Minimum Divergence Moment Based Binary Response Models:

Estimation and Inference

Ron Mittelhammer, George Judge, Douglas Miller, and N. Scott Cardell

Washington State University, Pullman, WA, 99164

University of California Berkeley, Berkley, CA, 94720

University of Missouri, Columbia, Mo, 65211

Salford Systems, Inc. San Diego, CA, 92108

Ron C. Mittelhammer is Regents Professor of Economic Sciences and Statistics, and Director of the School of Economic Sciences at Washington State University, School of Economic Sciences, PO Box 646210, Pullman, WA, 99164, (email: mittelha@wsu.edu), George G. Judge is Professor in the Graduate School, 207 Giannini Hall, University of California, Berkeley, Berkeley, CA, 94720 (e-mail: judge@are.berkeley.edu), Douglas Miller is Assistant Professor of Economics in the Department of Economics at the University of Missouri, Columbia, Missouri, 65211 (e-mail: millerdou@missouri.edu), and N. Scott Cardell is Director of Research, Salford Systems, Inc., San Diego, CA, 92108, scardell@gocougs.wsu.edu. The authors gratefully acknowledge the helpful and substantive comments of Marian Grendar, Guido Imbens, Art Owen, Paul Ruud, Kenneth Train, and David Wolpert.

Minimum Divergence Moment Based Binary Response Models:

Estimation and Inference

ABSTRACT

This paper introduces a new class of estimators based on minimization of the Cressie-Read (CR) power divergence measure for binary choice models, where neither a parameterized distribution nor a parameterization of the mean is specified explicitly in the statistical model. By incorporating sample information in the form of conditional moment conditions and estimating choice probabilities by optimizing a member of the set of divergence measures in the CR family, a new class of nonparametric estimators evolves that requires less a priori model structure than conventional parametric estimators such as probit or logit. Asymptotic properties are derived under general regularity conditions and finite sampling properties are illustrated by Monte Carlo sampling experiments. Except for some special cases in which the general regularity conditions do not hold, the estimators have asymptotic normal distributions, similar to conventional parametric estimators of the binary choice model. The sampling experiments focus on the mean square errors in the choice probability predictions and the probability derivatives with respect to the response variable values. The simulation results suggest that estimators within the CR class are more robust than conventional methods of estimation across varying probability distributions underlying the Bernoulli process. The size and power of test statistics based on the asymptotics of the CR-based estimators exhibit behavior similar to those based on conventional parametric methods. Overall, the new class of nonparametric estimators for the binary response model is a promising and potentially more robust alternative to the parametric methods often used in empirical practice.

Keywords: nonparametric binary response models and estimators, conditional moment equations, finite sample bias and precision, squared error loss, response variables, Cressie-Read statistic, information theoretic methods

AMS 1991 Classification Primary 62E20

JEL Classifications: C10, C2

Minimum Divergence Moment-Based Binary Response Models:

Estimation and Inference

1. INTRODUCTION

Discrete choice behavioral models occupy a significant niche in social science research, and especially in theoretical economics. As in any inference situation where attempts are made to learn from a sample of data, the analyst is confronted with numerous choices in the design of the information recovery process. Two important decisions relate to selecting the estimation criterion and how to represent the sample information. In this paper we investigate the statistical implications of these and other choices leading to a solution, in the binary case, of this decision problem under uncertainty. Our contributions include a new class of nonparametric alternatives to modeling binary response that nests the behavior of some prominent parametric estimators while exhibiting robustness relative to the actual, and generally unknown, statistical model underlying the generation of binary responses.

In binary response models it is assumed that on trial $i = 1, 2, \ldots, n,$ one of two alternatives is observed to occur for the independent binary random variables $\{Y_1, \ldots, Y_n\}$ having p_i, $i = 1, ..., n$, as their respective probabilities of success. In empirical applications the data sampling process for the binary random variable Y_i is generally specified in terms of a latent variable, Y_i^*, as

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

where $Y_i \equiv I\left(Y_i^* > 0\right)$, i = 1,...,n are independent Bernoulli random variables, $I(A)$ is an indicator function that takes the value 1 when condition (A) is true and takes the value 0 otherwise, and $\mathbf{x}_{i.}$, i = 1,..., n are independent outcomes of a $(1 \times k)$ random vector of response variables.

Given (1.1), the value of p_i is defined by

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

where $G(\bullet)$ is the cumulative distribution function (CDF) of the noise term ε_i^* of the latent variable equation (1.1), and $G_*(\bullet)$ is the complement of this CDF. In instances where the parametric family of probability density functions underlying the binary response model is known, the parametric functional form of $G\left(\mathbf{x}_{i.}\boldsymbol{\beta}\right)$ is then also known and one can fully define the log-likelihood function and utilize the traditional maximum likelihood (ML) approaches such as logit or probit as a basis for estimation and inference. If the particular choice of the parametric functional form for the distribution happens to be correct, then the usual ML properties of consistency, asymptotic normality and efficiency hold (McFadden 1974, 1984 and Train 2003). However, in reality there is normally some ambiguity relative to the correct behavioral model and thus uncertainty exists regarding the underlying data sampling process and how best to proceed with model specification, estimation , and inference (for further discussion, see Cosslett 1983, Maddala 1983, Ichimura 1993, Klein and Spady 1993, and McCullough and Nelder 1995). Given this estimation and inference problem we assume that the distribution of ε_i^* *is not* based on, or restricted to, a parametric family and suggest a range of nonparametric estimators to recover estimates of the choice probabilities as well as the corresponding derivatives of these probabilities with respect to the response variables. Our choice of estimation

criterion is based on the Cressie-Read family of power divergence measures, and we represent sample information in a general nonparametric way based on sample moments devoid of a finite parametric structure.

The organization of the paper is as follows: In Section 2 a class of data based nonparametric estimators based on moment conditions is defined and motivated. Section 3 provides details of the functional solutions for the Bernoulli probabilities and derivations of the marginal probability effects for the various estimators. Asymptotic sampling properties of the estimators are developed in Section 4. Monte Carlo sampling results are presented in Section 5 to indicate the finite sample performance of the nonparametric estimators. Finally, in Section 6, the sampling implications of our formulations are discussed and possible extensions of the methodology are suggested. Proofs of the theorems are given in the Appendices.

2. GENERAL NONPARAMETRIC ESTIMATION FORMULATIONS

Regarding alternatives to traditional binary likelihood and quasi-likelihood based formulations, two primary questions basic to estimation are how to represent the sample information and what estimation criterion should be used. We represent the sample information in the form of conditional moments that link the empirical sample observations to the Bernoulli probabilities underlying the binary responses. The estimation criterion is information theoretic in nature and utilizes the Cressie-Read family of power divergence statistics as the estimation objective function (Cressie and Read 1984 and Read and Cressie 1988).

2.1 Representation of Sample Information

Given a particular discrete choice problem and an observed sample of data, an approach often followed in specifying the statistical model is to assume that the data outcomes follow a particular distributional form that depends on a certain fixed set of unknown parameters. Two popular distributions within the binary response modeling context are the probit and logit distributions. If one actually has such specific parameterized knowledge, rather than only a feasible set of distributions implied by an underlying conceptual response model, then maximum likelihood estimation is a viable way to proceed. However in many cases little is know other than the observed sample of data originates from the basic model

$$\mathbf{Y} = \mathbf{p} + \boldsymbol{\varepsilon} \tag{2.1}$$

where $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, which implies only that the binary random vector **Y** has some mean vector **p.** If through a process of interpretation the analyst makes use of a conceptual model (e.g., random utility) as the basis for discrete response outcomes, the statistical model usually involves known covariate information in the form of associated response variables, or instruments, **Z**. Conditioning on what is known, we represent this information in terms of orthogonality relationships in the form of moment conditions

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

Without additional assumptions, this represents all of the sample and a priori information actually available for estimating the unknowns in the model.

Given (2.2), and the condition $E\left[Y_i - p_i \mid \mathbf{x}_{i.}\right] = 0, \ \forall i$, a natural candidate for inclusion in the **Z** matrix is the $(n \times k)$ matrix **X** relating to (1.1). Additional candidates for inclusion in **Z** might be powers and cross products of the nonconstant columns of the **X** matrix, because

$E\left[Y_i - p_i \mid \mathbf{x}_{i.}^{\tau}\right] = 0$ and/or $E\left[Y_i - p_i \mid \mathbf{x}_{i.}^{\tau} \odot \mathbf{x}_{j.}^{\delta}\right] = 0$, where $\odot$ denotes the Hadamard product and τ and δ are positive integers . If $\mathbf{p}$ were given an explicit parametric functional form, say as $\mathbf{p} = \mathbf{G}(\mathbf{x}\boldsymbol{\beta})$ with $\mathbf{G}(\bullet)$ being some cumulative distribution function, then these moment equations could form the basis for empirical moments of the form $n^{-1}\mathbf{Z}'\left(\mathbf{Y} - \mathbf{G}(\mathbf{x}\boldsymbol{\beta})\right) = \mathbf{0}$ and nonlinear generalized method of moments (GMM) might be used to estimate the unknown parameter vector. However in the context of (2.2), $\mathbf{G}(\bullet)$ is neither assumed known nor explicitly specified. Thus 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

$$n^{-1}\mathbf{z}'(\mathbf{y} - \mathbf{p}) = \mathbf{0} \tag{2.3}$$

cannot possibly be used in isolation to identify the Bernoulli probabilities since, regardless of their number, $\mathbf{p} = \mathbf{y}$ solves the set of moment constraints. In a solution context, for this inverse problem, there are more unknowns than estimating equations. Consequently, (2.3) is in the form of an ill-posed inverse problem and the system of equations is substantially underdetermined regarding a unique interior solution for the probability vector **p**.

2.2 Divergence-Based Estimation Criteria

As a basis for estimating the underdetermined **p**-vector in (2.3), we adopt the Cressie-Read (CR) family of power divergence measures (Cressie and Read, 1984) as our estimation objective function, which for the *i*th Bernoulli probability takes the form

$$I\left(\begin{pmatrix} p_{i1} \\ p_{i2} \end{pmatrix}, \begin{pmatrix} q_{i1} \\ q_{i2} \end{pmatrix}, \gamma\right) = \frac{1}{\gamma(\gamma+1)} \sum_{j=1}^{2} p_{ij} \left[\left(\frac{p_{ij}}{q_{ij}}\right)^{\gamma} - 1\right] \tag{2.4}$$

where $P(y_i = 1) = p_{i1}$, $P(y_i = 0) = p_{i2}$, and $\mathbf{p}_{i\cdot} = \begin{bmatrix} p_{i1} & p_{i2} \end{bmatrix}$ and $\mathbf{q}_{i\cdot} = \begin{bmatrix} q_{i1} & q_{i2} \end{bmatrix}$ represent the subject and reference distributions of the CR divergence measure, respectively. The choice of $\gamma \in (-\infty, \infty)$ determines an entire family of measures of divergence between the discrete **p** and **q** probability distributions. Henceforth, in regard to (2.4), it is tacitly understood that when $\gamma = 0$ or -1, the right hand side of (2.4) is defined in terms of its continuous limit as $\gamma \to 0$ or -1.

