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 makes use of information theoretic methods, in the form of the Cressie-Read (CR) family of divergence measures, 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 that includes the conventional logit model. The resulting new class of statistical models and estimators requires minimal \textit{a priori} model structure and non-sample information, and provides the basis for a range of model and estimator extensions.

\textbf{Keywords}: semiparametric binary response models and estimators, conditional moment equations, squared error loss, Cressie-Read statistic, information theoretic methods, minimum power divergence

AMS 1991 Classification Primary 62E20
JEL Classifications: C10, C2

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 (email: 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

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 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 CDF’s 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 \( i = 1, 2, \ldots, n \), one of two alternatives is observed to occur for each independent binary random variable \( Y_i, i = 1, 2, \ldots, n \), having its respective probability \( p_i, i = 1, \ldots, n \), 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, \( Y_i^* \):

\[
Y_i^* = x_i \beta + \epsilon_i^*
\]

where \( Y_i \equiv I(Y_i^* > 0), i = 1, \ldots, n \), 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 \( x_i, i = 1, \ldots, n \), are independent outcomes of a \( (1 \times k) \) random vector of response variables.

Here, and elsewhere, the linear index \( x_i \beta \) can be replaced by the more general functional notation \( m(x_i, \beta) \) 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

\[
p_i = P(Y_i = 1) = P(\epsilon_i^* > -x_i \beta) = 1 - G(-x_i \beta) = G_x(-x_i \beta)
\]

---

\(^1\) 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, \( X_i \) denotes the \( i^{th} \) row of \( X \), and \( X_j \) denotes the \( j^{th} \) column. Observed outcomes or fixed values are denoted by lower case letters. Exceptions to these conventions include \( \epsilon \) being an outcome of random \( \epsilon \), and \( \hat{\beta} \) being an outcome of random \( \hat{\beta} \).
where \( G(\cdot) \) is the CDF of the noise term \( \varepsilon^*_i \) in latent variable equation (1.1) and \( G_0(\cdot) \) 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 \( G(x, \beta) \) 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 \( \beta \) and the choice probabilities \( G(x, \beta) \). 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, 1984 and Train 2003).

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 \( \varepsilon^*_i \) 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; in Section 3, 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, properties of the CDFs are identified, and graphs of the shapes of these distributions are presented; and, in Section 5, moments of the minimum power divergence class of distributions are developed. Section 6 utilizes the class of CDFs 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. Asymptotic sampling properties of the estimators are demonstrated and, in Section 7, Monte Carlo sampling results are reported to illustrate the finite sampling performance of the estimators. Finally, in Section 8, the implications of the formulations are discussed and possible extensions of the methodology are suggested.

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 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 in fact the 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

\[
Y = p + \varepsilon, \quad \text{where } E(\varepsilon) = 0 \quad \text{and} \quad p \in \mathbb{R}^n(0,1)
\]  

(2.1)

The specification in (2.1) implies only that the expectation of \( Y \) is some mean vector of Bernoulli probabilities \( p \), and that outcomes of \( Y \) can be decomposed into their means and noise terms. If \( Y \) is a vector of binary random variables, the specification in (2.1) 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 (2.1) depend on the values of response variables \( Z \) through some general conditional expectation relationship, as

\[
E(Y | Z) = p(Z) = [p_1(Z_1), p_2(Z_2), \ldots, p_n(Z_n)]', \quad \text{with the conditional orthogonality condition}
\]

\[
E[Z'(Y - p(Z)) | Z] = 0 \quad \text{therefore implied. An application of the double expectation theorem yields the unconditional orthogonality result}
\]

\[
E[Z'(Y - p(Z))] = 0.
\]  

(2.2)

We emphasize that both the conditional and unconditional moment relationships are generally true subject only to the requirement that the Bernoulli probabilities have some functional relationship with the response variables in \( Z \), expressed via a conditional expectation relationship \( E(Y | Z) \). In fact, there is essentially no risk of model misspecification at this point given that some regression relationship exists between \( Y \) and \( Z \). This represents a very basic
level of information for estimating the unknown Bernoulli probabilities and is no more stringent than a fully nonparametric regression representation of binary response. Adding any additional functional and/or statistical characteristics to the model specification would require additional sample and/or non-sample information beyond this basic level, and is often information that is substantially more uncertain and, when used, is most often simply assumed rather than truly known.

