariance matrices of $\hat{\boldsymbol{\beta}}$ *and* $\tilde{\boldsymbol{\beta}}$, and $\Sigma_{\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{\beta}}}$ denotes the asymptotic covariance matrix between the elements of

$\hat{\boldsymbol{\beta}}$ *and* $\tilde{\boldsymbol{\beta}}$. The minimum asymptotic risk choice of α is characterized by the solution of $\frac{d\rho\left(\bar{\boldsymbol{\beta}},\boldsymbol{\beta}\right)}{d\alpha}=0$, which coincides with the solution of

$$\alpha\left(tr\Sigma_{\hat{\boldsymbol{\beta}}}+tr\Sigma_{\tilde{\boldsymbol{\beta}}}+\boldsymbol{\mu}'\boldsymbol{\mu}-2tr\Sigma_{\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{\beta}}}\right)-tr\Sigma_{\tilde{\boldsymbol{\beta}}}-\boldsymbol{\mu}'\boldsymbol{\mu}+tr\Sigma_{\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{\beta}}}=0 \tag{3.9}$$

and leads to the optimal weight

$$\alpha_{*}=\frac{tr\Sigma_{\tilde{\boldsymbol{\beta}}}+\boldsymbol{\mu}'\boldsymbol{\mu}-tr\Sigma_{\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{\beta}}}}{tr\Sigma_{\hat{\boldsymbol{\beta}}}+tr\Sigma_{\tilde{\boldsymbol{\beta}}}+\boldsymbol{\mu}'\boldsymbol{\mu}-2tr\Sigma_{\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{\beta}}}}=1-\frac{tr\Sigma_{\hat{\boldsymbol{\beta}}}-tr\Sigma_{\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{\beta}}}}{tr\Sigma_{\hat{\boldsymbol{\beta}}}+tr\Sigma_{\tilde{\boldsymbol{\beta}}}+\boldsymbol{\mu}'\boldsymbol{\mu}-2tr\Sigma_{\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{\beta}}}}. \tag{3.10}$$

Given the solution in (3.10), the minimum risk combination of the two estimators in (3.7) can be expressed as

$$\bar{\boldsymbol{\beta}}_{*}=\hat{\boldsymbol{\beta}}-\frac{tr\Sigma_{\hat{\boldsymbol{\beta}}}-tr\Sigma_{\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{\beta}}}}{tr\Sigma_{\hat{\boldsymbol{\beta}}}+tr\Sigma_{\tilde{\boldsymbol{\beta}}}+\boldsymbol{\mu}'\boldsymbol{\mu}-2tr\Sigma_{\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{\beta}}}}\left(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}\right) \tag{3.11}$$

Since $AE\left[\|\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}\|\right]=tr\Sigma_{\hat{\boldsymbol{\beta}}}+tr\Sigma_{\tilde{\boldsymbol{\beta}}}+\boldsymbol{\mu}'\boldsymbol{\mu}-2tr\Sigma_{\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{\beta}}}$, where $AE(\bullet)$ denotes asymptotic expectation, and if we substitute $\|\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}\|$ for its asymptotic expected value (Judge and Bock, p. 175), we may rewrite (3.11) as

$$\bar{\boldsymbol{\beta}}_{*}\approx\hat{\boldsymbol{\beta}}-\frac{a}{\|\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}\|^{2}}\left(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}\right) \tag{3.12}$$

where $a=tr\Sigma_{\hat{\boldsymbol{\beta}}}-tr\Sigma_{\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{\beta}}}$, which is in the form of a Stein shrinkage estimator where $\hat{\boldsymbol{\beta}}$ is shrunk toward the alternative estimator $\tilde{\boldsymbol{\beta}}$.

3.2 Comments on Sampling Characteristics and Asymptotic Risk Performance

We note that (3.11) will *always* exhibit asymptotic quadratic risk behavior that is at least as good as the base estimator $\hat{\boldsymbol{\beta}}$. The approximate version, $\bar{\boldsymbol{\beta}}_{*}$, defined in (3.12) can be shown to be first order equivalent to (3.11), and thus to the first order of approximation, will also exhibit quadratic risk behavior that is at least as good as the base estimator. Moreover, this first order superiority continues to hold if a consistent estimator is used to replace 'a' in the numerator of the numerator in (3.12). These results follow from related results on asymptotic risk performance of combining estimators (similar in development to (3.7)-(3.12)) presented in Judge and Mittelhammer (2003) for the case of

a linear model data sampling context, and by Mittelhammer and Judge (2003) in the context of linear structural equation estimators.

