Journal of Choice Modelling, 3(2), pp 1-28
www.jocm.org.uk

Bernoulli Regression Models: Revisiting the Specification of Statistical Models with Binary Dependent Variables

Jason S. Bergtold[1,*] Aris Spanos[2,†] Eberechukwu Onukwugha[3,∓]

[1]Department of Agricultural Economics, Kansas State University, Manhattan, KS 66506-4011
[2]Department of Economics, Virginia Tech, Blacksburg, VA
[3]Pharmaceutical Health Services Research, University of Maryland School of Pharmacy, Baltimore, MD

Received 5 June 2008, received version revised 12 June 2009, accepted 22 June 2009

Abstract

The latent variable and generalized linear modelling approaches do not provide a systematic approach for modelling discrete choice observational data. Another alternative, the probabilistic reduction (PR) approach, provides a systematic way to specify such models that can yield reliable statistical and substantive inferences. The purpose of this paper is to re-examine the underlying probabilistic foundations of conditional statistical models with binary dependent variables using the PR approach. This leads to the development of the Bernoulli Regression Model, a family of statistical models, which includes the binary logistic regression model. The paper provides an explicit presentation of probabilistic model assumptions, guidance on model specification and estimation, and empirical application.

Keywords: Bernoulli Regression Model, Generalized Linear Models, Latent Variable Models, Logistic Regression, Model Specification, Probabilistic Reduction Approach

* Corresponding author, T: +785-532-0984, F: + 785-532-6925, bergtold@ksu.edu
† T: +540-231-7981, aris@vt.edu
∓ T: +410-706-8981, F: +410-706-5394, eonukwug@rx.umaryland.edu

Licensed under a Creative Commons Attribution-Non-Commercial 2.0 UK: England & Wales License
http://creativecommons.org/licenses/by-nc/2.0/uk/

ISSN 1755-5345

1 Introduction

The theoretical specification of discrete choice models with binary dependent variables has developed using latent variable methods, such as the random utility model, to explain human behaviour (Marschak 1960; McFadden 1984; Train 2003). The latent variable approach specifies a theoretical (structural) model based on substantive information whose stochastic structure is given via a stochastic error term. In the random utility model, the distribution of the stochastic error term determines the functional form and interpretation of the choice probabilities being modelled (Train 2003). Thus, the underlying statistical properties of these models are dependent upon a potentially unverifiable distributional assumption, about which the modeller may have no *a priori* knowledge. If the distributional assumption concerning the stochastic error term is incorrect, then the estimable model obtained may be misspecified and the parameter estimates inconsistent (Coslett 1983).

The generalized linear modelling (GLM) framework (Nelder and Wedderburn 1972) provides an alternative approach to specifying discrete choice models. This approach specifies a purely statistical model without regard to potential theoretical requirements (Powers and Xie 2000). A binary choice model can be specified with appropriate choice of link function, which allows for the estimation of a conditional mean that is linear in the explanatory variables, facilitating the interpretation of parameter estimates (Hardin and Hilbe 2007). As a statistical modelling framework, this approach can ignore the probabilistic structure of the explanatory variables by focusing primarily on the distribution of the dependent variable, which affects the choice of link function. Ignoring the statistical information related to the probabilistic structure of the explanatory variables can leave the statistical model misspecified (e.g. Kay and Little 1986), resulting in inconsistent estimates and unreliable inferences.