Given (2.2), and the assumption that the latent variable representation of the Bernoulli outcomes in (1.1) applies, a natural candidate for the elements of the $Z$ matrix is the $(n \times k)$ matrix $X$ associated with (1.1). We proceed by letting $X$ 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 $p = G(x \beta)$ with $G(\cdot)$ being some cumulative distribution function, then the moment equations can be specified as $E\left(X'(Y - G(X \beta))\right) = 0$. Empirical representations of these moments, as $n^{-1}\left(x'(y - G(x \beta))\right) = 0$, 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 (2.2), $G(\cdot)$ is neither assumed known nor explicitly specified so that a GMM approach to estimating the binary response model using moments of the type (2.2) is not possible. Moreover, it is clear that the empirical moments

$$n^{-1}\left(x'(y - p(x))\right) = 0 \quad (2.3)$$

cannot possibly be used in isolation to identify the Bernoulli probabilities since, regardless of their number, $p(x) = y$ always solves the set of moment constraints. In addition, there are more unknowns than estimating equations. Consequently, the system of equations (2.3) 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 (2.1) and the representation of sample information in the form of empirical moments (2.3), 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)

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

subject to:

\[
\sum_{j=1}^{2} p_{ij} = 1, \quad p_{ij} \geq 0, \quad \forall i, j
\]

\[
\sum_{i=1}^{2} q_{ij} = 1, \quad q_{ij} \geq 0, \quad \forall i, j
\]

\[
n^{-1} (x'(y - p)) = 0
\]

The Bernoulli process underlying the binary outcomes for each observation is characterized by the probabilities \( \{ p_{i1}, p_{i2} \} \), where \( E(y_i | x_i) = p_{i1} \), \( \forall i \), and, in general, \( E(Y | x) = p \). The parenthetical component \( \frac{1}{\gamma(\gamma+1)} \sum_{j=1}^{2} p_{ij} \left[ \left( \frac{p_{ij}}{q_{ij}} \right)^{\gamma} - 1 \right] \) in the estimation objective function refers to the CR power divergence of the Bernoulli probability distribution \( \{ p_{i1}, p_{i2} \} \) from some respective reference Bernoulli distributions \( \{ q_{i1}, q_{i2} \} \). Regarding the interpretation of the CR
divergence measure itself, this parenthetical component is proportional to the weighted average deviation of \( \left( \frac{p}{q} \right)^\gamma \) from 1, the weights being the Bernoulli probabilities \( \{p_{ii}, p_{12}\} \), and the corresponding outcomes of the probability ratio being averaged are \( \left( \frac{p_{ii}}{q_{ii}} \right)^\gamma \) and \( \left( \frac{p_{12}}{q_{12}} \right)^\gamma \). The CR divergence is strictly convex in the \( p_{ij} \)'s, and assumes an unconstrained unique global minimum when \( p_{ij} = q_{ij}, \forall i \text{ and } j \).

The constraints (3.2) and (3.3) 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 (3.4) is the empirical implementation of the moment condition \( E \left( X' (Y - p) \right) = 0 \). 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, \( X \).

### 3.1. Identifying the Class of CDFs Underlying \( p \)

Henceforth defining \( p_i \equiv p_{ii} \) and \( q_i \equiv q_{ii} \), the divergence minimization problem defined in (3.1)-(3.4) can be characterized in Lagrange form as

\[
L(p, \lambda) = \sum_{i=1}^{\gamma} \left[ \frac{1}{\gamma (\gamma + 1)} \left[ p_{ii} \left( \frac{p_{ii}}{q_{ii}} \right)^\gamma + (1 - p_i) \left( \frac{1 - p_{ii}}{1 - q_{ii}} \right)^\gamma - 1 \right] \right] + \lambda' x' (y - p) \tag{3.5}
\]

subject to:

\[
0 \leq p_i, q_i \leq 1, \forall i. \tag{3.6}
\]
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 $p_i$, 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

$$\frac{\partial L}{\partial p_i} = 0 \implies \begin{cases} \left( \frac{p_i}{q_i} \right)^\gamma - \left( \frac{1-p_i}{1-q_i} \right)^\gamma - x_i \lambda \gamma = 0 \text{ for } \gamma \neq 0 \\ \ln \left( \frac{p_i}{q_i} \right) - \ln \left( \frac{1-p_i}{1-q_i} \right) - x_i \lambda = 0 \text{ for } \gamma = 0 \end{cases}$$

(3.7)

where $x_i$ is used to denote the $i^{th}$ row of the matrix $x$.

