Bernoulli Regression Models: Re-examining Statistical Models with Binary Dependent Variables

Jason Bergtold[1] and Aris Spanos[2]

[1] Agricultural Economist, Soils Dynamic Research Unit, Agricultural Research Service, United States Department of Agriculture, 411 S. Donahue, Auburn, AL 36832. Phone: (334) 844-0864. Email: jbergtold@ars.usda.gov.

[2] Professor, Department of Economics, Virginia Tech, 3016 Pamplin Hall, Virginia Tech, Blacksburg, VA 24061. Phone: (540) 231-7981. Email: aris@vt.edu.

Selected Paper prepared for presentation at the American Agricultural Economics Association Annual Meeting, Providence, Rhode Island, July 24 – 27, 2005

Bernoulli Regression Models:

Re-examining Statistical Models with Binary Dependent Variables

Abstract

The classical approach for specifying statistical models with binary dependent variables in econometrics using latent variables or threshold models can leave the model misspecified, resulting in biased and inconsistent estimates as well as erroneous inferences. Furthermore, methods for trying to alleviate such problems, such as univariate generalized linear models, have not provided an adequate alternative for ensuring the statistical adequacy of such models. The purpose of this paper is to re-examine the underlying probabilistic foundations of statistical models with binary dependent variables using the probabilistic reduction approach to provide an alternative approach for model specification. This re-examination leads to the development of the Bernoulli Regression Model. Simulated and empirical examples provide evidence that the Bernoulli Regression Model can provide a superior approach for specifying statistically adequate models for dichotomous choice processes.

Keywords: Bernoulli Regression Model, logistic regression, generalized linear models, discrete choice, probabilistic reduction approach, model specification

Bernoulli Regression Models:

Re-examining Statistical Models with Binary Dependent Variables

1. Introduction

The evolution of conditional statistical models with binary dependent variables has led to two interrelated approaches on how to specify these models. Powers and Xie (2000) refer to these two approaches as the *latent variable* or theoretical approach and the *transformational* or statistical approach. The latent variable approach assumes the existence of an underlying continuous latent or unobservable stochastic process giving rise to the dichotomous choice process being observed. The transformational approach views the observed dichotomous choice process as inherently categorical and uses transformations of the observed data to derive an operational statistical model (Powers and Xie, 2000).

The latent variable approach assumes the existence of a latent stochastic process $\left\{Y_i^* \mid \mathbf{X}_i = \mathbf{x}_i, i = 1,..., N\right\}$, where $Y_i^* = g\left(\mathbf{X}_i;\theta\right) + \varepsilon_i$, the functional form of $g\left(.;.\right)$ is derived from theory, $\mathbf{X}_i$ is a *k*-dimensional vector of explanatory (random) variables, θ is a set of estimable parameters, ε_i is IID and $E\left(\varepsilon_i\right) = 0$ (Maddala, 1983). Given that Y_i^* is not directly observable, what is observed is another variable, Y_i, related to Y_i^*, such that $Y_i = \mathbf{1}\left(Y_i^* > \lambda\right)$, for some $\lambda \in \mathbf{R}$, where $\mathbf{1}(.)$ is the indicator function. A statistical model is then specified based on the following probabilistic framework:

$$\mathbf{P}\left(Y_i = 1\right) = \mathbf{P}\left(Y_i^* > \lambda\right) = \mathbf{P}\left(g\left(\mathbf{X}_i;\theta\right) + \varepsilon_i > \lambda\right) = F\left(g\left(\mathbf{X}_i;\theta\right) - \lambda\right),$$

where $F(.)$ is the cumulative density function (cdf) of ε_i and assumed to be symmetric.[1] If ε_i is distributed IID extreme value, then $F(.)$ is the logistic cdf, and if ε_i is IID normal, then $F(.)$ is the normal cdf, giving rise to the binary logistic and probit regression models, respectively (Train, 2003). Given that the underlying latent process used to specify the model is unobservable, the assumptions concerning the distribution of ε_i and the functional form of $g(\mathbf{X}_i;\theta)$ cannot be directly verified. If these assumptions are wrong, then the estimable model obtained is misspecified and the parameter estimates inconsistent (Coslett, 1983).

The transformational approach follows the theory of univariate generalized linear models developed by Nelder and Wedderburn (1972), except these models arise from conditional rather than marginal distributions. Following Fahrmeir and Tutz (1994), let $\{Y_i \mid \mathbf{X}_i = \mathbf{x}_i, i = 1,..., N\}$ be an independent stochastic process, such that $Y_i \mid \mathbf{X}_i = \mathbf{x}_i \sim bin(p_i,1)$. Now consider the linear predictor $\eta_i = \beta' t(\mathbf{x}_i)$, where $t(\mathbf{x}_i)$ is a $(S \times 1)$ vector of transformations of $\mathbf{x}_i$ and β is a $(S \times 1)$ vector of parameters. It is assumed that the linear predictor is related to p_i via the inverse of a known one-to-one, sufficiently smooth response function, i.e. $\eta_i = l(p_i)$, where $l(.)$ is referred to as the *link function*. Hence, the transformational approach attempts to specify a transformation or functional form for the link function to obtain an operational statistical model. If $l(.)$ is the logistic (probit) transformation, then this approach gives rise to the traditional binary logistic (probit) regression model. In fact, the inverse of any one-to-one, sufficiently

[1] When $F(.)$ is symmetric:

$$\mathbf{P}(g(\mathbf{X}_i;\theta)+\varepsilon_i > \lambda) = \mathbf{P}(\varepsilon_i > \lambda - g(\mathbf{X}_i;\theta)) = 1 - F(\lambda - g(\mathbf{X}_i;\theta)) = F(g(\mathbf{X}_i;\theta) - \lambda).$$

smooth cdf could provide a proper model under this approach. Kay and Little (1987) but the functional form of $p_i = l^{-1}(\eta) = g(\eta)$ is in fact determined by the probabilistic structure of the observed data. That is, the functional form of $p_i = E(Y_i \mid \mathbf{X}_i = \mathbf{x}_i)$ is dependent upon the functional form of $f(Y_i \mid \mathbf{X}_i = \mathbf{x}_i)$ and can be specified using $f(\mathbf{X}_i \mid Y_i = j)$ for $j = 0,1$ (see also Arnold and Press, 1989). Again the wrong choice of $l(.)$ or η will leave the model misspecified and the parameter estimates inconsistent.

In an attempt to deal with these functional form issues, the purpose of this paper is to re-examine the underlying probabilistic foundations of conditional statistical models with binary dependent variables using the probabilistic reduction approach developed by Spanos (1986,1995,1999). This examination leads to a formal presentation of the Bernoulli Regression Model (BRM), a family of statistical models, which includes the binary logistic regression model. This paper provides a more complete extension of the work done by Kay and Little (1987). Other issues addressed include, specification and estimation of the model, as well as, using the BRM for simulation.

2. The Probabilistic Reduction Approach and Bernoulli Regression Model

The probabilistic reduction approach is based on re-interpreting the De Finetti representation theorem as a formal way of reducing the joint distribution of all observable random variables involved into simplified products of distributions by imposing certain probabilistic assumptions (Spanos, 1999). This decomposition provides a formal and intuitive mechanism for constructing statistical models, with the added benefit of identifying the underlying probabilistic assumptions of the statistical model being examined.

A statistical model is defined as a set of probabilistic assumptions that adequately capture the systematic information in the observed data in a parsimonious and efficient way. The primary goal of the probabilistic reduction approach is to obtain statistically adequate models, where the "adequacy of a statistical model is judged by the appropriateness of the [probabilistic] assumptions (making up the model) in capturing the systematic information in the observed data (Spanos, 1999; p.544)."

Let $\{Y_i, i = 1,..., N\}$ be a stochastic process defined on the probability space $(S, \Im, P(.))$, where $Y_i \sim bin(p,1)$ (Bernoulli), $E(Y_i) = p$ and $Var(Y_i) = p(1-p)$ for $i = 1,..., N$. Furthermore, let $\{\mathbf{X}_i = (X_{1,i},..., X_{K,i}), i = 1,..., N\}$ be a vector stochastic process defined on the same probability space with joint density function $f(\mathbf{X};\psi_2)$, where ψ_2 is an appropriate set of parameters. Furthermore, assume that $E(X_{k,i}^2) < \infty$ for $k = 1,..., K$ and $i = 1,..., N$, making $\{Y_i, i = 1,..., N\}$ and each $\{\mathbf{X}_{k,i}, i = 1,..., N\}$, $k = 1,..., K$, elements of $L_2(\mathbf{R}^N)$, the space of all square integrable stochastic processes over $\mathbf{R}^N$. The joint density function of the joint vector stochastic process $\{(Y_i, \mathbf{X}_i), i = 1,..., N\}$ takes the form:

$$f(Y_1,..., Y_N, \mathbf{X}_1,..., \mathbf{X}_N; \phi), \qquad (1)$$

where ϕ is an appropriate set of parameters.