As an example, the GLM approach lacks specific guidance on the functional form of the predictor past the linearity property, resulting in various specifications which may not always be appropriate or complete (Arnold *et al*. 1999). The relationship between logistic regression and discriminant analysis provides guidance on when the linearity property is satisfied using the log ratio of the conditional distributions of the explanatory variables conditional on the dependent variable. Linearity in the explanatory variables is satisfied when: (i) the explanatory variables are conditionally distributed multivariate normal with constant covariance matrix; or (ii) the explanatory variables are independent with conditional distributions from the simple exponential family [1] (Anderson 1972; Cox and Snell 1989; Kay and Little 1987). Linearity in the parameters is satisfied for the above cases when: (i) there are multivariate dichotomous variates following a log linear model; (ii) the explanatory variables are a combination of a multivariate normal distribution and a multivariate Bernoulli distribution; or (iii) the explanatory variables are independent and from the simple exponential family (Kay and Little 1987). While these functional specifications have been recognized they have been relegated to the realm of possible variable transformations for maintaining linearity (e.g. Hosmer and Lemeshow 2000). Arnold *et al*. (1999) state that "many of the logistic regression models discussed in the applied literature are questionable in the light of these observations" (p. 134). That is, although the above specification issues have been raised in the

[1] See equation (2).

literature, in applied settings the linearity property is not checked or is potentially ignored due to overriding theoretical considerations.

Taking account of the probabilistic structure of the explanatory variables as done by Kay and Little (1987) can help ensure that the functional form of the binary choice model is statistically adequate. The probabilistic reduction (PR) approach is a systematic approach to model specification of statistical models with binary dependent variables that takes into account all of the pertinent statistical information for model specification. The primary goal of this approach is to obtain statistically adequate models, where the adequacy of a statistical model is judged by the validity of the probabilistic model assumptions vis-a-vis the observed data (Spanos 1999). Structural or theoretical models can then be embedded into the statistical model whose adequacy has been secured. A significant advantage of the probabilistic reduction approach is the ability to explicitly identify the probabilistic model assumptions of the statistical model being estimated. A significant weakness in discrete choice modelling is the absence of these explicit assumptions in the specification of statistical models with binary dependent variables in the literature. The PR approach provides cohesive guidance on statistical model development.

In an attempt to provide a systematic statistical modelling approach that can be used for empirical modelling, 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. This examination leads to a formal presentation of the Bernoulli Regression Model, a family of statistical models, which includes the binary logistic regression model. The primary contributions of the paper include: (1) a systematic approach for derivation of the Bernoulli Regression Model from a proper joint distribution using the PR approach; (2) direct presentation of the underlying (testable) statistical assumptions of the Bernoulli Regression Model; (3) specification of Bernoulli Regression Models including cases that violate the linearity property; and (4) empirical application of the Bernoulli Regression Model. The PR approach provides a cohesive story concerning the statistical specification of the BRM and provides a proper framework for specifying statistically adequate binary choice models.

The remainder of the paper is organized as follows. The next section examines the statistical approaches used for model specification of binary dependent variables in the present literature. Section 3 formally presents the Bernoulli Regression Model with probabilistic model assumptions, while section 4 examines model specification, estimation and simulation. Sections 5 and 6 provide an empirical application of the Bernoulli Regression Model and concluding remarks, respectively.

2 Generalized Linear Models and Statistical Specification

The primary statistical approach for specifying statistical models with binary dependent variables was developed by Nelder and Wedderburn (1972). The univariate GLM approach historically has been associated with modelling experimental data, allowing the systematic component of the model, e.g. the unconditional mean, to be embedded in a linear structure even when the underlying conditional mean of the dependent variable is nonlinear in the parameters. A review of the literature (e.g. Fahrmeir and Tutz 2001) on model specification where the data are non-experimental and arise from a conditional distribution, i.e. $f_{Y|\mathbf{X}}(Y_i \mid \mathbf{X}_i;\beta)$, shows a lack of recognition about the role the probabilistic structure of the

conditioning variables plays in the specification of the conditional mean or regression function.

To demonstrate, consider the presentation of the GLM approach presented by Fahrmeir and Tutz (2001). Let $Y_i \mid \mathbf{X}_i = \mathbf{x}_i \sim Bin(1, p_i)$ and the linear predictor $\eta_i = \beta^T \mathbf{x}_i$, where $\mathbf{x}_i$ is a $(K \times 1)$ vector of covariates and β is a $(K \times 1)$ 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 $g(p_i) = \eta_i$, referred to as the link function. Let the inverse link function be given by $h(\eta_i) = p_i$. The GLM approach specifies a mapping relating the conditional mean p_i to the linear predictor η_i to capture the dependence between Y_i and $\mathbf{X}_i$ and obtain an operational statistical model. If $g(.)$ is the logistic (probit) transformation, then this approach gives rise to the traditional binary logistic (probit) regression model.