When $\gamma \leq 0$, the solutions are strictly interior to the inequality constraints and the inequality constraints are nonbinding. Accounting for the inequality constraints in (3.5) when $\gamma > 0$, the first-order condition in (3.7) and complementary slackness allows $p_i$ to be expressed as the following function of $x_\lambda$:

$$p_i(x_\lambda) = \arg_{p_i} \left[ \left( \frac{p_i}{q_i} \right)^\gamma - \left( \frac{1-p_i}{1-q_i} \right)^\gamma = x_\lambda \gamma \right] \text{ for } \gamma < 0 \text{ and } x_\lambda \in \mathbb{R}$$

$$= \arg_{p_i} \left[ \ln \left( \frac{p_i}{q_i} \right) - \ln \left( \frac{1-p_i}{1-q_i} \right) = x_\lambda \right] \text{ for } \gamma = 0 \text{ and } x_\lambda \in \mathbb{R}$$

(3.8)
A unique solution for $p_i(x,\lambda)$ necessarily exists by the strict monotonicity of either

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

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 $\gamma$, 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 $p_i(x,\lambda)$ in the argument $x,\lambda$ for $p_i \in (0,1)$ allows $p_i(x,\lambda)$ to be interpreted as a CDF on the appropriate support for $x,\lambda$. We also underscore for later use that the inverse CDFs clearly do exist in closed form, as is immediately obvious from the expressions in (3.8).

Explicit closed form solutions for the CDFs exist for $\gamma$ values that include -1, 0, and 1 integer values, as follows:

$$p_i(x,\lambda | \gamma = -1) = \begin{cases} 
0 & \frac{1}{2} + \frac{[\frac{(x,\lambda)^2}{2} + (4q_i - 2)(x,\lambda) + 1]^2 - 1}{2x,\lambda} 
\end{cases} \text{if } x,\lambda \neq 0 \text{ else } 0 \quad (3.9)$$

$$p_i(x,\lambda | \gamma = 0) = \frac{q_i \exp(x,\lambda)}{1 - q_i + q_i \exp(x,\lambda)} \quad (3.10)$$

$$p_i(x,\lambda | \gamma = 1) = \begin{cases} 
1 & \frac{1}{q_i + q_i(1-q_i)x,\lambda} \quad \text{for } x,\lambda \in \left[ \frac{q_i^{-\gamma}}{1-q_i^{-\gamma}}, q_i^{-\gamma} \right] 
0 & \frac{1}{q_i + q_i(1-q_i)x,\lambda} \quad \text{for } x,\lambda \in \left[ -\frac{1-q_i^{-\gamma}}{1-q_i^{-\gamma}}, q_i^{-\gamma} \right] 
\end{cases} \quad (3.11)$$

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 (3.10) coincides with the
the usual *logistic* binary choice model if the reference distribution is such that \( q_i = .5 \), and \( \lambda \) is interpreted as the coefficients weighting the response variables in a linear index formulation. The CDF in (3.11) subsumes the *linear probability model* when \( \lambda \) is again interpreted as the coefficients that weight the response variables in a linear index formulation. Such interpretations for \( \lambda \) will be further motivated ahead.

4. Properties of the MPD-Class of Probability Distributions for \( \gamma = 1, 0, -1 \)

In the previous section, a family of probability distributions was identified for each value of the power parameter, \( \gamma \), in the definition of the CR-power divergence statistic. For a CR family within the Class, member distributions are parameterized by the value of the reference probability value, \( q_i \). In this section, we discuss properties of some explicit families of MPD probability distributions, corresponding to the historically prominent \( \gamma = -1, 0, 1 \) cases. These distributions exist in closed form, and begin to illustrate the substantial variety and flexibility of distributions within the Class.

4.1. MPD(\( \gamma = 1 \))

The family of CDFs generated by this MPD specification is given by

\[
F(w; q) = \begin{cases} 
1 & \text{for } w \geq q^{-1} \\
(q + q(1-q)w) & \text{for } w \in (-q^{-1}, q^{-1}) \\
0 & \text{elsewhere}
\end{cases}
\]

(4.1)

and the functional form of the associated PDF is

\[
f(w; q) = \frac{dF(w; q)}{dw} = \begin{cases} 
q(1-q) & \text{for } w \in (-q^{-1}, q^{-1}) \\
0 & \text{elsewhere}
\end{cases}
\]

(4.2)

The PDF is a *uniform* density in the interval \((-q^{-1}, q^{-1})\).

MinimumPowerDivergentCDF_RM-4-21-08.doc
The mean and variance of the distribution can be found using the standard moment formulae for uniform distributions based on the endpoints of the support of the random variable. Specifically,

\[ E(W) = \frac{-(1-q)^{-1} + q^{-1}}{2} = \frac{.5-q}{q(1-q)} \] (4.3)

\[ Var(W) = \frac{\left(q^{-1}+(1-q)^{-1}\right)^2}{12} = \frac{1}{12q^2(1-q)^2} \] (4.4)

Thus, the PDF is centered at zero when \( q = .5 \) but has a negative or positive mean if \( q > .5 \) or \( q < .5 \), respectively. The mean and the variance of the PDF approach infinity if either \( q \to 0 \) or \( q \to 1 \), and the variance takes its smallest value of .75 when \( q = .5 \).