The reference distribution representing the set of probabilities $\mathbf{q}_{i\cdot} = \begin{bmatrix} q_{i1} & q_{i2} \end{bmatrix}$ is effectively the base probability distribution from which divergence is measured. The CR statistic achieves its smallest divergence value of 0 uniquely when the subject distribution equals the reference distribution. This follows directly from an application of Jensen's inequality (Lin, 1990) to (2.4), which implies that $Eg(\eta) = E\left(\frac{1}{\gamma(\gamma+1)}\left[(\eta)^{\gamma} - 1\right]\right) \geq g\left(E(\eta)\right) = g(1) = 0$, where $g(\eta)$ is strictly convex in η and the expectation in this case is being taken with respect to the distribution implied by the p_{ij}'s (see Ullah, p. 145). One choice for the reference distribution is the uniform distribution, which is consistent with the notion that outcomes 0 and 1 are equally likely, *a priori*, and for expository purposes we henceforth adopt this reference distribution specification. Consequently, the CR statistic may be written in the form

$$I_*\left(\begin{pmatrix} p_{i1} \\ p_{i2} \end{pmatrix}, \gamma\right) = \frac{1}{\gamma(\gamma+1)} \sum_{j=1}^{2} p_{ij}\left[\left(2p_{ij}\right)^{\gamma} - 1\right]. \tag{2.5}$$

Given that $p_{i1} + p_{i2} = 1$, the CR(p_i,γ) statistic, with $p_i = P(y_i = 1)$, can be equivalently written as

$$CR(p_i, \gamma) = \frac{1}{\gamma(\gamma+1)}\left[2^{\gamma}\left(p_i^{\gamma+1} + (1-p_i)^{\gamma+1}\right) - 1\right]. \tag{2.6}$$

The CR divergence measure is extended to a set of n binary responses by summing over the n responses, leading to the *generalized* CR statistic

$$CR(\mathbf{p},\gamma)=\frac{1}{\gamma(\gamma+1)}\left[2^{\gamma}\sum_{i=1}^{n}\left(p_i^{\gamma+1}+(1-p_i)^{\gamma+1}\right)-n\right]. \tag{2.7}$$

The family of divergence measures includes Owen's (2000) empirical likelihood ($\gamma=-1$), Kullback-Leibler's discrepancy or Shannon's (1948) entropy ($\gamma=0$), and the log-Euclidean discrepancy measure ($\gamma=1$).

2.3 The General Estimation Problem

Using the information available in the form of moment equations (2.3), consider the problem of choosing the binary response probabilities so as to minimize the generalized CR divergence of the Bernoulli probabilities,

$$\min_{\mathbf{p}}\{CR(\mathbf{p},\gamma)\}=\min_{\mathbf{p}}\left\{\frac{1}{\gamma(\gamma+1)}\left[2^{\gamma}\sum_{i=1}^{n}\left(p_i^{\gamma+1}+(1-p_i)^{\gamma+1}\right)-n\right]\right\} \tag{2.8}$$

subject to:

$$n^{-1}\mathbf{z}'(\mathbf{y}-\mathbf{p})=\mathbf{0}\ \text{ and }\ \mathbf{p}\in\underset{i=1}{\overset{n}{\times}}[0,1] \tag{2.9}$$

The fundamental motivation for this constrained estimation objective is to choose Bernoulli probability distributions as close to their respective reference distributions as possible while satisfying the sample data-based moment conditions. Closeness is measured in terms of the CR(γ) metric and CR(γ) is used as shorthand for the CR-measure determined by the given value of γ. As such, the problem (2.8)-(2.9) can be viewed as a nonparametric variant of a shrinkage-type estimator in which the estimates of the choice probabilities adapt to satisfy the data-based

constraint information while being shrunk toward their reference distributions (e.g. see Judge and Bock, (1978) for a parametric context).

3. SOLUTIONS FOR PROBABILITIES AND MARGINAL EFFECTS

Regarding the solution to the minimization problem in (2.8) – (2.9), the Lagrangian form of the optimization problem is given by

$$L(\mathbf{p},\boldsymbol{\lambda}) = \frac{1}{\gamma(\gamma+1)}\left[2^{\gamma}\sum_{i=1}^{n}\left(p_i^{\gamma+1}+(1-p_i)^{\gamma+1}\right)-n\right]+\boldsymbol{\lambda}'\mathbf{z}'(\mathbf{y}-\mathbf{p}) \tag{3.1}$$

where it is understood that the inequality constraints $\mathbf{p}\in\times_{i=1}^{n}[0,1]$ are also enforced, and we eliminate the superfluous multiplicative factor n^{-1} from the definition of the sample moment conditions in (2.9). The first derivatives of the Lagrangian with respect to the probability terms are given, $\forall i$, by

$$\frac{\partial L}{\partial p_i} = \left\{\begin{array}{l}\frac{2^{\gamma}}{\gamma}\left(p_i^{\gamma}-(1-p_i)^{\gamma}\right)-\mathbf{z}[i,.]\boldsymbol{\lambda}\\ \ln(p_i)-\ln(1-p_i)-\mathbf{z}[i,.]\boldsymbol{\lambda}\end{array}\right\} = 0 \text{ for } \gamma\left\{\begin{array}{l}\neq 0\\ =0\end{array}\right\}. \tag{3.2}$$

These first derivatives are necessary conditions that apply whenever the respective inequality constraint on p_i is not binding. Accounting for the Kuhn-Tucker complementary slackness conditions relating to the inequality constraints on $\mathbf{p}$ leads to the following solution for p_i, expressed as a function of $\boldsymbol{\lambda}$:

$$
\begin{aligned}
p_i(\boldsymbol{\lambda}) &= \arg_{p_i}\left[\left(p_i^{\gamma}-(1-p_i)^{\gamma}\right)=\frac{\mathbf{z}[i,.]\boldsymbol{\lambda}\gamma}{2^{\gamma}}\right] \quad \text{for } \gamma<0 \\
&= \arg_{p_i}\left[\ln(p_i)-\ln(1-p_i)=\mathbf{z}[i,.]\boldsymbol{\lambda}\right] \text{ for } \gamma=0 \\
&= \left\{\begin{array}{c} 1 \\ \left(\arg_{p_i}\left[\left(p_i^{\gamma}-(1-p_i)^{\gamma}\right)=\frac{\mathbf{z}[i,.]\boldsymbol{\lambda}\gamma}{2^{\gamma}}\right]\right) \\ 0 \end{array}\right\} \text{for } \gamma>0 \text{ and } \frac{\mathbf{z}[i,.]\boldsymbol{\lambda}\gamma}{2^{\gamma}} \left\{\begin{array}{c} \geq 1 \\ \in(-1,\,1) \\ \leq -1 \end{array}\right\}
\end{aligned}
\tag{3.3}
$$

A unique solution for $p_i(\boldsymbol{\lambda})$ necessarily exists by the monotonicity of either

$\eta(p_i)=\left(p_i^{\gamma}-(1-p_i)^{\gamma}\right)$ or $\eta(p_i)=\left(\ln(p_i)-\ln(1-p_i)\right)$ in p_i for $\gamma\neq 0$ or $\gamma=0$, respectively.

3.1 Solving for p(λ) when $\gamma=-1,\,0,$ and 1

Explicit closed form solutions exist for **p(λ)** when γ assumes one of the historically prominent values of -1, 0, or 1. When $\gamma=-1$, multiply both sides of the bracketed equality in (3.3) by $p_i(1-p_i)$ and solve the resultant second order polynomial in p_i for the appropriate root contained in the (0,1) interval to obtain

$$
p_i(\boldsymbol{\lambda})=\left\{\begin{array}{c} \left(.5+\frac{\left[\left(\mathbf{z}[i,.]\boldsymbol{\lambda}\right)^2+1\right]^{.5}-1}{2\mathbf{z}[i,.]\boldsymbol{\lambda}}\right) \\ .5 \end{array}\right\} \; \textit{if} \;\; \mathbf{z}[i,.]\boldsymbol{\lambda} \left\{\begin{array}{c} \neq \mathbf{0} \\ = \mathbf{0} \end{array}\right\} \tag{3.4}
$$

For $\gamma=0$, exponentiation of both sides of the bracketed equality in (3.3), and then solving for $p_i(\boldsymbol{\lambda})$ yields

$$
p_i(\lambda)=\frac{\exp\left(\mathbf{z}[i,.]\boldsymbol{\lambda}\right)}{1+\exp\left(\mathbf{z}[i,.]\boldsymbol{\lambda}\right)}. \tag{3.5}
$$

The functional form for p_i in (3.5) is the same as in the logistic binary choice model.

When $\gamma = 1$, the bracketed equality in (3.3) is a linear equation in p_i that can be straightforwardly solved to yield

$$p_i(\boldsymbol{\lambda}) = \left\{ \begin{array}{c} 1 \\ \left(\frac{1}{2} + \frac{\mathbf{z}[i,.]\boldsymbol{\lambda}}{4} \right) \\ 0 \end{array} \right\} \text{ for } \frac{1}{2} + \frac{\mathbf{z}[i,.]\boldsymbol{\lambda}}{4} \left\{ \begin{array}{c} \geq 1 \\ \in (0, 1) \\ \leq 0 \end{array} \right\}. \tag{3.6}$$

3.2 Solving for p($\boldsymbol{\lambda}$) Unconditional on γ

All of the preceding solutions for **p** in Section 3.1 are conditional on a given choice of the CR power parameter γ. Since the γ parameter can be viewed as indexing functional forms contained within the CR family of divergence measures, one might also consider solving for the choice of γ when estimating the Bernoulli probabilities in the binary response model.

Minimizing the CR statistic with respect to γ in the context of (2.8)-(2.9) can be interpreted as choosing the divergence measure, among all of the divergence measures within the CR family, that results in the assignment of the minimum divergence value to any given pair of *subject* probability distribution, **p**, and *reference* probability distribution, **q**, for any given moment constraint specification. As such, the influence of divergence measure choice on the subsequent choice of which **p**-distribution to be paired with the **q**-distribution in any estimation of the probabilities **p** is minimized. Thus, the γ-unconditional CR approach can be thought of as choosing the least influential or most neutral divergence measure to be applied to any pair of subject and reference distributions in the CR family.

The γ-unconditional approach to estimating the binary response probabilities is defined by the solution to the following extremum problem:

$$\min_{\mathbf{p},\gamma}\left\{CR(\mathbf{p},\gamma)\right\} = \min_{\mathbf{p},\gamma}\left\{\frac{1}{\gamma(\gamma+1)}\left[2^{\gamma}\sum_{i=1}^{n}\left(p_i^{\gamma+1}+(1-p_i)^{\gamma+1}\right)-n\right]\right\} \tag{3.7}$$

subject to:

$$n^{-1}\mathbf{z}'(\mathbf{y}-\mathbf{p}) = \mathbf{0} \text{ and } \mathbf{p} \in \underset{i=1}{\overset{n}{\times}}[0,1]. \tag{3.8}$$

Rewriting the CR objective function as

$$CR(\mathbf{p},\gamma) = \sum_{i=1}^{n} Q(p_i,\gamma), \text{ where } Q(p_i,\gamma) \equiv \frac{2^{\gamma}\left(p_i^{\gamma+1}+(1-p_i)^{\gamma+1}\right)-1}{\gamma(\gamma+1)} \tag{3.9}$$

note that the functions $Q(p_i,\gamma)$, $i = 1,..., n$ are convex in γ. Moreover,

$$\gamma^{*}(t) = \arg\min_{\gamma}\left\{Q(t,\gamma)\right\} \in [1.472,\ 1.5] \text{ for } t \in [0,\ 1]. \tag{3.10}$$

The upper bound value of $\gamma = 1.5$ is obtained when $t \to .5$, while the lower bound value of $\gamma = 1.472$ is obtained when either $t \to 0$ or $t \to 1$. It follows that the solution for γ in the estimation problem in (3.7) and (3.8) is restricted to this very narrow band of possible solution values. In discussing this option, we henceforth adopt the solution $\gamma = 1.5$ as the approximate γ-unconstrained solution to the CR problem.

