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

Ron Mittelhammer and George Judge

Washington State University
and
University of California, Berkeley

ABSTRACT

The Cressie-Read (CR) family of power divergence measures is used to identify a new class of statistical models and estimators for competing explanations of the data in binary choice models. A large flexible class of cumulative distribution functions and associated probability density functions emerge that subsumes the conventional logit model, and forms the basis for a large set of estimation alternatives to traditional logit and probit methods. Asymptotic properties of estimators are identified, and sampling experiments are used to provide a basis for gauging the finite sample performance of the estimators in this new class of statistical models.

Keywords: binary choice models and estimators, conditional moment equations, squared error loss, Cressie-Read statistic, information theoretic methods, minimum power divergence

AMS 1991 Classification Primary 62E20, 62G08, 62F10

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.

1. INTRODUCTION

Traditionally, the estimation and inference approaches used in empirical analyses of binary choice converts a fundamentally ill-posed inverse problem into a well-posed one that can be analyzed via conventional parametric statistical methods. The typical distributional choice in empirical analyses of binary choice models (BCMs) has been either the probit or logit cumulative distribution function (CDF), followed by maximum likelihood estimation and inference applied to the specific parametric statistical model chosen. The negative statistical consequences associated with following traditional parametric estimation and inference approaches when the assumed statistical model, such as the probit or logit, is suspect are well-known.

In attempts to mitigate model misspecification issues, a wide and varied collection of semiparametric models and estimators for the BCM have arisen in the literature (e.g., Ichimura (1993); Klein and Spady (1993); Gabler, Laisney, and Lechner (1993); Gozalo and Linton (1994); Manski (1975); Horowitz (1992); Han(1987); Cosslet (1983); Wang and Zhou (1993)). These semiparametric methods tend to be of either the "regression-estimating equation" or "maximum likelihood" variety, each utilizing some form of nonparametric estimate of the probability that the binary choice variable takes the value $y_i = 1$, conditional on the outcome of explanatory or response variables. On the basis of asymptotic performance comparisons among those semiparametric estimators for which asymptotics are tractable and well-developed, it is found that many of the estimators do not achieve $\sqrt{n}$ consistency. For those that do, the estimator of Klein and Spady is a dominating estimator in the sense of achieving the semiparametric efficiency bound. However, because 1) intricate regularity assumptions are often necessary to achieve semiparametric performance results, 2) unknown population distributions and/or distribution scores must be replaced with stochastic approximations when forming "optimal" estimating equations or likelihood functions underlying the definition of estimators, 3) one must often choose values of tuning, bandwidth, and/or trimming parameters in the specification of estimators, and 4) finite sample performance can be quite variable relative to what asymptotic results suggests about performance in infinite samples, it is generally not possible to definitively rule out many of the semiparametric alternatives when considering empirical analyses of data underlying BCMs.

In this paper we investigate information-theoretic (IT) methods for addressing both model specification uncertainty and econometric estimation aspects of BCM analyses.[1] The approach leads to a wide and flexible new class of CDFs whose members represent the full set of possible probability distributions that are both consistent with a nonparametric specification of the binary choice model and that are minimally power divergent from any reference distributions for the Bernoulli probabilities. The corresponding statistical models and associated estimators for the BCM represent a large set of alternatives to traditional logit and probit methods as well as existent semiparametric methods of analysis. While the class of statistical models is derived from one of the most general representations of the BCM possible, even with few assumptions relating to econometric model structure the regularity conditions required for establishing $\sqrt{n}$ consistency as well as asymptotic normality are not stringent, and the estimators are tractable to calculate even for large sample sizes. Consistent estimates of the asymptotic covariance matrices of the estimators are straightforwardly determined, enabling the usual asymptotic chi-square tests of hypotheses vis-à-vis functions of model parameters. The new class of probability distributions is interesting in its own right, and has potential for use in a wide array of statistical models outside of the BCM context.

1.1. The Parametric Model Base

Assume that, on trial $i = 1, 2, \ldots, n,$ one of two alternatives is observed to occur for each of the independent binary random variables $Y_i,\ i = 1, 2, \ldots, n,$ with $p_i,\ i = 1, ..., n$ representing the probabilities of observing successes ($y_i = 1$) on each respective trial. The value of p_i is represented as

$$p_i = P\left(y_i = 1 \mid \mathbf{x}_i\right) = F\left(\mathbf{x}_i\boldsymbol{\beta}\right) \tag{1.1}$$

[1] The recent intriguing work of Smith (2007) applies an information theoretic formulation to conditional moment-based models with specification and estimation objectives similar to those of the current paper. Smith's formulation is related to the traditional context of empirical likelihood (EL) methodology in which sample observations are weighted differentially by what amounts to empirical probability or convexity weights. Thus his work uses a locally-weighted form of the divergence criterion. In contrast the formulation in this paper is not of the typical EL variety, and rather applies divergence measures to probability distributions within, as opposed to across sample observations.

where $F(\bullet)$ is some cumulative probability distribution function (CDF), and $\mathbf{x}_i$, i = 1,..., n, are independent outcomes of a $(1\times k)$ random vector of explanatory variables[2]. When the parametric family of probability density functions underlying the binary choice model is assumed known (e.g., logit or probit), one can define the specific functional form of the log-likelihood and utilize traditional maximum likelihood (ML) approaches as a basis for estimation and inference relative to the unknown $\boldsymbol{\beta}$ and the response probabilities $F(\mathbf{x}_i\boldsymbol{\beta})$. If the particular choice of the parametric functional form for the distribution is also *correct*, then the usual ML properties of consistency, asymptotic normality, and efficiency hold (McFadden (1974, 1984), Train (2003)). See Green (2007) for a review of conventional binary choice models and a literature review.

In contrast to historical empirical practice, we assume that the CDF in (1.1) is neither based on, nor restricted to, the conventional logit and probit parametric families and suggest a new and flexible class of statistical models, CDFs, and associated estimators and inference procedures that can be used to recover estimates of the Bernoulli probabilities. Sample information is represented in a very general and robust way through nonparametric conditional expectations or regression functions, and sample moments based on them. The new class of CDFs that results is based on the minimum power divergence (MPD) principle for estimating the nonparametric regression functions. Estimation based on this new class of CDFs is implemented by either solving an extremum problem involving the Cressie-Read family of power divergence measures, or by applying familiar Maximum Likelihood or Nonlinear Least Squares methods.

1.2 Organization of the Paper

In section two, a nonparametric regression representation of the binary choice model is formulated and used to define a generally applicable conditional moments representation of sample information. In section three, a minimum power divergence criterion is applied to identify a new class of CDFs whose members are consistent with the nonparametric specification of the binary choice model and are minimally power

[2] We adopt the convention that capital letters denote random variables and lower case letters denote observed values or outcomes. An exception will be the use of ε to denote a random variable, and e to denote an outcome of the random variable, and the use of F and f to denote a CDF and PDF respectively.

divergent from reference distributions for the Bernoulli probabilities. General properties of this class of CDFs are established. In section 4 the minimum power divergence principle is applied directly to estimate unknowns in the binary choice models, followed by the development of asymptotic sampling properties of the estimators together with hypothesis testing procedures. Maximum likelihood and nonlinear least squares methods are suggested in section 5 to internalize the choice of the optimal MPD-distribution with which to represent conditional Bernoulli probabilities in the binary choice model. Section 6 provides Monte Carlo sampling results to illustrate the finite sampling performance of the estimators. Promising new directions for research and applications based on this new class of statistical models are delineated in the concluding section of the paper.

2. **Nonparametric Regression Representation of Binary Choices**

Seeking to minimize the use of model specification information that the applied econometrician generally does not possess, we begin by assuming that the $n \times 1$ vector of Bernoulli random variables, **Y**, can be modeled by the universally applicable stochastic representation

$$\mathbf{Y} = \mathbf{p} + \boldsymbol{\varepsilon}, \text{ where } E(\boldsymbol{\varepsilon}) = \mathbf{0} \text{ and } \mathbf{p} \in \overset{n}{\underset{i=1}{\times}} (0,1) \tag{2.1}$$

The specification in (2.1) implies only that the expectation of the random vector **Y** is some mean vector of Bernoulli probabilities $\mathbf{p}$, and that outcomes of **Y** are decomposed into their means and noise terms.

The Bernoulli probabilities in (2.1) are assumed to depend on the values of explanatory variables contained in the $(n \times k)$ matrix **X,** whose i^{th} row is $\mathbf{X}_i$, through some general conditional expectation or regression relationship

$\mathbf{E}(\mathbf{Y} \mid \mathbf{X}) = \mathbf{p}(\mathbf{X}) = \left[\mathrm{p}_1(\mathbf{X}_1) \mid \mathrm{p}_2(\mathbf{X}_2) \mid \ldots \mid \mathrm{p}_n(\mathbf{X}_n) \right]'$, where the conditional orthogonality condition $E\left[\mathbf{X}'(\mathbf{Y} - \mathbf{p}(\mathbf{X})) \mid \mathbf{X} \right] = \mathbf{0}$ is implied. It should be emphasized that the functional form of the relationship $\mathbf{p}(\mathbf{X})$ is *not* assumed known and is left unspecified at this point, underscoring the substantial generality and nonparametric nature of this model assumption. An application of the double expectation theorem leads to the unconditional orthogonality condition

$$E\left[\mathbf{X}'\left(\mathbf{Y}-\mathbf{p}\left(\mathbf{X}\right)\right)\right]=\mathbf{0}. \tag{2.2}$$

The information employed to this point represents a minimum set of statistical model assumptions for representing the unknown Bernoulli probabilities in the binary choice model. This formulation indicates that only some general regression relationship exists between regressor variables **X** and **Y**.