4.2. \( MPD(\gamma = 0) \)

In the \( \gamma = 0 \) case, the family of CDFs generated by the MPD approach is

\[ F(w; q) = \frac{q \exp(w)}{(1-q) + q \exp(w)} \] (4.5)

The functional form of the associated PDF in this case is

\[ f(w; q) = \frac{dF(w; q)}{dw} = \frac{q(1-q)\exp(w)}{\left[(1-q) + q \exp(w)\right]^2} \text{ for } w \in (-\infty, \infty) \] (4.6)

and, when the reference distribution is uniform \((q = .5)\), the distribution specializes to the familiar standard logistic probability density function. The mean of the distribution can be defined as

\[ E(W) = \int_{-\infty}^{\infty} w \frac{q(1-q)\exp(w)}{\left[(1-q) + q \exp(w)\right]^2} dw = \ln(1-q) - \ln(q) \] (4.7)

with variance

\[ Var(W) = \int_{-\infty}^{\infty} \left(w - (\ln(1-q) - \ln(q))\right)^2 \frac{q(1-q)\exp(w)}{\left[(1-q) + q \exp(w)\right]^2} dw = \frac{\pi^2}{3} \quad \forall q \in [0,1] \] (4.8)
Thus, the distribution is centered at zero when $q = .5$ and has a negative or positive mean when $q > .5$ or $q < .5$, respectively. The mean approaches infinity if either $q \to 0$ or $q \to 1$. The variance is constant throughout the range of choices for $q$ and is identical to the variance of the standard logistic distribution, $\left( \pi^2 / 3 \right)$.

The third central moment of this set of distributions is zero and the densities are symmetric about their mean values. Some illustrative plots of the densities for a range of $q$’s are given in Figure 4.1.

**Figure 4.1.** PDFs for $\gamma = 0$ and $q = .9, .7, .3$ and .1.

4.3. $MPD(\gamma = -1)$

In the $\gamma = -1$ case, the family of CDFs generated by the MPD approach is

$$F(w; q) = .5 + \frac{\left[w^2 + (4q - 2)w + 1\right]^5 - 1}{2w}$$ (4.9)

where $F(0; q) = .5$. It follows that the functional form of the PDF is
This PDF has no finite moments of any order and, in particular, has neither a mean nor a variance.

Some illustrative graphs of this family of distribution are given in Figure 4.2. Except when \( q = .5 \), the densities are either skewed left or skewed right depending on whether \( q < .5 \) or \( q > .5 \), respectively.

**Figure 4.2.** PDFs for \( \gamma = -1, q = .9, .7, .5, .3 \) and .1

5. General Properties of the MPD-Class of Probability Distributions

The existence of moments in the MPD-Class of distributions and their values depend on the \( \gamma \) 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 cannot be represented in closed form. Nevertheless, after transformation, probabilities and expectations are relatively straightforward to represent and develop when they exist.

Consider, for example, the general definition of the expectation of \( g(W) \) with respect to a density in the MPD-Class. Treating the probabilities as implicit functions of \( w \) and then collecting probability derivative terms, the differentiation of (3.8) with respect to \( w = x, \lambda \) implies the following general representation of probability densities for nonzero values of \( \gamma \) (the case of \( \gamma = 0 \) can be handled explicitly, as discussed in Section 4.2):

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

(5.1)

where \( \Upsilon(q, \gamma) \) denotes the appropriate support of the density function, which, as indicated in (3.8), depends on \( q \) and \( \gamma \) if \( \gamma > 0 \), and \( \Upsilon(q, \gamma) = \mathbb{R} \) otherwise. Expectations are then defined as

\[
E(g(W)) = \int_{w \in \Upsilon(q, \gamma)} g(w) \left( q^{-\gamma} F(w; q, \gamma)^{\gamma-1} + (1-q)^{-\gamma} (1 - F(w; q, \gamma))^{\gamma-1} \right)^{-1} dw
\]

(5.2)

Making a change of variables in (5.2) via the transformation \( p = F(w; q, \gamma) \) so that

\[
w = F^{-1}(p; q, \gamma) \quad \text{and} \quad \frac{\partial w}{\partial p} = \frac{\partial F^{-1}(p; q, \gamma)}{\partial p},
\]

where \( F^{-1}(p; q, \gamma) \) denotes the inverse function associated with the CDF, it follows that the expectation in (5.2) can be represented as

\[
E(g(W)) = \int_{0}^{1} g\left(F^{-1}(p; q, \gamma)\right) dp
\]

(5.3)

Note that this involves the closed form inverse CDF function given by

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

(5.4)
which follows directly from (5.1). When $g(W)$ is such that its expectation exists, (5.3) can be represented in general as

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

(5.5)

Next, it is useful to partition the class of densities into two subsets depending on the sign of $\gamma$ and consider moment representations.