The solution for $p_i(\boldsymbol{\lambda})$ derived from (3.3) with $\gamma = 1.5$ is

$$p_i(\boldsymbol{\lambda}) = \left\{\begin{array}{c} 1 \\ .5+\dfrac{\left(\text{sign}\left(\mathbf{z}[i,.]\boldsymbol{\lambda}\right)\sqrt{1-4w^2}\right)}{2} \\ 0 \end{array}\right\} \text{ for } \gamma > 0 \text{ and } \frac{1.5\left(\mathbf{z}[i,.]\boldsymbol{\lambda}\right)}{2^{1.5}} \left\{\begin{array}{c} \geq 1 \\ \in(-1,\ 1) \\ \leq -1 \end{array}\right\} \tag{3.11}$$

where $\mathrm{w} = \cos\left((1/3)\arccos\left(1 - .5625\left(\mathbf{z}[i,.]\boldsymbol{\lambda}\right)^2\right)\right) - (1/2)$.

3.3 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 the analysis. Estimates of these marginal probability effects, represented by $\partial p_i / \partial x_{ij}$ for the *j*th response variable and the *i*th binary response probability, are straightforwardly defined in the case of the fully parametric logit and probit models:

$$\text{Logit:}\quad \frac{\partial \hat{p}_i}{\partial x_{ij}} = \frac{\exp\left(\mathbf{x}_{i.}\hat{\mathbf{b}}\right)}{\left[1+\exp\left(\mathbf{x}_{i.}\hat{\mathbf{b}}\right)\right]^2}\hat{b}_j \tag{3.12}$$

$$\text{Probit:}\quad \frac{\partial \hat{p}_i}{\partial x_{ij}} = \phi\left(\mathbf{x}_{i.}\hat{\mathbf{b}}\right)\hat{b}_j, \tag{3.13}$$

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

In the case of the CR estimators, marginal probability effects can be derived by differentiating the appropriate definition of $p_i(\boldsymbol{\lambda})$ with respect to the instruments used in defining the moment conditions underlying the estimation problem. Assuming henceforth that **z** = **x**, the derivatives are given as follows:

$$\gamma < 0 \qquad \frac{\partial \hat{p}_i}{\partial z_{ij}} = \frac{\hat{\lambda}_{\mathrm{j}}}{2^{\gamma}\left(\hat{p}_i^{\gamma-1} + (1-\hat{p}_i)^{\gamma-1}\right)} \tag{3.14}$$

$$\gamma = 0 \qquad \frac{\partial \hat{p}_i}{\partial z_{ij}} = \frac{\exp\left(\mathbf{z}_{i.}\hat{\boldsymbol{\lambda}}\right)}{\left[1+\exp\left(\mathbf{z}_{i.}\hat{\boldsymbol{\lambda}}\right)\right]^2}\hat{\lambda}_{\mathrm{j}}. \tag{3.15}$$

$$\gamma > 0 \qquad \frac{\partial \hat{p}_i}{\partial z_{ij}} = \left\{ \begin{pmatrix} 0 \\ \left(\dfrac{\hat{\lambda}_j}{2^{\gamma}\left(\hat{p}_i^{\gamma-1} + (1-\hat{p}_i)^{\gamma-1} \right)} \right) \\ 0 \end{pmatrix} \right\} \text{ for } \frac{\mathbf{z}[i,.]\hat{\boldsymbol{\lambda}}\boldsymbol{\gamma}}{2^{\gamma}} \left\{ \begin{matrix} \geq 1 \\ \in (0,1) \\ \leq 0 \end{matrix} \right\} \tag{3.16}$$

The derivative in (3.15) is recognized as identical in functional form to the logit derivative defined in (3.12). Moreover, the solution for $\hat{\boldsymbol{\lambda}}$ and the logit estimate of $\hat{\mathbf{b}}$ are in fact identical because the first order conditions to both estimation problems coincide, and thus the marginal effect in (3.16), and indeed all other aspects of the solution to the CR(0) problem, are identical to the logit solution.

In comparing the derivative relationships between the parametric probit and logit models, and the models based on any of the CR formulations, it is apparent that the latter do not depend on estimates of the value of a fixed parameter vector $\boldsymbol{\beta}$, but instead depend on the estimated value of the Lagrange multiplier vector $\boldsymbol{\lambda}$ associated with the moment constraints.

4. ASYMPTOTIC PROPERTIES OF CR(γ)-BASED ESTIMATORS

As one important basis for evaluating estimation performance, in this section we develop asymptotic properties of the class of minimum divergence estimators for the binary response model. We also indicate how the asymptotic properties of the estimators can be used to define inference procedures. The proofs of the lemmas and theorems are presented in Appendix A. Consistent with the proof appendix, $\boldsymbol{\lambda}_0$ is defined by $\boldsymbol{\lambda}_0 = \arg\limits_{\boldsymbol{\lambda}} (E(\mathbb{Z}'(\mathrm{Y} - \mathrm{p}(\boldsymbol{\lambda})) = 0)$, where $\mathbb{Z}'$ denotes a random column vector whose *iid* outcomes represent the response variable observations, and the definition of $\mathrm{p}(\boldsymbol{\lambda})$ is based on (3.3) with the *i* subscripts suppressed and

$\mathbf{z}[i,.]$ replaced by $\mathbb{Z}$. In the case of the CR estimators, we also define $\boldsymbol{\beta} = \left(|\gamma| / 2^{\gamma}\right)\boldsymbol{\lambda}_0$ when $\gamma \neq 0$ and $\boldsymbol{\beta} = \boldsymbol{\lambda}_0$ otherwise,

4.1 Consistency and Asymptotic Distributions

The asymptotic results will utilize the following basic assumptions:

1. Each observation $(y_i, \mathbf{z}_{i\bullet}')$ is an iid random realization of the random row vector $(Y, \mathbb{Z})$ with some unknown distribution. The random variable Y is a zero-one binary variable.
2. $E(\mathbb{Z})$ and $E(\mathbb{Z}'\mathbb{Z})$ exist and are finite.
3. $E(\mathbb{Z}'\mathbb{Z})$ is a nonsingular matrix, and if $\gamma > 0$ then $E(\mathbb{Z}'\mathbb{Z} \mid -1 < \mathbb{Z}\boldsymbol{\beta} < 1)$ is a nonsingular matrix, with $\mu = P(-1 < \mathbb{Z}\boldsymbol{\beta} < 1) > 0$.

Under assumptions 1, 2 and 3, $\boldsymbol{\lambda}_0$ must be uniquely defined and finite. The role of assumption 3 is to guarantee that $\hat{\boldsymbol{\lambda}}$ is asymptotically identified, as well as to ensure the existence of a well-defined asymptotic distribution.

The following three theorems identify the consistency of the CR-based estimators, and also identify applicable asymptotic normal distributions, including the asymptotic covariance matrix of the estimators.

Theorem 1: Given assumptions 1-3, $\text{plim}(\hat{\boldsymbol{\lambda}}) = \boldsymbol{\lambda}_0$, so that $\hat{\boldsymbol{\lambda}}$ consistently estimates $\boldsymbol{\lambda}_0$.

Theorem 2: Given assumptions 1-3 and $\gamma < 1$, $\sqrt{n}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0) \overset{a}{\sim} N(\mathbf{0}, \boldsymbol{\Omega}_0^{-1}\boldsymbol{\Psi}\boldsymbol{\Omega}_0^{-1})$, where

$$\Omega_0 = E\left(\frac{\mathbb{Z}'\mathbb{Z}}{2^{\gamma}(\mathbf{p}(\boldsymbol{\lambda}_0)^{\gamma-1} + (1-\mathbf{p}(\boldsymbol{\lambda}_0))^{\gamma-1})}\right) \text{ and } \boldsymbol{\Psi} = E((Y - \mathbf{p}(\boldsymbol{\lambda}_0))^2\mathbb{Z}'\mathbb{Z}). \tag{4.1}$$

Theorem 3: Given assumptions 1-3, $\gamma \geq 1$, and $P(|\mathbb{Z}\boldsymbol{\beta}|=1)=0$, $\sqrt{n}(\hat{\boldsymbol{\lambda}}-\boldsymbol{\lambda}_0) \overset{a}{\sim} N(\mathbf{0},\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Psi}\boldsymbol{\Omega}_0^{-1})$, where $\boldsymbol{\Psi}$ is as defined in (4.1) and

$$\boldsymbol{\Omega}_0 = E\left(I(-1<\mathbb{Z}\boldsymbol{\beta}<1)\left(\frac{\mathbb{Z}'\mathbb{Z}}{2^{\gamma}(\mathbf{p}(\boldsymbol{\lambda}_0)^{\gamma-1}+(1-\mathbf{p}(\boldsymbol{\lambda}_0))^{\gamma-1})}\right)\right) = \mu E\left(\frac{\mathbb{Z}'\mathbb{Z}}{2^{\gamma}(\mathbf{p}(\boldsymbol{\lambda}_0)^{\gamma-1}+(1-\mathbf{p}(\boldsymbol{\lambda}_0))^{\gamma-1})}\right). \quad (4.2)$$

When $\gamma \geq 1$ and $P(|\mathbb{Z}\boldsymbol{\beta}|=1)>0$ the situation is more complex. While $\sqrt{n}(\hat{\boldsymbol{\lambda}}-\boldsymbol{\lambda}_0)$ does have an asymptotic distribution, it need not be in the normal parametric family.

Given that $\hat{\boldsymbol{\Lambda}} \overset{p}{\rightarrow} \boldsymbol{\lambda}_0$, the estimated Lagrange multipliers can be interpreted as estimates of some underlying fixed Lagrange multiplier value as sample size increases without bound. Thus, at least *in a limiting sense*, the Lagrange multipliers in the CR-method can be interpreted analogously to the parameters in parametric models, and the CR approach can thus be viewed as a semiparametric method *asymptotically*.

4.2 Implications for Inference

The preceding asymptotic results suggest a relatively straightforward method of constructing estimates for the asymptotic covariance matrices of $\hat{\boldsymbol{\lambda}}$. A consistent estimator of $\boldsymbol{\Psi}$ can be defined for $\gamma < 1$ by

$$\hat{\boldsymbol{\Psi}} = n^{-1}\sum_{i=1}^{n} \hat{\varepsilon}_i^2 \mathbf{Z}_{i.}'\mathbf{Z}_{i.} \text{ where } \hat{\varepsilon}_i = \mathrm{Y}_i - p_i(\hat{\boldsymbol{\lambda}}). \quad (4.3)$$

Define $\hat{\boldsymbol{\Omega}}_0$ when $\gamma < 1$ by

$$\hat{\boldsymbol{\Omega}}_0 = n^{-1}\sum_{i=1}^{n} \hat{\omega}_i \mathbf{Z}_{i.}'\mathbf{Z}_{i.} \text{ where } \hat{\omega}_i = \left(2^{\gamma}\left(p_i(\hat{\boldsymbol{\lambda}})^{\gamma-1}+\left(1-p_i(\hat{\boldsymbol{\lambda}})\right)^{\gamma-1}\right)\right)^{-1}. \quad (4.4)$$

Then a consistent estimator of the asymptotic covariance matrix of $\sqrt{n}(\hat{\boldsymbol{\Lambda}}-\boldsymbol{\lambda}_0)$ is defined by $\widehat{Cov}\left(\sqrt{n}(\hat{\boldsymbol{\Lambda}}-\boldsymbol{\lambda}_0)\right)=\hat{\boldsymbol{\Omega}}_0^{-1}\hat{\boldsymbol{\Psi}}\hat{\boldsymbol{\Omega}}_0^{-1}$. For the case where $\gamma \geq 1$, the preceding formulae still apply, but the summation is then taken over only those observations for which $p_i(\hat{\boldsymbol{\lambda}}) \in (0,1)$.

Confidence intervals and hypothesis tests relating to the values of $\boldsymbol{\lambda}_0$ can be conducted based on the usual t-type statistics and Wald asymptotic chi-square methods. The significance of individual response variables can be evaluated by testing the associated value of the Lagrange multiplier of the moment constraint constructed from the response variable. Confidence intervals and hypothesis tests on marginal effects can be obtained based on the asymptotic distribution derived from the functional mapping of $\hat{\boldsymbol{\lambda}}$ to the value of the derivative $\left((\partial \hat{p}_i)/(\partial z_{ij})\right)$ of interest, or based on bootstrapping.