3. Minimum Power Divergence Class of CDFs for the Binary Choice Model

Given sampled binary outcomes from (2.1), representation of sample information in the form of the $k<n$ empirical moments, $n^{-1}\mathbf{x}'\left(\mathbf{y}-\mathbf{p}\right)=\mathbf{0},$ connects the data space to the parameter space. At this stage only an infinite feasible set of probabilities are identified and in order to proceed an estimation criterion is needed to address the undetermined nature of $(n-k)$ of the elements in **p**. In this context instead of restricting the feasible set by some ad hoc functional rule, we determine the Bernoulli probabilities by minimizing the generalized Cressie-Read (CR) power divergence measure (Cressie and Read (1984); Read and Cressie (1988); Mittelhammer, et al., (2000)). In particular, for given γ and $q_i\in\left(0,1\right)$, $i=1,...,n$,

$$\min_{p_i,i=1,...,n}\ \sum_{i=1}^{n}\frac{\left(p_i\left(\left(\frac{p_i}{q_i}\right)^{\gamma}-1\right)+\left(1-p_i\right)\left(\left(\frac{1-p_i}{1-q_i}\right)^{\gamma}-1\right)\right)}{\gamma\left(\gamma+1\right)} \tag{3.1}$$

$$\text{s.t. } n^{-1}\left(\mathbf{x}'\left(\mathbf{y}-\mathbf{p}\right)\right)=\mathbf{0} \text{ and } \quad p_i\in\left(0,1\right),\ i=1,...,n \tag{3.2}$$

The summand in the estimation objective function (3.1) refers to the CR power divergence of the Bernoulli probabilities $\left\{p_i,\ 1-p_i\right\}$ from some given *reference* Bernoulli distribution $\left\{q_i,\ 1-q_i\right\}$. We address the specification of the reference distribution ahead.

The constraints in (3.2) represent the empirical implementation of the moment condition $E\left(\mathbf{X}'\left(\mathbf{Y}-\mathbf{p}\right)\right)=\mathbf{0}$ as well as conditions for the $p_i's$ to be interpretable as probabilities. There may be additional sample and/or nonsample information about the data sampling processes that is known and, if so, this type of information can be imposed in the constraint set of the MPD problem. The overall implication of the extremum formulation (3.1) – (3.2) is that the value of **p** is chosen from among the infinite number

of solutions, consistent with the sample moment equations $n^{-1}\mathbf{x}'(\mathbf{y}-\mathbf{p})=\mathbf{0}$, so as to be *minimally divergent* from the reference distribution **q.** Divergence is measured by the CR power divergence statistic. If **q** satisfies the moment conditions, so that $n^{-1}\mathbf{x}'(\mathbf{y}-\mathbf{q})=\mathbf{0}$, then $\mathbf{p}=\mathbf{q}$. Otherwise, MPD is a shrinkage-type estimator, where the solution for **p** is as minimally divergent from **q** as the sample data, in the form of moment constraints, will allow (see Pardo (2006) for a recent discussion of the use of minimum divergence measures for estimation in some statistical model contexts). This estimation approach frees the analyst from the necessity of defining a particular fully-specified parametric distributional structure underlying the Bernoulli probabilities and thus reduces the likely possibility of statistical model misspecification. Furthermore, this nonparametric formulation utilizes very general sample information along with reference probabilities that the Bernoulli probabilities will emulate as closely as the sample information permits. In the context of statistical model uncertainty the objective is to provide an estimation procedure that i) approximates the true underlying data sampling process well, ii) has good estimation and inference sampling performance and iii) improves upon traditional parametric approaches.

3.1 The Class of CDF's Underlying p

The Lagrange form of the divergence minimization problem (3.1)-(3.2), for given γ and $q_i \in (0,1)$, $i=1,...,n$, is

$$L(\mathbf{p},\boldsymbol{\lambda})=\sum_{i=1}^{n}\frac{\left(p_i\left(\left(\frac{p_i}{q_i}\right)^{\gamma}-1\right)+(1-p_i)\left(\left(\frac{1-p_i}{1-q_i}\right)^{\gamma}-1\right)\right)}{\gamma(\gamma+1)}+\boldsymbol{\lambda}'\mathbf{x}'(\mathbf{y}-\mathbf{p}) \tag{3.3}$$

$$\text{s.t.}\quad p_i \in (0,1),\ i=1,...,n \tag{3.4}$$

where the premultiplier n^{-1} on the moment constraints is henceforth suppressed. The $p_i's$ can be expressed as functions of the explanatory variables and Lagrange multipliers by solving first order conditions with respect to **p**, that are adjusted by the complementary slackness conditions of Kuhn-Tucker (1951) theory in the event that

inequality constraints are binding. The first-order conditions with respect to the p_i values in the problem imply

$$\frac{\partial L}{\partial p_i} = \mathbf{0} \Rightarrow \left\{ \begin{array}{c} \left(\left(\frac{p_i}{q_i} \right)^{\gamma} - \left(\frac{1-p_i}{1-q_i} \right)^{\gamma} \right) - \mathbf{x}_i \boldsymbol{\lambda} \gamma \\ \left(\ln\left(\frac{p_i}{q_i} \right) - \ln\left(\frac{1-p_i}{1-q_i} \right) \right) - \mathbf{x}_i \boldsymbol{\lambda} \end{array} \right\} = 0 \text{ for } \gamma \left\{ \begin{array}{c} \neq 0 \\ = 0 \end{array} \right\} \tag{3.5}$$

When $\gamma \leq 0$, the solutions are strictly interior to the inequality constraints so that the inequality constraints are nonbinding. Accounting for the inequality constraints in (3.5) when $\gamma > 0$, the first-order condition in (3.5) and the complementary slackness conditions allows p_i to be expressed as the following function of $\mathbf{x}_i \boldsymbol{\lambda}$

$$\begin{aligned} p\left(\mathbf{x}_i \boldsymbol{\lambda}; q_i, \gamma\right) &= \arg_{p_i} \left[\left(\left(\frac{p_i}{q_i} \right)^{\gamma} - \left(\frac{1-p_i}{1-q_i} \right)^{\gamma} \right) = \mathbf{x}_i \boldsymbol{\lambda} \gamma \right] \qquad \text{for } \gamma \neq 0 \\ &= \arg_{p_i} \left[\ln\left(\frac{p_i}{q_i} \right) - \ln\left(\frac{1-p_i}{1-q_i} \right) = \mathbf{x}_i \boldsymbol{\lambda} \right] \qquad \text{for } \gamma = 0 \qquad (3.6) \\ &= \left\{ \begin{array}{c} 1 \\ \left(\arg_{p_i} \left[\left(\left(\frac{p_i}{q_i} \right)^{\gamma} - \left(\frac{1-p_i}{1-q_i} \right)^{\gamma} \right) \right] = \mathbf{x}_i \boldsymbol{\lambda} \gamma \right) \\ 0 \end{array} \right\} \text{for } \gamma > 0 \text{ and } \mathbf{x}_i \boldsymbol{\lambda} \left\{ \begin{array}{c} \geq \gamma^{-1} q_i^{-\gamma} \\ \in \left(-\gamma^{-1}\left(1-q_i\right)^{-\gamma},\ \gamma^{-1} q_i^{-\gamma} \right) \\ \leq -\gamma^{-1}\left(1-q_i\right)^{-\gamma} \end{array} \right\} \end{aligned}$$

A unique solution for $p\left(\mathbf{x}_i \boldsymbol{\lambda}; q_i, \gamma\right)$ necessarily exists by the continuity and strict monotonicity of either $\eta\left(p_i\right) = \left(\left(\frac{p_i}{q_i} \right)^{\gamma} - \left(\frac{1-p_i}{1-q_i} \right)^{\gamma} \right)$ or $\eta\left(p_i\right) = \ln\left(\frac{p_i}{q_i} \right) - \ln\left(\frac{1-p_i}{1-q_i} \right)$ in $p_i \in (0,1)$, for $\gamma \neq 0$ or $\gamma = 0$, respectively. Because of this strict monotonicity, p_i is a monotonically increasing function of $\mathbf{x}_i \boldsymbol{\lambda}$, and p_i is also bounded between 0 and 1. This implies that the $p\left(\mathbf{x}_i \boldsymbol{\lambda}; q_i, \gamma\right)$ functions can be legitimately interpreted as cumulative probability distribution functions (CDFs) on the respective supports for $\mathbf{x}_i \boldsymbol{\lambda}$. The class of distributions defined via (3.6) characterize the unique set of distributions that are

consistent with the nonparametric representation of the conditional moments and that are minimally divergent from any choice of reference distributions, $\mathbf{q}$.

The CDFs in (3.6) are defined only as *implicit* functions of $\mathbf{x}_i\boldsymbol{\lambda}$, for almost all γ. However, numerical representations of the functional relationship between p_i and $\mathbf{x}_i\boldsymbol{\lambda}$ may be determined rather straightforwardly because of the strict monotonicity of p_i in $\mathbf{x}_i\boldsymbol{\lambda}$. Explicit closed form solutions for the CDFs exist for $p(\mathbf{x}_i\boldsymbol{\lambda}; q_i, \gamma)$ on a measure zero set of γ values that includes the set of all integers. For example

$$p(\mathbf{x}_i\boldsymbol{\lambda}; q_i, -1) = \left\{ \begin{array}{c} \left(.5 + \dfrac{\left[(\mathbf{x}_i\boldsymbol{\lambda})^2 + (4q_i - 2)(\mathbf{x}_i\boldsymbol{\lambda}) + 1 \right]^{.5} - 1}{2\mathbf{x}_i\boldsymbol{\lambda}} \right) \\ .5 \end{array} \right\} \text{ if } \mathbf{x}_i\boldsymbol{\lambda} \left\{ \begin{array}{c} \neq \mathbf{0} \\ = \mathbf{0} \end{array} \right. \tag{3.7}$$

$$p(\mathbf{x}_i\boldsymbol{\lambda}; q_i, 0) = \frac{q_i \exp(\mathbf{x}_i\boldsymbol{\lambda})}{(1 - q_i) + q_i \exp(\mathbf{x}_i\boldsymbol{\lambda})} \tag{3.8}$$

$$p(\mathbf{x}_i\boldsymbol{\lambda}; q_i, 1) = \left\{ \begin{array}{c} 1 \\ \left(q_i + q_i(1 - q_i)\mathbf{x}_i\boldsymbol{\lambda} \right) \\ 0 \end{array} \right\} \text{ for } \mathbf{x}_i\boldsymbol{\lambda} \left\{ \begin{array}{c} \geq q_i^{-\gamma} \\ \in \left(-(1 - q_i)^{-\gamma},\ q_i^{-\gamma} \right) \\ \leq -(1 - q_i)^{-\gamma} \end{array} \right\}. \tag{3.9}$$

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. When $\gamma = 0$ and the reference distribution is such that $q_i = .5$, the functional form for p_i in (3.8) coincides with the standard logistic binary choice model. When $\gamma = 1$, the CDF in (3.9) subsumes the linear probability model. We underscore for future reference that the entire class of *inverse* CDFs exist in closed form.