All of the systematic (and probabilistic) information contained in the vector stochastic process $\{(Y_i, \mathbf{X}_i), i = 1,..., N\}$ is captured by the Haavelmo Distribution, which is represented by the joint density function given by equation (1). Based on a weaker version of De Finetti's representation theorem, by specifying a set of reduction assumptions from three broad categories:

(D) Distributional, **(M)** Memory/Dependence, and **(H)** Heterogeneity,

concerning the vector stochastic process $\{(Y_i, \mathbf{X}_i), i = 1,..., N\}$, the modeler can reduce the Haavelmo distribution or joint density function into an operational form, giving rise to an operational statistical model and probabilistic model assumptions. By specifying particular reduction assumptions, the modeler is essentially partitioning the space of all possible statistical models into a family of operational models (indexed by the parameter space) (Spanos, 1999).

Assuming that the joint vector stochastic process $\{(Y_i, \mathbf{X}_i), i = 1,..., N\}$ is independent (I) and identically distributed (ID), the joint distribution given by equation (1) can be reduced (decomposed) in the following manner:

$$f(Y_1,...,Y_N, \mathbf{X}_1,..., \mathbf{X}_N; \phi) \overset{I}{=} \prod_{i=1}^{N} f_i(Y_i, \mathbf{X}_i; \varphi_i) \overset{ID}{=} \prod_{i=1}^{N} f(Y_i, \mathbf{X}_i; \varphi), \qquad (2)$$

where φ_i and φ are appropriate sets of parameters. The last component of the reduction in equation (2) can be further reduced so that:

$$f(Y_1,...,Y_N, \mathbf{X}_1,..., \mathbf{X}_N; \phi) \overset{IID}{=} \prod_{i=1}^{N} f(Y_i, \mathbf{X}_i; \varphi) = \prod_{i=1}^{N} f(Y_i \mid \mathbf{X}_i; \psi_1) \cdot f(\mathbf{X}_i; \psi_2), \quad (3)$$

where ψ_1 and ψ_2 are appropriate sets of parameters.

It is the reduction in (3) that provides us with the means to define an operational statistical model with binary dependent variables. For the reduction in equation (3) to give rise to a proper statistical model, it is necessary that the joint density function $f(Y_i, \mathbf{X}_i; \varphi)$ exist. The existence of $f(Y_i, \mathbf{X}_i; \varphi)$ is dependent upon the compatibility of the conditional density functions, $f(Y_i \mid \mathbf{X}_i; \psi_1)$ and $f(\mathbf{X}_i \mid Y_i; \eta_1)$ (where η_1 is an appropriate set of parameters) (Arnold and Castillo, 1999), i.e.

$$f(Y_i \mid \mathbf{X}_i; \psi_1) \cdot f(\mathbf{X}_i; \psi_2) = f(\mathbf{X}_i \mid Y_i; \eta_1) \cdot f(Y_i; p) = f(Y_i, \mathbf{X}_i; \varphi), \qquad (4)$$

where $f(Y_i; p) = p^{Y_i}(1-p)^{1-Y_i}$.

Arnold *et* al. (1999;p. 17) state that a sufficient condition for the compatibility of $f(Y_i \mid \mathbf{X}_i; \psi_1)$ and $f(\mathbf{X}_i \mid Y_i; \eta_1)$ is that the ratio:

$$\frac{f(Y_i = 1 \mid \mathbf{X}_i; \psi_1) \cdot f(\mathbf{X}_i \mid Y_i = 0; \eta_1)}{f(Y_i = 0 \mid \mathbf{X}_i; \psi_1) \cdot f(\mathbf{X}_i \mid Y_i = 1; \eta_1)}$$

does not depend on $\mathbf{Z}_i$. Thus, using equation (4), the above ratio must be equal to $\frac{p}{1-p}$, implying the following condition must be met:

$$\frac{f(\mathbf{X}_i \mid Y_i = 1; \eta_1)}{f(\mathbf{X}_i \mid Y_i = 0; \eta_1)} \cdot \frac{f(Y_i = 1; p)}{f(Y_i = 0; p)} = \frac{f(Y_i = 1 \mid \mathbf{X}_i; \psi_1)}{f(Y_i = 0 \mid \mathbf{X}_i; \psi_1)} \cdot \frac{f(\mathbf{X}_i; \psi_2)}{f(\mathbf{X}_i; \psi_2)}. \qquad (5)$$

Assume that $f(Y_i \mid \mathbf{X}_i; \psi_1)$ is a conditional Bernoulli density function with the following functional form:

$$f(Y_i \mid \mathbf{X}_i; \psi_1) = g(\mathbf{X}_i; \psi_1)^{Y_i} \left[1 - g(\mathbf{X}_i; \psi_1)\right]^{1-Y_i}, \qquad (6)$$

where $g(\mathbf{X}_i; \psi_1): \mathbf{R}^K \times \Theta_1 \rightarrow [0,1]$ and $\psi_1 \in \Theta_1$, the parameter space associated with ψ_1.[2] The density function specified by equation (6) satisfies the usual properties of a density function, i.e. following the properties of the Bernoulli density function (see Spanos, 1999):

(i) $f(Y_i \mid \mathbf{X}_i; \psi_1) \geq 0$ for $Y_i = 0,1$ and $\mathbf{X}_i = \mathbf{x}_i \in \mathbf{R}^K$,

(ii) $\sum_{Y_i = 0,1} f(Y_i \mid \mathbf{X}_i; \psi_1) = g(\mathbf{X}_i; \psi_1) + (1 - g(\mathbf{X}_i; \psi_1)) = 1$, and

[2] This choice of functional form is based upon a similar functional form used by Cox and Wermuth (1992).

(iii) $F(b \mid \mathbf{X}_i;\psi_1) - F(a \mid \mathbf{X}_i;\psi_1) = \begin{cases} 0 & \text{if } a < 0 \text{ and } b < 0 \\ 1 - g(\mathbf{X}_i;\psi_1) & \text{if } a < 0 \text{ and } 0 \le b < 1 \\ g(\mathbf{X}_i;\psi_1) & \text{if } 0 < a < 1 \text{ and } b \ge 1, \text{ for } (a,b) \in \mathbf{R}, \\ 1 & \text{if } a \le 0 \text{ and } b \ge 1 \\ 0 & \text{if } a > 1 \text{ and } b > 1 \end{cases}$

where (i) follows from the nonnegativity of $g(\mathbf{X}_i;\psi_1)$ and $F(.\mid \mathbf{X}_i;\psi_1)$ represents the cumulative conditional Bernoulli density function, which takes the following functional form:

$$F(z \mid \mathbf{X}_i;\psi_1) = \begin{cases} 0 & \text{for } z < 0 \\ 1 - g(\mathbf{X}_i;\psi_1) & \text{for } 0 \le z < 1 \\ 1 & \text{for } z \ge 1 \end{cases}.$$

Substituting equation (6) into (5) and letting $\pi_j = p^j(1-p)^{1-j}$ for $j = 0,1$ gives:

$$\frac{f(\mathbf{X}_i \mid Y_i = 1;\eta_1)}{f(\mathbf{X}_i \mid Y_i = 0;\eta_1)} \cdot \frac{\pi_1}{\pi_0} = \frac{g(\mathbf{X}_i;\psi_1)}{1 - g(\mathbf{X}_i;\psi_1)} \cdot \frac{f(\mathbf{X}_i;\psi_2)}{f(\mathbf{X}_i;\psi_2)}, \tag{7}$$

which implies that:

$$g(\mathbf{X}_i;\psi_1) = \frac{\pi_1 \cdot f(\mathbf{X}_i \mid Y_i = 1;\eta_1)}{\pi_0 \cdot f(\mathbf{X}_i \mid Y_i = 0;\eta_1) + \pi_1 \cdot f(\mathbf{X}_i \mid Y_i = 1;\eta_1)}. \tag{8}$$

Given the general properties of density functions and that $\pi_j \in (0,1)$ for $j = 0,1$, the range of $g(\mathbf{X}_i;\psi_1)$ is $[0,1]$, justifying the assumption that $g(\mathbf{X}_i;\psi_1): \mathbf{R}^K \times \Theta_1 \to [0,1]$.