The linearity property of the predictor η_i in the GLM specification arises when (i) η_i is linear in the explanatory variables or (ii) η_i is linear in the parameters. The first condition implies the second when satisfied, and the second condition allows for nonlinear terms in the explanatory variables (e.g. interaction, logarithmic and polynomial terms). For logistic regression, Kay and Little (1987) introduce the inverse conditional distribution as a way to specify the functional form of the predictor η_i (or index function) based on the log odds:

$$\ln \frac{\mathbf{P}(Y_i = 1 \mid \mathbf{X}_i = \mathbf{x}_i)}{\mathbf{P}(Y_i = 0 \mid \mathbf{X}_i = \mathbf{x}_i)} = \ln \frac{f_{\mathbf{X}|Y=1}(\mathbf{X}_i; \theta_1)}{f_{\mathbf{X}|Y=0}(\mathbf{X}_i; \theta_0)} + \ln\left(\frac{p}{1-p}\right) = \beta^T \mathbf{x}_i, \qquad (1)$$

where $f_{\mathbf{X}|Y}(\mathbf{X}_i; \theta_j)$ is the inverse conditional distribution and $p = \mathbf{P}(Y_i = 1)$. Kay and Little (1987) show that the linearity property will be satisfied when the inverse conditional distributions take certain functional forms. That is, many different models for **X** given Y will imply different functional forms for the log odds given by equation (1).

For the univariate case, when the inverse conditional distribution belongs to the exponential family defined by the distribution:

$$f_{\mathbf{X}|Y=j}(\mathbf{X}_i; \theta_j) = B(\theta_j) h(\mathbf{X}_i) \exp\left\{\sum_{m=1}^{M} R_m(\mathbf{X}_i) Q_m(\theta_j)\right\}, \theta_j = (\theta_{1,j}, ..., \theta_{M,j}) \quad (2)$$

which includes the beta, binomial, gamma, normal and Poisson distributions, the linear predictor will be linear in the parameters. For the multivariate case, specification of the linear predictor follows those examples mentioned in the introduction, which are linear in the variables and/or parameters. Kay and Little (1987) use sequential conditioning to derive alternative models, but conclude that larger multivariable cases become increasingly more difficult to specify. These multivariable cases are dealt with later in the current paper.

While Kay and Little show that the inverse conditional distribution provides useful information in specifying the logistic regression model, they do not extend their methodology back to the joint distribution to provide a unifying framework for deriving the conditional mean function. The joint distribution appears to be a more natural starting point for motivating and specifying conditional statistical models. In

fact, the inverse conditional distribution contains the probabilistic information about the explanatory variables needed to arrive at a properly specified model. Furthermore, the derivation of the Bernoulli Regression Model next shows how the more general framework developed here can give rise to logistic regression models with index functions that are nonlinear in the parameters.

3 Bernoulli Regression Models (The Probabilistic Reduction Approach)

The PR approach views statistical models as a parameterization of the joint distribution of all the observable (vector) stochastic processes involved. It is the importance accorded to the joint distribution that leads to recognition of the role of inverse conditional distributions in providing relevant statistical information for model specification. A statistical model is chosen so as to render the observed data a 'truly typical realization' of the process defined by the statistical model. A significant advantage of the PR approach is that it explicitly provides the underlying probabilistic model assumptions upon which the statistical adequacy of the estimated model and associated inferences depend. These assumptions can then be tested to check for statistical adequacy.