5. SAMPLING EXPERIMENTS

This sampling experiment section is designed with several purposes in mind. To avoid the sometimes artificial applied example in a theoretical paper, we have fashioned the sample design to reflect the elements of a discrete choice problem often found in practice. The results of the sampling experiments illustrate the finite sample properties of various $\mathrm{CR}(\gamma)$ in comparison to the traditional probit and logit estimators. Finally, carrying through the experiments suggests how an application can be implemented and the results interpreted.

We implement the following version of the binary response model, where the latent variable equation and binary outcomes are given by:

$$Y_i = I\left(Y_i^* > 0\right) \text{ and } Y_i^* = \beta_0 + 1\mathrm{x}_{i1} - 1\,\mathrm{x}_{i2} + 2\mathrm{x}_{i3} - 2\,\mathrm{x}_{i4} + \varepsilon_i^* \quad \text{for i = 1,...,}n, \tag{5.1}$$

the x_{ij}'s are iid outcomes from a Uniform(0,2) distribution, and the value of the intercept β_0 is increased incrementally, beginning at zero, to simulate various probabilities of the Bernoulli probability $p_i = P(y_i = 1)$, $\forall i$. Specifications of the binary response model such as (5.1) are typical of a wide range of empirical analyses in the literature, where the latent variable is modeled as a linear index in response variables that affect the probabilities that either of the dichotomous outcomes is chosen. The distributions allow for substantially different noise outcome characteristics that include a symmetric bell-shaped sampling distribution ($N(0,9)$), a highly right-skewed sampling distribution (Exponential(4)), and a very fat-tailed, U-shaped symmetric distribution (Beta(.5,.5)) with compact support, where the latter two distributions are centered to have zero means.

For each of the distribution, and intercept value combinations, we estimated the logit, probit, and four alternative $\mathrm{CR}(\gamma)$ estimators associated with $\gamma = \{-1, 0, 1, 1.5\}$. Note the scenario involving the normal noise distribution is the case where probit is actually the correctly specified maximum likelihood estimator. For the Monte Carlo results reported here, the sample size was set to n = 300, and all Monte Carlo calculations are based on 5,000 sample repetitions. At this number of repetitions, all of the empirically calculated averages representing estimates of the associated MSEs are very accurately determined, where standard errors of the mean calculations are typically < .0001 in magnitude. Sampling results for selected scenarios are displayed by figures in Appendix B and discussed in narrative form in Sections 5.2 - 5.4. Results for CR(1.5) are suppressed in both the discussion and graphs because they were virtually indistinguishable from the CR(1) case. The CR(0) and ML logit results are numerically identical under all scenarios, and thus the results for this estimator are reported together in the discussion

and figures. Graphs of other scenarios (n = 75, 150) and tables reporting the data underlying the figures are available electronically from the authors.

5.1 Estimation Performance Measures

In terms of sampling performance we concentrate on two measures. The first is the empirical mean square error (MSE) $MSE(\hat{\mathbf{p}}) = n^{-1}\sum_{i=1}^{n}(p_i - \hat{p}_i)^2$ in predicting the binary response probabilities across sample observations. We also calculate the empirical MSE associated with predicting the values of the marginal effects of changes in response variable values on the binary response probabilities, which were defined across all $k = 4$ explanatory variables and on all sample observations as $MSE(\partial\hat{\mathbf{p}}/\partial\mathbf{x}) = (kn)^{-1}\sum_{j=1}^{k}\sum_{i=1}^{n}(\partial p_i/\partial x_{ij} - \partial\hat{p}_i/\partial x_{ij})^2$.

5.2 Probability Prediction Performance

The empirical results for probability prediction performance are presented graphically in Figure 1 of Appendix B. Regarding relative performance in predicting $P(y=1)$ across sample scenarios, it is evident that in the *non-normal* sampling situations, the CR(1) estimator is superior in MSE to all of the remaining estimators, especially for higher values of $P(y=1)$. The probit is next in terms of empirical risk performances, followed closely by the logit-CR(0) measure. This illustrates the negative performance implications of assuming an incorrect underlying data sampling process in the parametric model approach. The CR(-1) measure performs worse than its competitors, and when the average value of $P(y=1)$ is relatively high, the predictive risk performance of the CR(-1) measure degrades substantially.

For the case of the *normal* sampling distribution, CR(-1) is notably superior in predictive MSE except for the highest average values of $P(y=1)$. The logit-CR(0) and probit estimators are

hardly distinguishable, and are risk-wise the next best method of predicting probabilities. The CR(1) estimator performs least well under the normal sampling distribution. However, the differences between all of the estimation methods becomes quite small as average $P(y=1)$ increases.

5.3 Estimation of Probability Derivatives

The empirical MSE for estimation of the vector of probability derivatives with respect to the vector **Z** is displayed graphically in Figure 2. For the *non-normal* sampling cases, CR(1) is the best estimator of probability derivatives in terms of MSE. Sampling performances are very similar for the Beta distribution case, with CR(1) having a modest advantage in the Exponential distribution case. The next closest competitor is probit, followed closely by logit-CR(0). The relative advantage of CR(1) is small in the Beta case, except for high average $P(y=1)$. The advantage of CR(1) is more substantial in the Exponential case, for average $P(y=1)$. This suggests that the relative MSE advantage may be affected by the degree of skewness in the underlying noise distribution.

In the case of sampling from a *normal* distribution, the probit and logit-CR(0) estimators are superior to the CR(1) estimator. However, the CR(-1) estimator is superior to all of the estimators for the smaller average values of $P(y=1)$, but its advantage dissipates for high average $P(y=1)$.

5.4 Hypothesis Testing

The inference performance associated with the probit, logit, and alternative CR approaches to estimating the binary response model was illustrated by analyzing their sampling performance in

testing various hypotheses relating to the parameters or Lagrange multipliers of the model. Specifically, the hypothesis $H_o : \boldsymbol{\iota}'\boldsymbol{\lambda} = \mathrm{r}$ can be tested based on a t-ratio and the asymptotic covariance results identified in Section 4, where $\boldsymbol{\lambda}$ is replaced by the parameter vector $\boldsymbol{\beta}$ in the case of the probit and logit formulations and $\boldsymbol{\iota}$ is a conformable vector defining linear combinations of the Lagrange multipliers to be tested. The specific form of the t-ratio is given by

$$\mathrm{t} = \frac{n^{1/2}\left(\boldsymbol{\iota}'\hat{\boldsymbol{\lambda}} - \mathrm{r}\right)}{\left(\boldsymbol{\iota}'\hat{\boldsymbol{\Omega}}_0^{-1}\hat{\boldsymbol{\Psi}}\hat{\boldsymbol{\Omega}}_0^{-1}\boldsymbol{\iota}\right)^{1/2}}. \tag{5.2}$$

The particular results presented here concerns testing the value of the "intercept term", which is the Lagrange multiplier, say λ_1, associated with the unit vector in the **Z** matrix. The empirical power functions of these tests were simulated based on setting $r = E\left(\hat{\Lambda}_1\right)$ to assess the size of the test. A grid of r values up to $\pm$ 3.5 standard deviations above and below $E\left(\hat{\Lambda}_1\right)$, in half-standard deviations units, was used to assess the power of the tests. The true mean of Λ_1 was calculated as the sample moment of the entire set of $\hat{\lambda}_1$ outcomes that were observed for the particular sampling scenario being analyzed.

Results are displayed for the three different sampling distributions, and for the lowest and highest average values of $P\left(y = 1\right)$, in Figure 3. The critical values of the test were based on the asymptotically valid standard normal distribution, and the target size of the test was set equal to .05. All of the methods behaved similarly at all levels of $P\left(y = 1\right)$, although the CR(-1) estimator still exhibited inferior behavior in terms of size and power in comparison to the others. The CR(-1) estimator's size is the least accurate of all of the methods and the test exhibits

notable bias for some of the sampling scenarios. The power function characteristics are particularly poor for high average values of $P(y=1)$.

Comparing the power functions of the remaining methods, the power characteristics are all quite similar. In particular, the sizes of the test are all close to the target level of .05 and the power, for deviations away from the true null hypothesis, are also all quite comparable. In the case of sampling from the normal distribution, the probit method appears to have a modest advantage in terms of size, lack of bias, and overall power, although the CR(1) method has power characteristics that are quite competitive with the probit approach. The CR(0)-logit approach appears to be modestly inferior in performance in comparison to CR(1) and probit, especially for high $P(y=1)$.

Overall, the CR(1) method appears to offer good power function performance when the parametric model (probit) is correct, and affords a degree of robustness and performs somewhat better than the parametric specifications when the parametric model specification is incorrect.

6. SUMMARY AND CONCLUSIONS

In this paper, we represent sample information about binary response outcomes nonparametrically via general moment conditions, $\mathrm{E}\left[\mathbf{Z}'(\mathbf{Y}-\mathbf{p})\right]=\mathbf{0}$. We then use the Cressie-Read (CR) family of divergence measures, CR(γ) for $\gamma\in(-\infty,\infty)$, to solve for the unknown Bernoulli probabilities, **p**, and the corresponding marginal effects on **p** of changes in the response variable values, which are underdetermined by the moment conditions alone. The solved values of the probabilities, as well as the derivatives of the probabilities with respect to the response variables, are functions of the sample data through data-determined Lagrange multipliers and are not dependent on a fixed set of parameters. The properties of consistency and

asymptotic normality for the family of CR(γ) estimators were analytically demonstrated. The nonparametric CR(γ) estimators were shown via Monte Carlo simulations to have finite sample performance comparable to their parametric counterparts. For each of the MC scenarios examined, there was always at least one member of the CR class of estimators that was MSE-superior to the parametric estimation methods.

In parametric estimation there are usually two competing goals–efficiency, when the DSP is correctly specified, and robustness when it is not. Members of the CR-family of estimators performed well in comparison to the usual parametric estimators of the binary response model relative to both estimation goals.

A number of important issues relating to the CR approach to estimation and inference remain. These include generalizing the uniform reference distribution, $\boldsymbol{q}$, to take account, in any particular applied problem, of known or estimable characteristics of the Bernoulli probabilities; developing a minimum divergence median-based estimator and contrasting it to Manski's maximum score estimator (Manski 1975 and Horowitz 1992); generalizing the model to include endogenous response variables; pursuing other moment-formulations as a basis for representing the sample information; and developing a basis under model uncertainty for combining estimators of the type examined in this paper to obtain a minimum loss-based estimator (Judge and Mittelhammer 2004; Qin 2000).

Finally, we note that Grunwald and David (2004) have established a close relationship between the maximum entropy principle (CR(0)) and the problem of minimizing worst case expected loss. In the important case where the class of distributions is described by mean value constraints, the challenge is to extend their results to our CR context of **p** and **q** distributions and the distance measures of relative entropy and divergence.