A more intuitive and practical choice for $g(\mathbf{X}_i;\psi_1)$ can be found by using results from Kay and Little (1987). Using the identity $f(.) = \exp(\ln f(.))$ and after rearranging some terms, $g(\mathbf{X}_i;\psi_1)$ becomes:

$$g(\mathbf{X}_i;\psi_1) = \frac{\exp\{h(\mathbf{X}_i;\eta_1)\}}{1 + \exp\{h(\mathbf{X}_i;\eta_1)\}} = [1 + \exp\{-h(\mathbf{X}_i;\eta_1)\}]^{-1}, \tag{9}$$

where $h(\mathbf{X}_i;\eta_1) = \ln\left(\frac{f(\mathbf{X}_i \mid Y_i = 1;\eta_1)}{f(\mathbf{X}_i \mid Y_i = 0;\eta_1)}\right) + \kappa$ and $\kappa = \ln(\pi_1) - \ln(\pi_0)$. Written as the composite function, $g(h(\mathbf{X}_i;\eta_1))$, $g(.)$ represents the logistic cumulative density function (the transformation function) and $h(.;.)$ represents the traditional index function. Equation (9) illustrates the functional relationship between ψ_1 and η_1 (i.e. $\psi_1 = G(\eta_1)$), as well.[3]

The conditional distribution $f(Y_i \mid \mathbf{X}_i;\psi_1)$ allows the modeler to define a statistical generating mechanism (SGM), which is viewed as an idealized representation of the true underlying data generating process (see Spanos, 1999). The SGM is usually characterized by a set of conditional moment functions of $f(Y_i \mid \mathbf{X}_i;\psi_1)$, such as the regression function:

$$Y_i = E(Y_i \mid \mathbf{X}_i = \mathbf{x}_i) + u_i, \tag{10}$$

where $E(Y_i \mid \mathbf{X}_i = \mathbf{x}_i)$ represents the systematic component and u_i the nonsystematic component (the error term). The orthogonal decomposition in equation (10) arises when $\{Y_i, i = 1,..., N\}$ and $\{X_{k,i}, i = 1,..., N\}$ are elements of L_2 for $k = 1,..., K$ (see Small and McLeish, 1994 and Spanos, 1999). The SGM can contain higher order conditional moment functions when they capture systematic information in the data. These can be specified using u_i, in the following manner:

$$u_i^s = E(u_i^s \mid \mathbf{X}_i = \mathbf{x}_i) + v_{i,s}, \tag{11}$$

[3] Note, that in some cases one is able to reparametricize $h(\mathbf{x}_i;\eta_1)$, so that $h(\mathbf{x}_i;\eta_1) = h(\mathbf{x}_i;G(\eta_1)) = h(\mathbf{x}_i;\psi_1)$. In other cases, $\psi_1 = \eta_1$ (see section 3 for examples).

where *s* denotes the s^{th} order conditional moment function. When $s = 2,3 \text{ or } 4$, equation (11) represents the skedastic (conditional variance), clitic (conditional skewness) and kurtic (conditional kurtosis) functions, respectively.

Given that $Var(Y_i \mid \mathbf{X}_i = \mathbf{x}_i) < \infty$, the stochastic process $\{Y_i \mid \mathbf{X}_i = \mathbf{x}_i, i = 1,..., N\}$ can be decomposed orthogonally giving rise to the following regression function:

$$Y_i = E(Y_i \mid \mathbf{X}_i = \mathbf{x}_i) + u_i = g(\mathbf{x}_i; \psi_1) + u_i = [1 + \exp\{-h(\mathbf{x}_i; \eta_1)\}]^{-1} + u_i, \qquad (12)$$

where the last inequality follows by substituting in equation (9). The distribution of the error term, u_i, is given by:

u_i	$1 - g(\mathbf{X}_i;\psi_1)$	$-g(\mathbf{X}_i;\psi_1)$
$f(u_i)$	$g(\mathbf{X}_i;\psi_1)$	$1-g(\mathbf{X}_i;\psi_1)$

where $E(u_i) = 0$ and $Var(u_i) = g(\mathbf{X}_i;\psi_1)(1 - g(\mathbf{X}_i;\psi_1))$. If $\mathbf{X}_i$ is discrete then $f(u_i)$ will be discrete, but if $\mathbf{X}_i$ is continuous then $f(u_i)$ will be a multimodal distribution. For example, consider the univariate case when $X_i \mid Y_i = j \sim N(0.6 + 0.6Y_i, 1)$, then $f(u_i)$ has the (simulated) bimodal distribution in Figure 1.

Equation (12) represents the SGM for a family of statistical models known as the Bernoulli Regression Model, which is more formally specified in Table 1.[4] The first three model assumptions, i.e. distributional, functional form and heteroskedasticity, arise from the derivations provided above. The homogeneity and independence assumptions are a result of the IID reduction assumptions made about the joint vector stochastic process $\{(Y_i, \mathbf{X}_i), i = 1,..., N\}$.

[4] The conditional variance (or skedastic function) and higher order moment functions are not included in the SGM because they are specified in terms of the conditional mean, $g(\mathbf{x}_i;\psi_1)$

The regression function given by equation (12) is similar to the traditional binary logistic regression model, but above derivations show that it arises naturally from the joint density function given by equation (1), suggesting it as an obvious candidate for modeling discrete choice processes when the dependent variable is distributed Bernoulli(p). Another important observation is that the functional forms for both $g(\mathbf{X}_i;\psi_1)$ and $h(\mathbf{X}_i;\eta_1)$ are both dependent upon the functional form of $f(\mathbf{X}_i \mid Y_i;\eta_1)$ and in turn the joint distribution of Y_i and $\mathbf{X}_i$.

3. Model Specification

Kay and Little (1987) provide the necessary specifications for $h(\mathbf{x}_i;\eta_1)$ when $X_{k,i}$, $k = 1,...,K$ (the explanatory variables) have distributions from the simple exponential family and are independent conditional on Y_i of each other. When these conditions are not met, the model specification becomes more complex. Kay and Little (1987) provide examples involving sets of random variables with multivariate Bernoulli and normal distributions, but due to the complexity of dealing with multivariate distributions they advocate accounting for any dependence between the explanatory variables by including cross-products of transformations (based on their marginal distributions) of the explanatory variables. This paper builds on the model specification work initiated by Kay and Little (1987).

An initial issue concerning specification of BRMs is that $f(\mathbf{X}_i \mid Y_i;\eta_1)$ is not usually known and for many cases cannot be readily derived.[5] A potential alternative is to assume that:

[5] For help with such derivations, work by Arnold, Castillo and Sarabia (1999) may be of assistance.

$$f(\mathbf{X}_i \mid Y_i;\eta_1) = f(\mathbf{X}_i;\eta_1(Y_i)). \tag{13}$$

In this sense, one is treating the moments of the conditional distribution of $\mathbf{X}_i$ given Y_i as functions of Y_i. That is $\eta_1(Y_i = j) = \eta_{1,j}$ for $j = 0,1$. Lauritzen and Wermuth (1989) use a similar approach to specify conditional Gaussian distributions, and Kay and Little (1987) use this approach to specify the logistic regressions models in their paper (see also Tate, 1954 and Oklin and Tate, 1961).

Table 2 provides the functional forms for $g(x_i;\psi_1)$ needed to obtain a properly specified BRM with one explanatory variable for a number of different conditional distributions of the form $f(X_i;\eta_{1,j})$. Following Kay and Little (1987), all of the cases examined in Table 2 have index functions that are linear in the parameters. Examples of conditional distributions that give rise to nonlinear index functions include when $f(X_i;\eta_{1,j})$ is distributed F, extreme value or logistic. In such cases, one option is to explicitly specify $f(X_i;\eta_{1,j})$ and estimate the model using equation (9), which can be difficult numerically due to the inability to reparametricize the model, leaving both $\eta_{1,0}$ and $\eta_{1,1}$ in $h(\mathbf{x}_i;\eta_{1,j})$. Another option is to transform X_i so that it has one of the conditional distributions specified in Table 2. To illustrate this latter approach, consider the following example.

Example 1: Let $f(X_i;\eta_{1,j})$ be a conditional Weibull distribution of the form:

$$f(X_i;\eta_1) = \frac{\gamma \cdot X_i^{\gamma-1}}{\alpha_j^{\gamma}} \exp\left\{-\left(\frac{X_i}{\alpha_j}\right)^{\gamma}\right\}, \tag{14}$$

where $(\alpha_j, \gamma) \in \mathbf{R}_+^2$ and $X_i > 0$. That is $X_i \mid Y_i = j \sim W(\alpha_j, \gamma)$. If $X_i \sim W(\alpha, \gamma)$ then $X_i^{\gamma} \sim Exp(\alpha)$ (i.e. exponential). Thus, $X_i^{\gamma} \mid Y_i = j \sim Exp(\alpha_j)$, and using the results from Table 2:

$$g(x_i; \psi_1) = \left[1 + \exp\left\{-\beta_0 - \beta_1 x_i^{\gamma}\right\}\right]^{-1}, \qquad (15)$$

where $\beta_0 = \left[\kappa + \gamma \ln\left(\frac{\alpha_0}{\alpha_1}\right)\right]$ and $\beta_1 = \left(\frac{1}{\alpha_0}\right)^{\gamma} - \left(\frac{1}{\alpha_1}\right)^{\gamma}$.

