A Minimum Power Divergence Class of CDFs and Estimators for the Binary Choice Model

Ron Mittelhammer and George Judge®

Washington State University and University of California, Berkeley

ABSTRACT

This paper uses information theoretic methods to introduce a new class of probability distributions and estimators for competing explanations of the data in the binary choice model. No explicit parameterization of the function connecting the data to the Bernoulli probabilities is stated in the specification of the statistical model. A large class of probability density functions emerges including the conventional logit model. The new class of statistical models and estimators requires minimal *a priori* model structure and non-sample information, and provides a range of model and estimator extensions. An empirical example is included to reflect the applicability of these methods.

Key words: *Semiparametric Binary Estimators, Conditional Moment Equations, Squared Error Loss, Cressie-Read Statistic, Information Theoretic Methods*
AMS 1991 Classification Primary 62E20
JEL Classifications: C10, C2

1. INTRODUCTION

Much in the theory and practice of econometrics involves subject matter theories and data that are partial and incomplete. This is especially true as it relates to discrete choice behavioral-random utility models where *i*) the underlying economic theory is based on an abstract mathematical structure that identifies axiomatically its impact on behavior, and *ii*) parametric statistical models are often used to obtain solutions to finite discrete pure and noisy non-parametric ill-posed inverse problems. The non-parametric restriction avoids using information that the researcher usually does not possess and the inverse problem results because one must use indirect observations to recover the structure connecting the data to the unobservable choice probabilities. The problems are ill-posed or under-determined because, without assumptions, there are more unknowns than data points and thus there is insufficient information to solve the problem uniquely. This results in the common situation where a function must be inferred despite insufficient information and only a feasible set of solutions is specified.

Pursuing estimation and inference as it relates to discrete choice behavior (DCB), a generation of econometricians has, with the aid of assumptions and parametric-model oriented structures, used probit or logit cumulative distribution functions (CDFs) to convert the basic ill-posed

® Ron C. Mittelhammer, Regents Professor of Economic Sciences and Statistics, Washington State University, Pullman, WA, 99164, (email: mittelha@wsu.edu).
George G. Judge, Professor in the Graduate School, 207 Giannini Hall, University of California, Berkeley, Berkeley, CA, 94720 (e-mail: judge@are.berkeley.edu).
The authors gratefully acknowledge the helpful and substantive comments of Martin Burda, Marian Grendar, Guido Imbens, Joanne Lee, Arthur Lewbel, Art Owen, Paul Ruud, Kenneth Train, and David Wolpert, on concepts related to this paper.

inverse problem into a well-posed one that can be analyzed via conventional parametric statistical methods. While this may have made DCB models amenable to traditional estimation and inference procedures, questions arise about the appropriate parametric statistical model choice. Recognizing, in a DCB model context, the statistical problems associated with using traditional parametric estimation and inference procedures when the statistical model is suspect, we focus on information-theoretic methods (Cover and Thomas, 2006) that acknowledge inherent model and data uncertainty and allow for the possibility of a wide class of legitimate CDFs underlying the statistical model of the data sampling process, with corresponding estimators for the unknowns of the model. Enlarging the set of legitimate CDFs and concomitant estimation and inference rules, for DCB problems, is the topic of this paper.

1.1. The Parametric and Semi-Parametric Base

In binary response models, it is assumed that, on trial i=1,2,...,n, one of two alternatives is observed to occur for each independent binary random variable Y_i, i=1,2,...,n, having its respective probability p_i, i=1,2,...,n, of success[1]. In empirical applications, the data sampling process for the binary random variable Y_i is often specified as a function of the latent variable, Y_i^*:

$$Y_i^* = \mathbf{x}_{i.}\boldsymbol{\beta} + \varepsilon_i^* \tag{1.1}$$

where $Y_i \equiv I\left(Y_i^* > 0\right)$, i = 1,...,n, are independent Bernoulli random variables, $I(A)$ is an indicator function that takes the value "1" when condition (A) is true and takes the value "0" otherwise, and $\mathbf{x}_{i.}$, i = 1,..., n, are independent outcomes of a $(1\times k)$ random vector of response variables. Here, and elsewhere, the linear index $\mathbf{x}_{i.}\boldsymbol{\beta}$ can be replaced by the more general functional notation $m(\mathbf{x}_{i.},\beta)$ if the effect of the response variables on the latent variable is thought to be nonlinear.

Given (1.1), the value of p_i is

$$p_i = P(y_i = 1) = P\left(e_i^* > -\mathbf{x}_{i.}\boldsymbol{\beta}\right) = 1 - G\left(-\mathbf{x}_{i.}\boldsymbol{\beta}\right) = G_*\left(-\mathbf{x}_{i.}\boldsymbol{\beta}\right) \tag{1.2}$$

where $G(\cdot)$ is the CDF of the noise term ε_i^* in latent variable equation (1.1) and $G_*(\cdot)$ is the complement of this CDF. When the parametric family of probability density functions underlying the binary response model is assumed known, the parametric functional form of $G(\mathbf{x}_{i.}\boldsymbol{\beta})$ is also known. Therefore, one can fully define the log-likelihood function and utilize the traditional maximum likelihood (ML) approaches of logit or probit as a basis for estimation and inference relative to the unknown $\boldsymbol{\beta}$ and the choice probabilities $G(\mathbf{x}_{i.}\boldsymbol{\beta})$. If the particular choice of the parametric functional form for the distribution is correct, then the usual ML properties of consistency, asymptotic normality, and efficiency hold (McFadden, 1974; McFadden, 1984 and Train, 2003).

[1] A scalar random variable is denoted by X or Y. Multivariate random variables (vector or matrix) are denoted by a bold capital letter **X** or **Y**. A subscripted index on a vector indicates particular row or column elements of the vector. For example, $\mathbf{X}_i$ denotes the i^{th} row of **X**, and $\mathbf{X}_{.i}$ denotes the j^{th} column. Observed outcomes or fixed values are denoted by lower case letters. Exceptions to these conventions include e being an outcome of random ε, and $\hat{\mathrm{b}}$ being an outcome of random $\hat{\beta}$

In reality, there is most often substantial ambiguity surrounding the "correct" behavioral model. Thus uncertainty exists regarding the underlying data sampling process and how best to proceed with model specification, estimation, and inference. This has led to the creation of semi-parametric methods suggested by Cosslett (1983), Maddala (1983), Ichimura (1993), Klein and Spady (1993), and McCullough and Nelder (1995). However, these methods usually rely on a restricted set of assumptions and their resultant conditional nature is apparent. We assume that the distribution of ε_i^* *is neither* based on, nor restricted to, the conventional logit and probit parametric family and suggest a range of CDF's and empirical estimators to recover estimates of the choice probabilities and corresponding derivatives with respect to the response variables. Sample information is represented in a nonparametric way through sample moments. This class of CDFs is based on the minimum power divergence (MPD) principle derived from the Cressie-Read family of divergence measures.

1.2. Topical Map

The organization of the paper is as follows: in Section 2, a nonparametric representation of the binary response model is formulated in terms of conditional moments. Section 3 defines a wide class of CDFs is defined whose members *i*) are consistent with a nonparametric specification of the binary response model, *ii*) satisfy moment conditions involving the response variables and binary outcomes, and *iii*) are minimally power divergent from reference distributions for the Bernoulli probabilities. In Section 4, the class of MPD CDFs is used in an application of the Minimum Power Divergence Principle to define a new class of estimators for the unknown Bernoulli probabilities and their derivatives with respect to the response variables, and asymptotic sampling properties of the estimators are noted. An illustrative numerical example of the application of the methodology is presented in section 5. Finally, in section 6 implications of the formulations are discussed, and a number of possible estimator variants and extensions are noted for future research.