3.2 Properties of the MPD-Class of Probability Distribution Functions

We use the notation $MPD(q, \gamma)$ to denote a specific member of the MPD-class of distributions identified by particular values of *q* and *γ*. A vast array of symmetric and skewed probability density functions are contained within the MPD-Class of PDFs. To illustrate the range of possibilities, graphs of some members of the Class are presented in

Figures 3.1 and 3.2. These graphs do much to suggest the widely varying distributional possibilities as γ and q takes on different values.

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

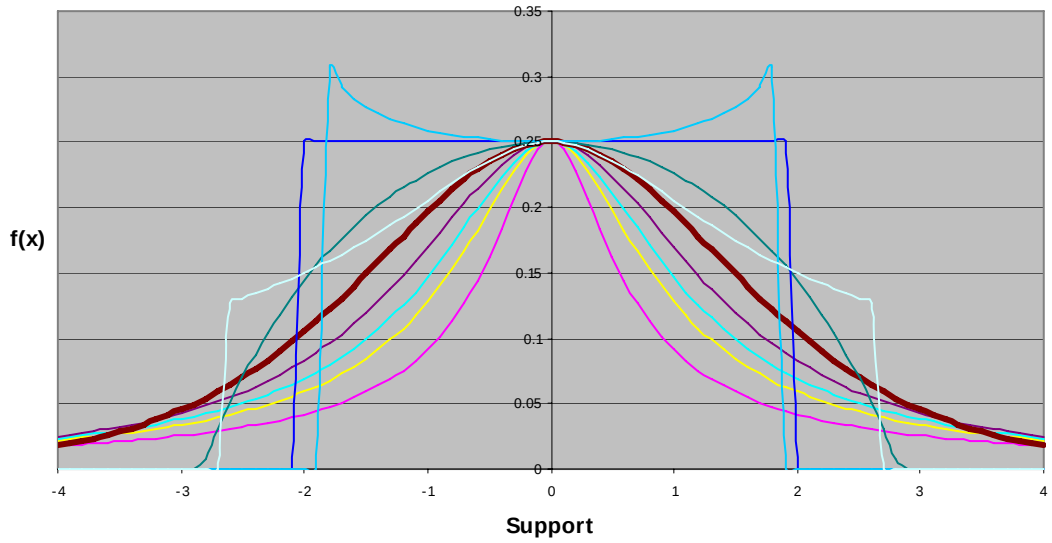
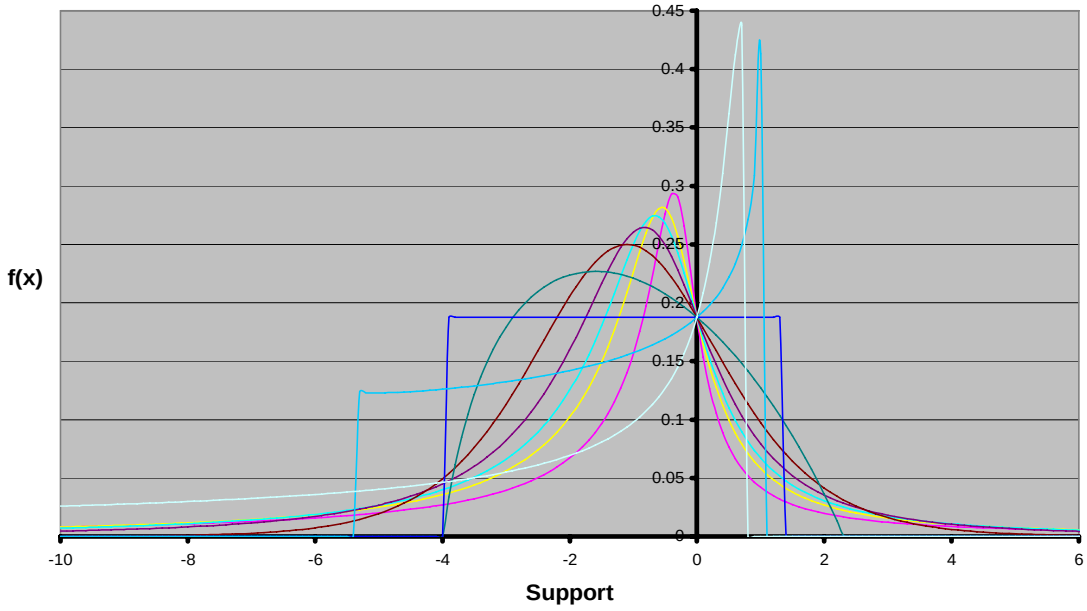


Figure 3.2. PDFs for q = .75 and γ = -3, -1.5, -1, -.5, 0, .5, 1, 1.5, and 3



The existence of moments in the MPD-Class of distributions and their values depend on the γ parameter. Regarding the representation of moments and any

expectations taken with respect to a distribution in the MPD-Class, it is generally more straightforward, from a computational standpoint, to transform the integrals involved via the inverse probability integral transform. Except for a set of $\gamma's$ of *measure zero*, this follows because the CDFs and the probability density functions in the MPD-Class are only defined implicitly and not in closed form. Nevertheless, after transformation, probabilities and expectations are straightforward to represent and develop.

The case of $\gamma = 0$ is a limiting case that defines a family of logistic distributions, and can be handled explicitly. For other γ cases, consider the general definition of the expectation of $g(W)$ with respect to $MPD(q,\gamma)$. Treating the probabilities as implicit functions of w and then collecting probability derivative terms, the differentiation of (3.6) with respect to $w = \mathbf{x}_i\lambda$ implies the following general representation of probability densities for nonzero γ,

$$f(w;q,\gamma) = \frac{1}{q^{-\gamma}F(w;q,\gamma)^{\gamma-1} + (1-q)^{-\gamma}(1-F(w;q,\gamma))^{\gamma-1}} \text{ for } w \in \Upsilon(q,\gamma) \quad (3.10)$$

where $F(w;q,\gamma)$ denotes the cumulative distribution function, and $\Upsilon(q,\gamma)$ denotes the appropriate support of the density function. As indicated in (3.6), this support depends on q and γ if $\gamma > 0$, and $\Upsilon(q,\gamma) = R$ otherwise. Expectations may be then defined as

$$E(g(W)) = \int_{w\in\Upsilon(q,\gamma)} \frac{g(w)}{q^{-\gamma}F(w;q,\gamma)^{\gamma-1} + (1-q)^{-\gamma}(1-F(w;q,\gamma))^{\gamma-1}} dw \quad (3.11)$$

Making a change of variables via the transformation $p = F(w;q,\gamma)$ so that $w = F^{-1}(p;q,\gamma)$ and $\frac{\partial w}{\partial p} = \frac{\partial F^{-1}(p;q,\gamma)}{\partial p}$, where $F^{-1}(p;q,\gamma)$ denotes the inverse function associated with the CDF, it follows that the expectation in (3.11) can be represented as

$$E(g(W)) = \int_0^1 g(F^{-1}(p;q,\gamma))dp \quad (3.12)$$

Note that (3.12) involves the *closed form* inverse CDF function given by

$$w = F^{-1}\left(p;q,\gamma\right) = \gamma^{-1}\left(\left(\frac{p_i}{q_i}\right)^{\gamma} - \left(\frac{1-p_i}{1-q_i}\right)^{\gamma}\right) \text{ for } p \in (0,1) \tag{3.13}$$

When $g(W)$ is such that its expectation exists, (3.12) can be represented in general as

$$E\left(g(W)\right) = \int_0^1 g\left(\gamma^{-1}\left(\left(\frac{p_i}{q_i}\right)^{\gamma} - \left(\frac{1-p_i}{1-q_i}\right)^{\gamma}\right)\right)dp \tag{3.14}$$

Moments of all orders exist for densities in the MPD-Class when $\gamma > 0$. This follows immediately from the fact that the integrand in

$$E\left(W^{\delta}\right) = \int_0^1 \left(\gamma^{-1}\left(\left(\frac{p}{q}\right)^{\gamma} - \left(\frac{1-p}{1-q}\right)^{\gamma}\right)\right)^{\delta} dp \tag{3.15}$$

is bounded for each positive integer-valued δ, finite $q \in (0,1)$, and each finite positive-valued γ. The means of the probability densities are given by evaluating the integral (3.15) for $\delta = 1$, resulting in

$$E(W) = \frac{q^{-\gamma} - (1-q)^{-\gamma}}{\gamma(\gamma+1)} \tag{3.16}$$

The second moment around the origin is obtained by solving (3.15) when $\delta = 2$, resulting in

$$E\left(W^2\right) = \gamma^{-2}\left[\frac{q^{-2\gamma} + (1-q)^{-2\gamma}}{1+2\gamma} - 2q^{-\gamma}(1-q)^{-\gamma} B(\gamma+1,\gamma+1)\right] \tag{3.17}$$

where $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ and $\Gamma(\alpha) = \int_0^{\infty} w^{\alpha-1}e^{-w}dw$ are the well-known Beta and Gamma functions, respectively. The variance of the distribution then follows by subtracting the square of (3.16) from (3.17).