APPENDIX A. MOTIVATION AND PROOFS OF ASYMPTOTIC RESULTS

A.1 Preliminaries

Given data $(\mathbf{y},\mathbf{z})$ where $\mathbf{y}$ is an n-vector, and $\mathbf{z}$ is an n by k matrix, $\mathbf{z}_{i\bullet}$ denotes the ith row of $\mathbf{z}$ and $\mathbf{z}_{\bullet k}$ denotes the kth column of $\mathbf{z}$, the minimum discrepancy estimator can be defined as:

$$\hat{\mathbf{p}} = \begin{cases} \underset{\mathbf{p}:\mathbf{z}'(\mathbf{y}-\mathbf{p})=0;0\le p_i\le 1,all\ i}{\operatorname{argmin}} \left(\dfrac{1}{\gamma(\gamma+1)} \left(2^{\gamma} \sum_{i=1}^{n} \left(p_i^{\gamma+1} + (1-p_i)^{\gamma+1} \right) - n \right) \right), & \gamma \neq 0,-1 \\ \underset{\mathbf{p}:\mathbf{z}'(\mathbf{y}-\mathbf{p})=0;0\le p_i\le 1,all\ i}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left((p_i \ln(p_i) + (1-p_i)\ln(1-p_i)) + \ln(2) \right) \right), & \gamma = 0 \\ \underset{\mathbf{p}:\mathbf{z}'(\mathbf{y}-\mathbf{p})=0;0\le p_i\le 1,all\ i}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left(\dfrac{-1}{2}(\ln(p_i) + \ln(1-p_i)) - \ln(2) \right) \right), & \gamma = -1 \end{cases} \tag{A.1}$$

The case in (A.1) for $\gamma \neq 0,-1$ may be restated as

$$\hat{\mathbf{p}} = \underset{\mathbf{p}:\mathbf{z}'(\mathbf{y}-\mathbf{p})=0;0\le p_i\le 1,all\ i}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left(\frac{2^{\gamma}\left(p_i^{\gamma+1} + (1-p_i)^{\gamma+1} \right) - 1}{\gamma(\gamma+1)} \right) \right), \quad \gamma \neq 0,-1. \tag{A.2}$$

Now define

$$Q(q,\gamma) = \begin{cases} \dfrac{2^{\gamma}\left(q^{\gamma+1} + (1-q)^{\gamma+1} \right) - 1}{\gamma(\gamma+1)} & \gamma \neq 0,-1 \\ q\ln(q) + (1-q)\ln(1-q) + \ln(2) & \gamma = 0 \\ \dfrac{-1}{2}\left(\ln(q) + \ln(1-q) \right) - \ln(2), & \gamma = -1 \end{cases} \tag{A.3}$$

and let

$$CR(\mathbf{p},\gamma) = \sum_{i=1}^{n} Q(p_i,\gamma) \tag{A.4}$$

where $\hat{\mathbf{p}} = \underset{\mathbf{p}:\mathbf{z}'(\mathbf{y}-\mathbf{p})=0;0\le p_i\le 1,all\ i}{\operatorname{argmin}} (CR(\mathbf{p},\gamma))$. Henceforth, we will sometimes use the abbreviation $CR(\mathbf{p})$ and $Q(p_i)$, or CR and Q, in place of $CR(\mathbf{p},\gamma)$ and $Q(p_i,\gamma)$. Differentiating $Q(p_i)$ yields:

$$\frac{dQ}{dp_i} = \begin{cases} \frac{2^\gamma}{\gamma}\left(p_i^\gamma - (1-p_i)^\gamma \right) & \gamma \neq 0 \\ \ln(p_i) - \ln(1-p_i) & \gamma = 0 \end{cases} \tag{A.5}$$

and

$$\frac{d^2Q}{dp_i^2} = 2^\gamma \left(p_i^{\gamma-1} + (1-p_i)^{\gamma-1} \right). \tag{A.6}$$

Therefore, $CR = \sum_{i=1}^{n} Q\left(p_i \right)$ is globally convex in $\mathbf{p}$ and any local minimum of CR must be the unique global minimum, which applies to the constrained CR as well.

The Lagrangian for the problem defined in (A.1) is:

$$L(\mathbf{p}, \boldsymbol{\lambda}) = CR(\mathbf{p}, \gamma) + (\mathbf{z}\boldsymbol{\lambda})'(\mathbf{y} - \mathbf{p}) = \sum_{i=1}^{n} \left(Q(p_i, \gamma) + \mathbf{z}_{i\bullet}\boldsymbol{\lambda}(y_i - p_i) \right). \tag{A.7}$$

Solving the Lagrangian problem implies:

$$\frac{dQ}{dp_i} = \mathbf{z}_{i\bullet}\boldsymbol{\lambda}. \tag{A.8}$$

For $0 \leq p_i \leq 1$, dQ/dp_i is a strictly monotone function of the $p_i's$. If we define $\boldsymbol{\beta}$ by $\boldsymbol{\beta} = \left(|\gamma| / 2^\gamma \right)\boldsymbol{\lambda}$ for $\gamma \neq 0$, $\boldsymbol{\beta} = \boldsymbol{\lambda}$ for $\gamma = 0$, and $\Xi(q)$ by:

$$\Xi(q) = \begin{cases} \text{sign}(\gamma)\left(q^\gamma - (1-q)^\gamma \right) & \gamma \neq 0 \\ \ln(q) - \ln(1-q) & \gamma = 0 \end{cases} \tag{A.9}$$

this leads to the result:

$$\Xi(p_i) = \mathbf{z}_{i\bullet}\hat{\mathbf{b}} \tag{A.10}$$

Note that for n observations, k variables and $\mathbf{z}'\mathbf{z}$ of full rank, $\mathbf{z}'(\mathbf{y} - \mathbf{p}) = 0$ defines an $n-k$ dimensional subspace. For $\gamma \leq 0$, dCR/dp_i grows without bound as p_i approaches 0 or 1. Therefore, for $\gamma \leq 0$, if there is any interior solution to the constraint equation, the solution to

the complete problem must be an interior solution to the constraint equation. If $\mathbf{p}$ and $\mathbf{q}$ are two solutions to the constraint equation and $\mathbf{p}$ contains one or more zeros or ones while $\mathbf{q}$ contains none, then $\exists \delta, 0 < \delta \leq 1 : CR(\delta\mathbf{q} + (1-\delta)\mathbf{p}, \gamma) < CR(\mathbf{p}, \gamma)$. For $\gamma \leq 0$ the solution can be simply characterized because dQ / dp_i tends to $\pm\infty$ at the boundaries. Thus, for any finite v and any $\gamma \leq 0$, the equation $\Xi(q) = v$ has a unique interior solution. Let $p(v;\gamma) = p(v) \equiv \Xi^{-1}(v)$ denote that solution. For $\gamma > 0$, $\Xi(0) = -1$, $\Xi(1) = 1$ and thus when $|v| > 1$, the boundary constraint $0 \leq p_i \leq 1$ is binding. Therefore, for $\gamma > 0$:

$$p(v) = p(v,\gamma) = \begin{cases} 1 & v > 1 \\ \Xi^{-1}(v) & -1 \leq v \leq 1 \\ 0 & v < -1 \end{cases} \tag{A.11}$$

Note that $\Xi(q)$ is well defined and monotone for $0 \leq q \leq 1$ thus $\Xi^{-1}(v)$ is well defined for $-1 \leq v \leq 1$. In order to deal with the indeterminacy caused by the fact that, for $\gamma > 0$, $p(v) = 1$, for all $v \geq 1$ and $p(v) = 0$, for all $v \leq -1$ define $M(\mathbf{z}, \mathbf{b}, n, \gamma)$ by:

$$M(\mathbf{z}, \mathbf{b}, n, \gamma) = \begin{cases} \dfrac{1}{n} \displaystyle\sum_{\substack{i=1 \\ i:-1<\mathbf{z}_i\mathbf{b}<1}}^{n} \mathbf{z}_{i\bullet}'\mathbf{z}_{i\bullet} & \gamma > 0 \\ \dfrac{1}{n}\mathbf{z}'\mathbf{z} & \gamma \leq 0 \end{cases} \tag{A.12}$$

For $\mathbf{v}$ a vector of length n let $\mathbf{p}(\mathbf{v}) = \left(\mathbf{p}(v_1), \ldots, \mathbf{p}(v_n)\right)$. Thus the problem can be restated as:

$$\hat{\mathbf{b}} = \underset{\mathbf{b}:\mathbf{z}'(\mathbf{y}-\mathbf{p}(\mathbf{z}\mathbf{b}))=0}{\operatorname{argmin}} \left(CR(\mathbf{p}, \gamma)\right). \tag{A.13}$$

If there is any finite $\mathbf{b}$ that solves $\mathbf{z}'(\mathbf{y} - \mathbf{p}(\mathbf{z}\mathbf{b})) = 0$ and for which $M(\mathbf{z}, \mathbf{b}, n, \gamma)$ is nonsingular, then

$$\hat{\mathbf{b}} = \underset{\mathbf{b}}{\arg} \left(\mathbf{z}'(\mathbf{y} - \mathbf{p}(\mathbf{z}\mathbf{b})) = 0\right), \tag{A.14}$$

and $\mathbf{p}(\mathbf{z}\hat{\mathbf{b}})$ is the unique interior solution to (A.1). Note that it is impossible for there to be a single unique interior solution to the constraint equation, $\mathbf{z}'(\mathbf{y}-\mathbf{p})=0$. For example, if $\mathbf{p}$ is an interior solution, then so is $\alpha\mathbf{p}+(1-\alpha)\mathbf{y}$ for all α, $0<\alpha<1$. Thus it is the restriction to the single index function $\mathbf{p}(\mathbf{zb})$ that makes the solution unique.

Assume that the observations of $(y_i,\mathbf{z}_{i\bullet})$ are iid random realizations of random variables $(Y,\mathbb{Z})$ with some unknown distribution. Let

$$\boldsymbol{\beta}=\boldsymbol{\beta}(\gamma)=\underset{\mathbf{b}}{\arg}\,(E(\mathbb{Z}'(Y-\mathrm{p}(\mathbb{Z}\mathbf{b};\gamma))=0) \tag{A.15}$$

Define $\varphi(\mathbb{Z})=\mathbf{p}(\mathbb{Z}\boldsymbol{\beta})$ and $\varphi_i=\varphi(\mathbf{z}_{i\bullet})=\mathbf{p}(\mathbf{z}_{i\bullet}\boldsymbol{\beta})$. Let $\rho(\mathbb{Z})=E(Y\mid\mathbb{Z})$ and $\rho_i=\rho(\mathbf{z}_{i\bullet})=E(Y\mid\mathbb{Z}=\mathbf{z}_{i\bullet})$. Consequently, $\rho(\mathbb{Z})$ is the true probability and φ is an approximation to the true probabilities based on the metric $CR(\mathbf{p},\gamma)$ and the measure $F(\rho)$. Note that, in general, $\rho_i\neq\varphi_i$. The $\boldsymbol{\beta}$ vector is the analog to the true coefficient vector in parametric statistics.

In order to structure the problem for solution and to facilitate the exposition, define:

$$\Delta(\mathbf{b})=\frac{1}{n}\sum_{i=1}^{n}(p(\mathbf{z}_{i\bullet}'\mathbf{b})-p(\mathbf{z}_{i\bullet}'\boldsymbol{\beta}))\mathbf{z}_{i\bullet}'=\frac{1}{n}\sum_{i=1}^{n}(p(\mathbf{z}_{i\bullet}'\mathbf{b})-\varphi(\mathbf{z}_{i\bullet}'))\mathbf{z}_{i\bullet}'=\frac{1}{n}\sum_{i=1}^{n}(p(\mathbf{z}_{i\bullet}'\mathbf{b})-\varphi_i)\mathbf{z}_{i\bullet}', \tag{A.16}$$

and

$$\mathbf{u}=\frac{1}{n}\sum_{i=1}^{n}(y_i-p(\mathbf{z}_{i\bullet}'\boldsymbol{\beta}))\mathbf{z}_{i\bullet}'=\frac{1}{n}\sum_{i=1}^{n}(y_i-\varphi(\mathbf{z}_{i\bullet}'))\mathbf{z}_{i\bullet}'=\frac{1}{n}\sum_{i=1}^{n}(y_i-\varphi_i)\mathbf{z}_{i\bullet}'. \tag{A.17}$$

Redefining equation (A.13), the complete solution to the original problem is

$$\hat{\mathbf{b}}=\underset{\mathbf{b}}{\arg}\,(\Delta(\mathbf{b})=\mathbf{u}),\quad \hat{\mathbf{p}}=\mathbf{p}(\mathbf{z}\hat{\mathbf{b}}). \tag{A.18}$$

A.2 Proofs of Asymptotic Results

The following lemma will be used in the derivation of the asymptotic distributions of the semiparametric estimators.