2. MODEL OF BINARY RESPONSE

We seek the class of CDFs that is congruent with basic and generally applicable conditions relating to the binary response model. These conditions include *i*) a generally applicable nonparametric statistical model specification of the Bernoulli outcomes reflecting signal and noise components, *ii*) a simple orthogonality condition between response variables and the noise component, and *iii*) minimum divergence between members of the CDF class and any possible reference distribution for the Bernoulli probabilities underlying the binary response model. The resultant class of CDFs contains a flexible collection of CDFs that subsumes the logistic distribution as a special case, and the overall approach provides an alternative statistical rationale for the specification of a logit model of binary response.

2.1. Nonparametric Representation of Binary Responses and Conditional Moments

Seeking to minimize the invocation of model specification information that the researcher usually does not possess, we begin by assuming that the vector of Bernoulli random variables, **Y**, adheres to the very general statistical model

$$\mathbf{Y} = \mathbf{p} + \boldsymbol{\varepsilon}, \text{ where } E(\boldsymbol{\varepsilon}) = \mathbf{0} \text{ and } \mathbf{p} \in \overset{n}{\underset{i=1}{\times}} (0,1) \tag{2.3}$$

The specification in (2.3) implies only that the expectation of **Y** is some mean vector of Bernoulli probabilities $\mathbf{p}$, and that outcomes of **Y** can be decomposed into their means and

noise terms. The noise term outcomes each have dichotomous support, taking the value $-p_i$ with probability $(1-p_i)$ and $1-p_i$ with probability p_i. If $\mathbf{Y}$ is a vector of binary random variables, the specification in (2.3) is in fact *always true*.

Next, thinking in the context of an economic model such as the random utility model, it is assumed that the Bernoulli probabilities in (2.3) depend on the values of response variables $\mathbf{Z}$ through some general conditional expectation relationship that is given by $\mathbf{E}(\mathbf{Y}\,|\,\mathbf{Z}) = \mathbf{p}(\mathbf{Z}) = \left[\mathbf{p_1}(\mathbf{Z_1}), \mathbf{p_2}(\mathbf{Z_2}), ..., \mathbf{p_n}(\mathbf{Z_n})\right]'$, with the conditional orthogonality condition $E\left[\mathbf{Z}'\left(\mathbf{Y}-\mathbf{p}(\mathbf{Z})\right)|\,\mathbf{Z}\right] = \mathbf{0}$ therefore implied. An application of the double expectation theorem yields the unconditional orthogonality result

$$E\left[\mathbf{Z}'\left(\mathbf{Y}-\mathbf{p}(\mathbf{Z})\right)\right] = \mathbf{0}. \tag{2.4}$$

We emphasize that both the conditional and unconditional moment relationships are true subject only to the very general requirement that the Bernoulli probabilities have some functional regression relationship with the response variables in $\mathbf{Z}$, expressed via a conditional expectation relationship $\mathbf{E}(\mathbf{Y}\,|\,\mathbf{Z})$. In fact, there is essentially no risk of model misspecification at this point given that *some* regression relationship exists between $\mathbf{Y}$ and $\mathbf{Z}$.

We note that under the general regression specification $\mathbf{E}(\mathbf{Y}\,|\,\mathbf{Z}) = \mathbf{p}(\mathbf{Z})$, it also follows that $E\left[\mathbf{g}(\mathbf{Z})'\left(\mathbf{Y}-\mathbf{p}(\mathbf{Z})\right)\right] = \mathbf{0}$, for any function $\mathbf{g}(\mathbf{Z})$ for which the expectation exists. This is a natural reflection of the fact that the residuals of a general nonparametric regression relationship are orthogonal in expectation to any measurable function of the regressors. We note later that an important implication of this fact is that the functional form of the index used in the CDF that underlies the Bernoulli probabilities of the binary response model is determined by the choice of $\mathbf{g}(\mathbf{Z})$. Also, while the regression relationship $\mathbf{E}(\mathbf{Y}\,|\,\mathbf{g}(\mathbf{Z})) = \mathbf{p}(\mathbf{g}(\mathbf{Z}))$ implies $E\left[\mathbf{g}(\mathbf{Z})'\left(\mathbf{Y}-\mathbf{p}(\mathbf{g}(\mathbf{Z}))\right)\right] = \mathbf{0}$, the moment condition $E\left[\mathbf{Z}'\left(\mathbf{Y}-\mathbf{p}(\mathbf{g}(\mathbf{Z}))\right)\right] = \mathbf{0}$, does not necessarily follow because of the conditioning on the more restricted space of $\mathbf{Z}$-outcomes. The regression relationship presumes specific additional knowledge relating to functional form. For now, we proceed under the assumption of the more general nonparametric regression assumption, and the moment condition (2.4). This will eventually lead us to the ubiquitous linear index model within the binary choice framework.

The preceding model assumptions represent a very basic level of information for estimating the unknown Bernoulli probabilities that is no more stringent than a fully nonparametric regression representation of binary response. Adding functional and/or statistical characteristics to the model specification would require additional sample and/or non-sample information. This type of information is substantially more uncertain and, when used, is most often simply assumed rather than truly known.

Given (2.4), and the assumption that the latent variable representation of the Bernoulli outcomes in (1.1) applies, a natural candidate for the elements of the $\mathbf{Z}$ matrix is the $(n \times k)$ matrix $\mathbf{X}$ associated with (1.1). We proceed by letting $\mathbf{X}$ denote response variables that affect

the values of the binary response probabilities, but it is not necessary that the genesis of the **X** variables be based on the latent variable representation (1.1).

If the probabilities **p** could be given an explicit parametric functional form, say as $\mathbf{p}=\mathbf{G}(\mathbf{x}\boldsymbol{\beta})$ with $G(\cdot)$ being some cumulative distribution function, then the moment equations can be specified as $E\left(\mathbf{X}'\left(\mathbf{Y}-\mathbf{G}\left(\mathbf{X}\boldsymbol{\beta}\right)\right)\right)=\mathbf{0}$. Empirical representations of these moments, as $n^{-1}\left(\mathbf{x}'\left(\mathbf{y}-\mathbf{G}\left(\mathbf{x}\boldsymbol{\beta}\right)\right)\right)=\mathbf{0}$, could form the basis for a nonlinear generalized method of moments (GMM) approach to estimating the unknown parameter vector and Bernoulli probabilities. However, in the context of (2.4), $G(\cdot)$ is neither assumed known nor explicitly specified so that a GMM approach to estimating the binary response model using moments of the type (2.4) is not possible. Moreover, it is clear that the empirical moments

$$n^{-1}\left(\mathbf{x}'\left(\mathbf{y}-\mathbf{p}\left(\mathbf{x}\right)\right)\right)=\mathbf{0} \tag{2.5}$$

cannot possibly be used in isolation to identify the Bernoulli probabilities since, regardless of their number, $\mathbf{p}(\mathbf{x})=\mathbf{y}$ always solves the set of moment constraints. In addition, there are more unknowns than estimating equations since all that has been assumed to this point is that $\mathbf{p}(\mathbf{x})$ is an unknown nonparametric vector function varying over a function space, and thus currently consists of *n* fully unknown values. Consequently, the system of equations (2.5) is substantially underdetermined and will not provide a unique interior solution for the probability vector **p**. We seek an extremum basis for choosing among the infinite number of solutions for **p**.