Distributions in the MPD-Class with $\gamma \le -1$, do not have moments defined of any order because the integral in (3.15) is divergent for any choice of $\delta \ge 1$. Moments do exist for values of $\gamma \in (-1,0)$, but only a finite number of moments exist and how high an order of moment exists depends on the value of the parameter γ. If $\gamma > -1$ the mean of the distribution exists, and its functional representation in terms of γ and q is

precisely the same as in (3.16). If $\gamma > -\frac{1}{2}$, the second moment about the origin, and thus the variance, exist and the first two moments have exactly the same functional forms as in (3.16) and (3.17), respectively. In general, the moment of order δ will exist provided that $\gamma > -\delta^{-1}$, in which case it will be identical in functional form to the corresponding moment in the subclass of MPD-Class distributions for which $\gamma > 0$.

4. MPD Solutions as Estimators of Binary Choice Models

The expansive and flexible set of probability distributions in the MPD-Class provides a corresponding basis for estimation of the unknown binary choice probabilities and $\boldsymbol{\lambda}$. In this section we examine the use of the MPD solutions for $\mathbf{p}$ and $\boldsymbol{\lambda}$ directly as a basis for estimating the true underlying binary choice probabilities, which we denote henceforth as $\widehat{MPD}(q,\gamma)$ estimators, indicating MPD solutions for given q and γ values.

Returning to the model of binary choice outlined in Section 2, consider the minimum power divergence extremum problem depicted by the Lagrange multiplier specification in (3.3)-(3.4). Solving the first order conditions with respect to $\mathbf{p}$ results in Bernoulli probabilities that are expressed as functions of the sample data and the Lagrange multipliers $\boldsymbol{\lambda}$. 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 for any statistical model based on an MPD distribution. The divergence-minimizing estimate of $\boldsymbol{\lambda}$ can be determined by substituting the functional representation of $p(\mathbf{x}_i\boldsymbol{\lambda})$ into the first order conditions with respect to $\boldsymbol{\lambda}$, so that the optimal $\boldsymbol{\lambda}$ solves the (scaled) sample moment equations

$$\boldsymbol{\lambda}_{\text{MPD}} = \mathbf{arg}_{\boldsymbol{\lambda}} \left\{ \mathbf{x}'\left(\mathbf{y} - \mathbf{p}\left(\mathbf{x}\boldsymbol{\lambda}\right)\right) = \mathbf{0} \right\} \tag{4.1}$$

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 one basis for evaluating the estimation performance of $\widehat{MPD}(q,\gamma)$, we establish asymptotic properties of the solutions. In this discussion, it will be useful to represent the first order conditions of (3.3) with respect to the Lagrange multipliers, for random $\mathbf{X}$, as

$$\mathbf{G}_n(\boldsymbol{\lambda}) = n^{-1}\sum_{i=1}^{n}\mathbf{X}_i'\left(\mathrm{F_c}\left(\mathbf{X}_i\boldsymbol{\beta}\right) - p\left(\mathbf{X}_i\boldsymbol{\lambda}\right) + \varepsilon_i\right) = n^{-1}\sum_{i=1}^{n}\mathbf{G}_{ni}(\boldsymbol{\lambda}) = \mathbf{0} \tag{4.2}$$

where $Y_i \equiv \mathrm{F_c}\left(\mathbf{X}_i\boldsymbol{\beta}\right) + \varepsilon_i$.

4.1. Consistency

For consistency of $\widehat{MPD}(q,\gamma)$ for $\boldsymbol{\beta}$, we make the following basic assumptions:

Assumption 1. *The observations* $(y_i, \mathbf{x}_i)$, $i = 1,...,n$, *are iid random realizations of the random row vector* $(\mathrm{Y}, \mathbb{X})$.

Assumption 2. $\mathbf{G}_n(\boldsymbol{\beta}) \to \mathbf{0}$ *with probability 1.*

Assumption 3. *All* $\mathbf{G}_n(\boldsymbol{\lambda})$ *are continuously differentiable with probability 1 in a neighborhood N of* $\boldsymbol{\beta}$, *and the associated Jacobians* $\dfrac{\partial\mathbf{G}_n(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}}$ *converge uniformly to a nonstochastic limit* $\dfrac{\partial\mathbf{G}(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}}$ *that is nonsingular at* $\boldsymbol{\lambda} = \boldsymbol{\beta}$.

If *iid* random sampling is in fact the sampling mechanism utilized for generating sample data, then assumption 1 is satisfied by definition. A sufficient condition for assumption 2 to be satisfied is that $MPD(q,\gamma)$ be appropriately specified to represent the functional form of the true underlying CDF, $F\left(\mathbf{x}_i\boldsymbol{\beta}\right)$. This condition is akin to correctly specifying the functional form of the probability distribution in a maximum likelihood (ML) estimation problem. In this event,

$$E\left(\mathbf{G}_{ni}(\boldsymbol{\beta})\right) \equiv E\left(\mathbf{X}_i'\varepsilon_i\right) = \mathbf{0} \quad \text{and} \quad \mathbf{G}_{ni}(\boldsymbol{\beta}),\ i = 1,...,n, \text{ are iid} \tag{4.3}$$

imply that

$$\mathbf{G}_n(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\mathbf{G}_{ni}(\boldsymbol{\beta}) \overset{wp1}{\to} \mathbf{0} \tag{4.4}$$

by Kolmogorov's strong law of large numbers (Serfling (1980, p. 27)), resulting in the applicability of assumption 2.

Regarding assumption 3, note that the gradient of $\mathbf{G}_n(\boldsymbol{\lambda})$ is given by

$$\frac{\partial \mathbf{G}_n(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = -n^{-1}\sum_{i=1}^{n} \frac{\partial p(\mathbf{X}_i\boldsymbol{\lambda})}{\partial \mathbf{X}_i\boldsymbol{\lambda}} \mathbf{X}_i'\mathbf{X}_i = -n^{-1}\sum_{i=1}^{n} f(\mathbf{X}_i\boldsymbol{\lambda})\mathbf{X}_i'\mathbf{X}_i \tag{4.5}$$

where $f(\bullet)$ denotes a probability density function in the MPD class of distributions. It is apparent that the continuous differentiability of $\mathbf{G}_n(\boldsymbol{\lambda})$ depends on the continuity of $\frac{\partial p(z)}{\partial z} = f(z)$. This follows from the functional definition of $f(z)$ in (3.10) and the fact that, except on the boundaries of the supports of the distributions indicated in (3.6), $\frac{\partial p(z)}{\partial z}$ is continuous everywhere when $\gamma \le 0$ and continuous except on an event having probability 0 when $\gamma > 0$. Moreover, because $F(w;q,\gamma) \in [0,1]$, MPD densities are all bounded, as $f(z) < \xi < \infty$. Thus $f(\mathbf{X}_i\boldsymbol{\lambda})\mathbf{X}_i'\mathbf{X}_i < \xi\mathbf{X}_i'\mathbf{X}_i$, so $E\sup_{\lambda}\left(f(\mathbf{X}_i\boldsymbol{\lambda})\mathbf{X}_{ij}\mathbf{X}_{ik}\right) < \xi E\left(\mathbf{X}_{ij}\mathbf{X}_{ik}\right) < \infty$, $\forall i, j$. Therefore $\frac{\partial \mathbf{G}_n(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}}$ converges uniformly to $\frac{\partial \mathbf{G}(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} \equiv E\left(f(\mathbf{X}_1\boldsymbol{\lambda})\mathbf{X}_1'\mathbf{X}_1\right)$, which will be nonsingular at $\boldsymbol{\lambda} = \boldsymbol{\beta}$ if $E\left(\mathbf{X}_1'\mathbf{X}_1 \middle| f(\mathbf{X}_1\boldsymbol{\lambda}) > 0\right)$ is nonsingular.

Theorem 1. *Under assumptions 1 – 3, the* $\widehat{MPD}(q,\gamma)$ *estimator* $\hat{\boldsymbol{\lambda}} = \arg_{\boldsymbol{\lambda}}\left[\mathbf{G}_n(\boldsymbol{\lambda}) = \mathbf{0}\right]$ *is a consistent estimator of* $\boldsymbol{\beta}$.

Proof*:* The assumptions imply the regularity conditions shown by Yuan and Jennrich (1998) to be sufficient for $\hat{\boldsymbol{\lambda}} = \arg_{\boldsymbol{\lambda}}\left[\mathbf{G}_n(\boldsymbol{\lambda}) = \mathbf{0}\right] \xrightarrow{a.s.} \boldsymbol{\beta}$, and thus for $\hat{\boldsymbol{\lambda}}$ to be a (strongly) consistent estimator of $\boldsymbol{\beta}$.

Hence, if the data-generating process governing the data outcomes adheres to the regularity conditions specified and the model distribution is appropriately specified, the $\widehat{MPD}(q,\gamma)$ will consistently estimate $\boldsymbol{\beta}$. If the model distribution is not specified correctly, $\widehat{MPD}(q,\gamma)$ will generally be inconsistent. This result is similar to the case of a

misspecified ML estimation problem, where convergence occurs but to a value other than $\boldsymbol{\beta}$. Consistency of $\mathbf{p}_{\text{MPD}} = \mathbf{p}\left(\mathbf{x}\boldsymbol{\lambda}_{\text{MPD}}\right)$ follows immediately from the continuity of $\mathbf{p}\left(\mathbf{x}\boldsymbol{\lambda}\right)$ in $\boldsymbol{\lambda}$.