Let $\{Y_i, i=1,...,N\}$ be a stochastic process defined on the probability space $(S, \Im, P(.))$, where $Y_i \sim Bin(1,p)$ (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 $f_{\mathbf{x}}(\mathbf{X}_i;\theta)$, where θ is an appropriate set of parameters. Then S is the support of Y and $\mathbf{X}$ and $\Im$ is the sigma field generated by the vector stochastic process $\{(Y_i, \mathbf{X}_i), i=1,...,N\}$. Let the joint density function of the joint vector stochastic process $\{(Y_i, \mathbf{X}_i), i=1,...,N\}$ take the form:

$$f(Y_1,...,Y_N, \mathbf{X}_1,...,\mathbf{X}_N;\phi) \tag{3}$$

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 joint distribution given by equation (3). Based on a set of reduction assumptions from three broad categories:

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

the modeller can reduce the joint distribution into an operational parameterization defining a statistical model.

Assuming that the joint vector stochastic process $\{(Y_i, \mathbf{X}_i), i=1,...,N\}$ is IID, the joint distribution given by equation (3) can be reduced as follows:

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

$$= \prod_{i=1}^{N} f_{Y|\mathbf{X}}(Y_i \mid \mathbf{X}_i;\beta) \cdot f_{\mathbf{X}}(\mathbf{X}_i;\theta) \tag{4}$$

where φ_i , φ, β and θ are appropriate sets of parameters. For the reduction in equation (4) to give rise to a proper statistical model, it is necessary that the joint density function $f(Y_i, \mathbf{X}_i; \varphi)$ exist. This existence is dependent upon the compatibility of the conditional density function, $f_{Y|\mathbf{X}}(Y_i \mid \mathbf{X}_i; \beta)$ and its inverse conditional distribution $f_{\mathbf{X}|Y}(\mathbf{X}_i; \theta_j)$, where θ_j is an appropriate set of parameters and a function of $Y_i = j$ (Arnold *et al*. 1999).[2] That is:

$$f_{Y|\mathbf{X}}(Y_i \mid \mathbf{X}_i; \beta) \cdot f_{\mathbf{X}}(\mathbf{X}_i; \theta) = f_{\mathbf{X}|Y}(\mathbf{X}_i; \theta_j) \cdot f_Y(Y_i; p) = f(Y_i, \mathbf{X}_i; \varphi), \qquad (5)$$

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

Using a result from Arnold *et al*. (1999, p. 17), a sufficient condition for the compatibility of $f_{Y|\mathbf{X}}(Y_i \mid \mathbf{X}_i; \beta)$ and $f_{\mathbf{X}|Y}(\mathbf{X}_i; \theta_j)$ is that the ratio:

$$\frac{f_{Y|\mathbf{X}}(Y_i = 1 \mid \mathbf{X}_i; \beta) \cdot f_{\mathbf{X}|Y=0}(\mathbf{X}_i; \theta_0)}{f_{Y|\mathbf{X}}(Y_i = 0 \mid \mathbf{X}_i; \beta) \cdot f_{\mathbf{X}|Y=1}(\mathbf{X}_i; \theta_1)} \qquad (6)$$

does not depend on $\mathbf{X}_i$.[3] Using equation (5), the above ratio must be equal to $\frac{p}{1-p}$, implying that the following condition must be met:

$$\frac{f_{\mathbf{X}|Y=1}(\mathbf{X}_i; \theta_1)}{f_{\mathbf{X}|Y=0}(\mathbf{X}_i; \theta_0)} \cdot \frac{f_Y(Y_i = 1; p)}{f_Y(Y_i = 0; p)} = \frac{f_{Y|\mathbf{X}}(Y_i = 1 \mid \mathbf{X}_i; \beta)}{f_{Y|\mathbf{X}}(Y_i = 0 \mid \mathbf{X}_i; \beta)} \cdot \frac{f_{\mathbf{X}}(\mathbf{X}_i; \theta)}{f_{\mathbf{X}}(\mathbf{X}_i; \theta)}. \qquad (7)$$

The last ratio in condition (7), $f_{\mathbf{x}}(\mathbf{X}_i; \theta) / f_{\mathbf{x}}(\mathbf{X}_i; \theta)$ = 1 and drops out. Given condition (7) follows from condition (6), condition (7) is a sufficient condition for the compatibility of the conditional and inverse conditional distributions. Thus, condition (7) is sufficient for the existence of the joint distribution given by equation (3) and in turn the underlying statistical model being developed.