3. MINIMUM POWER DIVERGENCE CDFS FOR THE BINARY RESPONSE MODEL

Given the sample binary outcome representation (2.3) and the representation of sample information in the form of empirical moments (2.5), we consider a criterion for determining a class of CDFs that is both consistent with these representations and minimally divergent from any reference distributions for the Bernoulli probabilities underlying the binary outcomes. In this context, consider the determination of Bernoulli probabilities by minimizing some member of the family of generalized Cressie-Read (CR) power divergence measures, (Cressie and Read, 1984; Read and Cressie, 1988; Mittelhammer et al., 2000)

$$\min_{p_{ij}'s}\left\{\sum_{i=1}^{n}\left(\frac{1}{\gamma(\gamma+1)}\sum_{j=1}^{2}p_{ij}\left[\left(\frac{p_{ij}}{q_{ij}}\right)^{\gamma}-1\right]\right)\right\} \tag{3.6}$$

subject to:

$$\sum_{j=1}^{2}\mathrm{p}_{ij}=1,\ \mathrm{p}_{ij}\geq 0,\ \forall \mathrm{i},\mathrm{j} \tag{3.7}$$

$$\sum_{j=1}^{2}\mathrm{q}_{ij}=1,\ \mathrm{q}_{ij}\geq 0,\ \forall \mathrm{i},\mathrm{j} \tag{3.8}$$

$$n^{-1}\left(\mathbf{x}'\left(\mathbf{y}-\mathbf{p}\right)\right)=\mathbf{0} \tag{3.9}$$

The Bernoulli process underlying the binary outcomes for each observation is characterized by the probabilities $\{\mathrm{p}_{i1},\mathrm{p}_{i2}\}$, where $E\left(y_i\mid\mathbf{x}_{i.}\right)=p_{i1},\ \forall i$, and, in general, $E\left(\mathbf{Y}\mid\mathbf{x}\right)=\mathbf{p}$. The

parenthetical component $\left(\frac{1}{\gamma(\gamma+1)}\sum_{j=1}^{2} p_{ij}\left[\left(\frac{p_{ij}}{q_{ij}}\right)^{\gamma}-1\right]\right)$ in the estimation objective function refers to the CR power divergence of the Bernoulli probability distribution $\{p_{i1}, p_{i2}\}$ from any respective *reference* Bernoulli distributions $\{q_{i1}, q_{i2}\}$. Regarding interpretation of the CR divergence measure, this parenthetical component is proportional to the weighted average deviation of $\left(\frac{p}{q}\right)^{\gamma}$ from 1, with the Bernoulli probabilities $\{p_{i1}, p_{i2}\}$ being the weights. The corresponding probability ratios being averaged are $\left(\frac{p_{i1}}{q_{i1}}\right)^{\gamma}$ and $\left(\frac{p_{i2}}{q_{i2}}\right)^{\gamma}$. The CR divergence measure is strictly convex in the p_{ij}'s, and assumes an unconstrained unique global minimum when $p_{ij} = q_{ij}$, $\forall i$ and j.

The constraints (3.7) and (3.8) are necessary conditions required for the p_{ij}'s and q_{ij}'s to be interpreted as probabilities, and for the collection of these probabilities to represent proper probability distributions. The constraint (3.9) is the empirical implementation of the moment condition $E\left(\mathbf{X}'\left(\mathbf{Y}-\mathbf{p}\right)\right)=\mathbf{0}$. There may be additional sample and/or nonsample information about the data sampling processes that is known and, if so, this type of constraint can be imposed in the constraint set. However, as argued above, the constraints in the estimation problem defined above represent a minimalist set of data and model specification information to impose on the behavior of dichotomous outcomes under the assumption that the Bernoulli probabilities underlying the problem are functionally related to some set of response variables, **X**.

Regarding the genesis of the reference probabilities, they could originate from a number of sources. For one, they could be specified *a priori* if prior information were available on the values of the Bernoulli probabilities, and in the case of pure ignorance, one could specify these prior probabilities as the discrete uniform distribution $p_{i1} = p_{i2} = .5$, $\forall i$. Another source could be a simple sample mean estimate of $P(y=1)$ which would represent a sample-based estimate of the Bernoulli probabilities under ignorance of the effect of regressor conditioning (i.e., an unconditional estimate of the Bernoulli probabilities). Alternatively, the reference probabilities could originate from an application of an alternative data-based method of estimating the conditional probabilities, such as Maximum Likelihood (e.g., logit, probit), a linear probability model, or a nonparametric Kernel-density approach. In any case, the reference probabilities would then be viewed as a base set of probability estimates, and the MPD estimates would be the values of the probabilities that are least divergent from the base probabilities while satisfying the data constraints, those being (2.5) in the current discussion.

3.1. Identifying the Class of CDFs Underlying **p**

Henceforth defining $p_i \equiv p_{i1}$ and $q_i \equiv q_{i1}$, the divergence minimization problem in (3.6)-(3.9) can be characterized in Lagrange form as

$$L(\mathbf{p},\boldsymbol{\lambda}) = \sum_{i=1}^{n}\left(\frac{1}{\gamma(\gamma+1)}\left[p_i\left(\frac{p_i}{q_i}\right)^{\gamma} + (1-p_i)\left(\frac{1-p_i}{1-q_i}\right)^{\gamma} - 1\right]\right) + \boldsymbol{\lambda}'\mathbf{x}'(\mathbf{y}-\mathbf{p}) \qquad (3.10)$$

subject to: $$0 \le p_i, q_i \le 1,\ \forall i\,. \qquad (3.11)$$

The premultiplier n^{-1} on the moment constraints is suppressed because of its superfluity to the optimal solution. The representations of the p_i's as functions of the response variables and Lagrange multipliers can be defined by solving first order conditions with respect to **p**, appropriately adjusted by the complementary slackness conditions of Kuhn-Tucker theory in the event that inequality constraints are binding. The first-order conditions with respect to the p_i values in the problem imply

$$\frac{\partial L}{\partial p_i} = \mathbf{0} \Rightarrow \left\{ \begin{array}{l} \left(\left(\frac{p_i}{q_i}\right)^{\gamma} - \left(\frac{1-p_i}{1-q_i}\right)^{\gamma} \right) - \mathbf{x}_i\boldsymbol{\lambda}\gamma \\ \left(\ln\left(\frac{p_i}{q_i}\right) - \ln\left(\frac{1-p_i}{1-q_i}\right) \right) - \mathbf{x}_i\boldsymbol{\lambda} \end{array} \right\} = 0 \text{ for } \gamma \left\{ \begin{array}{l} \neq 0 \\ = 0 \end{array} \right\} \qquad (3.12)$$

where $\mathbf{x}_i$ is used to denote the i^{th} row of the matrix **x.**

When $\gamma \le 0$, the solutions are strictly interior to the inequality constraints and the inequality constraints are nonbinding. Accounting for the inequality constraints in (3.10) when $\gamma > 0$, the first-order condition in (3.12) and complementary slackness allows p_i to be expressed as the following function of $\mathbf{x}_i\boldsymbol{\lambda}$:

$$\begin{aligned} p_i(\mathbf{x}_i\boldsymbol{\lambda}) &= \arg_{p_i}\left[\left(\left(\frac{p_i}{q_i}\right)^{\gamma} - \left(\frac{1-p_i}{1-q_i}\right)^{\gamma} \right) = \mathbf{x}_i\boldsymbol{\lambda}\gamma \right] \quad \text{for } \gamma < 0 \text{ and } \mathbf{x}_i\boldsymbol{\lambda} \in \mathbf{R} \\ &= \arg_{p_i}\left[\ln\left(\frac{p_i}{q_i}\right) - \ln\left(\frac{1-p_i}{1-q_i}\right) = \mathbf{x}_i\boldsymbol{\lambda} \right] \quad \text{for } \gamma = 0 \text{ and } \mathbf{x}_i\boldsymbol{\lambda} \in \mathbf{R} \end{aligned} \qquad (3.13)$$

$$= \left\{ \begin{array}{c} 1 \\ \arg_{p_i}\left[\left(\left(\frac{p_i}{q_i}\right)^{\gamma} - \left(\frac{1-p_i}{1-q_i}\right)^{\gamma} \right) = \mathbf{x}_i\boldsymbol{\lambda}\gamma \right] \\ 0 \end{array} \right\} \text{for } \gamma > 0 \text{ and } \mathbf{x}_i\boldsymbol{\lambda} \left\{ \begin{array}{c} \ge \gamma^{-1} q_i^{-\gamma} \\ \in \left(-\gamma^{-1}(1-q_i)^{-\gamma},\ \gamma^{-1} q_i^{-\gamma} \right) \\ \le -\gamma^{-1}(1-q_i)^{-\gamma} \end{array} \right\}$$