4.2. Asymptotic Normality

Given that $\widehat{MPD}\left(q,\gamma\right)$ is consistent, asymptotic normality of the estimator of $\boldsymbol{\beta}$ is attained under the following additional assumption:

Assumption 4: $n^{1/2}\mathbf{G}_n\left(\boldsymbol{\beta}\right) \xrightarrow{d} N\left(\mathbf{0},\, \mathbf{V}\right)$

Theorem 2. *Under assumptions* $1-4$*, the* $\widehat{MPD}\left(q,\gamma\right)$ *estimator* $\hat{\boldsymbol{\lambda}} = \arg_{\boldsymbol{\lambda}}\left[\mathbf{G}_n\left(\boldsymbol{\lambda}\right) = \mathbf{0}\right]$ *is asymptotically normally distributed, with limiting distribution* $n^{1/2}\left(\hat{\boldsymbol{\lambda}} - \boldsymbol{\beta}\right) \xrightarrow{d} N\left(0, \mathrm{A}^{-1}\mathbf{V}\mathrm{A}^{-1}\right)$*, 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* $\mathbf{V} = E\left(F\left(\mathbf{X}_1\boldsymbol{\beta}\right)\left(1-F\left(\mathbf{X}_1\boldsymbol{\beta}\right)\right)\mathbf{X}_1'\mathbf{X}_1\right)$.

Proof: Upon establishing the appropriate definitions for A and V underlying the binary choice model specification, the assumptions 1-4 imply the regularity conditions shown by Yuan and Jennrich (1998) to be sufficient for the solution of the estimating equation to have the normal limiting distribution as defined.

Regarding the applicability of assumption 4 to the MPD-Estimation problem, note that

$$n^{1/2}\mathbf{G}_n\left(\boldsymbol{\beta}\right) = n^{-1/2}\sum_{i=1}^{n}\mathbf{G}_{ni}\left(\boldsymbol{\beta}\right) = n^{-1/2}\sum_{i=1}^{n}\mathbf{X}_i'\varepsilon_i \tag{4.6}$$

is a scaled sum of *iid* random vectors, each having a zero mean vector and a covariance matrix $\mathbf{Cov}\left(\mathbf{X}_1'\varepsilon_1\right) = E\left(F\left(\mathbf{X}_1\boldsymbol{\beta}\right)\left(1-F\left(\mathbf{X}_1\boldsymbol{\beta}\right)\right)\mathbf{X}_1'\mathbf{X}_1\right)$. Based on the multivariate version of the Lindberg-Levy Central Limit Theorem (Serfling, (1980, p. 28)), $n^{-1/2}\sum_{i=1}^{n}\mathbf{X}_i'\varepsilon_i \xrightarrow{d} N\left(\mathbf{0},\, E\left(F\left(\mathbf{X}_1\boldsymbol{\beta}\right)\left(1-F\left(\mathbf{X}_1\boldsymbol{\beta}\right)\right)\mathbf{X}_1'\mathbf{X}_1\right)\right)$. Consequently, as specified in

Theorem 2, the MPD-Estimator will follow the normal limiting distribution if $\mathbf{Cov}\left(\mathbf{X}_1'\varepsilon_1\right)$ is nonsingular.

The asymptotic normality of $\mathbf{p}_{\text{MPD}} = \mathbf{p}\left(\mathbf{x}\boldsymbol{\lambda}_{\text{MPD}}\right)$, including a representation of the asymptotic covariance matrix of the distribution, follows immediately from the fact that $\mathbf{p}\left(\mathbf{x}\boldsymbol{\lambda}\right)$ is continuously differentiable in $\boldsymbol{\lambda}$, allowing for an application of the delta method to derive the asymptotic results.

4.3. Asymptotic Inference

Based on the asymptotic results of the previous subsections, the usual hypotheses tests based on normal distribution theory hold in large samples. The principal issue in empirical application is how to define appropriate sample approximations to the covariance matrices associated with the asymptotic distributions. Given the definition of the covariance matrix in Theorem 2, a consistent estimator of the Jacobian matrix $\mathbf{A} = \dfrac{\partial \mathbf{G}\left(\boldsymbol{\lambda}\right)}{\partial \boldsymbol{\lambda}} \equiv E\left(f\left(\mathbf{X}_1\boldsymbol{\beta}\right)\mathbf{X}_1'\mathbf{X}_1\right)$ is defined by

$$\hat{\mathbf{A}} = n^{-1}\sum_{i=1}^{n} f\left(\mathbf{X}_i\hat{\boldsymbol{\lambda}}\right)\mathbf{X}_i'\mathbf{X}_i \tag{4.7}$$

and a consistent estimator of $\mathbf{V} = E\left(F\left(\mathbf{X}_1\boldsymbol{\beta}\right)\left(1-F\left(\mathbf{X}_1\boldsymbol{\beta}\right)\right)\mathbf{X}_1'\mathbf{X}_1\right)$ is defined by

$$\hat{\mathbf{V}} = n^{-1}\sum_{i=1}^{n} F\left(\mathbf{X}_i\hat{\boldsymbol{\lambda}}\right)\left(1-F\left(\mathbf{X}_i\hat{\boldsymbol{\lambda}}\right)\right)\mathbf{X}_i'\mathbf{X}_i\ . \tag{4.8}$$

It follows that a Wald-type statistic for testing the J linear restrictions $\mathrm{H_o}:\mathbf{C}\boldsymbol{\beta} = \mathbf{r}$ is given by

$$n\left(\mathbf{C}\hat{\boldsymbol{\lambda}} - \mathbf{r}\right)'\left[\mathbf{C}\mathrm{A}^{-1}\mathbf{V}\mathrm{A}^{-1}\mathbf{C}'\right]^{-1}\left(\mathbf{C}\hat{\boldsymbol{\lambda}} - \mathbf{r}\right) \xrightarrow{d} \chi_J^2 \text{ under } \mathrm{H_o}\,. \tag{4.9}$$

Hypotheses relating to the value of $p\left(\mathbf{z}\boldsymbol{\beta}\right)$, where $\mathbf{z}$ is a row vector of response variate values, can be based on an application of the delta method. This gives

$$p\left(\mathbf{z}\hat{\boldsymbol{\lambda}}\right) \overset{a}{\sim} N\left(p\left(\mathbf{z}\boldsymbol{\beta}\right),\ n^{-1}f\left(\mathbf{z}\hat{\boldsymbol{\lambda}}\right)^2\mathbf{z}\mathrm{A}^{-1}\mathbf{V}\mathrm{A}^{-1}\mathbf{z}'\right) \tag{4.10}$$

so that, given $\mathrm{H_o}: p\left(\mathbf{z}\boldsymbol{\beta}\right) = p_o$,

$$\frac{n\left(p\left(\mathbf{z}\hat{\boldsymbol{\lambda}}\right)-p_o\right)}{\left(f\left(\mathbf{z}\hat{\boldsymbol{\lambda}}\right)^2\,\mathbf{z}\hat{\mathrm{A}}^{-1}\hat{\mathbf{V}}\hat{\mathrm{A}}^{-1}\mathbf{z}'\right)^{1/2}}\overset{a}{\sim}N(0,1)\text{ under }\mathrm{H_o}\,. \tag{4.11}$$

5. ML and NLS Estimation of Binary Choice Based on the MPD Class of CDFs

The consistency and asymptotic normality of the MPD solutions given in the previous section rely on appropriate choices of q and γ in order to specify the appropriate MPD distribution that coincides with the underlying true data sampling distribution. The specification issue is fully analogous to the issues involved in considering a choice of either the normal or logistic distribution for probit or logit analysis, respectively. While the vast MPD class of CDFs provides the analyst with a rich and flexible population of distributions from which to choose a characterization of Bernoulli binary response probabilities, knowledge of the true functional form of the data sampling distribution remains a daunting requirement in empirical applications.

In this section we suggest ML and NLS approaches to estimating the binary choice model in which the optimal distributional choice from among all of the members of the MPD class is embedded in the estimation process. The approach results in a highly flexible and distributionally-robust approach to estimating binary choice models that can be consistent, asymptotically normal, and efficient even in the absence of knowledge of the one true functional form of the underlying sampling distribution.

5.1 ML-MPD Estimation

A maximum likelihood estimator of the binary choice model that is based on the flexible class of MPD-distributions can be defined through optimizing the following log-likelihood function for the observed sample observations:

$$\ell\left(\boldsymbol{\beta},q,\gamma\,|\,\mathbf{y},\mathbf{x}\right)=\sum_{i=1}^{n}\left[\,y_i\ln\left(p\left(\mathbf{x}_i\boldsymbol{\beta};q,\gamma\right)\right)+\left(1-y_i\right)\ln\left(1-p\left(\mathbf{x}_i\boldsymbol{\beta};q,\gamma\right)\right)\right] \tag{5.1}$$

It is instructive to consider maximizing the likelihood in two steps, first defining the profile likelihood function of $\boldsymbol{\beta}$ as

$$p\ell\left(\boldsymbol{\beta}\,|\,\mathbf{y},\mathbf{x}\right)\equiv\sup_{q,\gamma}\left\{\ell\left(\boldsymbol{\beta},q,\gamma\,|\,\mathbf{y},\mathbf{x}\right)\right\} \tag{5.2}$$

and then deriving the maximum likelihood estimator of $\boldsymbol{\beta}$ by maximizing the profile likelihood as

$$\hat{\boldsymbol{\beta}}_{ml} = \arg\sup_{\boldsymbol{\beta}} \left\{ p\ell\left(\boldsymbol{\beta} \mid \mathbf{y},\mathbf{x}\right) \right\} \equiv \arg\sup_{\boldsymbol{\beta}} \left\{ \left\{ \sup_{\gamma,q} \left\{ \ell\left(\boldsymbol{\beta}, q, \gamma \mid \mathbf{y},\mathbf{x}\right) \right\} \right\} \right\} \tag{5.3}$$

One can interpret the likelihood profiling step (5.2) as determining the optimal MPD distribution associated with any choice of the $\boldsymbol{\beta}$ vector, and the second ML step (5.3) as determining the overall optimal estimate of the parameters of the linear index argument in the CDF that determines the binary response probabilities.