If there is more than one explanatory variable, then a number of different approaches exist for model specification. The first approach is to explicitly specify the multivariate distribution $f(\mathbf{X}_i; \eta_{1,j})$. If $f(\mathbf{X}_i; \eta_{1,j})$ is multivariate normal with homogenous covariance matrix, then:

$$g(\mathbf{x}_i; \psi_1) = \left[1 + \exp\left(-\beta_0 - \sum_{k=1}^{K} \beta_k x_{k,i}\right)\right]^{-1}.$$

On the other hand, if the covariance matrix exhibits heterogeneity (based on Y_i), then:

$$g(\mathbf{x}_i; \psi_1) = \left[1 + \exp\left(-\beta_0 - \sum_{k=1}^{K} \beta_k x_{k,i} - \sum_{j=1}^{K} \sum_{l \geq j}^{K} \beta_{j,l} x_{j,i} x_{l,i}\right)\right]^{-1}$$

(Kay and Little, 1987). Kay and Little (1987) state there are a limited number of other multivariate distributions that exist in the literature which would give rise to readily estimable and tractable BRMs. Three additional multivariate distributions that do suffice, include the binomial, beta and gamma distributions. The following example presents the case for a conditional bivariate gamma distribution.

Example 2: Let $f(X_{1,i}, X_{2,i}; \eta_{1,j})$ be a conditional bivariate gamma distribution, of the form:

$$f\left(X_{1,i}, X_{2,i}; \eta_{1,j}\right) = \frac{\alpha_j \theta_{1,j} \theta_{2,j}}{\Gamma\left[\theta_{1,j}\right]\Gamma\left[\theta_{2,j}\right]} e^{-\alpha_j X_{2,i}} X_{1,i}^{\theta_{1,j}-1}\left(X_{2,i} - X_{1,i}\right)^{\theta_{2,j}-1},$$

where $\Gamma[.]$ is the gamma function, $X_{2,i} > X_{1,i} \geq 0$ and $\left(\alpha_j, \theta_{1,j}, \theta_{2,j}\right) \in \mathbf{R}_+^3$ (Spanos, 1999). Then:

$$g\left(x_{1,i}, x_{2,i}; \psi_1\right) = \left[1 + \exp\left\{-\beta_0 - \beta_1 x_{2,i} - \beta_2 \ln\left(x_{1,i}\right) - \beta_3 \ln\left(x_{2,i} - x_{1,i}\right)\right\}\right]^{-1},$$

where $\beta_0 = \left[\kappa + \ln\left(\frac{\alpha_1 \theta_{1,1} \theta_{2,1} \Gamma\left[\theta_{1,0}\right]\Gamma\left[\theta_{2,0}\right]}{\alpha_0 \theta_{1,0} \theta_{2,0} \Gamma\left[\theta_{1,1}\right]\Gamma\left[\theta_{2,1}\right]}\right)\right]$, $\beta_1 = \left(\alpha_0 - \alpha_1\right)$, $\beta_2 = \left(\theta_{1,1} - \theta_{1,0}\right)$ and $\beta_3 = \left(\theta_{2,1} - \theta_{2,0}\right)$.

Another approach for specifying a BRM when $K > 1$ is to decompose $f\left(\mathbf{X}_i; \eta_{1,j}\right)$ into a product of simpler conditional density functions. Following Kay and Little (1987), consider the case where the explanatory variables are independent of each other conditional on Y_i. Then, $f\left(\mathbf{X}_i; \eta_{1,j}\right) = \prod_{k=1}^{K} f\left(X_{k,i}; \eta_{1,k,j}\right)$, making the index function $h\left(\mathbf{x}_i; \eta_1\right) = \sum_{k=1}^{K} \ln\left(\frac{f\left(X_{k,i}; \eta_{1,k,1}\right)}{f\left(X_{k,i}; \eta_{1,k,0}\right)}\right) + \kappa$. The results in Table 2 then can be used to specify $h\left(\mathbf{x}_i, \eta_1\right)$ by specifying the (sub) index functions, $h\left(x_{k,i}; \eta_{1,k}\right) = \ln\left(\frac{f\left(X_{k,i}; \eta_{1,k,1}\right)}{f\left(X_{k,i}; \eta_{1,k,0}\right)}\right)$, (without κ) for each $X_{k,i}$. The difficulty here is assessing the conditional independence of the explanatory variables given Y_i, but results by Tate (1954) and Oklin and Tate (1961) may be some help.

If some or none of the explanatory variables are independent conditional on Y_i, then another approach for decomposing $f(\mathbf{X}_i;\eta_{1,j})$ is sequential conditioning (Spanos, 1999), i.e.

$$f(\mathbf{X}_i;\eta_{1,j}) = f(X_{1,i};\eta_{1,1,j})\prod_{k=2}^{K} f(X_{k,i} \mid X_{k-1,i},...,X_{1,i};\xi_{k,j}),$$

where $\xi_{k,j}$ is an appropriate set of parameters. Given the potential complexity of this approach, it can be combined with the previous approach to reduce the dimensionality and increase the tractability of the problem. To illustrate this alternative, consider the following example.

Example 3: Let

$$f(X_{1,i},X_{2,i},X_{3,i};X_{4,i};\eta_{1,j}) = f(X_{1,i},X_{2,i};\eta_{2,j})\cdot f(X_{3,i},X_{4,i};\eta_{3,j}),$$

where $X_{1,i}$ and $X_{2,i}$ are independent conditional on Y_i of $X_{3,i}$ and $X_{4,i}$. Now assume that (i) $X_{1,i}$ given $Y_i = j$ is distributed *bin*(1, ρ_j), (ii) $X_{2,i}$ given $X_{1,i} = l$ $(l = 0,1)$ and $Y_i = j$ is distributed exponential, i.e.:

$$f(X_{1,i};\xi_{1,j,l}) = \frac{1}{\theta_{j,l}}\exp\left\{-\frac{X_{1,i}}{\theta_{j,l}}\right\},$$

and (iii) $X_{3,i}$ and $X_{4,i}$ given $Y_i = j$ are jointly distributed bivariate beta, i.e.:

$$f(X_{3,i},X_{4,i};\eta_{3,j}) = \left(\frac{\Gamma(\alpha_j+\delta_j+\gamma_j)}{\Gamma(\alpha_j)\Gamma(\delta_j)\Gamma(\gamma_j)}\right)\left[X_{3,i}^{\alpha_j-1}\cdot X_{4,i}^{\delta_j-1}\cdot(1-X_{3,i}-X_{4,i})^{\gamma_j-1}\right],$$

where $X_{3,i} \geq 0$, $X_{4,i} \geq 0$ and $X_{3,i} + X_{4,i} \leq 1$ for $i = 1,...,N$; $(\alpha_j,\delta_j,\gamma_j) > 0$ for $j = 0,1$; and $\Gamma(.)$ is the gamma function (Spanos, 1999). Using these assumptions:

$$f\left(X_{1,j}, X_{2,j}; \eta_{2,j}\right) = f\left(X_{1,j}; \xi_{1,j,l}\right) \cdot f\left(X_{2,j}; \rho_j\right)$$
$$= \left[\frac{\rho_j}{\theta_{j,1}} \exp\left\{-\frac{X_{1,i}}{\theta_{j,1}}\right\}\right]^{X_{2,i}} \left[\frac{\left(1-\rho_j\right)}{\theta_{j,0}} \exp\left\{-\frac{X_{1,i}}{\theta_{j,0}}\right\}\right]^{1-X_{2,i}}$$

(see Kay and Little, 1987), implying that:

$$g\left(\mathbf{x}_i; \psi_1\right) = \left[1 + \exp\left\{-\beta_0 - \beta_1 x_{1.i} - \beta_2 x_{2,i} - \beta_3 x_{1,i} x_{2,i} - \beta_4 \ln\left(x_{3,i}\right)\right.\right.$$
$$\left.\left. - \beta_5 \ln\left(x_{4,i}\right) - \beta_6 \ln\left(1 - x_{3,i} - x_{4,i}\right)\right\}\right]^{-1}$$

where $\beta_0 = \left[\kappa + \left(\frac{\left(1-\rho_1\right)\theta_{0,0}}{\left(1-\rho_0\right)\theta_{1,0}}\right) + \ln\left(\lambda\right)\right]$, $\beta_1 = \left(\frac{1}{\theta_{0,0}} - \frac{1}{\theta_{1,0}}\right)$,

$\beta_2 = \left[\ln\left(\frac{\rho_1\theta_{0,1}}{\rho_0\theta_{1,1}}\right) + \ln\left(\frac{\left(1-\rho_1\right)\theta_{0,0}}{\left(1-\rho_0\right)\theta_{1,0}}\right)\right]$, $\beta_3 = \left(\frac{1}{\theta_{0,1}} - \frac{1}{\theta_{1,1}} - \frac{1}{\theta_{0,0}} + \frac{1}{\theta_{1,0}}\right)$, $\beta_4 = \alpha_1 - \alpha_0$,

$\beta_5 = \delta_1 - \delta_0$, $\beta_6 = \gamma_1 - \gamma_0$ and $\lambda = \frac{\Gamma\left(\alpha_1 + \delta_1 + \gamma_1\right)\Gamma\left(\alpha_0\right)\Gamma\left(\delta_0\right)\Gamma\left(\gamma_0\right)}{\Gamma\left(\alpha_0 + \delta_0 + \gamma_0\right)\Gamma\left(\alpha_1\right)\Gamma\left(\delta_1\right)\Gamma\left(\gamma_1\right)}$.