Regarding the finite sample performance of the combining estimator relative to the base estimator, given the absence of parametric distributional assumptions in this semiparametric framework, no analytical risk superiority result would appear tractable. For certain specialized parametric sampling distribution assumptions, such as the case of multivariate normality, it may be possible to derive some limited analytical risk comparisons, as in Judge and Mittelhammer (2003) and Mittelhammer and Judge (2003), where the risk superiority of the combining estimator was demonstrated under certain regularity conditions. Relating to finite sample behavior, we add that in these recent studies, extensive Monte Carlo experimentation was conducted amounting to 1,300 different sampling scenarios characterized by a variety of conditions on noise variance, collinearity, degree of parameter identification, and spanning normal, uniform, beta, and gamma sampling distributions. In these sampling experiments, the combining estimator exhibited quadratic risk superiority relative to the base estimator in the vast majority of the experiments analyzed. We conjecture that the same kinds of results would apply to the combined estimator proposed in section 3.1.

4. Alternative Multinomial Choice Models

The multinomial formulation that was presented heretofore is based exclusively on decision maker's attributes represented by $\mathbf{x}_i$, 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.

4.1 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:

$$y_{1ij} = G_j\left(\mathbf{w}_{ij}, \boldsymbol{\beta}\right) + \varepsilon_{ij} = p_{ij} + \varepsilon_{ij} \tag{4.1}$$

where $\mathbf{w}_{ij}$ now refers to a vector of observed attribute levels corresponding to alternative j and observation i. Note the formulation in (4.1) is consistent with utility maximization, as noted and motivated in Train (2003, p. 41). For expository purposes, we rewrite the information in (4.1) 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

$$\mathbf{y}_1 = \mathbf{p} + \boldsymbol{\varepsilon}. \tag{4.2}$$

Accounting for endogeneity in some of the attributes as before, we let $\mathbf{w} = \left[\mathbf{z}_1, \hat{\mathbf{y}}_2\right]$ be a matrix of dimension $\left(nJ \times \left(m_1 + 1\right) = nJ \times k\right)$. Then we can utilize the information contained in (4.2) in the form of the empirical moment constraint

$$\left(nJ\right)^{-1} \mathbf{w}'\left(\mathbf{y}_1 - \mathbf{p} - \boldsymbol{\varepsilon}\right) = \mathbf{0} \tag{4.3}$$

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

$$\left(nJ\right)^{-1} \mathbf{w}'\left(\mathbf{y}_1 - \mathbf{p}\right) = \mathbf{0} \tag{4.4}$$

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

$$\max_{\mathbf{p}} H\left(\mathbf{p}\right) = -\mathbf{p}' \ln\left(\mathbf{p}\right) \tag{4.5}$$

subject to the information-moment constraint

$$\mathbf{w}'\mathbf{y}_1 = \mathbf{w}'\mathbf{p} \tag{4.6}$$

and the n normalization (adding up) conditions

$$\left[\mathbf{1}_J' \otimes \mathbf{I}_n\right]\mathbf{p}=\mathbf{1}_n . \tag{4.7}$$

Note that maximization of (4.5) subject to the moment constraints (4.6) and the adding up-normalization conditions (4.7) is equivalent to minimization of the KL cross-entropy distance measure relative to a uniform reference distribution for each vector of probabilities $\left(p_{i1}, p_{i2} \ldots, p_{iJ}\right)$, 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 (4.5-4.7) 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

$$L = -\mathbf{p}'\ln\left(\mathbf{p}\right) + \boldsymbol{\lambda}'\left[\mathbf{w}'\left(\mathbf{y}_1 - \mathbf{p}\right)\right] + \boldsymbol{\tau}'\left[\mathbf{1}_n - \left[\mathbf{1}_J' \otimes \mathbf{I}_n\right]\mathbf{p}\right]. \tag{4.8}$$

The solution to this optimization problem is

$$\hat{p}_{ij} = \frac{\exp\left(-\mathbf{w}_{ij}'\hat{\boldsymbol{\lambda}}\right)}{\Omega_i\left(-\hat{\boldsymbol{\lambda}}\right)} = \frac{\exp\left(\mathbf{w}_{ij}'\hat{\boldsymbol{\beta}}\right)}{\Omega_i\left(\hat{\boldsymbol{\beta}}\right)} = \frac{\exp\left(\mathbf{w}_{ij}'\hat{\boldsymbol{\beta}}\right)}{\sum_{k=1}^{J}\exp\left(\mathbf{w}_{ik}'\hat{\boldsymbol{\beta}}\right)} \tag{4.9}$$