**5.1. Moments for $\gamma > 0$**

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

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

(5.6)

is bounded for each positive integer-valued $\delta$, finite $q \in (0,1)$, and each finite positive-valued $\gamma$. The means of the probability densities are given by evaluating the integral (5.6) for $\delta = 1$, resulting in

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

(5.7)

It is straightforward to verify that (4.3) is a special case of (5.7).

The second moment around the origin is obtained by solving (5.6) when $\delta = 2$, resulting in

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

(5.8)
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 (5.7) from (5.8) to yield

\[
\text{var}(W) = \left(1 + \gamma\right)^{-2} \left\{ \frac{\left( q^{-2\gamma} + (1-q)^{-2\gamma} \right)}{(1+2\gamma)} + \frac{2q^{-\gamma}(1-q)^{-\gamma}\left[1-(1+\gamma)^2 B(\gamma+1,\gamma+1)\right]}{\gamma^2} \right\}.
\] (5.9)

The functional representation of higher order moments becomes progressively more expansive. However, the evaluation of moments for any of the MPD-Class distributions with \( \gamma > 0 \) remains straightforward via the numerical calculation of the well-defined and well-behaved integral equation (5.6).

**5.2. Moments for \( \gamma < 0 \)**

Distributions in the MPD-Class with \( \gamma \leq -1 \) do not have moments defined of any order because the integral in (5.6) is divergent for any choice of \( \delta \geq 1 \). 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 \( \gamma \).

If \( \gamma > -1 \) the mean of the distribution exists, and its functional representation in terms of \( \gamma \) and \( q \) is precisely the same as in (5.7). If \( \gamma > -1 \) the second moment about the origin, and thus the variance, exist and have exactly the same functional forms as in (5.8) and (5.9), respectively. In general, the moment of order \( \delta \) 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 \( \gamma > 0 \).

To further illustrate the myriad of distributional characteristics contained within the MPD-Class of probability distributions, Figures 5.1 and 5.2 plot the PDFs associated with selected values of \( q_i \) and \( \gamma \).
**Figure 5.1.** PDFs for $q = .5$, $\gamma = -3$, -1.5, -1, -.5, 0, .5, 1, 1.5, and 3

![Figure 5.1. PDFs for $q = .5$, $\gamma = -3$, -1.5, -1, -.5, 0, .5, 1, 1.5, and 3](image1)

**Figure 5.2.** PDFs for $q = .75$, $\gamma = -3$, -1.5, -1, -.5, 0, .5, 1, 1.5, and 3

![Figure 5.2. PDFs for $q = .75$, $\gamma = -3$, -1.5, -1, -.5, 0, .5, 1, 1.5, and 3](image2)

6. **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.

6.1. MPDP Estimation

Returning to the model of binary response outlined in Section 2, consider the minimum power divergence problem depicted by the Lagrange multiplier optimization problem in (3.5)-(3.6). 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 (3.7)-(3.8), 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 \( \lambda \) is determined by substituting the functional representation of \( p_i(x,\lambda) \) into the first order conditions with respect to \( \lambda \), and then solving the equations:

\[
\lambda_{\text{MPD}} = \arg_x \left\{ x' \left( y - p(x\lambda) \right) = 0 \right\} \tag{6.1}
\]

for the value of \( \lambda \). The estimated value of \( p \) follows directly by substitution, as

\[
p_{\text{MPD}} = p(x\lambda_{\text{MPD}}) .
\]

As in all Lagrange-form optimization problems, \( \lambda \) 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 \( \lambda \) can be thought of as representing the “relative contribution” of each of the \( k \) data constraints to the minimized divergence value. The polar case \( \lambda_i = 0 \) would indicate that the \( i^{th} \) data constraint is non-binding and redundant, adding no informational value to that already contained in the reference distribution for those probabilities. It is not apparent from general Lagrange multiplier theory that \( \lambda \) can actually be interpreted as an estimate of the
parameter vector $\beta$ underlying the linear index representation of the Bernoulli probabilities depicted in (1.1)-(1.2). We motivate this interpretation next.

6.2 Interpreting $\lambda$ as an Estimator of $\beta$

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

$$L(\mathbf{p}, \lambda) = \sum_{i=1}^{n} \left( \frac{1}{\gamma(\gamma+1)} \left[ p_i \left( \frac{p_i}{q_i} \right)^\gamma + (1 - p_i) \left( \frac{1-p_i}{1-q_i} \right)^\gamma - 1 \right] \right) + \lambda \left[ n^{-1}(\mathbf{x}'(\mathbf{F}(\mathbf{x}\beta) - \mathbf{p})) \right], \quad 0 \leq p_i, q_i \leq 1, \forall i \quad (6.2)$$

The first-order conditions with respect to $\mathbf{p}$ would be precisely as indicated in (3.7), leading to the same representations of the optimal $\mathbf{p}$ expressed in terms of $\lambda$, i.e., the same $\mathbf{p}(\mathbf{x}\lambda)$ vector of probabilities represented by (3.8).