Kay and Little (1987) provide a number of similar examples involving discrete and continuous variables. If the decomposition of $f\left(\mathbf{X}_i; \eta_{1,j}\right)$ involved an unknown multivariate distribution conditional on Y_i of continuous variables, then it becomes considerably more difficult to derive the specification of $g\left(\mathbf{x}_i; \psi_1\right)$. Guidelines and results presented by Arnold, Castillo and Sarabia (1999) provide a means for attempting these specifications, and are beyond the current scope of this paper.

4.0 Model Estimation

In order to utilize all of the information present in the distribution of the sample, given by equation (1), the method of maximum likelihood should be used to estimate the parameters of the BRM (Spanos, 1999). Given the independence of the sample, the log-likelihood function for the logistic form of the BRM is:

$$\ln L(\varphi;(\mathbf{y},\mathbf{x})) = \sum_{i=1}^{N}\left[y_i \ln\left(g\left(h(\mathbf{x}_i;\psi_1)\right)\right) + (1-y_i)\ln\left(1-g\left(h(\mathbf{x}_i;\psi_1)\right)\right)\right], \qquad (16)$$

where $g(.)$ is the logistic cdf and $h(.,.)$ is written as a function of ψ_1, the parameters of interest. Now let $\partial\mathbf{h}_i$ denote the gradient of $h(\mathbf{x}_i;\psi_1)$ with respect to the vector ψ_1 (e.g. β), $\partial^2\mathbf{h}_i$ the Hessian, and $g'(.)$ the logistic probability density function. Then:

$$\frac{\partial \ln L(\varphi;(\mathbf{y},\mathbf{x}))}{\partial \psi_1} = \sum_{i=1}^{N}\left[\left(\frac{y_i - g\left(h(\mathbf{x}_i;\psi_1)\right)}{g\left(h(\mathbf{x}_i;\psi_1)\right)\left(1-g\left(h(\mathbf{x}_i;\psi_1)\right)\right)}\right) g'\left(h(\mathbf{x}_i;\psi_1)\right)\partial\mathbf{h}_i\right], \text{ and}$$

$$\frac{\partial^2 \ln L(\varphi,(\mathbf{y},\mathbf{x}))}{\partial \psi_1 \partial \psi_1'} = -\sum_{i=1}^{N}\left[\left(\frac{y_i - g\left(h(\mathbf{x}_i;\psi_1)\right)}{g\left(h(\mathbf{x}_i;\psi_1)\right)\left(1-g\left(h(\mathbf{x}_i;\psi_1)\right)\right)}\right)^2 \left(g'\left(h(\mathbf{x}_i;\psi_1)\right)\right)^2 (\partial\mathbf{h}_i)(\partial\mathbf{h}_i)^{\mathrm{T}}\right]$$
$$+\sum_{i=1}^{N}\left[\left(\frac{y_i - g\left(h(\mathbf{x}_i;\psi_1)\right)}{g\left(h(\mathbf{x}_i;\psi_1)\right)\left(1-g\left(h(\mathbf{x}_i;\psi_1)\right)\right)}\right)\left(g'\left(h(\mathbf{x}_i;\psi_1)\right)(\partial\mathbf{h}_i)(\partial\mathbf{h}_i)^{\mathrm{T}} + g\left(h(\mathbf{x}_i;\psi_1)\right)\left(\partial^2\mathbf{h}_i\right)\right)\right].$$

When $h(\mathbf{x}_i;\eta_1)$ is nonlinear in the parameters estimation becomes more difficult, because the likelihood function may no longer be globally concave and many computer routines only estimate logistic regression models with index functions linear in the parameters (Train, 2003). In these cases, the researcher may need to write their own code and use a number of different algorithms to estimate the model. The asymptotic properties of consistency and asymptotic normality of the MLE estimates follow if certain regularity conditions are satisfied (see Gourieroux, 2000 and Spanos, 1999).

5. Simulation

A significant benefit of using the probabilistic reduction approach for developing the BRM is that it provides a mechanism for randomly generating the vector stochastic process, $\{(Y_i,\mathbf{X}_i), i=1,...,N\}$ using the relationship given by equation (4) for simulations involving the BRM. The process involves performing two steps:

Step 1: Generate a realization of the stochastic process $\{Y_i, i = 1,..., N\}$ using a binomial random number generator.

Step 2: Using $f(\mathbf{X}_i; \eta_{1,j})$ generate a realization of the vector stochastic process, $\{\mathbf{X}_i, i = 1,..., N\}$ using appropriate random number generators with the parameters given by $\eta_{1,j} = \eta_{1,0}$ when $Y_i = 0$ and $\eta_{1,j} = \eta_{1,1}$ when $Y_i = 1$.

It should be noted that no a priori theoretical interpretation is imposed on the generation process, it is purely statistical in nature.[6] Furthermore, the parameters ψ_1 can be easily determined from the parameters $\eta_{1,j}$, via $\psi_1 = G(\eta_{1,1}, \eta_{1,0})$ when conducting simulations.

To illustrate, consider the BRM given in Example 1. Let $Y_i \sim bin(0.6,1)$ and $X_i \mid Y_i = j$ have a conditional Weibull distribution with $\alpha_0 = 1, \alpha_1 = 1.4$ and $\gamma = 3$. In this situation, the mapping $\psi_1 = G(\eta_{1,1}, \eta_{1,0})$ given in Example 1 gives $\beta_0 = -0.6040$, $\beta_1 = 0.6356$ and $\gamma = 3$ for the parameters of the regression function given by equation (15). A Monte Carlo simulation using the above two-step procedure for randomly generating a binary choice process was used to examine the asymptotic properties of the parameters β_0, β_1 and γ. A random sample of Y_i ($p = 0.6$) was generated 1000 times and then was used to generate X_i 100 times using equation (14) for $N = 50, 100, 250, 500, 1000, 2500$ and 5000. For each run, the regression equation given by equation (15) was estimated using the log likelihood function given by equation (16)

[6] This generation procedure is in contrast to procedures assuming the existence of an unobservable latent stochastic process (see Train, 2003).

and a derivative-free algorithm developed by Nelder and Mead (1965).[7] The results of the simulation are reported in Table 3. Given the convergence of the mean to the true value, the decreasing standard errors, and convergence of the skewness and kurtosis towards 0 and 3 respectively, as N increases, it would seem that there is evidence for concluding that β_0, β_1 and γ are consistent and asymptotically normal.

6. Empirical Example

Data was obtained from Al-Hmoud and Edwards (2004) from a study examining private sector participation in the water and sanitation sector of developing countries. Using there data a model was constructed examining this participation based on four explanatory factors. The dependent variable, *total private investment* (Y), was binary, taking a value of '1' if there was private investment in a given year and '0' otherwise. Of the four explanatory variables used in the model, two were binary and two were continuous. The two binary variables were *low renewable water resources* (X_3) and *government effectiveness* (X_4). The two continuous variables were *per capita GDP* (X_1) and *percent urban population growth* (X_2). The dataset contained 149 observations for 39 countries from 1996 to 2001, but data was not available for all countries for all years, resulting in an unbalanced panel (Al-Hmoud and Edwards, 2004).

Given that Y is distributed Bernoulli, a BRM was chosen to model private sector participation in developing countries in the water and sanitation sector. To examine how

[7] It was found that this algorithm provided the best convergence properties for the given problem. A potential problem with index functions nonlinear in the parameters is the difficulty algorithms using derivatives and Hessians may have in finding an optimal solution due to potentially highly nonlinear or large relatively flat regions of the objective surface.

to proceed with model specification, the sample conditional correlation matrix given Y was estimated using the sample correlation coefficients of the residuals from appropriate regressions of the explanatory variables on Y.[8] The sample conditional correlation matrix was:

$$\begin{pmatrix} 1.00 & -0.45 & 0.11 & 0.54 \\ -0.45 & 1.00 & -0.52 & -0.19 \\ 0.11 & -0.52 & 1.00 & 0.19 \\ 0.54 & -0.19 & 0.19 & 1.00 \end{pmatrix},$$

which provided no determination on how to decompose $f\left(X_{1,i}, X_{2,i}, X_{3,i}, X_{4,i}; \eta_{1,j}\right)$ into independent components. Thus, sequential conditioning was used to give:

$$f\left(X_{1,i}, X_{2,i}, X_{3,i}, X_{4,i}; \eta_{1,j}\right) = f\left(X_{1,i}, X_{2,i}; \eta'_{1,j,k,l}\right) \cdot f\left(X_{3,i}, X_{4,i}; \mathbf{q}_j\right), \quad (17)$$

where $\eta'_{1,j,k,l} = \eta_1\left(Y_i = j, X_{3,i} = k, X_{4,i} = l\right)$ and

$$f\left(X_{3,i}, X_{4,i}; \mathbf{q}_j\right) = q_{j,0,0}^{(1-X_{3,i})(1-X_{4,i})} q_{j,1,0}^{X_{3,i}(1-X_{4,i})} q_{j,0,1}^{(1-X_{3,i})X_{4,i}} q_{j,1,1}^{X_{3,i}X_{4,i}}, \quad (18)$$

where $q_{j,0,0} + q_{j,1,0} + q_{j,0,1} + q_{j,1,1} = 1$. That is, equation (18) is a multivariate Bernoulli($\mathbf{q}_j$) distribution conditional on $Y_i = j$.