A unique solution for $p_i(\mathbf{x}_i\boldsymbol{\lambda})$ necessarily exists by the strict monotonicity (and unboundedness if $\gamma \le 0$) of either

$$\eta(p_i) = \left(\left(\frac{p_i}{q_i}\right)^{\gamma} - \left(\frac{1-p_i}{1-q_i}\right)^{\gamma} \right) \text{ or } \eta(p_i) = \ln\left(\frac{p_i}{q_i}\right) - \ln\left(\frac{1-p_i}{1-q_i}\right)$$

in $p_i \in (0,1)$, for $\gamma \neq 0$ or $\gamma = 0$, respectively. The solution is implicit and does not exist in closed form except on a measure zero set for γ, but because of the strict monotonicity of $\eta(p_i)$ in p_i, the solution is straightforward to find via numerical methods. The strictly increasing nature of $p_i(\mathbf{x}_i\boldsymbol{\lambda})$ in the argument $\mathbf{x}_i\boldsymbol{\lambda}$ for $p_i \in (0,1)$ allows $p_i(\mathbf{x}_i\boldsymbol{\lambda})$ to be

interpreted as a CDF on the appropriate support for $\mathbf{x}_i\boldsymbol{\lambda}$. We also underscore for later use that the inverse CDFs clearly do exist in closed form, as is immediately obvious from the expressions in (3.13).

Explicit closed form solutions for the CDFs exist for all integer-valued γ, such as

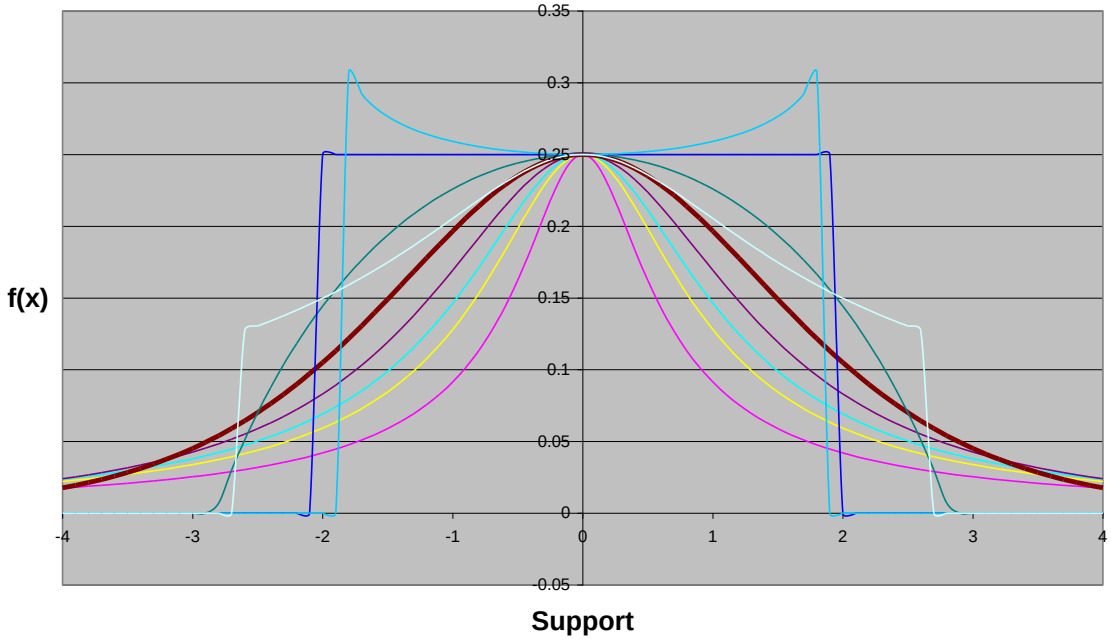
$$p_i\left(\mathbf{x}_i\boldsymbol{\lambda} \mid \gamma=-1\right)=\left\{\begin{array}{c}\left(.5+\dfrac{\left[\left(\mathbf{x}_i\boldsymbol{\lambda}\right)^2+\left(4q_i-2\right)\left(\mathbf{x}_i\boldsymbol{\lambda}\right)+1\right]^{.5}-1}{2\mathbf{x}_i\boldsymbol{\lambda}}\right) \\ .5\end{array}\right\} \text{ if } \mathbf{x}_i\boldsymbol{\lambda}\left\{\begin{array}{c}\neq \mathbf{0} \\ =\mathbf{0}\end{array}\right. \tag{3.14}$$

$$p_i\left(\mathbf{x}_i\boldsymbol{\lambda} \mid \gamma=\mathbf{0}\right)=\frac{q_i \exp\left(\mathbf{x}_i\boldsymbol{\lambda}\right)}{\left(1-q_i\right)+q_i \exp\left(\mathbf{x}_i\boldsymbol{\lambda}\right)}. \tag{3.15}$$

$$p_i\left(\mathbf{x}_i\boldsymbol{\lambda} \mid \gamma=\mathbf{1}\right)=\left\{\begin{array}{c}1 \\ \left(q_i+q_i\left(1-q_i\right)\mathbf{x}_i\boldsymbol{\lambda}\right) \\ 0\end{array}\right\} \text{ for } \mathbf{x}_i\boldsymbol{\lambda}\left\{\begin{array}{c}\geq q_i^{-\gamma} \\ \in\left(-\left(1-q_i\right)^{-\gamma},\, q_i^{-\gamma}\right) \\ \leq -\left(1-q_i\right)^{-\gamma}\end{array}\right\} \tag{3.16}$$