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 $\Omega_i\left(\hat{\boldsymbol{\beta}}\right)$ being a normalization factor. The unknown $\boldsymbol{\beta}$ that links the p_{ij} to the $\mathbf{w}_{ij}$ is the negative of the Lagrange multiplier vector that is chosen so that the optimum solution $\hat{p}_{ij}$ satisfies the constraints (4.6). The formulation in (4.9) is identical to the standard result for the maximum-utility motivated multinomial 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 (2.21)-(2.22), the information matrix of the current formulation can be derived where

$$I\left(\mathbf{p}_j\right)_{me} = \sum_{i=1}^{n} \frac{\Omega_i(\boldsymbol{\beta})}{\exp\left(\mathbf{w}_{ij}'\boldsymbol{\beta}\right)} \mathbf{1}(i)\mathbf{1}(i)' = \sum_{i=1}^{n} \frac{1}{p_{ij}} \mathbf{1}(i)\mathbf{1}(i)'. \tag{4.10}$$

and

$$\sum_{j=1}^{J}\left(\frac{\partial \mathbf{p}_j}{\partial \boldsymbol{\beta}}\right) \mathrm{I}\left(\mathbf{p}_j\right)_{\mathrm{ME}} \left(\frac{\partial \mathbf{p}_j}{\partial \boldsymbol{\beta}'}\right) = I(\boldsymbol{\beta})_{ME} = \sum_{i=1}^{n}\sum_{j=1}^{J} p_{ij}\left(\mathbf{w}_{ij} - \bar{\mathbf{w}}_i\right)\left(\mathbf{w}_{ij} - \bar{\mathbf{w}}_i\right)' = \mathrm{I}(\boldsymbol{\beta})_{\mathrm{ML}} \tag{4.11}$$

where $\bar{\mathbf{w}}_i = \sum_{j=1}^{J} p_{ij}\mathbf{w}_{ij}$. The inverse of the latter matrix represents the $(K \times K)$ information matrix for the estimator $\hat{\boldsymbol{\beta}}$, and the result in (4.11) demonstrates that the information matrix of the KL-maximum entropy approach and of the multinomial logit approach are again identical.

4.2 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-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 $\mathbf{w}_i$ 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 jth alternative, with an initial block reserved for attribute specific characteristics that interact with common parameters across alternatives. That is, redefine the $\mathbf{w}_i$ vectors to be $\mathbf{w}_i = \left[\mathbf{r}_i' \ \left[\mathbf{0}\ \mathbf{0}\ \ldots \mathbf{d}_{ij}'\ \mathbf{0}\ldots\mathbf{0}\right]\right]'$, where $\mathbf{r}_i'$ is a row vector of alternative-specific attributes for

the i[th] observation, $\mathbf{d}'_{ij}$ is a vector of decision maker-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 i[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 $\mathbf{w}'_i\beta$.

Any of the model variants can be combined using the methodology outlined in section 3. As noted in section 3.2, it would be expected that the combined estimator will exhibit asymptotic mean square error performance that is at least as good as either of the base estimators that are being combined.

5. Summary and Implications

Endogeneity is an important and common problem in a range of linear and nonlinear econometric models. Recognizing this, in this paper, our focus has been on binary choice models and how one may, in a semiparametric way, handle the estimation and inference problem under endogeneity. The estimators that are suggested 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. Empirical likelihood and exponential empirical likelihood distance measures along with relevant underlying moment conditions frame the estimation problem. A solution basis is demonstrated that permits the recovery of the unknown response coefficients and the corresponding marginal probabilities and defining sampling properties. Because there is usually uncertainty concerning the stochastic characteristics of the econometric model, estimation procedures are developed that permit combining alternative plausible-competing models-estimators. Asymptotic and finite sample characteristics of the combined estimator are discussed. Developing analytical and

Monte Carlo sampling results for the proposed estimators and applying them to real economic problems, are the next steps in the research process.

References

Ahn, H., H. Ichimura, H. 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 Royal Statistical Society of Series B **46**:440-464.

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

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*.

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, Kenneth, (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.