After taking account of the heterogeneity in the continuous explanatory variables, it was assumed that $X_{1,i}$ and $X_{2,i}$ were jointly distributed bivariate normal conditional on $Y_i = j$, $X_{3,i} = k$ and $X_{4,i} = l$ for $j,k,l = 0,1$, i.e.

$$f\left(X_{1,i}, X_{2,i}; \eta'_{1,j,k,l}\right) = (2\pi)^{-1} \left|\Sigma_{j,k,l}\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\mathbf{X}_i - \mu_{j,k,l}\right)' \Sigma_{j,k,l}^{-1}\left(\mathbf{X}_i - \mu_{j,k,l}\right)\right\}, \quad (19)$$

[8] For the binary explanatory variables, appropriate logistic regression models were estimated, while for the continuous explanatory variables normal linear regression models were used.

where $\mathbf{X}_i = \left(X_{1,i}, X_{2,i}\right)'$, $\mu_{j,k,l} = \left(\mu_{1,j,k,l}, \mu_{2,j,k,l}\right)'$ is a vector of conditional means, $\Sigma_{j,k,l}$ is the conditional covariance matrix, and $|.|$ signifies the determinant operator. Given equation (18), this implies that:

$$\begin{aligned} f\left(X_{1,i}, X_{2,i}, X_{3,i}, X_{4,i}; \eta_{1,j}\right) = & \left[q_{j,0,0} \cdot f\left(X_{1,i}, X_{2,i}; \eta'_{1,j,0,0}\right)\right]^{\left(1-X_{3,i}\right)\left(1-X_{4,i}\right)} \times \\ & \left[q_{j,1,0} \cdot f\left(X_{1,i}, X_{2,i}; \eta'_{1,j,1,0}\right)\right]^{X_{3,i}\left(1-X_{4,i}\right)} \times \\ & \left[q_{j,0,1} \cdot f\left(X_{1,i}, X_{2,i}; \eta'_{1,j,0,1}\right)\right]^{\left(1-X_{3,i}\right)X_{4,i}} \times \\ & \left[q_{j,1,1} \cdot f\left(X_{1,i}, X_{2,i}; \eta'_{1,j,1,1}\right)\right]^{X_{3,i}X_{4,i}}. \end{aligned} \tag{20}$$

Plugging equation (20) into $h\left(\mathbf{x}_i; \eta_1\right)$ and computing $\psi_1 = G\left(\eta_{1,j}\right)$:

$$\begin{aligned} h\left(\mathbf{x}_i; \psi_1\right) = & \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \beta_4 x_{4,i} + \beta_5 x_{1,i}^2 + \beta_6 x_{1,i} x_{2,i} + \beta_7 x_{2,i}^2 + \\ & \beta_8 x_{1,i} x_{3,i} + \beta_9 x_{2,i} x_{3,i} + \beta_{10} x_{1,i} x_{4,i} + \beta_{11} x_{2,i} x_{4,i} + \beta_{12} x_{3,i} x_{4,i} + \beta_{13} x_{1,i}^2 x_{3,i} + \\ & \beta_{14} x_{1,i} x_{2,i} x_{3,i} + \beta_{15} x_{2,i}^2 x_{3,i} + \beta_{16} x_{1,i}^2 x_4 + \beta_{17} x_{1,i} x_{2,i} x_{4,i} + \beta_{18} x_{2,i}^2 x_{4,i} + \\ & \beta_{19} x_{1,i} x_{3,i} x_{4,i} + \beta_{20} x_{2,i} x_{3,i} x_{4,i} + \beta_{21} x_{1,i}^2 x_{3,i} x_{4,i} + \beta_{22} x_{1,i} x_{2,i} x_{3,i} x_{4,i} + \\ & \beta_{23} x_{2,i}^2 x_{3,i} x_{4,i}, \end{aligned} \tag{21}$$

which when plugged into equation (9) provides an estimable BRM. If $\Sigma_{j,k,l} = \Sigma$, then all the terms involving $x_{1,i}^2, x_{1,i} x_{2,i}$ and $x_{2,i}^2$ would disappear, but this was not the case.[9]

Since the index function given by equation (21) is linear in the parameters, standard computer software packages with logistic regression models, were used to estimate the corresponding BRM. Estimation results for the logistic regression model using equation (21) and a more common specification found in the applied literature:

$$h\left(\mathbf{x}_i; \psi_1\right) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \beta_4 x_{4,i}, \tag{22}$$

[9] Results of these analyses using the data are available from the authors upon request.

are presented in Table 4. Misspecification testing results for the BRM using equation (21) indicated the presence of heterogeneity across years, so fixed effects (using dummy variables) for the years 1996-1999 were incorporated into both models.[10]

The two models were compared using a likelihood ratio test, with the null hypothesis being that the more common specification of the logit model using equation (22) with fixed effects across time was correct. The computed likelihood ratio test statistic was 69.3229 with an associated p-value of 0.0000, indicating that the more common formulation of the logistic regression model is misspecified. Further evidence that the BRM using equation (22) was superior to the more common specification of the logistic regression model is given by the higher R^2 values, higher within-sample prediction and lower mean square error.[11]

7. Conclusion

The latent variable approach and the transformational approach for specifying statistical models with binary dependent variables can result in statistical misspecification. Both approaches do not explicitly recognize that the functional form of $E(Y_i \mid \mathbf{X}_i = \mathbf{x}_i)$ depends on $f(\mathbf{X}_i \mid Y_i = j; \eta_1)$ and in turn the existence of $f(Y_i, \mathbf{X}_i; \varphi)$. Using the probabilistic reduction approach and results derived by Kay and Little (1987),

[10] A likelihood ratio test was conducted in a Fisher testing framework to examine the BRM without fixed effects across time (see Spanos, 1999). The null hypothesis was no fixed effects and the likelihood test statistic was 34.1369 with an association p-value of 0.00001, indicating no support for the null hypothesis. Heterogeneity across regions was tested as well, but no evidence of this type of heterogeneity was found.

[11] Additional misspecification tests for functional form and dependence indicated that the functional form was not misspecified, but there may exist temporal and/or spatial dependence in the data. These tests and results are available from the authors upon request and will be explored further in a future paper.

this relationship is formally defined to derive the Bernoulli Regression Model. While specification of these models can be difficult at times, examination of the sample conditional correlation matrix of the explanatory variables given Y_i can help determine plausible decompositions of $f(\mathbf{X}_i \mid Y_i = j; \eta_1)$ to arrive at operational BRMs. Furthermore, the model assumptions shown in Table 1 can be tested to verify that the BRM obtained is statistically adequate, thereby allowing the model to provide reliable statistical inferences and predictions. The theoretical and empirical examples provide evidence that the common use of logit and probit models with linear index functions both in the parameters and variables are suspect when the underlying model assumptions have not been verified.

The Bernoulli Regression Model can provide a parsimonious description of the probabilistic structure of conditional binary choice process being examined and imposes no a priori theoretical or ad hoc restrictions (or assumptions) upon the model, thereby providing a theory-free statistical model of the conditional binary choice process being examined. As noted by Spanos (1995), this freedom allows the modeler to conduct statistical inferences (if the statistical assumptions made about the underlying stochastic process are appropriate) that can be used to examine if the theory in question can account for the systematic information in the observed data.

References

1. Al-Hmoud, R.B. and J. Edwards. "A Means to an End: Studying the Existing Environment for Private Sector Participation in the Water and Sanitation Sector." Working Paper. Department of Economics and Geography, Texas Tech University, Lubbock Texas, 2004.
2. Arnold, B.C., E. Castillo and J.M. Sarabia. *Conditional Specification of Statistical Models*. New York, NY: Springer Verlag, 1999.
3. Arnold, B.C. and S.J. Press. "Compatible Conditional Distributions." *Journal of the American Statistical Association*. 84(March, 1989): 152 – 156.
4. Coslett, S.R. "Distribution-Free Maximum Likelihood Estimator of the Binary Choice Model." *Econometrica*. 51(May 1983): 765 – 782.
5. Cox, D.R. and N. Wermuth. "Response Models for Mixed Binary and Qualitative Variables. *Biometrika*. 79(1992): 441 – 461.
6. Fahrmeir, L. and G. Tutz. *Multivariate Statistical Modeling Based on Generalized Linear Models*. New York: Springer-Verlag, 1994.
7. Gourieroux, C. *Econometrics of Qualitative Dependent Variables*. Cambridge, UK: Cambridge University Press, 2000.
8. Kay, R. and S. Little. "Transformations of the Explanatory Variables in the Logistic Regression Model for Binary Data." *Biometrika*. 74(September, 1987): 495 – 501.
9. Keane, M.P. "Current Issues in Discrete Choice Modeling." *Marketing Letters*. 8(1997): 307 – 322.