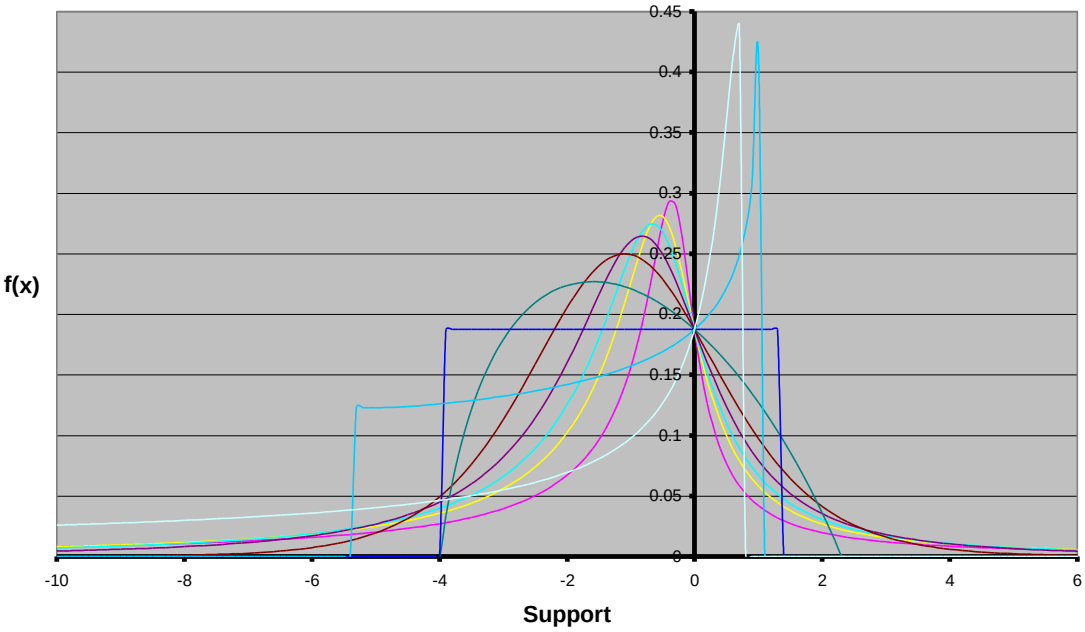
The integer values -1, 0, and 1 correspond, respectively, to the so-called Empirical Likelihood, Exponential Empirical Likelihood, and Log Euclidean Likelihood choices for measuring divergence via the Cressie-Read statistic. The functional form for p_i in (3.15) coincides with the usual standard *logistic* binary choice model if the reference distribution is such that $q_i = .5$. The CDF in (3.16) subsumes the *linear probability model*.

Figure 3.1 PDFs for q = .5, γ = -3, -1.5, -1, -.5, 0, .5, 1, 1.5, and 3



As an illustration of the myriad of distributional characteristics contained within the MPD-Class of probability distributions, Figures 3.1 and 3.2 contain plots of the PDFs associated with selected values of q_i and γ.

Figure 3.2 PDFs for q = .75, γ = -3, -1.5, -1, -.5, 0, .5, 1, 1.5, and 3



4. MINIMUM POWER DIVERGENCE PRINCIPLE FOR ESTIMATION AND INFERENCE

The expansive and flexible set of probability distributions in the MPD-Class provides a corresponding basis for applying the Minimum Power Divergence Principle (MPDP) to estimation and inference relative to the unknown binary response probabilities. In addition, the MPDP framework provides a basis for estimating the marginal effects of changes in the response variables and pursuing inference relating to hypotheses about the binary response process.

4.1. MPDP Estimation

In the context of the model of binary response outlined in Section 2, consider the minimum power divergence problem depicted by the Lagrange multiplier characterization in (3.10)-(3.11). A solution results in an estimator for the binary probabilities that, given the sample moment constraints, is minimally divergent from the reference distribution specified. This solution can, as delineated in the discussion relating to (3.12)-(3.13), be characterized in terms of functions of the Lagrange multipliers via solutions of first-order conditions.

It is possible to generate MPD-estimates of the Lagrange multipliers and then, in turn, produce MPD-estimates of the Bernoulli probabilities that are purely a function of the sample data. The divergence-minimizing estimate of $\boldsymbol{\lambda}$ is determined by substituting the functional representation of $p_i(\mathbf{x}_i\boldsymbol{\lambda})$ into the first order conditions with respect to $\boldsymbol{\lambda}$, and then solving the equations:

$$\boldsymbol{\lambda}_{\text{MPD}} = \arg_{\boldsymbol{\lambda}} \left\{ \mathbf{x}'\left(\mathbf{y} - \mathbf{p}\left(\mathbf{x}\boldsymbol{\lambda}\right)\right) = \mathbf{0} \right\} \tag{4.17}$$

for the value of $\boldsymbol{\lambda}$. The estimated value of $\mathbf{p}$ follows directly by substitution, as $\mathbf{p}_{\text{MPD}} = \mathbf{p}\left(\mathbf{x}\boldsymbol{\lambda}_{\text{MPD}}\right)$.

As in all Lagrange-form optimization problems, $\boldsymbol{\lambda}$ reflects the marginal change in the objective function with respect to a marginal change in the constraint equations. In the current context, the $k \times 1$ vector $\boldsymbol{\lambda}$ can be thought of as representing the "relative contribution" of each of the k data constraints to the minimized divergence value. The polar case $\lambda_i = 0$ indicates that the i^{th} data constraint is non-binding and redundant and adds no informational value to that already contained in the reference distribution for those probabilities. It is not apparent from general Lagrange multiplier theory that $\boldsymbol{\lambda}$ can actually be interpreted as an estimate of the parameter vector $\boldsymbol{\beta}$ underlying the linear index representation of the Bernoulli probabilities depicted in (1.1)-(1.2). We motivate this interpretation next.

4.2 Interpreting $\boldsymbol{\lambda}$ as an Estimator of $\boldsymbol{\beta}$

Suppose one could actually utilize the true conditional population moments in the MPD estimation problem as $n^{-1}\left(\mathbf{E}\left(\mathbf{x}'\left(\mathbf{y}-\mathbf{p}\right)\right)\right) = n^{-1}\left(\mathbf{x}'\left(\mathbf{F}\left(\mathbf{x}\boldsymbol{\beta}\right)-\mathbf{p}\right)\right) = \mathbf{0}$, where $\mathbf{F}\left(\mathbf{x}\boldsymbol{\beta}\right)$ is the actual CDF defining the Bernoulli probabilities. The Lagrange form of the problem would then be

$$L\left(\mathbf{p},\boldsymbol{\lambda}\right) = \sum_{i=1}^{n}\left(\frac{1}{\gamma\left(\gamma+1\right)}\left[p_i\left(\frac{p_i}{q_i}\right)^{\gamma} + \left(1-p_i\right)\left(\frac{1-p_i}{1-q_i}\right)^{\gamma} - 1\right]\right) + \boldsymbol{\lambda}'\left[n^{-1}\mathbf{x}'\left(\mathbf{F}\left(\mathbf{x}\boldsymbol{\beta}\right)-\mathbf{p}\right)\right],\ 0 \le p_i, q_i \le 1,\ \forall i \tag{4.18}$$

The first-order conditions with respect to $\mathbf{p}$ would be precisely as indicated in (3.12), leading to the same representations of the optimal $\mathbf{p}$ expressed in terms of $\boldsymbol{\lambda}$, i.e., the same $\mathbf{p}\left(\mathbf{x}\boldsymbol{\lambda}\right)$ vector of probabilities represented by (3.13).