It is known (Patefield (1977, 1985), and Murphy and van der Vaart (2000)) that the profile likelihood $p\ell\left(\boldsymbol{\beta} \mid \mathbf{y},\mathbf{x}\right)$ can be utilized in effectively the same way as an ordinary likelihood for purposes of defining an appropriate score function $\dfrac{\partial p\ell\left(\boldsymbol{\beta} \mid \mathbf{y},\mathbf{x}\right)}{\partial \boldsymbol{\beta}}$ and an information matrix representation of the asymptotic covariance matrix with respect to the ML estimator $\hat{\boldsymbol{\beta}}_{ml}$, $\left[-E \dfrac{\partial^2 p\ell\left(\boldsymbol{\beta} \mid \mathbf{y},\mathbf{x}\right)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \right]^{-1}$. The score and information matrix calculations are equivalent, with regard to the representation of the $\boldsymbol{\beta}$ aspect of the score and information, to the appropriate submatrix of the information matrix calculated from the full likelihood, as $\dfrac{\partial \ell\left(\boldsymbol{\theta} \mid \mathbf{y},\mathbf{x}\right)}{\partial \boldsymbol{\beta}}$ and $\left[-E \dfrac{\ell\left(\boldsymbol{\theta} \mid \mathbf{y},\mathbf{x}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right]^{-1}$ respectively, where $\boldsymbol{\theta} \equiv \left[\boldsymbol{\beta}, q, \gamma \right]'$. The asymptotic covariance matrix for $\hat{\boldsymbol{\beta}}_{ml}$ together with the asymptotic normality of the ML estimator, $n^{1/2}\left(\hat{\boldsymbol{\beta}}_{ml} - \boldsymbol{\beta}\right) \xrightarrow{L} N\left(\mathbf{0}, \left[-E \dfrac{\partial^2 p\ell\left(\boldsymbol{\beta} \mid \mathbf{y},\mathbf{x}\right)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \right]^{-1} \right)$, can then be used for hypothesis testing and confidence region generation based on the asymptotic chi-square distributions of the usual Wald, Lagrange Multiplier, or Likelihood Ratio test statistics, analogous to traditional probit or logit contexts but rooted in a substantially

more robust sampling distribution model[3]. In empirical applications, an estimate of the asymptotic covariance of $\hat{\boldsymbol{\beta}}$ would be defined by $-\left[\frac{\partial^2 p\ell(\boldsymbol{\beta}\,|\,\mathbf{y},\mathbf{x})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right]^{-1}_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{ml}}$.

5.2 NLS-MPD Estimation

A nonlinear least squares estimator of the binary choice model that is based on the flexible class of MPD-distributions can be defined through minimizing the following sum of squared errors (SSE) function:

$$SSE(\boldsymbol{\beta},q,\gamma\,|\,\mathbf{y},\mathbf{x}) = \sum_{i=1}^{n}\left[y_i - \left(p(\mathbf{x}_i\boldsymbol{\beta};q,\gamma)\right)\right]^2 \tag{5.4}$$

Given the heteroskedastic nature of the Bernoulli random variables, whereby the variance of the i$^{\text{th}}$ Bernoulli trial is given by $p_i(1-p_i)$, one might consider pursuing a heteroskedasticity-adjusted SSE function to seek gains in asymptotic efficiency as:

$$SSE_w(\boldsymbol{\beta},q,\gamma\,|\,\mathbf{y},\mathbf{x}) = \sum_{i=1}^{n}\frac{\left[y_i - \left(p(\mathbf{x}_i\boldsymbol{\beta};q,\gamma)\right)\right]^2}{\left[p(\mathbf{x}_i\boldsymbol{\beta};q,\gamma)\right]\left[1-p(\mathbf{x}_i\boldsymbol{\beta};q,\gamma)\right]} \tag{5.5}$$

However, it is well known (e.g., Pagan and Ullah, 1999) that the first order conditions for minimizing (5.5) are precisely the same as that of maximizing the likelihood in (5.1), and thus we proceed by focusing on the simpler consistent but potentially less efficient estimator defined by minimizing (5.4).

Given that the uncorrected SSE is used as the estimation objective, the nonlinear least squares estimator will then have an asymptotic distribution whose covariance matrix reflects this fact, as $\hat{\boldsymbol{\beta}} \overset{a}{\sim} N\left(\boldsymbol{\beta}, \left[\nabla_{\boldsymbol{\beta}}\mathbf{p}'\nabla_{\boldsymbol{\beta}}\mathbf{p}\right]^{-1}\left[\nabla_{\boldsymbol{\beta}}\mathbf{p}'\boldsymbol{\Psi}\nabla_{\boldsymbol{\beta}}\mathbf{p}\right]\left[\nabla_{\boldsymbol{\beta}}\mathbf{p}'\nabla_{\boldsymbol{\beta}}\mathbf{p}\right]^{-1}\right)$ where

[3] As in all applications of maximum likelihood, as well as the nonlinear least squares application in section 5.2, issues of regularity conditions arise for the asymptotics to apply. Under appropriate boundedness assumptions relating to X, and given the boundedness and continuous differentiability of the MPD class of distributions, extremum estimator asymptotics apply along the lines of Hansen (1982), Newey (1991), and van der Vaart (1998).

$\nabla_{\boldsymbol{\beta}}\mathbf{p} \equiv \left[\dfrac{\partial \mathbf{p}(\mathbf{x}\boldsymbol{\beta};\mathrm{q},\gamma)}{\partial \boldsymbol{\beta}'}\right]$ and $\boldsymbol{\Psi}$ is a diagonal covariance matrix of the binary observations whose i[th] diagonal entry equals $\left[p(\mathbf{x}_i\boldsymbol{\beta};q,\gamma)\right]\left[1-p(\mathbf{x}_i\boldsymbol{\beta};q,\gamma)\right]$, the Bernoulli variance for the i[th] observation. In applications, the unknown parameters would be replaced by their NLS estimates, and the resulting estimate of the covariance matrix could be used to define appropriate test and confidence region-generating statistics.

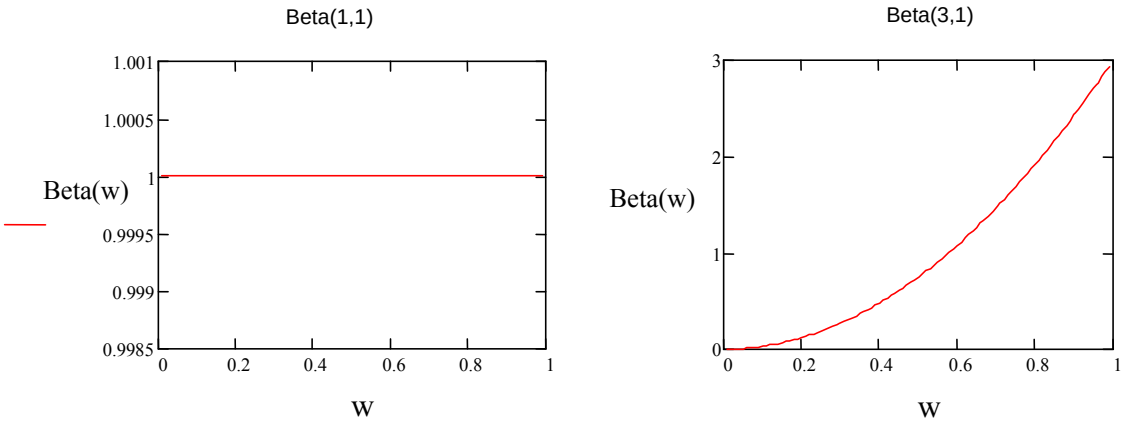
6. Sampling Performance

To investigate the finite sample estimation performance of the $\widehat{MPD}(q,\gamma)$, ML-MPD, and NLS-MPD approaches, the results of Monte Carlo experiments are reported in this section. The sampling experiments were designed so that the sampling distribution underlying observed $p_i = P(y_i = 1)$ values achieve targeted mean and variability levels, and map one-to-one with covariate values, representing a covariate population data sampling process (DSP).

*6.1 Sampling Desig*n

Random samples of size n = 100, 250, and 500 of $P(y_i = 1)$ values were sampled from a Beta distribution $B(a,b)$ that had $b = 1$ and a set to achieve targeted unconditional mean $P(y_i = 1)$ levels of .5 or .75, i.e., $a = \dfrac{b}{E(Y)(1-E(Y))}$ with $E(Y_i) = .5 \; or \; .75$. The resulting two distributions of $P(y_i = 1)$ values are uniform, centered at .5, or right-skewed and centered at .75, respectively. Graphs of the two population distributions are presented in Figure 6.1 below. The two distribution alternatives represent a situation in which all values of P(y = 1) are equally likely to be observed, and a situation in which it is substantially more probable to observe values of P(y = 1) that are greater than .5 than less ($\int_{.5}^{1} Beta(w;3,1)\,dw = .875$).

Figure 6.1. Population Distributions for P(y = 1)



A linear index representation of the Bernoulli probabilities is formed as

$$P\left(y_i = 1\right) = F\left(\beta_0 + \beta_1 x_i\right) \quad \text{for i = 1 ,..., } n, \tag{6.1}$$

where $F(\bullet)$ is the cumulative distribution or link function underlying the Bernoulli probabilities, and the x_i's are chosen so that $x_i = \left(F^{-1}\left(P\left(y_i = 1\right)\right) - \beta_0\right)/\beta_1$. The values of the parameters are set to $\beta_0 = 1$ and $\beta_1 = 2$. Explicit functional forms for the binary model link function $F(\bullet)$ include two MPD distributions, given by MPD(q=.5, $\gamma = -1$) and MPD(q=.75, $\gamma = 1.5$), as well as a N(0,1) distribution. The former distribution is a symmetric distribution associated with the "empirical likelihood" choice of $\gamma = -1$, has the real line for its support, and has substantially fatter tails then the N(0,1) distribution. The latter distribution has finite support, is heavily skewed to the left, and has a density value that increases at an increasing rate as the argument of the link function increases. The standard normal distribution is the link function that is optimal for the Probit model of binary choice. These three different link functions are illustrated in Figure 6.2.