10. Lauritzen, S.L. and N. Wermuth. “Graphical Models for Association Between Variables, Some Which Are Qualitative and Some Quantitative.” *Annals of Statistics*. 17(1989): 31 – 57.
11. Maddala, G.S. *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge, UK: Cambridge University Press, 1983.
12. Nelder, J.A. and R. Mead. “A Simplex Method for Function Minimization.” *Computer Journal*. 7(1965): 308 – 313.
13. Nelder, J.A. and R.W.M. Wedderburn. “Generalized Linear Models.” *Journal of the Royal Statistical Society, Series A (General)*. 3(1972): 370 – 384.
14. Oklin, I. and R.F. Tate. “Multivariate Correlation Models with Mixed Discrete and Continuous Variables.” *The Annals of Mathematical Statisitcs*. 32(June, 1961): 448 – 465.
15. Powers, D.A. and Y. Xie. *Statistical Methods for Categorical Data Analysis*. San Deigo, CA: Academic Press, 2000.
16. Small, C.G. and D.L. McLeish. *Hilbert Space Methods in Probability and Statistical Inference*. New York: John Wiley and Sons, Inc., 1994.
17. Spanos, A. “On Theory Testing In Econometrics: Modeling with Nonexperimental Data.” *Journal of Econometrics*. 67(1995): 189 – 226.
18. Spanos, A. *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*. Cambridge, UK: Cambridge University Press, 1999.
19. Spanos, A. *Statistical Foundations of Econometrics Modeling*. Cambridge, UK: Cambridge University Press, 1986.

20. Tate, R.F. “Correlation Between a Discrete and a Continuous Variable. Point-Biserial Correlation.” *The Annals of Mathematical Statistics*. 25(September, 1954): 603 – 607.
21. Train, K.E. *Discrete Choice Methods with Simulation*. Cambridge, UK: Cambridge University Press, 2003.

Table 1: Bernoulli Regression Model

SGM: $Y_i = g(\mathbf{x}_i;\psi_1) + u_i,\ i = 1,...,N$,

where (i)

u_i	$1-g(\mathbf{X}_i;\psi_1)$	$-g(\mathbf{X}_i;\psi_1)$
$f(u_i)$	$g(\mathbf{X}_i;\psi_1)$	$1-g(\mathbf{X}_i;\psi_1)$

(ii) $E(u_i) = 0$; and

(iii) $Var(u_i) = g(\mathbf{X}_i;\psi_1)(1-g(\mathbf{X}_i,\psi_1))$.

Assumptions

Distributional:	$Y_i \mid \mathbf{X}_i = \mathbf{x}_i \sim bin(g(\mathbf{x}_i,\psi_1),1)$, (conditional Bernoulli).
Functional Form:	$E(Y_i \mid \mathbf{X}_i = \mathbf{x}_i) = g(\mathbf{x}_i;\psi_1) = [1+\exp\{-h(\mathbf{x}_i;\eta_1)\}]$, where $h(\mathbf{x}_i;\eta_1) = \ln\left[\frac{f(\mathbf{X}_i \mid Y_i = 1;\eta_1)}{f(\mathbf{X}_i \mid Y_i = 0;\eta_1)}\right] + \kappa$ and $\psi_1 = G(\eta_1)$.
Heteroskedasticity:	$Var(Y_i \mid \mathbf{X}_i = \mathbf{x}_i) = g(\mathbf{x}_i;\psi_1)(1-g(\mathbf{x}_i;\psi_1))$.
Homogeneity:	$\psi_1 = G(\eta_1)$ is not a function of $i = 1,...,N$.
Independence:	$\{Y_i \mid \mathbf{X}_i = \mathbf{x}_i, i = 1,...,N\}$ is an independent stochastic process.

Table 2: Specification of $g(x_i;\eta_1)$ with one explanatory variable and conditional distribution, $f(X_i;\eta_{1,j})$, for $j = 0,1$.

Distribution of X_i given Y_i	$f(X_i;\eta_{1,j}) =$ [2]	$g(x_i;\psi_1) =$
Beta[1]	$\frac{X_i^{\alpha_j-1}(1-X_i)^{\gamma_j-1}}{\mathbf{B}[\alpha_j,\gamma_j]}$, where $(\alpha_j,\gamma_j)\in\mathbf{R}_+^2$ and $0\le X_i\le 1$.	$[1+\exp\{\beta_0+\beta_1\ln(x_i)+\beta_2\ln(1-x_i)\}]^{-1}$, where $\beta_0=\left[\kappa+\ln\left(\frac{\mathbf{B}[\alpha_0,\gamma_0]}{\mathbf{B}[\alpha_1,\gamma_1]}\right)\right]$, $\beta_1=(\alpha_1-\alpha_0)$ and $\beta_2=(\gamma_1-\gamma_0)$.
Binomial[1]	$\binom{n}{X_i}\theta_j^{X_i}(1-\theta_j)^{n-X_i}$, where $0<\theta_j<1$, $X_i=0,1$ and $n=1,2,3,...$	$[1+\exp\{-\beta_0-\beta_1 x_i\}]^{-1}$, where $\beta_0=\left[\kappa+n\ln\left(\frac{1-\theta_1}{1-\theta_0}\right)\right]$ and $\beta_1=\ln\left(\frac{\theta_1}{\theta_0}\right)-\ln\left(\frac{1-\theta_1}{1-\theta_0}\right)$.
Chi-square	$\frac{2^{-\frac{v_j}{2}}}{\Gamma\left[\frac{v_j}{2}\right]}X_i^{\frac{v_j-2}{2}}\exp\left\{-\frac{X_i}{2}\right\}$, where $v=1,2,3,...$ and $x\in\mathbf{R}_+$.	$[1+\exp\{-\beta_0-\beta_1 x_i\}]^{-1}$, where $\beta_0=\left[\kappa+\left(\frac{v_0-v_1}{2}\right)\ln(2)+\ln\left(\frac{\Gamma\left[\frac{v_0}{2}\right]}{\Gamma\left[\frac{v_1}{2}\right]}\right)\right]$ and $\beta_1=\frac{v_1-v_0}{2}$.
Exponential	$\frac{1}{\theta_j}\exp\left\{-\frac{X_i}{\theta_j}\right\}$, where $\theta_j\in\mathbf{R}_+$ and $X_i\in\mathbf{R}_+$.	$[1+\exp\{-\beta_0-\beta_1 x_i\}]^{-1}$, where $\beta_0=\left[\ln\left(\frac{\theta_0}{\theta_1}\right)+\kappa\right]$ and $\beta_0=\left(\frac{1}{\theta_0}-\frac{1}{\theta_1}\right)$.
Gamma[1]	$\frac{1}{\gamma_j\Gamma[\alpha_j]}\left(\frac{X_i}{\gamma_j}\right)^{\alpha_j-1}\exp\left\{-\frac{X_i}{\gamma_j}\right\}$, where $(\alpha_j,\gamma_j)\in\mathbf{R}_+^2$ and $X_i\in\mathbf{R}_+$.	$[1+\exp\{\beta_0+\beta_1 x_i+\beta_2\ln(x_i)\}]^{-1}$, where $\beta_0=\left[\kappa+\ln\left(\frac{\gamma_0\Gamma[\alpha_0]}{\gamma_1\Gamma[\alpha_1]}\right)+(\alpha_0-1)\ln(\gamma_0)-(\alpha_1-1)\ln(\gamma_1)\right]$, $\beta_1=\left(\frac{1}{\gamma_0}-\frac{1}{\gamma_1}\right)$ and $\beta_2=(\alpha_1-\alpha_0)$.
Geometric	$\theta_j(1-\theta_j)^{X_i-1}$, where $0\le\theta_j\le 1$ and $X_i=1,2,3,...$	$[1+\exp\{-\beta_0-\beta_1 x_i\}]^{-1}$, where $\beta_0=\left[\kappa+\ln\left(\frac{\theta_1}{\theta_0}\right)-\ln\left(\frac{1-\theta_1}{1-\theta_0}\right)\right]$ and $\beta_1=\ln\left(\frac{1-\theta_1}{1-\theta_0}\right)$.

Table 2 continued.