Now suppose further that the probability model is specified correctly in the sense that the MPD-distribution matches the functional form of the true underlying probability distribution $\mathbf{F}\left(\mathbf{x}\boldsymbol{\beta}\right)$. The first order conditions with respect to $\boldsymbol{\lambda}$ imply that

$$\mathbf{H}\left(\boldsymbol{\beta},\boldsymbol{\lambda}\right) \equiv n^{-1}\mathbf{x}'\left(\mathbf{F}\left(\mathbf{x}\boldsymbol{\beta}\right) - \mathbf{F}\left(\mathbf{x}\boldsymbol{\lambda}\right)\right) = \mathbf{0} \tag{4.19}$$

in which case it is apparent that one solution for $\boldsymbol{\lambda}$ is given directly by $\boldsymbol{\lambda} = \boldsymbol{\beta}$. That this is the unique solution to the problem follows from the Implicit Function Theorem, which can be used to demonstrate that (4.19) determines $\boldsymbol{\lambda}$ as a function of $\boldsymbol{\beta}$ in the neighborhood of $\boldsymbol{\beta}$. By this theorem, if the Jacobian of the k constraints in (4.19) with respect to the $k \times 1$ vector $\boldsymbol{\lambda}$ is nonsingular when evaluated at $\boldsymbol{\lambda} = \boldsymbol{\beta}$, then such a functional relationship exists. The Jacobian is given by

$$\left.\frac{\partial \mathbf{H}\left(\boldsymbol{\beta},\boldsymbol{\lambda}\right)}{\partial \boldsymbol{\lambda}}\right|_{\boldsymbol{\beta}=\boldsymbol{\lambda}} = -n^{-1}\mathbf{x}'\left(\mathbf{f}\left(\mathbf{x}\boldsymbol{\beta}\right) \quad \mathbf{x}\right) = -n^{-1}\sum_{i=1}^{n} f\left(\mathbf{x}_i\boldsymbol{\beta}\right)\mathbf{x}_i\mathbf{x}_i' \tag{4.20}$$

where $\mathbf{f}\left(\mathbf{x}\boldsymbol{\beta}\right)$ is the vector of true underlying probability density function values. The Jacobian is negative definite and thus nonsingular under the mild assumption that there are k or more rows of $\mathbf{x}$ that are not linearly independent and for which $f\left(\mathbf{x}_i\boldsymbol{\beta}\right) > 0$. It follows that

$\boldsymbol{\lambda}$ equals $\boldsymbol{\beta}$ in the solution to (4.19). Using observable sample moments in place of unobservable population moments, as $n^{-1}\mathbf{x}'\left(\mathbf{Y}-\mathbf{F}\left(\mathbf{x}\boldsymbol{\lambda}\right)\right)=n^{-1}\mathbf{x}'\left(\mathbf{F}\left(\mathbf{x}\boldsymbol{\beta}\right)-\mathbf{F}\left(\mathbf{x}\boldsymbol{\lambda}\right)+\boldsymbol{\varepsilon}\right)=\mathbf{0}$, the solution for $\boldsymbol{\lambda}$ is then interpretable as a random variable estimating $\boldsymbol{\beta}$ which, for example, is consistent under familiar regularity conditions that include $n^{-1}\mathbf{x}'\boldsymbol{\varepsilon}\xrightarrow{p}\mathbf{0}$.

It can be shown under general regularity conditions[2] that the estimator of $\hat{\boldsymbol{\lambda}}$ is asymptotically normally distributed. The limiting distribution of the estimator is defined by

$$n^{1/2}\left(\hat{\boldsymbol{\lambda}}-\boldsymbol{\beta}\right)\xrightarrow{d}N\left(0,\mathrm{A}^{-1}\mathbf{V}\mathrm{A}^{-1}\right),$$

where $\mathbf{A}=\dfrac{\partial\mathbf{G}\left(\boldsymbol{\lambda}\right)}{\partial\boldsymbol{\lambda}}\equiv E\left(f\left(\mathbf{X}_{1.}\boldsymbol{\beta}\right)\mathbf{X}_{1.}'\mathbf{X}_{1.}\right)$ and $\mathbf{V}=E\left(F\left(\mathbf{X}_{1.}\boldsymbol{\beta}\right)\left(1-F\left(\mathbf{X}_{1.}\boldsymbol{\beta}\right)\right)\mathbf{X}_{1.}'\mathbf{X}_{1.}\right)$. This result enables all of the usual hypothesis testing methodology that is reliant an asymptotic normal distribution theory.

4.3. Estimating the Marginal Probability Effects of Changes in Response Variables

In empirical work, the effect that changes in response variables have on the probabilities of the discrete choices being realized is often a focal point of analysis. Estimates of these marginal probability effects, represented by $\partial p_i/\partial x_{ij}$ for the j^{th} response variable and the i^{th} binary response probability, are straightforwardly defined in the case of the fully parametric logit and probit models as

$$\text{Logit: }\frac{\partial\hat{p}_i}{\partial x_{ij}}=\frac{\exp\left(\mathbf{x}_{i.}\hat{\boldsymbol{\beta}}\right)}{\left[1+\exp\left(\mathbf{x}_{i.}\hat{\boldsymbol{\beta}}\right)\right]^2}\hat{\boldsymbol{\beta}}_j \tag{4.21}$$

$$\text{Probit: }\frac{\partial\hat{p}_i}{\partial x_{ij}}=\phi\left(\mathbf{x}_{i.}\hat{\boldsymbol{\beta}}\right)\hat{\beta}_j \tag{4.22}$$

where $\phi(\cdot)$ is the standard normal probability density function.

In the case of the MPD-Class of estimators, marginal probability effects, derived by differentiating the appropriate definition of $p_i\left(\mathbf{x}\boldsymbol{\lambda}\right)$ with respect to the response variables used in defining the linear index, yield the following:

$$\gamma<0:\ \frac{\partial p_i}{\partial x_{ij}}=\frac{\lambda_{\mathrm{j}}}{\left(q_i^{-\gamma}p_i^{\gamma-1}+\left(1-q\right)^{-\gamma}\left(1-p_i\right)^{\gamma-1}\right)} \tag{4.23}$$

$$\gamma=0:\ \frac{\partial p_i}{\partial x_{ij}}=\frac{q_i\left(1-q_i\right)^{-1}\exp\left(\mathbf{x}_{i.}\boldsymbol{\lambda}\right)}{\left[1+q_i\left(1-q_i\right)^{-1}\exp\left(\mathbf{x}_{i.}\boldsymbol{\lambda}\right)\right]^2}\lambda_{\mathrm{j}} \tag{4.24}$$

$$\gamma>0:\ \frac{\partial p_i}{\partial x_{ij}}=\left\{\begin{array}{c}0\\ \left(\dfrac{\lambda_{\mathrm{j}}}{\left(q_i^{-\gamma}p_i^{\gamma-1}+\left(1-q\right)^{-\gamma}\left(1-p_i\right)^{\gamma-1}\right)}\right)\\ 0\end{array}\right\}\text{ for }p_i\left\{\begin{array}{c}\geq 1\\ \in(0,1)\\ \leq 0\end{array}\right\} \tag{4.25}$$

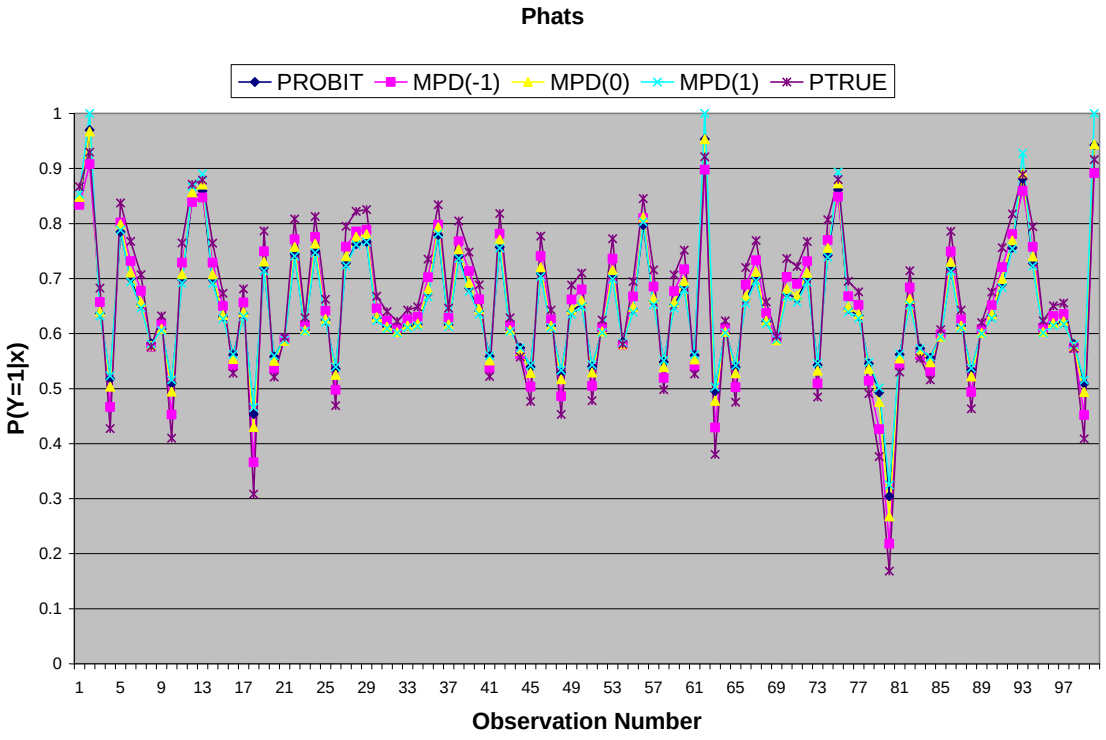
[2] Theorems and proofs of these results can obtained from the authors upon request.