Assume that $f_{Y|\mathbf{X}}(Y_i \mid \mathbf{X}_i; \beta)$ is a conditional Bernoulli probability mass function with the following functional form:

---

[2] This is the standard approach adopted in the statistics literature for viewing conditional distributions when the conditioning variable is binary (e.g. Kay and Little 1987; Lauritzen and Wermuth 1989; Oklin and Tate 1961; Scrucca and Weisberg 2004; Tate 1954; and Warner 1963).

[3] This follows from Theorem 1.2 in Arnold *et al*. (1999, p. 8). In this case, the theorem states that the condition: $f_{Y|\mathbf{X}}(Y_i \mid \mathbf{X}_i; \beta) / f_{\mathbf{X}|Y}(\mathbf{X}_i; \theta_j) = u(Y_i) \cdot v(\mathbf{X}_i)$ must hold for the two conditional distributions $f_{Y|\mathbf{X}}(Y_i \mid \mathbf{X}_i; \beta)$ and $f_{\mathbf{X}|Y}(\mathbf{X}_i; \theta_j)$ to be compatible and ensure that $f(Y_i, \mathbf{X}_i; \varphi)$ exists. The function $u(Y_i)$ is the marginal distribution of Y_i and $v(\mathbf{X}_i)$ is the inverse multivariate distribution of $\mathbf{X}_i$. Letting $d_j = f_{Y|\mathbf{X}}(Y_i = j \mid \mathbf{X}_i; \beta) / f_{\mathbf{X}|Y=j}(\mathbf{X}_i; \theta_j)$ for $j = 0,1$, then $d_1 / d_0 = [u(Y_i = 1) \cdot v(\mathbf{X}_i)] / [u(Y_i = 0) \cdot v(\mathbf{X}_i)] = p/(1-p)$, where d_1 / d_0 is given by condition (6). It follows from this result that the ratio d_1 / d_0 does not depend on $\mathbf{X}_i$ and for the given specification of Y_i and $\mathbf{X}_i$ this ratio is constant.

$$f_{Y|\mathbf{X}}(Y_i \mid \mathbf{X}_i;\beta) = h(\mathbf{X}_i;\beta)^{Y_i}\left[1-h(\mathbf{X}_i;\beta)\right]^{1-Y_i}, \tag{8}$$

where $h(\mathbf{X}_i;\beta):\mathbf{R}^K \times \Theta_\beta \to [0,1]$ and $\beta \in \Theta_\beta$, the parameter space associated with β. The mass function specified by equation (8) can be shown to satisfy the usual properties of a density function. Substituting equation (8) into (7) and $f_Y(Y_i;p) = p^{Y_i}(1-p)^{1-Y_i}$:

$$\frac{f_{\mathbf{X}|Y=1}(\mathbf{X}_i;\theta_1)}{f_{\mathbf{X}|Y=0}(\mathbf{X}_i;\theta_0)} \cdot \frac{p}{(1-p)} = \frac{h(\mathbf{X}_i;\beta)}{1-h(\mathbf{X}_i;\beta)}, \tag{9}$$

which implies that:

$$h(\mathbf{X}_i;\beta) = \frac{p \cdot f_{\mathbf{X}|Y=1}(\mathbf{X}_i;\theta_1)}{p \cdot f_{\mathbf{X}|Y=1}(\mathbf{X}_i;\theta_1) + (1-p) \cdot f_{\mathbf{X}|Y=0}(\mathbf{X}_i;\theta_0)}.^{4} \tag{10}$$

Given the general properties of density functions, the range of $h(\mathbf{X}_i;\beta)$ is $[0,1]$, justifying the assumption that $h(\mathbf{X}_i;\beta):\mathbf{R}^K \times \Theta_\beta \to [0,1]$. Thus, $h(\mathbf{X}_i;\beta)$ provides a general specification for the conditional mean of a conditional binary stochastic process.