The random outcomes of the binary choices are generated according to their respective Bernoulli probability distributions as

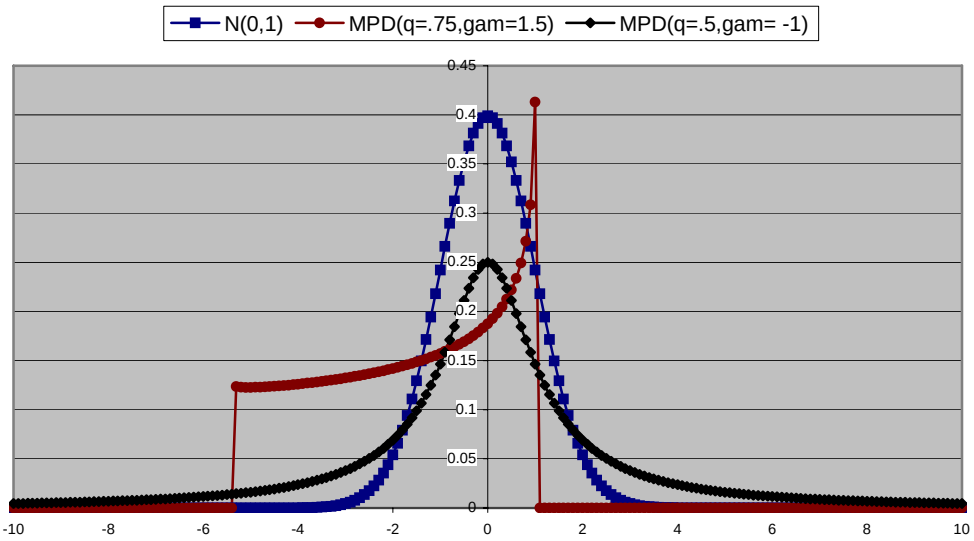
$$Y_i \sim Bernoulli\left(F\left(\beta_0 + \beta_1 x_i\right)\right),\ i = 1,...,n. \tag{6.2}$$

This implies a regression-type relationship of the form

$$Y_i = F\left(\beta_0 + \beta_1 x_i\right) + \varepsilon_i,\ i = 1,...,n \tag{6.3}$$

and acts as the definition of the residual outcome of ε_i.

Figure 6.2. Link Function Distributions: N(0,1), MPD(q=.5,gam= -1), MPD(q=.75,gam=1.5)



The measure of estimation performance used is the expected squared prediction error defined by $\int_X \left(\hat{p}(x) - F(\beta_0 + \beta_1 x)\right)^2 dF(\beta_0 + \beta_1 x)$, where $\hat{p}(x)$ denotes the probability prediction from an $\widehat{MPD}(q,\gamma)$, ML-MPD, NLS-MPD, or Probit estimator. Empirically, the measure is calculated as $n^{-1}\sum_{i=1}^{n}\left(\hat{p}(x_i) - F(\beta_0 + \beta_1 x_i)\right)^2$.

All sampling results are based on 1,000 repetitions, and calculations were performed using Aptech Systems' GAUSS 8.0 software. At this number of repetitions, all empirically-calculated expectations of the performance measures are very accurate, with standard errors of mean calculations typically .0001 or less in magnitude. We note that sampling results for the $\widehat{MPD}(.5,0)$ estimator necessarily produces results identical to those of the standard ML logit estimator.

6.2. Sampling Results

The probability prediction MSE results for the ML-MPD and NLS-MPD estimators, the Probit estimator, and the two MPD estimators $\widehat{MPD}(.5,-1)$ and $\widehat{MPD}(.5,0)$ are displayed in Figure 6.3 for the MPD(q=.5, $\gamma = -1$) DSP and in Figure 6.4 for the MPD(q=.75, $\gamma = 1.5$) DSP . Given these sampling distribution specifications for

the link functions, the $\widehat{MPD}(.5,-1)$ estimator is specified correctly for the former DSP, and neither $\widehat{MPD}(q,\gamma)$ estimator is specified correctly in the case of the latter DSP. In implementing the ML-MPD and NLS-MPD estimators, the feasible set of distributions examined was defined by $\gamma \in [-2,2]$ and $q \in [.1, .9]$.

It is apparent that the Probit estimator, a Quasi-ML in this application, is strongly dominated by all of the alternative estimators when the DSP is $MPD(q=.5,\gamma=-1)$.

Figure 6.3. MSE for P(y = 1) Predictions, n = 100, 250, 500, DSP = MPD(.5,-1)

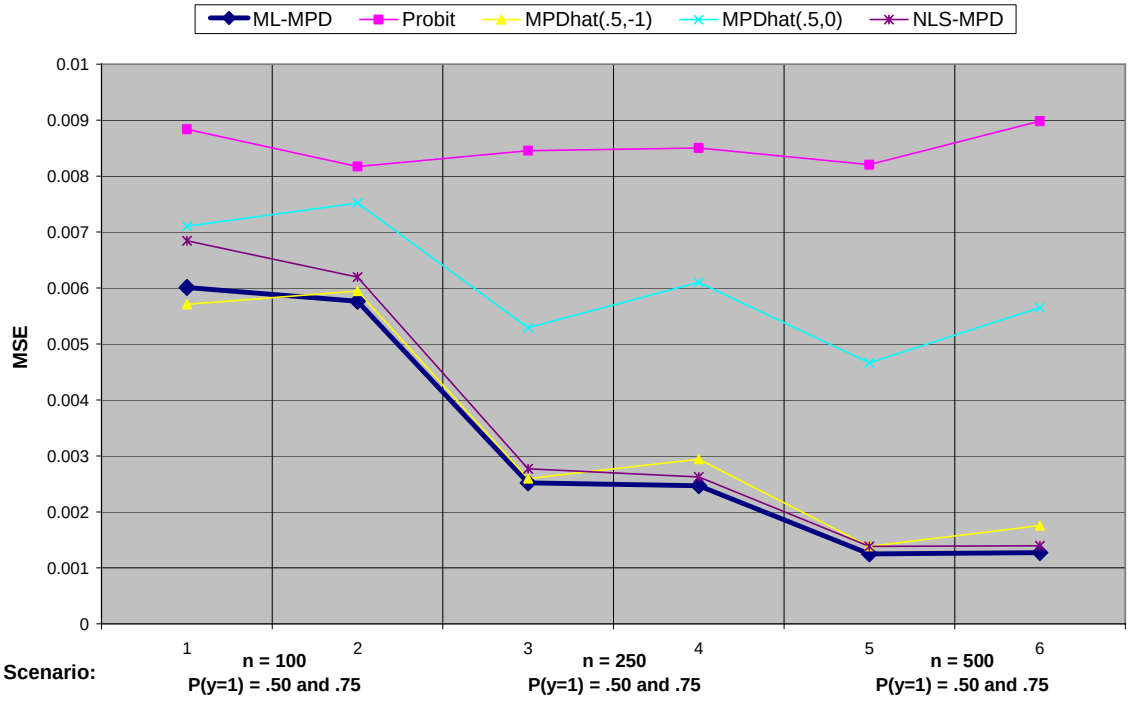
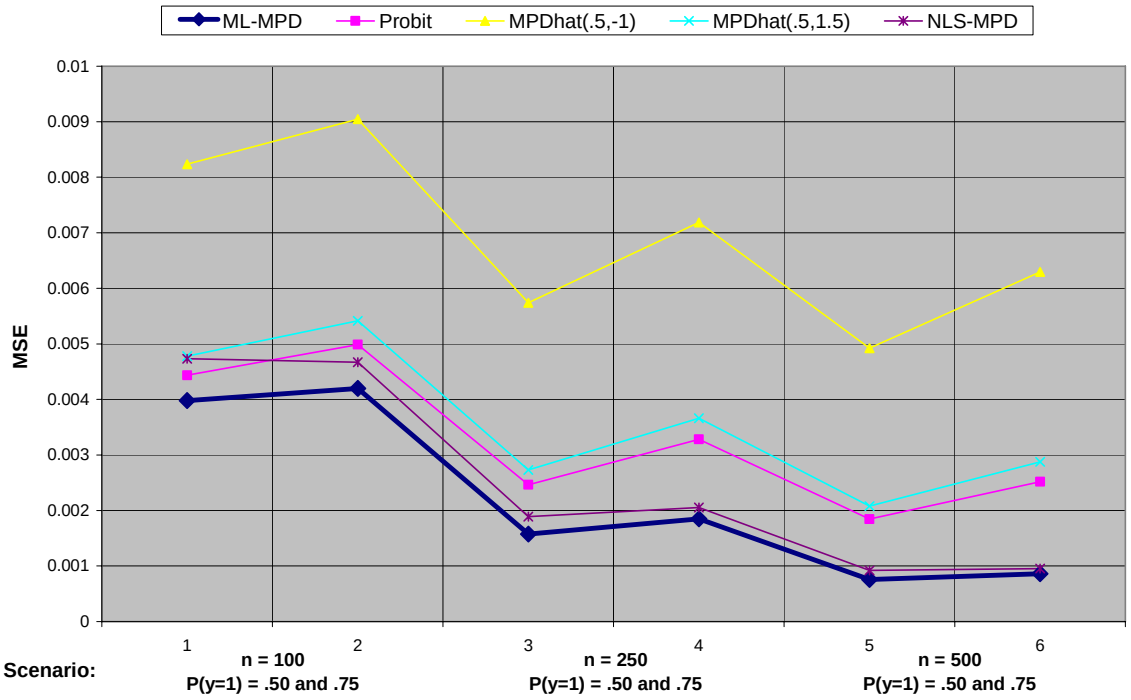


Figure 6.4. MSE for P(y = 1) Predictions, n = 100, 250, 500, DSP = MPD(.75, 1.5)