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 $\mathbf{F}(\mathbf{x}\beta)$. The first order conditions with respect to $\lambda$ imply that

$$\mathbf{H}(\beta, \lambda) = n^{-1}\mathbf{x}'(\mathbf{F}(\mathbf{x}\beta) - \mathbf{F}(\mathbf{x}\lambda)) = 0 \quad (6.3)$$

in which case it is apparent that one solution for $\lambda$ is given directly by $\lambda = \beta$. That this is the unique solution to the problem follows from the Implicit Function Theorem, which can be used to demonstrate that (6.3) determines $\lambda$ as a function of $\beta$ in the neighborhood of $\beta$. By this theorem, if the Jacobian of the $k$ constraints in (6.3) with respect to the $k \times 1$ vector $\lambda$ is nonsingular when evaluated at $\lambda = \beta$, then such a functional relationship exists. The Jacobian is given by
\[
\frac{\partial H(\beta, \lambda)}{\partial \lambda} \bigg|_{\beta=\lambda} = -n^{-1} x'(f(x\beta) - x) = -n^{-1} \sum_{i=1}^{n} f'(x_i\beta) x_i x_i'
\]  
(6.4)

where \( f(x\beta) \) 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 \( x \) that are not linearly independent and for which \( f(x,\beta) > 0 \). It follows that \( \lambda = \beta \) in the solution to (6.3). Using observable sample moments in place of unobservable population moments, as \( n^{-1} x'(Y - F(x\lambda)) = n^{-1} x'(F(x\beta) - F(x\lambda) + \epsilon) = 0 \), the solution for \( \lambda \) is then interpretable as a random variable estimating \( \beta \) which, for example, is consistent under familiar regularity conditions that include \( n^{-1} x' \epsilon \overset{p}{\rightarrow} 0 \). In Section 6.4, we demonstrate formally that the MPD estimator is consistent and asymptotically normally distributed under general regularity conditions.

6.3. 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 \( \frac{\partial p_i}{\partial x_{ij}} \) 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

Logit: \( \frac{\partial \hat{p}_i}{\partial x_{ij}} = \frac{\exp(x_i\hat{\beta})}{\left[ 1 + \exp(x_i\hat{\beta}) \right]^2} \hat{\beta}_j \)  
(6.5)

Probit: \( \frac{\partial \hat{p}_i}{\partial x_{ij}} = \phi(x_i\hat{\beta}) \hat{\beta}_j \)  
(6.6)

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 \( p_i(x, \lambda) \) with respect to the response variables used in defining the linear index, yield the following:

\[
\gamma < 0: \quad \frac{\partial p_i}{\partial x_j} = \frac{\lambda_j}{q_i^\gamma p_i^{\gamma-1} + (1-q)^\gamma (1-p_i)^{\gamma-1}} \\
\gamma = 0: \quad \frac{\partial p_i}{\partial x_j} = \frac{q_i (1-q_i)^{-1} \exp(x_i, \lambda)}{1 + q_i (1-q_i)^{-1} \exp(x_i, \lambda)} \lambda_j \\
\gamma > 0: \quad \frac{\partial p_i}{\partial x_j} = \begin{cases} 0 & \text{for } p_i \in (0,1) \\ \lambda_j & q_i^\gamma p_i^{\gamma-1} + (1-q)^\gamma (1-p_i)^{\gamma-1} \end{cases}
\]

The derivative in (6.8) is recognized as being identical in functional form to the logit derivative defined in (6.5) when \( q_i = .5 \). Moreover, the solution for \( \lambda_{\text{MPD}} \) and the logit estimate of \( \hat{\beta} \) are in fact identical when \( q_i = .5 \) since the first-order conditions to both estimation problems coincide.

6.4. Asymptotics and Inference

As one basis for evaluating estimation performance, we demonstrate general asymptotic properties of the MPD-Class of estimators and indicate how the asymptotic properties of the estimators can be used to define inference procedures. In the ensuing discussion, it will be useful to represent the first order conditions of (6.2) with respect to the Lagrange multipliers as

\[
G_{\epsilon}(\lambda) = n^{-1} \sum_{i=1}^{n} X_i' \left( F(X_i, \beta) - p_i(X_i, \lambda) + \epsilon_i \right) = n^{-1} \sum_{i=1}^{n} g_i(\lambda) = 0.
\]

6.4.1. Consistency

For the consistency of the MPD estimator \( \lambda \) of \( \beta \), we make the following basic assumptions:
**Assumption 1.** The observations \( (y_i, x_i), i = 1, \ldots, n \), are iid random realizations of the random row vector \( (Y, X) \).