The derivative in (4.24) is recognized as being identical in functional form to the logit derivative defined in (4.21) when $q_i = .5$. Moreover, the solution for $\boldsymbol{\lambda}_{\text{MPD}}$ and the logit estimate of $\hat{\mathbf{b}}$ are in fact identical when $q_i = .5$ since the first-order conditions to both estimation problems coincide.

5. AN ILLUSTRATIVE NUMERICAL EXAMPLE

With an eye toward applicability, in this section we present a numerical illustration of the application of the MPD estimation approach. The example relates to a binary choice problem in which the dependent variable can be thought of as a binary variable (y=1) or (y=0) for a particular alternative (e.g., a commodity, service, candidate, policy initiative, or a payoff in an ultimatum game) is chosen. For this illustration, it is assumed that there is one principal explanatory factor, x, that affects the propensity for choosing the alternative. The outcome of the binary choice is then represented in the form $Y = P(x) + \varepsilon$, where $P(x) = P(y = 1 | x)$. The sample size is n = 100, and the actual data used in this example is presented in the Appendix. Note that the sample of data was generated from a sampling process that drew a random sample of 100 outcomes of $P(y = 1 | x)$ from a Beta(a,b) process, where a=6 and b=3. This resulted in a mean probability value of 2/3. Based on an $MPD(\gamma = -1, q = .5)$ distribution and a linear index of the form $-1 + x$, the probability outcomes were rationalized in terms of the calculated x-values, and probabilities generated from an $MPD(\gamma = -1, q = .5)$ evaluated at the argument value $-1 + x$.

Figure 5.3 Probability Predictions from Binary Response Model Estimators



The MPD estimation problem (3.10)-(3.11) was solved for $\gamma = -1, 0, 1$, and for $q = .5$. The ubiquitous probit estimator was also estimated for comparison purposes. Note in the current

context, the MPD estimator with $\gamma = -1$ is fully consistent with the true data sampling process, and the probit estimator is a quasi-maximum likelihood estimator in this application. The solution to the Lagrangian optimization problem resulted in predictions of $P(y=1|x)$ that are displayed in Figure 5.3, and were calculated based on either the standard normal distribution for the probit estimator, or using (3.14)-(3.16) in the case of the MPD estimators.

The binary response model estimators all produce estimates of the $P(y=1|x)$'s that are in a reasonable neighborhood of their true values, but in the large majority of cases, the MPD estimator based on the distribution $MPD(\gamma=-1, q=.5)$, which is consistent with the data sampling process, produced the higher quality probability predictions. This observation is borne out numerically in Table 5.1, which presents root mean square error (RMSE) statistics for the predictions of the y outcomes and the probabilities, and also presents the % of correct predictions of the y outcomes. It is apparent, especially from the RMSE of $\hat{p}'s$ and the percentage of correct predictions of y outcomes, that the MPD estimator based on $MPD(\gamma=-1, q=.5)$ was the superior estimator in this empirical application.

Y-hat (RMSE)	P-hat (RMSE)	Correct Y Prediction %
0.460667524	0.055663499	88%
0.458269651	0.02880366	94%
0.45984704	0.047025515	90%
0.461830529	0.061114776	86%

Table 5.1 Goodness of Fit Statistics

Regarding the estimates of the linear index coefficients, both the estimated coefficients and the true coefficients underlying the data sampling process are displayed in Table 5.2. These estimates were derived by either applying the maximum likelihood principle in the case of the probit, or else are the values of the Lagrange multipliers in the solution to (3.10)-(3.11). It is again apparent in this application that the MPD estimator based on $MPD(\gamma=-1, q=.5)$ produced estimates of the coefficients that were closer to their true values.

PROBIT	MPD(-1)	MPD(0)	MPD(1)	True B
-0.144530056	-0.64960787	-0.330920414	-0.1761258	-1
0.536319369	1.463482151	0.98324884	0.76295307	2

Table 5.2 Estimated and True Coefficients of the Linear Index

The mean value of the marginal effects of a change in x on the values of $P(y=1|x)$, calculated based on (4.21)-(4.25), are equal to .1898, .2503, .2105 and .1850 for the probit, and the three MPD estimators *for* $\gamma=-1,0,1$ respectively. The true mean value of these marginal effects is given by .3110. Again the MPD estimator with $\gamma=-1$ is superior in this example.

6. IMPLICATIONS, PROBLEMS, AND EXTENSIONS

In this paper, we represent sample information underlying binary choice outcomes through general moment conditions, $\mathrm{E}\left[\mathbf{Z}'(\mathbf{Y}-\mathbf{p})\right]=\mathbf{0}$. We then use the Cressie-Read (CR) family of divergence measures, CR(γ) for γ∈(-∞,∞), to identify a class of CDFs and to solve for the unknown Bernoulli probabilities, **p**. As a result, a large MPD class of corresponding

estimators of the unknown choice probabilities emerges. The solved values of the probabilities are functions of the sample data through data-determined Lagrange Multipliers and are thus not represented in terms of a fixed set of parameters. Estimation implications of this formulation are presented.

A number of important issues relating to the MPD approach in the discrete choice model context remain, and point to a number of problems and extensions of the MPD estimator that can and should be explored in future work. The problems and extensions fall into three general categories, and include: i) the specification of measurable functions of $\mathbf{Z}$ used to define the moment condition $E\left[\mathbf{g}(\mathbf{Z})'\left(\mathbf{Y}-\mathbf{p}(\mathbf{Z})\right)\right]=\mathbf{0}$, ii) the choice of the value of the power parameter γ in the definition of the Cressie-Read power divergence statistic, and iii) the choice of the reference distribution, $\mathbf{q}$, for the Bernoulli choice probabilities.