A more intuitive and practical choice for $h(\mathbf{X}_i;\beta)$ can be found by using the identity $f(.) = \exp(\ln f(.))$, and after rearranging some terms:

$$h(\mathbf{X}_i;\beta) = \frac{\exp\{\eta(\mathbf{X}_i;\beta)\}}{1+\exp\{\eta(\mathbf{X}_i;\beta)\}} = \left[1+\exp\{-\eta(\mathbf{X}_i;\beta)\}\right]^{-1}, \tag{11}$$

where $\eta(\mathbf{X}_i;\beta) = \ln\left(\frac{f_{\mathbf{X}|Y=1}(\mathbf{X}_i;\theta_1)}{f_{\mathbf{X}|Y=0}(\mathbf{X}_i;\theta_0)}\right) + \kappa$ and $\kappa = \ln\left(\frac{p}{1-p}\right)$. Written as the composite function, $h(\eta(\mathbf{X}_i;\beta))$, $h(.)$ represents the logistic cumulative density function (the transformation function or inverse link function) and $\eta(.;.)$ represents the traditional index function (or predictor). The relationship in equation (11) will usually involve a reparameterization of the form $\beta = \beta(\theta_j, j = 0,1)$, where $\beta(.):\Theta_\theta \to \Theta_\beta$ and Θ_θ is the parameter space associated with θ_j, $j = 0,1$.

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

[4] The same formula can be derived from the relationship between discriminant analysis and logistic regression using Bayes formula (see Cox and Snell 1989), but this approach does not highlight the importance of the joint distribution when considering other forms of dependence or heterogeneity.

$$Y_i = E(Y_i \mid \mathbf{X}_i = \mathbf{x}_i) + u_i \,. \tag{12}$$

The SGM can contain higher order conditional moment functions when they capture additional systematic information in the data.

The regression function for the conditional stochastic process $\{Y_i \mid \mathbf{X}_i = \mathbf{x}_i, i = 1,..., N\}$ is:

$$Y_i = h(\mathbf{x}_i;\beta) + u_i = [1 + \exp\{-\eta(\mathbf{x}_i;\beta)\}]^{-1} + u_i \,.^{5} \tag{13}$$

The distribution of the error term u_i is given by (Maddala 1983):

u_i	$1-h(\mathbf{x}_i;\beta)$	$-h(\mathbf{x}_i;\beta)$
$f_u(u_i)$	$h(\mathbf{X}_i;\beta)$	$1-h(\mathbf{X}_i;\beta)$

where $E(u_i) = 0$ and $Var(u_i) = h(\mathbf{X}_i;\beta)(1-h(\mathbf{X}_i;\beta))$. If $\mathbf{X}_i$ is discrete, then $f_u(u_i)$ will be discrete; but if $\mathbf{X}_i$ is continuous, then $f_u(u_i)$ will be multimodal or a mixture distribution.

Equation (13) represents the SGM for a family of statistical models known as the Bernoulli Regression Model, which is summarized with probabilistic model assumptions in Table 1. The first three model assumptions, i.e. distributional, functional form and heteroskedasticity, arise from the assumed probability mass function given by equation (8). The homogeneity and independence assumptions follow from the IID reduction assumptions (i.e. see condition (4)) made about the joint vector stochastic process $\{(Y_i, \mathbf{X}_i), i = 1,..., N\}$. The independence assumption concerning the joint vector stochastic process $\{(Y_i, \mathbf{X}_i), i = 1,..., N\}$ implies that the conditional stochastic process $\{Y_i \mid \mathbf{X}_i = \mathbf{x}_i, i = 1,..., N\}$ is independent over *i*. Each of the assumptions is briefly discussed to emphasize the strengths of using the PR approach

Distributional: The distributional assumption is by nature conditional Bernoulli given the stochastic process $\{Y_i \mid \mathbf{X}_i = \mathbf{x}_i, i = 1,..., N\}$ is a conditional binary choice process. There is no need to test this assumption, as it is assured by inspection of the observed data.