**Assumption 2.** \( G_n(\beta) \to 0 \) with probability 1.

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

Applying assumptions 1-3 to the MPD-Estimator imposes conditions on the data-generating process underlying data outcomes as well as the specification of the model of binary outcomes. 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 the family of MPD-distributions be appropriately specified to encompass the functional form of the true underlying CDF, \( F(\cdot) \). This condition is akin to correctly specifying the functional form of the probability distribution underlying a maximum likelihood (ML) estimation problem. In this event,

\[
E \left( g_i(\beta) \right) \equiv E(X_i' \epsilon_i) = 0 \quad \text{and} \quad g_i(\beta), \ i = 1, \ldots, n, \ \text{are iid} \quad (6.11)
\]

imply that

\[
G_n(\beta) = n^{-1} \sum_{i=1}^{n} g_i(\beta) \overset{w^p}{\to} 0 \quad (6.12)
\]

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 \( G_n(\lambda) \) is given by

\[
\frac{\partial G_n(\lambda)}{\partial \lambda} = -n^{-1} \sum_{i=1}^{n} \frac{\partial p(X, \lambda)}{\partial X_i \lambda} X_i' X_i = -n^{-1} \sum_{i=1}^{n} f(X_i, \lambda) X_i' X_i \quad (6.13)
\]
where \( f(z) \) denotes a probability density function in the MPD class of distributions. It is apparent that the continuous differentiability of \( G_n(\lambda) \) depends on the continuity of

\[
\frac{\partial p(z)}{\partial z} = f(z).
\]

This follows from the functional definition of \( f(z) \) in (5.1) and that, except on the boundaries of the supports of the densities indicated in (3.8), \( \frac{\partial p(z)}{\partial z} \) is continuous everywhere when \( \gamma \leq 0 \) and continuous except on an event having probability 0 when \( \gamma > 0 \).

Moreover, in (4.12) implies that the MPD densities are all bounded, as \( f(z) < \xi < \infty \). Thus \( f(X, \lambda)X_i'X_j < \xi X_i'X_j \), so \( E_{\lambda} \left( f(X, \lambda)X_jX_j \right) < \xi E \left( X_jX_j \right) < \infty \), \forall i, j \). Therefore

\[
\frac{\partial G_n(\lambda)}{\partial \lambda} \text{ converges uniformly to } \frac{\partial G(\lambda)}{\partial \lambda} = E \left( f(X, \lambda)X_iX_i \right),
\]

which will be nonsingular at \( \lambda = \beta \) if \( E \left( X_iX_i \middle| f(X, \lambda) > 0 \right) \) is nonsingular.

**Theorem 1.** Under assumptions 1 – 3, the MPD-Estimator \( \hat{\lambda} = \text{arg}_{\lambda} \left[ G_n(\lambda) = 0 \right] \) is a consistent estimator of \( \beta \).

**Proof:** The assumptions imply the regularity conditions shown by Yuan and Jennrich (1998) to be sufficient for \( \hat{\lambda} = \text{arg}_{\lambda} \left[ G_n(\lambda) = 0 \right] \overset{a.s.}{\rightarrow} \beta \), and thus for \( \hat{\lambda} \) to be a (strongly) consistent estimator of \( \beta \).

Hence, if the data-generating process governing the data outcomes adheres to the conditions identified above and the model distribution is appropriately specified, the MPD-Estimator will consistently estimate \( \beta \). If the model distribution is not specified correctly, the MPD-Estimator 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 \( \beta \).
6.4.2. Asymptotic Normality

Given that the MPD-Estimator is consistent, asymptotic normality of the estimator of $\beta$ is implied if the following additional assumption is made:

**Assumption 4:** $n^{1/2} G_n(\beta) \xrightarrow{d} N(0, V)$

**Theorem 2.** Under assumptions 1 – 4, the MPD-Estimator $\hat{\lambda} = \arg\min G_n(\lambda) = 0$ is asymptotically normally distributed, with the limiting distribution $n^{1/2}(\hat{\lambda} - \beta) \xrightarrow{d} N(0, A^{-1} V A^{-1})$, where $A = \frac{\partial G(\lambda)}{\partial \lambda} \equiv E\left(f(X_1\beta)X_1'X_1\right)$ and $V = E\left(F\left(X_1\beta\right)(1 - F\left(X_1\beta\right))X_1'X_1\right)$.