Logarithmic	$\alpha_j \left(\theta_j^{X_i} \right)$, where $\alpha_j = -\left[\ln\left(1-\theta_j\right)\right]^{-1}$, $0 < \theta_j < 1$ and $\mathrm{X_i} = 1,2,3,...$	$\left[1+\exp\left\{-\beta_0 - \beta_1 x_i\right\}\right]^{-1}$, where $\beta_0 = \left[\kappa + \ln\left(\frac{\alpha_1}{\alpha_0}\right)\right]$ and $\beta_1 = \ln\left(\frac{\theta_1}{\theta_0}\right)$.
Log-Normal	$\frac{1}{X_i} \cdot \frac{1}{\sigma_j\sqrt{2\pi}} \exp\left\{-\frac{\left(\ln\left(X_i\right)-\mu_j\right)^2}{2\sigma_j^2}\right\}$, where $\mu_j \in \mathbf{R}$, $\sigma_j^2 \in \mathbf{R}_+$ and $X_i \in \mathbf{R}$.	$\left[1+\exp\left\{\beta_0 + \beta_1 \ln(x_i) + \beta_2 \left(\ln(x_i)\right)^2\right\}\right]^{-1}$, where $\beta_0 = \left[\kappa + \ln\left(\frac{\sigma_0}{\sigma_1}\right) + \left(\frac{\mu_0^2}{2\sigma_0^2} - \frac{\mu_1^2}{2\sigma_1^2}\right)\right]$, $\beta_1 = \left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_0}{\sigma_0^2}\right)$ and $\beta_2 = \left(\frac{1}{2\sigma_0^2} - \frac{1}{2\sigma_1^2}\right)$.
Normal[1]	$\frac{1}{\sigma_j\sqrt{\pi}} \exp\left\{-\frac{1}{2\sigma_j^2}\left(X_i-\mu_j\right)^2\right\}$, where $\mu_j \in \mathbf{R}$, $\sigma_j^2 \in \mathbf{R}_+$ and $X_i \in \mathbf{R}$.	$\left[1+\exp\left\{\beta_0 + \beta_1 x_i + \beta_2 x_i^2\right\}\right]^{-1}$, where $\beta_0 = \left[\kappa + \ln\left(\frac{\sigma_0}{\sigma_1}\right) + \left(\frac{\mu_0^2}{2\sigma_0^2} - \frac{\mu_1^2}{2\sigma_1^2}\right)\right]$, $\beta_1 = \left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_0}{\sigma_0^2}\right)$ and $\beta_2 = \left(\frac{1}{2\sigma_0^2} - \frac{1}{2\sigma_1^2}\right)$.
Pareto	$\theta_j x_0^{\theta_j} X_i^{-\theta_j - 1}$, where $\theta_j \in \mathbf{R}_+$, $x_0 > 0$ and $X_i \geq x_0$.	$\left[1+\exp\left\{-\beta_0 - \beta_1 \ln\left(x_i\right)\right\}\right]^{-1}$, where $\beta_0 = \left[\kappa + \ln\left(\frac{\theta_1}{\theta_0}\right) + \left(\theta_1 - \theta_0\right)\ln\left(x_0\right)\right]$ and $\beta_1 = \left(\theta_0 - \theta_{1.}\right)$
Poisson[1]	$\frac{e^{-\theta_j}\theta_j^{X_i}}{X_i!}$, where $\theta_j > 0$ and $X_i = 1,2,3,...$	$\left[1+\exp\left\{-\beta_0 - \beta_1 x_i\right\}\right]^{-1}$, where $\beta_0 = \left[\kappa + \theta_0 - \theta_{\cdot}\right]$ and $\beta_1 = \ln\left(\frac{\theta_1}{\theta_0}\right)$.

[1] *Source*: Kay and Little (1987).

[2] *Source*: Spanos (1999). $\mathbf{B}[\ ]$ represents the beta function and $\Gamma[\ ]$ represents the gamma function.

Table 3: Monte Carlo Simulation Results for Example 1

Parameter	Number of Observations (*N*)	Mean	Standard Deviation	Skewness	Kurtosis	Minimum	Maximum
			True Value = -0.6040				
β_0	N = 50	-1.3818	2.8420	-6.2629	51.0895	-30.5758	0.8559
	N = 100	-1.1167	2.2610	-7.7925	78.1452	-29.5084	0.4964
	N = 250	-0.6592	0.4495	-1.9983	11.9573	-4.3456	0.2839
	N = 500	-0.6179	0.2515	-0.5588	3.8758	-1.6702	0.0973
	N = 1000	-0.6212	0.1785	-0.5503	3.7557	-1.4291	-0.1987
	N = 2500	-0.6101	0.1131	-0.2670	2.9358	-0.9543	-0.2819
	N = 5000	-0.6085	0.0789	-0.0278	3.1540	-0.9131	-0.3112
			True Value = 0.6356				
β_1	N = 50	1.3637	2.9204	6.0719	48.3619	0.0000	31.1302
	N = 100	1.1422	2.3355	7.4744	72.7571	0.0001	29.9172
	N = 250	0.6832	0.4959	2.2396	13.0258	0.0005	4.8237
	N = 500	0.6435	0.2791	0.8397	4.5379	0.0514	2.0420

Table 3 continued.

Parameter	Number of Observations (*N*)	Mean	Standard Deviation	Skewness	Kurtosis	Minimum	Maximum
β_1	N = 1000	0.6506	0.1992	0.7351	4.1469	0.1581	1.6016
	N = 2500	0.6421	0.1269	0.3445	2.9474	0.2739	1.0701
	N = 5000	0.6376	0.0895	0.0660	2.9984	0.3223	0.9763
				True Value = 3.0			
γ	N = 50	4.6698	4.3463	2.4179	11.6444	-6.6156	36.2235
	N = 100	4.1471	3.5111	2.6295	13.0030	0.0824	28.0070
	N = 250	3.5300	1.7017	2.5781	16.4192	0.4513	17.7591
	N = 500	3.2363	0.9155	1.2591	6.8497	1.1333	9.1500
	N = 1000	3.0825	0.5811	0.6526	4.3230	1.6281	6.1177
	N = 2500	3.0361	0.3655	0.2861	2.9450	2.1125	4.2341
	N = 5000	3.0250	0.2609	0.3726	3.2808	2.2855	4.1462

Table 4: Estimation Results for the Empirical BRM and Traditional Logit Models

Variable	BRM using Equation (21)	Traditional Logit using Equation (22)
	Coefficient Estimate (Standard Error)[1]	Coefficient Estimate (Standard Error)[1]
Intercept	-24.6608	-1.6690
	(17.2947)	(0.7677)
Dummy 1996	-5.8496	-2.5791
	(1.3295)	(0.6876)
Dummy 1997	-3.8065	-1.9291
	(1.0811)	(0.6536)
Dummy 1998	-2.8882	-1.7267
	(1.0222)	(0.6620)
Dummy 1999	-1.9274	-1.0273
	(0.9624)	(0.6615)
$X_{1,i}$	0.0041	0.0004
	(0.0060)	(0.0001)
$X_{2,i}$	12.7504	0.4738
	(7.8883)	(0.1436)
$X_{3,i}$	24.2342	1.0485
	(18.5156)	(0.4706)
$X_{4,i}$	-27.9374	0.5495
	(75.4572)	(0.4884)
$X_{1,i}^2$	0.0000	---
	(0.0000)	
$X_{1,i}X_{2,i}$	-0.0019	---
	(0.0014)	
$X_{2,i}^2$	-1.3945	---
	(0.8552)	
$X_{1,i}X_{3,i}$	-0.0067	---
	(0.0097)	
$X_{2,i}X_{3,i}$	-11.5153	---
	(8.0715)	
$X_{1,i}X_{4,i}$	0.0024	---
	(0.0255)	
$X_{2,i}X_{4,i}$	9.4477	---
	(32.8429)	
$X_{3,i}X_{4,i}$	14.8636	---
	(76.3755)	
$X_{1,i}^2X_{3,i}$	0.0000	---
	(0.0000)	
$X_{1,i}X_{2,i}X_{3,i}$	0.0010	---
	(0.0016)	
$X_{2,i}^2X_{3,i}$	1.6699	---
	(0.9339)	

Table 4 continued.

Variable	BRM using Equation (21)	Traditional Logit using Equation (22)
	Coefficient Estimate (Standard Error)[1]	**Coefficient Estimate (Standard Error)[1]**
$X_{1,i}^2 X_{4,i}$	-0.0000 (0.0000)	---
$X_{1,i} X_{2,i} X_{4,i}$	0.0022 (0.0058)	---
$X_{2,i}^2 X_{4,i}$	-0.9815 (3.5851)	---
$X_{1,i} X_{3,i} X_{4,i}$	0.0053 (0.0268)	---
$X_{2,i} X_{3,i} X_{4,i}$	-0.5565 (33.2265)	---
$X_{1,i}^2 X_{3,i} X_{4,i}$	-0.0000 (0.0000)	---
$X_{1,i} X_{2,i} X_{3,i} X_{4,i}$	-0.0033 (0.0059)	---
$X_{2,i}^2 X_{3,i} X_{4,i}$	-0.5530 (3.6362)	---
Other Statistics		
Log-Likelihood	-45.3512	-80.0127
McFadden's Pseudo R^2	0.5466	0.2001
Estrella's R^2	0.6543	0.2590
Percent Correctly Predicted	87.25	67.79
Mean Square Error	3.7833	5.5505

[1] The standard errors are calculated using the estimate of the asymptotic information matrix.

Figure 1: Simulated Density Plot for a Bernoulli Regression Model with One Explanatory Variable Conditionally Distributed Normal Given $Y_i = j$.