Lemma 1: Given assumptions 1 and 2, $\sqrt{n}\mathbf{U} \overset{a}{\sim} N(\mathbf{0}, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi} = E((Y - p(\mathbb{Z}\boldsymbol{\beta}))^2 \mathbb{Z}'\mathbb{Z})$.

Proof of Lemma 1: Let $\mathbf{w}_i = (y_i - \varphi_i)\mathbf{z}'_{i\bullet}$ and $\mathbf{W} = (Y - \varphi(\mathbb{Z}))\mathbb{Z}'$. Given assumptions 1 and 2 and the definition of φ_i, $\mathbf{w}_i$ is an iid realization of the random vector $\mathbf{W}$ which has a mean of zero and finite variance covariance matrix, $\text{var}(\mathbf{W}) = \boldsymbol{\Psi}$. Thus

$$\sqrt{n}\mathbf{U} = \sqrt{\mathbf{n}}\mathbf{W} \underset{\sim}{\mathbf{a}} \mathbf{N}(\mathbf{0}, \boldsymbol{\Psi}). \tag{A.19}$$

Proof of Theorem 1:

If one were to make the highly parametric assumption that $\exists \mathbf{b} : \rho(\mathbb{Z}) \equiv p(\mathbb{Z}\mathbf{b})$, then it would be trivial to show that $\boldsymbol{\beta} = \underset{\mathbf{b}}{\arg}\left(\rho(\mathbb{Z}) = p(\mathbb{Z}\mathbf{b})\right)$. Thus $\varphi(\mathbb{Z})$ would be the same as $\rho(\mathbb{Z})$ and the rest of the demonstration would be straightforward. Instead the framework here is nonparametric and, therefore, the best one can hope for is convergence to $\varphi(\mathbb{Z})$. The function $\varphi(\mathbb{Z})$ can be thought of as a best approximation function under the implicit metric defined by the CR function and the distribution of $(Y, \mathbb{Z})$. Note that for $\gamma > 0$ it is possible that $\varphi(\mathbb{Z})$ may take the value of 0 or 1 even if $\rho(\mathbb{Z})$ never does. Thus, in general, $P(Y = 1 - \varphi(\mathbb{Z})) > 0$ and in the analysis data set it may happen that $y_i = 1 - \hat{p}(\mathbf{z}_{i\bullet})$.

For $\gamma = 0$ the model reduces to logit and the result is well known, so we focus on $\gamma \neq 0$. Recall that $\hat{\boldsymbol{\beta}}$ solves:

$$\Delta(\mathbf{b}) = \frac{1}{n}\sum_{i=1}^{n}(p(\mathbf{z}_{i\bullet}\mathbf{b}) - p(\mathbf{z}_{i\bullet}\boldsymbol{\beta}))\mathbf{z}'_{i\bullet} = \frac{1}{n}\sum_{i=1}^{n}(y_i - p(\mathbf{z}_{i\bullet}\boldsymbol{\beta}))\mathbf{z}'_{i\bullet} = \mathbf{u}. \tag{A.20}$$

Note that $p(\mathbf{z}_{i\bullet}\boldsymbol{\beta}) \equiv \varphi(\mathbf{z}_{i\bullet})$, and $p(\mathbf{z}_{i\bullet}\boldsymbol{\beta})$ is used here to emphasize dependence on $\boldsymbol{\beta}$. Also,

$E(\mathbb{Z}'(Y-\varphi(\mathbb{Z}))) = E(\mathbb{Z}'(Y-\rho(\mathbb{Z}))) = \mathbf{0}$ and $E(\mathbb{Z}'(Y-\rho(\mathbb{Z}))\,|\,\mathbb{Z}) = \mathbf{0}$; but $E(\mathbb{Z}'(Y-\varphi(\mathbb{Z}))\,|\,\mathbb{Z}) \neq 0$.

Differentiating (A.11) yields $|\gamma|(p(v)^{\gamma-1} + (1-p(v))^{\gamma-1})dp = dv$. Thus,

$$\frac{dp(v)}{dv} = \frac{1}{|\gamma|(p(v)^{\gamma-1} + (1-p(v))^{\gamma-1})}, \tag{A.21}$$

and

$$\frac{\partial \Delta(\mathbf{b})}{\partial \mathbf{b}} = \sum_{i=1}^{n} \frac{\mathbf{z}_{i\bullet}'\mathbf{z}_{i\bullet}}{|\gamma|(p(\mathbf{z}_{i\bullet}\mathbf{b})^{\gamma-1} + (1-p(\mathbf{z}_{i\bullet}\mathbf{b}))^{\gamma-1})} = H(\mathbf{b}). \tag{A.22}$$

Note that, in general, for a vector $\mathbf{q}$, $|\mathbf{q}| = \max_{\mathbf{v}:\mathbf{v}'\mathbf{v}=1}(\mathbf{v}'\mathbf{q})$. Now consider $\mathbf{v}'\Delta(\boldsymbol{\beta} + a\mathbf{v})$. For any $\mathbf{v}$, $\mathbf{v}'\Delta(\hat{\boldsymbol{\beta}}) = \mathbf{v}'\mathbf{u}$. Equating $\boldsymbol{\beta} + a\mathbf{v}$ and $\hat{\boldsymbol{\beta}}$ yields, $a = |\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|$ and $\mathbf{v} = \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) / \left(\left|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right|\right)$. Then

$$\frac{d\mathbf{v}'\Delta(\boldsymbol{\beta} + a\mathbf{v})}{da} = \sum_{i=1}^{n} \frac{\mathbf{v}'\mathbf{z}_{i\bullet}'\mathbf{z}_{i\bullet}\mathbf{v}}{|\gamma|(p(\mathbf{z}_{i\bullet}(\boldsymbol{\beta} + a\mathbf{v}))^{\gamma-1} + (1-p(\mathbf{z}_{i\bullet}(\boldsymbol{\beta} + a\mathbf{v})))^{\gamma-1})}. \tag{A.23}$$

Let $\xi_L(\mathbf{M})$ denote the smallest eigenvalue of a square matrix $\mathbf{M}$. Define

$$k(\delta) = \min_{\mathbf{b}:|\mathbf{b}-\beta|<\delta} \left(\xi_L \left(E \left(\frac{\mathbb{Z}'\mathbb{Z}}{|\gamma|(p(\mathbb{Z}\mathbf{b})^{\gamma-1} + (1-p(\mathbb{Z}\mathbf{b}))^{\gamma-1})} \right) \right) \right) \tag{A.24}$$

Let k be a scalar for which $E(\mathbb{Z}'\mathbb{Z}\,|\,|\mathbb{Z}| < k)$ is positive definite and $P(|\mathbb{Z}| < k) > 0$. Given assumptions 1 and 2 it is clear that such a k exists and it follows that:

$$k(\delta) \geq \min_{\substack{\mathbf{v}:|\mathbf{v}|\leq k \\ \mathbf{b}:|\mathbf{b}-\beta|\leq\delta}} \left(\frac{1}{|\gamma|(p(\mathbf{v}\mathbf{b})^{\gamma-1} + (1-p(\mathbf{v}\mathbf{b}))^{\gamma-1})} \right) P(|\mathbb{Z}| < k)\, \xi_L\left(E\left(\mathbb{Z}'\mathbb{Z}\,||\,\mathbb{Z}| < k\right)\right) > 0 \tag{A.25}$$

At this point we need to consider probability statements that relate to the full data matrix. Therefore, let $\mathbf{Y}$ and $\mathbf{Z}$ represent respectively the random vector and the random matrix of which $\mathbf{y}$ and $\mathbf{z}$ are realizations. Thus $(\mathbf{Y}_i, \mathbf{Z}_{i\bullet})$ is an iid random vector with the same

distribution as $(Y,\mathbb{Z})$. It follows from equation (B.25) that, for all $\varepsilon>0$, $\delta>0$, $\exists n_1(\varepsilon,\delta)$ such that for all $n>n_1(\varepsilon,\delta)$:

$$P\left(\min_{\mathbf{b}:|\mathbf{b}-\beta|\le\delta}\left(\xi_L\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbf{Z}_{i\bullet}'\mathbf{Z}_{i\bullet}}{|\gamma|(p(\mathbf{Z}_{i\bullet}\mathbf{b})^{\gamma-1}+(1-p(\mathbf{Z}_{i\bullet}\mathbf{b}))^{\gamma-1})}\right)\right)\le\frac{k(\delta)}{2}\right)<\varepsilon \tag{A.26}$$

Considering Δ as a function of $\mathbf{Z}$ rather that $\mathbf{z}$ and combining Equations (B.23) and (B.26) for all $n>n_1(\varepsilon,\delta)$ yields

$$P\left(\min_{\mathbf{b}:|\mathbf{b}-\beta|\ge\delta}\left(\left|\frac{1}{n}\Delta(\mathbf{b})\right|\right)<\frac{\delta k(\delta)}{2}\right)<\varepsilon. \tag{A.27}$$

Write $\mathbf{b}$ in the form $\mathbf{b}=\beta+a\mathbf{v}$ where a is a scalar and $\mathbf{v}$ a K-vector with $|\mathbf{v}|=1$.

Lemma 1 proves that $\sqrt{n}\mathbf{U}=n^{-1/2}\sum_{i=1}^{n}(Y_i-p(\mathbf{Z}_{i\bullet}\beta))\mathbf{Z}_{i\bullet}'$ is a mean zero asymptotically normal random variable. Thus, for any $\mathbf{v}$, $\mathbf{v}'\mathbf{U}$ is a mean zero asymptotically normal random variable. Therefore,

$$\exists n_2(\varepsilon,k):\forall(n>n_2(\varepsilon,k))\,P\left(\left|\frac{1}{n}\sum_{i=1}^{n}(y_i-p(\mathbf{Z}_{i\bullet}\beta))\mathbf{Z}_{i\bullet}'\right|>k\right)<\varepsilon. \tag{A.28}$$

Let $n(\varepsilon,\delta)=\max(n_1(\frac{\varepsilon}{2},\delta),n_2\left(\left(\varepsilon/2\right),\left(\left(k\left(\delta\right)\right)/3\right)\right)$. Thus, for all $n>n(\varepsilon,\delta)$,

$$P\left(\min_{\mathbf{b}:|\mathbf{b}-\beta|\ge\delta}\left(\left|\frac{1}{n}(\Delta(\mathbf{b})-\mathbf{U})\right|\right)>\frac{k(\delta)}{6}\right)>1-\varepsilon. \tag{A.29}$$

We have already shown that a feasible solution exists. Therefore, for all $n>n(\varepsilon,\delta)$:

$$P(|\hat{\beta}-\beta|<\delta)>1-\varepsilon \tag{A.30}$$

Thus $\text{plim}(\hat{\beta})=\beta$, from which $\text{plim}(\hat{\Lambda})=\lambda_0$ from the definition $\beta=\left(|\gamma|/2^{\gamma}\right)\lambda$.