**Proof:** The preceding assumptions 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 indicated. □

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

$$n^{1/2} G_n(\beta) = n^{-1/2} \sum_{i=1}^{n} g_i(\beta) = n^{-1/2} \sum_{i=1}^{n} X_i' \varepsilon_i$$

(6.14)

is a scaled sum of iid random vectors, each having a zero mean vector and a covariance matrix

$${\text{Cov}}(X_i' \varepsilon_i) = E\left(F\left(X_1\beta\right)(1 - F\left(X_1\beta\right))X_1'X_1\right)$$. Based on the multivariate version of the Lindberg-Levy Central Limit Theorem (Serfling, 1980, p. 28),

$$n^{-1/2} \sum_{i=1}^{n} X_i' \varepsilon_i \xrightarrow{d} N\left(0, E\left(F\left(X_1\beta\right)(1 - F\left(X_1\beta\right))X_1'X_1\right)\right)$$. Consequentially, as specified in Theorem 2, the MPD-Estimator will follow the normal limiting distribution if $${\text{Cov}}(X_i' \varepsilon_i)$$ is nonsingular.

6.4.3. Asymptotic Inference

Based on the asymptotic results of the previous subsections, all 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 $A = \frac{\partial G(\lambda)}{\partial \lambda} \equiv E \left( f(X_i, \beta) X_i' X_i \right)$ is defined by

$$\hat{A} = n^{-1} \sum_{i=1}^{n} f(X_i, \hat{\lambda}) X_i' X_i.$$  \hfill (6.15)

and a consistent estimator of $V = E \left( F(X_i, \beta) \left(1 - F(X_i, \beta)\right) X_i' X_i \right)$ is defined by

$$\hat{V} = n^{-1} \sum_{i=1}^{n} F(X_i, \hat{\lambda}) \left(1 - F(X_i, \hat{\lambda})\right) X_i' X_i.$$ \hfill (6.16)

It follows that a Wald-type statistic for testing the $J$ linear restrictions $H_0 : C\beta = r$ is given by

$$n \left( C\hat{\lambda} - r \right)' \left( C A^{-1} V A^{-1} C' \right)^{-1} \left( C\hat{\lambda} - r \right) \overset{d}{\to} \chi^2_J \text{ under } H_0.$$ \hfill (6.17)

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

$$p(z\hat{\lambda}) \overset{d}{\sim} N \left( p(z\beta) , \, n^{-1} f(z\hat{\lambda})^2 z A^{-1} V A^{-1} z' \right)$$ \hfill (6.18)

so that, given $H_0 : p(z\beta) = p_0$,

$$n \left( p(z\hat{\lambda}) - p_0 \right) \left( f(z\hat{\lambda})^2 z A^{-1} \hat{V} A^{-1} z' \right)^{-1/2} \overset{d}{\sim} N(0,1) \text{ under } H_0.$$ \hfill (6.19)

7. Summary, Implications, and Extensions

In this paper, we represent sample information underlying binary choice outcomes through general moment conditions, $E \left[ Z'(Y - p) \right] = 0$. We then use the Cressie-Read (CR) family of divergence measures, $CR(\gamma)$ for $\gamma \in (-\infty, \infty)$, 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 and inference implications of this formulation are analytically assessed.

A number of important issues relating to the MPD approach to estimation and inference in the discrete choice model context remain. One of these includes 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 Sections 4 and 5. Another important issue concerns extending the MPD formulation to handle endogeneity of some \( X \)’s by considering moment conditions of the form \( E\left[ Z'(Y - p(X)) \right] = 0 \), where \( Z \) is not necessarily a function of \( X \) (Judge, et al., 2006).

A third 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

\[
L(p, \lambda) = \sum_{i=1}^{n} \sum_{j=1}^{m} \left( \frac{1}{\gamma (\gamma + 1)} \sum_{j=1}^{m} p_{ij} \left( \frac{p_{ij}}{q_{ij}} \right)^{\gamma} - 1 \right) + \sum_{j=1}^{m} \lambda_{j}' x'(y_j - p_j) + \sum_{i=1}^{n} \eta_i \left( \sum_{j=1}^{m} p_{ij} - 1 \right). \tag{8.1}
\]

The estimation and inference potential of this type of formulation is currently being investigated.

A final important issue concerns the use of probability distributions in the MPD-Class to form a basis for robust 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

\[
\ell(\beta, q, \gamma | y, x) = \sum_{i=1}^{n} \left( y_i \ln \left( 1 - F(-x_i; \beta, q, \gamma) \right) + (1 - y_i) \ln \left( F(-x_i; \beta, q, \gamma) \right) \right). \tag{8.2}
\]

Maximum likelihood estimation of the unknowns in the specification of the Bernoulli probabilities can proceed by maximizing (8.2). This formulation allows the vast array of CDFs in the MPD-Class to represent the Bernoulli probabilities underlying the discrete choice process via the selection of the \( q \) and \( \gamma \) values. These and other issues such as sampling experiments to gauge finite sample performance, are the subject of ongoing research.
REFERENCES