Functional Form: The functional form assumption has two components: the transformation (or link) and index (or predictor) functions. The functional form of the transformation function is derived mathematically from the general functional form for *h*(.) given by equation (10), arising naturally from the joint distribution given by equation (1). Thus, the logistic transformation function provides an obvious choice for modelling binary choice process and need not be tested empirically. Other

[5] The conditional variance $Var(Y_i \mid \mathbf{X}_i = \mathbf{x}_i) = h(\mathbf{x}_i;\beta)\cdot(1-h(\mathbf{x}_i;\beta))$ is bounded given the range of $h(\mathbf{X}_i;\beta)$ is $[0,1]$.

Table 1. Bernoulli Regression Model

SGM: $Y_i = h(\mathbf{x}_i;\beta) + u_i$, $i = 1,\ldots,N$,
where

u_i	$1-h(\mathbf{x}_i;\beta)$	$-h(\mathbf{x}_i;\beta)$
$f_u(u_i)$	$h(\mathbf{X}_i;\beta)$	$1-h(\mathbf{X}_i;\beta)$

Assumptions

Distributional: $Y_i \mid \mathbf{X}_i = \mathbf{x}_i \sim Bin(1, h(\mathbf{X}_i;\beta))$ (conditional Bernoulli).

Functional Form: $h(\mathbf{x}_i;\beta) = E(Y_i \mid \mathbf{X}_i = \mathbf{x}_i) = [1+\exp\{-\eta(\mathbf{x}_i;\beta)\}]$, where

$$\eta(\mathbf{x}_i;\beta) = \ln\left[\frac{f_{\mathbf{X}|Y=1}(\mathbf{X}_i;\theta_1)}{f_{\mathbf{X}|Y=0}(\mathbf{X}_i;\theta_0)}\right] + \kappa \text{ and } \beta = \beta(\theta_j, j = 0,1).$$

Heteroskedasticity: $Var(Y_i \mid \mathbf{X}_i = \mathbf{x}_i) = h(\mathbf{x}_i;\beta)(1-h(\mathbf{x}_i;\beta))$.

Homogeneity: β is not a function of the index $i = 1,\ldots,N$.

Independence: $\{Y_i \mid \mathbf{X}_i = \mathbf{x}_i, i = 1,\ldots,N\}$ is an independent stochastic process.

functional forms for the transformation function may not be as tractable or derivable from a proper joint distribution.[6]

The functional form of the index function is determined by the inverse conditional distribution $f_{\mathbf{X}|Y}(\mathbf{X}_i;\theta_j)$. Thus, the functional form assumption amounts to a distributional assumption concerning the functional form of the inverse conditional distribution. Specification of the index function will be discussed in detail in the next section. Graphical methods may be useful in helping to determine the needed covariates to include in the index function (see Scrucca and Weiserg 2004). Furthermore, a potential misspecification test for the functional form of the index function would be to test for the significance of RESET type terms of the fitted index function (e.g. $\hat{\eta}^2$ and $\hat{\eta}^3$) in the logistic regression as additional covariates using likelihood ratio or Wald test statistics.

[6] While the use of the logistic regression framework provides a tractable approach for specifying the Bernoulli Regression Model, this approach does not rule out the use of other transformation functions or the direct estimation of equation (10). Other transformation functions have been suggested (for examples see Aranda-Ordaz 1981; Aldrich and Nelson 1984; Greene 2003; Maddala 1983; and McFadden 1984). The other most common transformation function is the probit specification or normal cdf. The probit specification provides a unique problem in that the normal cdf cannot be expressed in a finite number of additions, multiplications, root extractions or subtractions, making the derivation of such a model from equation (10) a significant challenge (Hogg and Craig 1978).