Proof of Theorem 2:

Expand $\Delta(\mathbf{b})$ in a Taylor series around $\boldsymbol{\beta}$, where $\hat{\boldsymbol{\beta}} = \arg_{l}(\Delta(l) = \mathbf{U})$ is the minimum discrepancy estimator of $\boldsymbol{\beta}$, to obtain

$$\Delta(\hat{\boldsymbol{\beta}}) = \mathbf{0} + H(\boldsymbol{\beta}^*)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}), \tag{A.31}$$

where $\boldsymbol{\beta}^*$ is between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$. In general, different $\boldsymbol{\beta}^*$ points will be required to represent the different coordinate functions in $\Delta(\hat{\boldsymbol{\beta}})$. At the optimum, $\Delta(\hat{\boldsymbol{\beta}}) = \mathbf{U}$ and $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$; therefore $\boldsymbol{\beta}^* \xrightarrow{p} \boldsymbol{\beta}$, and

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{d}{=} \left(\frac{1}{n} H(\boldsymbol{\beta})\right)^{-1} \sqrt{n}\mathbf{U} \overset{d}{=} \boldsymbol{\Omega}^{-1}\sqrt{n}\mathbf{U}, \tag{A.32}$$

where $\overset{d}{=}$ denotes the equivalence of the limiting distributions. From Lemma 1, $\sqrt{n}\mathbf{U} \overset{d}{=} N(\mathbf{0}, \boldsymbol{\Psi})$, therefore, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{d}{=} N(\mathbf{0}, \boldsymbol{\Omega}^{-1}\boldsymbol{\Psi}\boldsymbol{\Omega}^{-1})$, from which $\sqrt{n}(\hat{\boldsymbol{\Lambda}} - \boldsymbol{\lambda}_0) \overset{a}{\sim} N(\mathbf{0}, \boldsymbol{\Omega}_0^{-1}\boldsymbol{\Psi}\boldsymbol{\Omega}_0^{-1})$ from the definition $\boldsymbol{\beta} = \left(|\gamma| / 2^{\gamma}\right)\boldsymbol{\lambda}$.

Note that $H(\mathbf{b})$ is a continuous function of $\mathbf{Z}$ for all $\mathbf{Z}$ and $\gamma < 1$. Furthermore, the concavity of the problem means that φ is still well defined. Thus Theorem 2 is valid for all $\gamma < 1$.

Proof of Theorem 3:

To proceed with $\gamma \geq 1$ we must consider the discontinuity in $H(\mathbf{b})$ that occurs at $\mathbf{z}_{i\bullet}\mathbf{b} = \pm 1$ for any i. Let $\aleph_c(\mathbf{b}, \delta, \mathbf{z}, n)$ denote the set of indices, i, $1 \leq i \leq n$ for which the interval $(\min_{\mathbf{v}:|\mathbf{v}-\mathbf{b}|<\delta}(\mathbf{z}_{i\bullet}\mathbf{v}), \max_{\mathbf{v}:|\mathbf{v}-\mathbf{b}|<\delta}(\mathbf{z}_{i\bullet}\mathbf{v}))$ does not contain ± 1; and $\aleph_d(\mathbf{b}, \delta, \mathbf{z}, n)$ denote the set of indices, i,

$1 \le i \le n$ for which the interval $(\min_{\mathbf{v}:|\mathbf{v}-\mathbf{b}|<\delta}(\mathbf{z}_{i\bullet}\mathbf{v}), \max_{\mathbf{v}:|\mathbf{v}-\mathbf{b}|<\delta}(\mathbf{z}_{i\bullet}\mathbf{v}))$ does contain ± 1. Now define Δ_c and Δ_d as:

$$\Delta_c(\mathbf{b}) = \frac{1}{n}\sum_{i\in\aleph_c(\beta,\delta,\mathbf{z},n)} (p(\mathbf{z}_{i\bullet}\mathbf{b}) - \varphi_i)\mathbf{z}'_{i\bullet}, \tag{A.33}$$

and

$$\Delta_d(\mathbf{b}) = \frac{1}{n}\sum_{i\in\aleph_d(\beta,\delta,\mathbf{z},n)} (p(\mathbf{z}_{i\bullet}\mathbf{b}) - \varphi_i)\mathbf{z}'_{i\bullet}. \tag{A.34}$$

Similarly define H_c:

$$H_c(\mathbf{b}) = \frac{\partial \Delta_c(\mathbf{b})}{\partial \mathbf{b}} = \sum_{i\in\aleph_c(\beta,\delta,\mathbf{z},n)} I(-1 < \mathbf{z}_{i\bullet}\boldsymbol{\beta} < 1)\frac{\mathbf{z}'_{i\bullet}\mathbf{z}_{i\bullet}}{|\gamma|(p(\mathbf{z}_{i\bullet}\mathbf{b})^{\gamma-1} + (1-p(\mathbf{z}_{i\bullet}\mathbf{b}))^{\gamma-1})}. \tag{A.35}$$

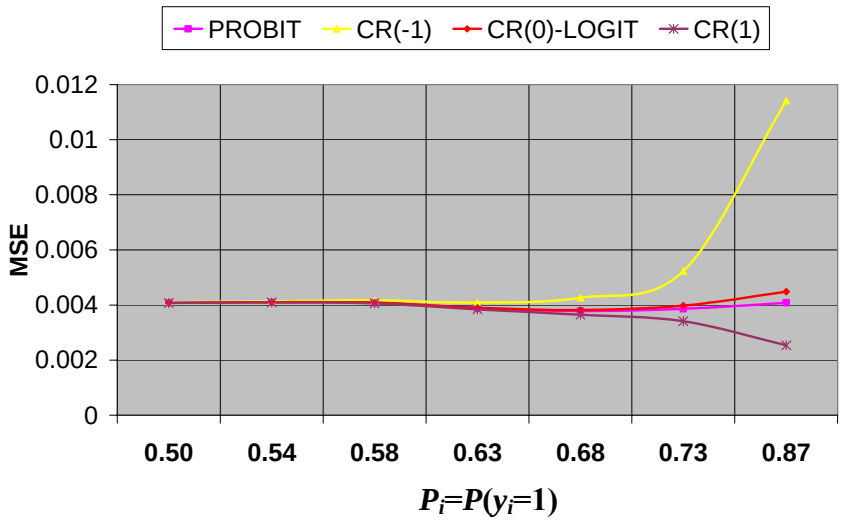
The assumption that $P(|\mathbb{Z}\boldsymbol{\beta}|=1)=0$ guarantees that $P(\pm 1 \in (\min_{\mathbf{v}:|\mathbf{v}-\mathbf{b}|<\delta}(\mathbb{Z}\mathbf{v}), \max_{\mathbf{v}:|\mathbf{v}-\mathbf{b}|<\delta}(\mathbb{Z}\mathbf{v})) \xrightarrow{p} 0$. Also note that, for $\gamma > 1$, the contribution of any specified observation i to Δ_c and Δ_d is $O(n^{-1}|\mathbf{z}_{i\bullet}|)$ and the contribution of any specified observation i to $n^{-1}H_c$ is $O(n^{-1}|\mathbf{z}_{i\bullet}|^2)$. Thus, given assumptions 1-3 and $P(|\mathbb{Z}\boldsymbol{\beta}|=1)=0$ as $\delta \to 0, \Delta_c \to \Delta, \Delta_d \to 0$, and $n^{-1}H_c \to \Omega$. Therefore:

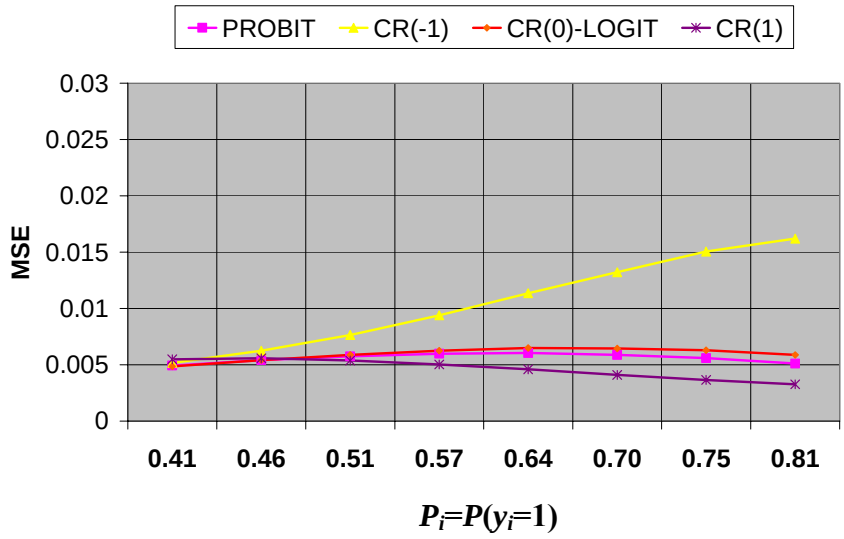
$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{d}{=} \left(\frac{1}{n}H_c(\boldsymbol{\beta})\right)^{-1}\sqrt{n}\left(\mathbf{U} - \Delta_d(\hat{\boldsymbol{\beta}})\right) \overset{d}{=} \left(\frac{1}{n}H(\boldsymbol{\beta})\right)^{-1}\sqrt{n}\mathbf{U} \overset{d}{=} \boldsymbol{\Omega}^{-1}\sqrt{n}\mathbf{U} \tag{A.36}$$

from which $\sqrt{n}(\hat{\boldsymbol{\Lambda}} - \boldsymbol{\lambda}_0) \overset{a}{\sim} N(\mathbf{0}, \boldsymbol{\Omega}_0^{-1}\boldsymbol{\Psi}\boldsymbol{\Omega}_0^{-1})$ from the definition $\boldsymbol{\beta} = (|\gamma|/2^{\gamma})\boldsymbol{\lambda}$.

APPENDIX B: FIGURES

Figure 1. MSE of Probability Predictions: Beta(.5,.5), Exp(4), N(0,9), n=300





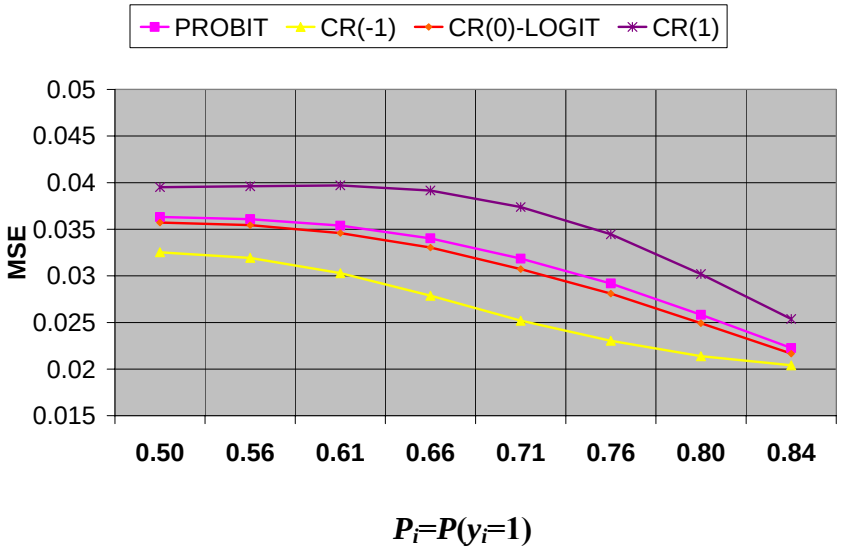
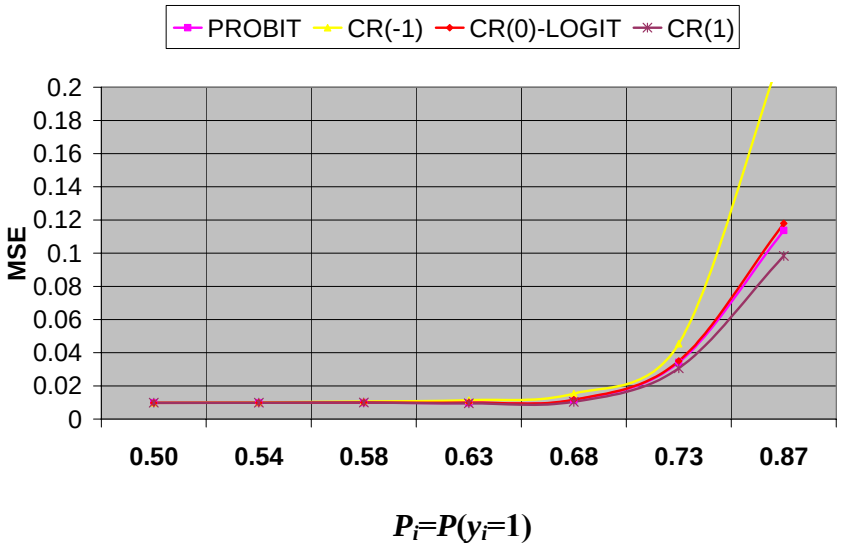
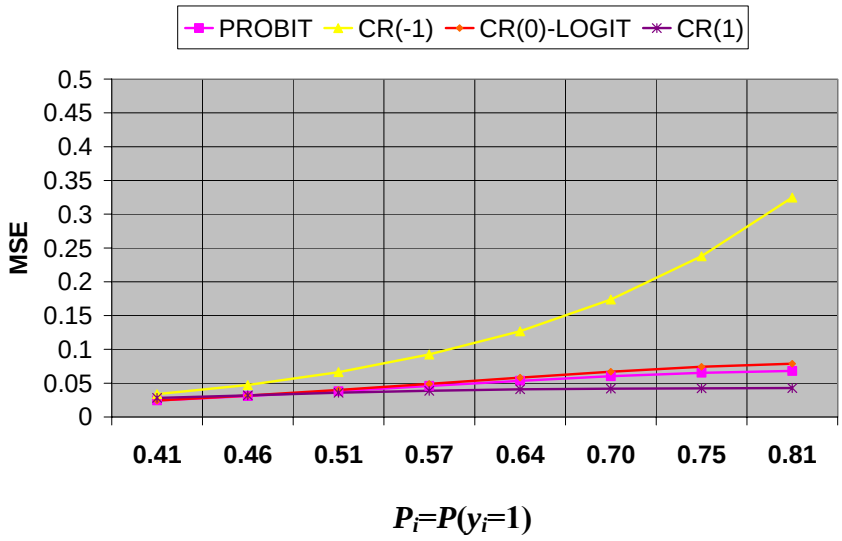