Regarding problem i), alternative moments of the form $E\left[\mathbf{g}(\mathbf{Z})'\left(\mathbf{Y}-\mathbf{p}(\mathbf{Z})\right)\right]=\mathbf{0}$ can be considered, lead to empirical moment constraints of the form $n^{-1}\mathbf{g}(\mathbf{Z})'\left(\mathbf{Y}-\mathbf{p}(\mathbf{Z})\right)=\mathbf{0}$. The MPD CDFs that arise when such moments are used is precisely of the form (3.13) with the linear index $\mathbf{z}\boldsymbol{\lambda}$ replaced by the generalized linear index $\mathbf{g}(\mathbf{Z})\boldsymbol{\lambda}$. This allows for substantial generality in the index specification associated with the CDF representing the Bernoulli probabilities and essentially leads to a generalized linear model form. An issue here is how to choose the precise functional form of the premultiplier $\mathbf{g}(\mathbf{Z})$. One general possibility would be to specify a flexible functional form, such as a polynomial in the $\mathbf{Z}$ variables. By Weirstrasse's theorem, any continuous index function of Z can be represented arbitrarily closely by a polynomial of sufficient degree. For example a quadratic approximation could be specified as $\mathbf{g}(\mathbf{Z};\boldsymbol{\alpha},\mathbf{H})=\boldsymbol{\alpha}'\mathbf{Z}+\mathbf{Z}'\mathbf{H}\mathbf{Z}$, and then the parameter vector $\boldsymbol{\alpha}$ and symmetric parameter matrix $\mathbf{H}$ could be estimated along with the other unknowns in the model specification. The generalized index function, together with the large class of MPD distributions, would make for a very general set of model specifications for the binary response probabilities.

Regarding problem ii), the use of probability distributions in the MPD-Class can form a basis for an application of the Maximum Likelihood principle, when estimating binary response probabilities and marginal effects of changes in the response variables on those probabilities, for example, consider representing the Bernoulli probabilities by the MPD-Class of distributions and characterizing the observed binary outcomes with the linear index representation

$$\ell(\boldsymbol{\beta},\mathbf{q},\gamma\,|\,\mathbf{y},\mathbf{x})=\sum_{i=1}^{n}\left(y_i\ln\left(1-F\left(-\mathbf{x}_i\boldsymbol{\beta};\mathbf{q},\gamma\right)\right)+(1-y_i)\ln\left(F\left(-\mathbf{x}_i\boldsymbol{\beta};\mathbf{q},\gamma\right)\right)\right) \qquad (6.26)$$

Maximum likelihood estimation of the unknowns in the specification of the Bernoulli probabilities can proceed by maximizing (6.26). This formulation allows the vast array of CDFs in the MPD-Class to represent the Bernoulli probabilities underlying the discrete choice process via selection of the $\mathbf{q}$ and γ values. Given the substantial flexibility exhibited by the family of MPD distributions, it is expected that the maximum likelihood estimator of the Bernoulli probabilities would be robust relative to the true underlying Bernoulli choice probability distribution.

Regarding problem *iii*), one can envision the use of the reference distribution, **q**, to take into account known or estimable characteristics of the Bernoulli probabilities in any particular applied problem. Illustrations of the distributional shape and scale impacts of different **q** reference distribution choices are given in Section 3. The use of E(Y), estimated by the sample mean of **Y** outcomes, could represent a situation where one was allowing for the fact that there may be no regression relationship between the **Z** and **Y** values, and thus the estimator was shrinking the probability estimates to the unconditional mean of Y. One could use the agnostic values $q_i = .5,\ \forall i$ to denote the proposition that the dichotomous outcomes of **Y** were equally likely a priori. Or an alternative estimate $q_i = \hat{P}\left(y_i = 1 | \mathbf{z}_i\right)$, perhaps in the form of a kernel density estimate, could be used as the reference distribution in the MPD estimator specification. In all of these variants, it would be expected that the closer the q_i's were to the true conditional Bernoulli probabilities, the better the MPD estimates of the Bernoulli probabilities would be.

There are other issues that represent extensions of the MPD estimation procedure to other estimation considerations, such as the potential endogeneity of some of the elements in **Z** by considering moment conditions that involve instruments, **W**, of the form $\mathrm{E}\left[\mathbf{W}'\left(\mathbf{Y} - \mathbf{p}\left(\mathbf{Z}\right)\right)\right] = 0$, where **W** is not necessarily a function of **Z** (Judge et al., 2006). Another consideration is the extension of the univariate distribution formulations of this paper to multivariate distributions. One such extension, which yields the multivariate logistic distribution as a special case, begins with a multinomial specification of the minimum power divergence estimation problem in Lagrange form as

$$L\left(\mathbf{p},\boldsymbol{\lambda}\right) = \sum_{i=1}^{n}\sum_{i=1}^{n}\left(\frac{1}{\gamma\left(\gamma+1\right)}\sum_{j=1}^{m} p_{ij}\left[\left(\frac{p_{ij}}{q_{ij}}\right)^{\gamma} - 1\right]\right) + \sum_{j=1}^{m}\boldsymbol{\lambda}_{\mathbf{j}}'\mathbf{x}'\left(\mathbf{y}_{\mathbf{j}} - \mathbf{p}_{\mathbf{j}}\right) + \sum_{i=1}^{n}\eta_i\left(\sum_{j=1}^{m} p_{ij} - 1\right).$$

The estimation and inference potential of this type of formulation needs to be investigated.

These and other problems and extensions of the MPD approach to estimation of the binary response model are the focus of our ongoing research and we hope the research of others.

REFERENCES

Cosslett, S.R. (1983). Distribution-Free Maximum Likelihood Estimation of the Binary Choice Model. *Econometrica*, 51, 765-782.

Cover, T.M. and G.A. Thomas (2006). Elements of Information Theory, New York: Wiley Interscience, 2nd edition.

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

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

Judge, G., R. Mittelhammer, and D. Miller (2006). "Estimating the link function in multinomial response models under endogeneity", (in: Jean-Paul Chavas -Ed., *Volume in Honor of Stanley Johnson*), University of California Press.

Klein, R.W. and R.H. Spady (1993). An Efficient Semiparametric Estimator for Binary Response Models. *Econometrica*, 61 (2), 387-421.

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. (1984). "Qualitative Response Models," (in: Z. Griliches and M. Intriligator -Ed., *Handbook of Econometrics 2*), Amsterdam, North Holland, 1395-1457.

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

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

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

Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*, New York: John Wiley & Sons.

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

DATA APPENDIX

X	Y	X	Y
2.085469	1	0.456264	1
3.775275	1	0.765417	1
0.920827	1	1.273079	0
0.351319	1	0.664946	1
1.739147	1	0.958574	1
1.249493	1	1.818404	1
0.998814	1	1.029649	1
0.655567	0	0.495795	0
0.784314	1	0.997569	0
0.313537	0	1.173193	1
1.234565	1	0.552919	1
2.152973	1	3.396665	1
2.275978	1	0.24603	0
1.233602	1	0.761646	1
0.893696	1	0.449999	0
0.556043	0	1.045188	1
0.91715	0	1.256925	1
0.050008	1	0.849541	0
1.352036	0	0.697049	1
0.542048	1	1.107927	0
0.69321	0	1.051658	1
1.493306	0	1.247389	0
0.775927	1	0.469169	1
1.525616	0	1.483798	0
0.862811	0	2.294302	1
0.437359	0	0.959389	1
1.402679	1	0.900694	0
1.600352	1	0.481789	1
1.627889	1	0.236919	0
0.877827	0	-0.68752	0
0.802608	0	0.559879	1
0.760061	0	1.023944	0
0.810798	1	0.61171	1
0.825372	1	0.53154	0
1.104247	1	0.724173	1
1.70496	1	1.349192	1
0.819931	1	0.810259	1
1.469409	1	0.426032	1
1.157592	1	0.753728	0
0.939004	1	0.899412	1
0.544254	1	1.193455	1
1.569075	1	1.562179	1
0.775076	1	2.472882	1
0.616227	1	1.399881	1
0.453326	0	0.76208	1
1.299014	1	0.830212	0
0.81028	1	0.843309	0
0.405274	0	0.64993	0
0.937509	1	0.311159	0
1.009584	1	3.202306	0