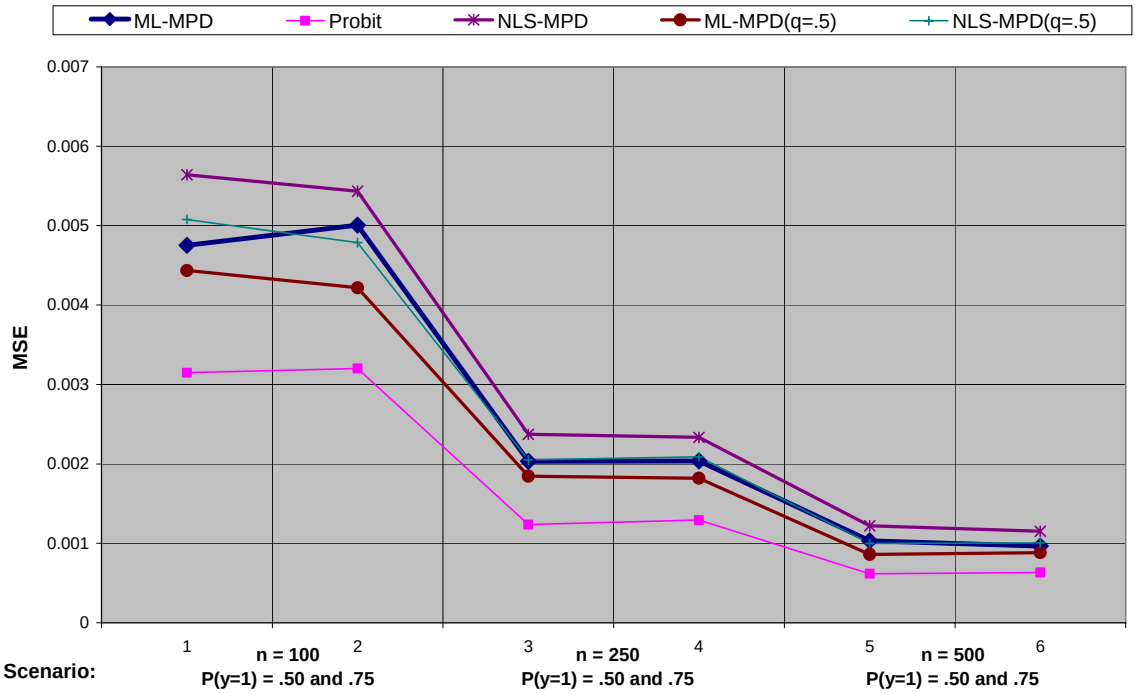


The correctly specified $\widehat{MPD}(.5,-1)$ estimator performs very well in MSE, but the ML-MPD estimator effectively equals or is superior to the $\widehat{MPD}(.5,-1)$ estimator across all scenarios. The NLS-MPD estimator also performs very well, nearly equaling the precision of the ML-MPD estimator except for the smallest sample size.

When the DSP is $MPD(q=.75,\gamma=1.5)$ the ML-MPD continues to be the clear estimator of choice, with the NLS-MPD estimator again being almost as precise especially for the larger sample sizes. The Probit estimator, a Quasi-ML estimator in this application, is again dominated, as are the $\widehat{MPD}(.5,-1)$ and $\widehat{MPD}(.5,0)$ estimators. In both this and the preceding sampling experiments the consistency of the ML-MPD and NLS-MPD estimators is illustrated in the figures as *n* increases.

Finally, in the case where the DSP is $N(0,1)$, and supposing the analyst is omniscient and chooses (the correct) probit model, it is not surprising that the probit-ML estimator would be the superior choice in MSE performance, as indicated in Figure 6.5.

Figure 6.5. MSE for P(y = 1) Predictions, n = 100, 250, 500, DSP = N(0,1)



The superiority of the Probit estimator diminishes relative to both the ML-MPD and NLS-MPD estimators as the sample size increases. Supposing that the analyst chose to model the binary choice probabilities by utilizing only symmetric distributions contained in the MPD-class (i.e., restricting q = .5), the relative superiority of the correctly specified Probit estimator is significantly diminished even for the smallest sample size, as indicated by the ML-MPD(q=.5) and NLS-MPD(q=.5) results in Figure 6.5. Overall, the comparisons in Figure 6.5 indicate both the precision gains that would occur by having correct prior information about distributional functional form, and also indicate the unavoidable cost of flexibility when, as is almost always the case in practice, one does not possess omniscience in the choice of the statistical model.

Overall, the sampling results illustrate the robustness attainable from utilizing the large and flexible class of MPD distributions for modeling binary choice coupled with a ML or NLS approach for choosing optimally among the member distributions within the class. The method mitigates imprecision due to lack of knowledge-misspecification of the true data sampling process underlying binary choices and competes well with generally unknown and unattainable correct choices of the data sampling process.

7. Summary and Extensions

Representing sample information underlying binary choice outcomes through moment conditions $\mathrm{E}\left[\mathbf{X}'\left(\mathbf{Y}-\mathbf{p}\right)\right]=\mathbf{0}$ based on generally applicable nonparametric conditional expectation or regression representations of the data sampling process, the Cressie-Read (CR) family of power divergence measures was used to identify a new class of statistical models and an associated large and varied class of associated CDFs for characterizing observations on binary choice outcomes. The unknown Bernoulli probabilities, **p,** expressed as functions of response variates, **x**, were solved for by implementing the minimum power divergence principle and represent the unique class of distributions that are both consistent with the moment conditions, and that are minimally divergent from any conceivable reference distribution for the Bernoulli probabilities. Estimation implications of this formulation were assessed analytically and by sampling experiments.

It is straight forward to extend the univariate distribution formulations of this paper to their multivariate counterparts. For example, one such extension, which subsumes the multivariate logistic distribution as a special case, begins with a multinomial specification of the minimum power divergence estimation problem in Lagrange form as

$$L\left(\mathbf{p},\boldsymbol{\lambda}\right)=\sum_{i=1}^{n}\sum_{i=1}^{n}\left(\frac{1}{\gamma\left(\gamma+1\right)}\sum_{j=1}^{m}p_{ij}\left[\left(\frac{p_{ij}}{q_{ij}}\right)^{\gamma}-1\right]\right)+\sum_{j=1}^{m}\boldsymbol{\lambda}_{\mathbf{j}}'\mathbf{x}'\left(\mathbf{y}_{\mathbf{j}}-\mathbf{p}_{\mathbf{j}}\right)+\sum_{i=1}^{n}\eta_{i}\left(\sum_{j=1}^{m}p_{ij}-1\right). \quad (8.1)$$

Solving first order conditions with respect to the p_{ij}'s leads to the standard multivariate logistic distribution when the reference distributions are uniform.

The analytical and sampling results for the new minimum power divergence class of statistical models and estimators represents a base for considering a range of important new problems relating to binary estimation and inference. One question concerns how to make use of the reference distribution, **q**, to take into account known or estimable characteristics of the Bernoulli probabilities in any particular applied problem, and through the minimum power divergence principle, incorporate that information into the estimation of probabilities and marginal effects. The recent work of Smith (2007) is suggestive of one method, based on a kernel density approach, for specifying values of

the reference distribution probabilities. Consideration of alternative nonparametric conditional moment formulations and their effect on efficiency of the resultant estimators provides another set of interesting research problems. These and other issues are the subject of ongoing research.

REFERENCES

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

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

Gabler, S., F. Laisney, and M. Lechner (1993): Seminonparametric Estimation of Binary Choice Models with an Application to Labor Force Participation. Journal of Business and Economic Statistics 11, 61-80.

Gazalo, P.L. and O. Linton (1994): Local Nonlinear Least Squares Estimation Using Parametric Information Nonparametrically. Discussion Paper No. 1075, Cowles Foundation, Yale University.

Green, W. (2007): Discrete Choice Modeling. Handbook of Econometrics. Vol. 2, Applied Econometrics, Part 4.2, Ed. T. Mills and K. Patterson, Palgrave, London.

Han, A.K. (1987): Nonparametric Analysis of a Generalized Regression Model. Journal of Econometrics 35, 303-316.

Horowitz, J.L. (1992): A Smoothed Maximum Score Estimator for the Binary Response Model. Econometrica 60, 505-531.

Ichimura, H. (1993): Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. Journal of Econometrics **5**8, 71-120.

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

Hansen, L.P. (1982): Large Sample Properties of Generalized Method of Moments Estimators, Econometrica, 50, 1029-1054.

H. W. Kuhn and A. W. Tucker. *Non-linear Programming*. In Proc. 2 Berkeley Symp. on Mathematical Statistics and Probability, pages 481-492. Univ. Calif. Press, 1951.

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.

Manski, C.F. (1975): The Maximum Score Estimation of the Stochastic Utility Model of Choice. Journal of Econometrics 3, 205-228.

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, eds. Handbook of Econometrics 2, Amsterdam, North Holland, pp 1395-1457.

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

Murphy, S.A. and A.W. van der Vaart (2000): On Profile Likelihood, Journal of the American Statistical Association, 95, 449-465.

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

Newey, W. (1991): Uniform Convergence in Probability and Stochastic Equicontinuity, Econometrica, 59, 1161-1167.

Pardo, L. (2006): Statistical Inference Based on Divergence Measures, Boca Raton: Chapman and Hall.

Pagan, A. and A. Ullah (1999): Nonparametric Econometrics, New York: Cambridge University Press.

Patefield, W.M. (1977): On the Maximized Likelihood Function, Sankhya 39, 92-96.

Patefield, W. M. (1985): Information from the Maximized Likelihood Function, Biometrika, 1985, 664-668.

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.

Smith, R. (2007): Efficient Theoretic Information Inference for Conditional Moment Restrictions, Journal of Econometrics 138, 430-460.

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

van der Vaart, A.W. (1998): Asymptotic Statistics, New York: Cambridge University Press.

Wang, W. and M. Zhou (1993): Iterative Least Squares Estimator of Binary Choice Models: Semiparametric Approach. Department of Statistics Technical Report No. 346, University of Kentucky.

Yuan, K. and R. I. Jennrichs (1998): Asymptotics of Estimating Equations Under Natural Conditions, Journal of Multivariate Analysis 65, 245-260.