Figure 2. MSE of dP/dZ: Beta(.5,.5), Exp(4), N(0,9), n=300




PROBIT
CR(-1)
CR(0)-LOGIT
CR(1)
MSE
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0.41
0.46
0.51
0.57
0.64
0.70
0.75
0.81
$P_i=P(y_i=1)$

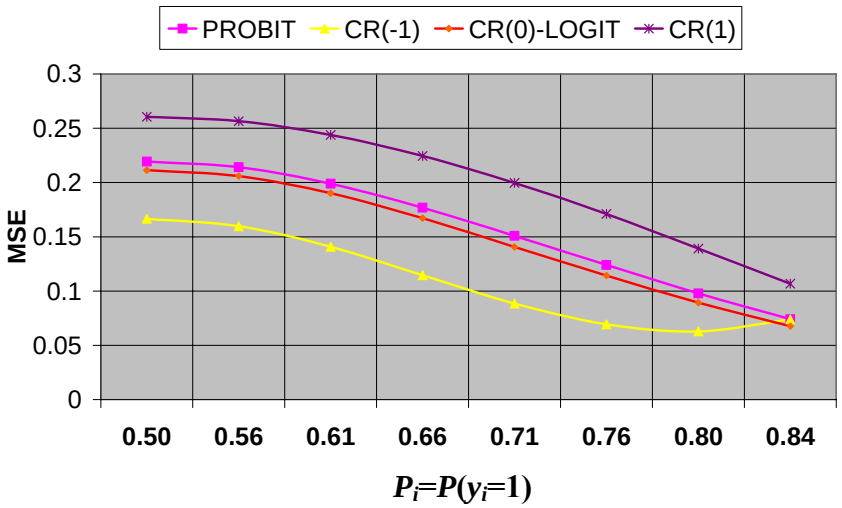
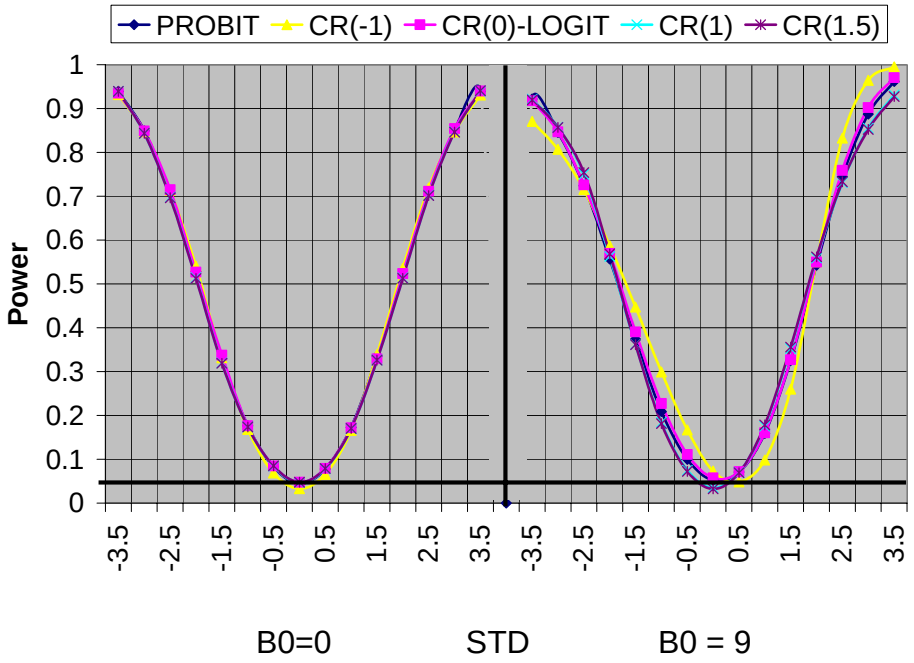
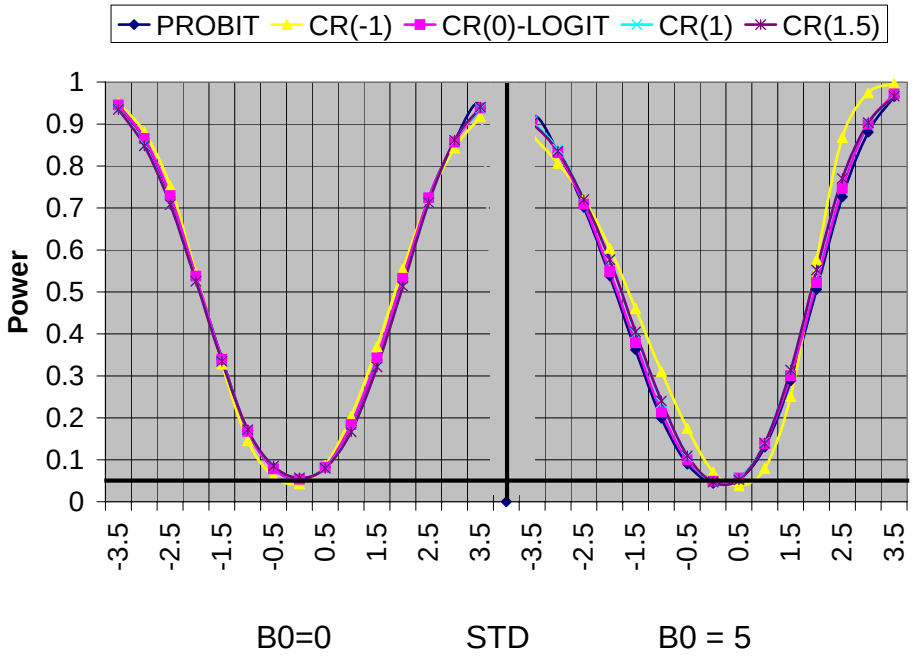
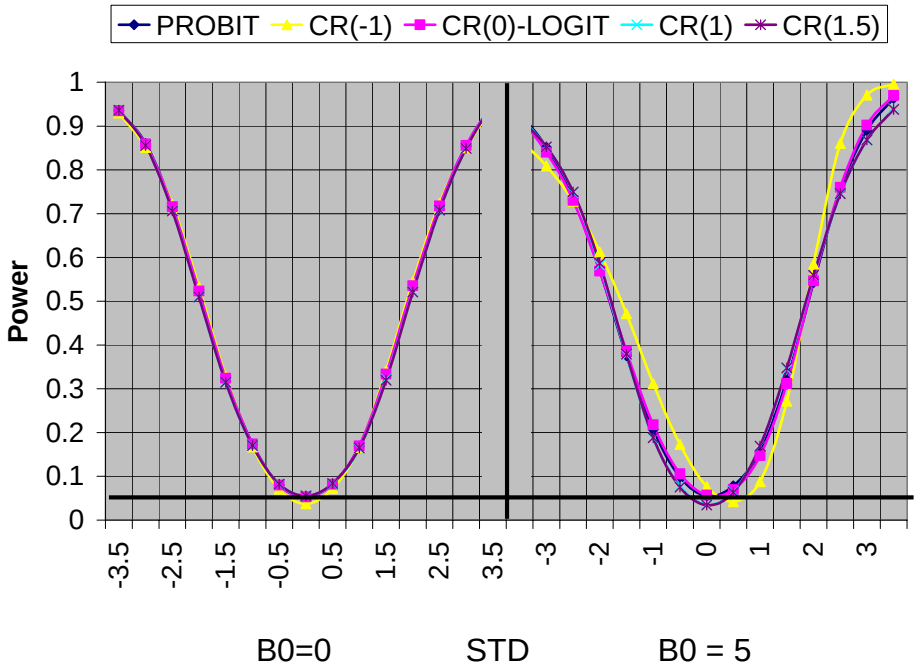

PROBIT
CR(-1)
CR(0)-LOGIT
CR(1)
MSE
0.3
0.25
0.2
0.15
0.1
0.05
0
0.50
0.56
0.61
0.66
0.71
0.76
0.80
0.84
$P_i=P(y_i=1)$

Figure 3: Power for $H_0 : \lambda_0 = r,\ r = E(\lambda_0) \pm k\, std(\lambda_0)$, Beta(.5,.5), Exp(4), N(0,9) and n = 300






PROBIT
CR(-1)
CR(0)-LOGIT
CR(1)
CR(1.5)
Power
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-3.5
-2.5
-1.5
-0.5
0.5
1.5
2.5
3.5
-3
-2
-1
0
1
2
3
B0=0
STD
B0 = 5

REFERENCES

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

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

Grunwald, P. & A. David (2004) Game Theory, Maximum Entropy, Minimum Discrepancy and Robust Bayesian Decision Theory. *Annals of Statistics* **32**, 1367-1433.

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* **58**, 71-120.

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

Judge, G. & R. Mittelhammer (2004) A Semiparametric Basis for Combining Estimation Problems under Quadratic Loss. *Journal of American Statistical Association* **99**, 479-487.

Judge, G., R. Mittelhammer, & D. Miller (2004) *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, in progress.

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

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

Lin, J. (1991) Divergence measures based on the Shannon entropy. *IEEE Trans. Inform. Theory,* **37**, 145-151.

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) Maximum Score Estimation of the Stochastic Utility Model of Choice. *Journal of Econometrics* **3**, 205-228.

McCullough, P. & 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.

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

Mittelhammer, R. & G. Judge (in press) (2004) Combining Estimators to Improve Structural Model Estimators to Improve Structural Model Estimation and Inference under Quadratic Loss. *Journal of Econometrics*.

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

Qin, J. (2000) Combining Parametric and Empirical Likelihood Data. *Biometrika* **87,** 484-490.

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

Shannon, C.E. (1948) A Mathematical Theory of Communication. *Bell Systems Technical Journal* **27**, 379-423 and 623-656.

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

Ullah, A. (1996) Entropy, Divergence and Distance Measures with Econometric Applications. *Journal of Statistical Planning and Inference* **49**, 137-162.