Heteroskedasticity: The BRM is inherently heteroskedastic, but the skedastic function need not be estimated, given that it is an explicit function of the conditional mean. This result is in contrast to the practice of correcting for heteroskedasticity in the literature.[7]

Homogeneity: The model as specified assumes that the parameters of the model, β, are not functions of the index *i*. If the observed data are panel or exhibit heterogeneity across groups or time (e.g. trends), then this assumption is violated and the model should be corrected for any departures. In this situation, the inclusion of appropriate fixed effects or estimation techniques, such as mixed logit (see Train 2003), may be helpful.

Corrections for heterogeneity in the conditional variance (skedastic) function may not be needed. The conditional variance function is explicitly a function of the conditional mean. The only parameters that would exhibit heterogeneity in the variance are β, which can be corrected for by properly modelling the conditional mean. Variance heterogeneity will arise due to heterogeneity exhibited by Y_i or $\mathbf{X}_i$. For Y_i, this is captured by shifts in the parameter *p*. For $\mathbf{X}_i$, the variance/covariance matrix of the inverse conditional distribution may exhibit heterogeneity, but this will be captured by the parameter vector β, as $\beta = \beta(\theta_j, j = 0,1)$ (i.e. a function of these variance/covariance parameters). Thus, the appropriateness of alternative variance correction methods may need to be re-examined.

Independence: While independence across space and time is assumed for the BRM specification in Table 1, it may be the case that the model exhibits temporal or spatial dependence. In this situation, the reduction assumption concerning dependence may be modified, which will give rise to an alternative specification of the model. For example, consider the case where Y_i is temporally dependent following a Markov(1) process and $\mathbf{X}_i$ is temporally independent. Then the reduction of the joint distribution takes the following form:

$$
\begin{aligned}
f(Y_1,\ldots,Y_N,\mathbf{X}_1,\ldots,\mathbf{X}_N;\phi) &\overset{M(1)}{=} f_1(Y_1,\mathbf{X}_1;\varphi_1)\prod_{i=2}^{N} f_i(Y_i,\mathbf{X}_i \mid Y_{i-1};\varphi_i) \\
&\overset{Stationarity}{=} f_1(Y_1,\mathbf{X}_1;\varphi_1)\prod_{i=2}^{N} f(Y_i,\mathbf{X}_i \mid Y_{i-1};\varphi) \\
&= f_1(Y_1,\mathbf{X}_1;\varphi_1)\prod_{i=2}^{N} f_{Y|\mathbf{X}}(Y_i \mid Y_{i-1},\mathbf{X}_i;\varphi)\cdot f_{\mathbf{X}}(\mathbf{X}_i \mid Y_{i-1};\theta)).
\end{aligned}
$$

Letting $f_Y(Y_{i-1};q_j)$ be the inverse conditional distribution of Y_{i-1} given $Y_i = j$, the index function in this situation takes the following form:

[7] Another potential issue concerning the skedastic function is over-dispersion, which can arise when there are repeated observations from the same respondent that are not independent and do not have the same likelihood of success. Correcting for this problem has been explored in the literature (e.g. Williams 1982).

$$\eta(\mathbf{X}_i;\beta)=\ln\left(\frac{f_{\mathbf{X}|Y=1}(\mathbf{X}_i\mid Y_{i-1};\theta_1)}{f_{\mathbf{X}|Y=0}(\mathbf{X}_i\mid Y_{i-1};\theta_0)}\right)+\ln\left(\frac{f_Y(Y_{i-1};q_0)}{f_Y(Y_{i-1};q_1)}\right)+\kappa.$$

This index function can be specified by postulating appropriate conditional distributions for each of the inverse conditional distributions. This example could potentially serve for modelling spatial dependence for the nearest neighbour, as well. Thus, the BRM framework has flexibility for taking account of alternative forms of dependence. The dependence assumption used for model specification in this framework can be tested. This will help ensure that the observed data provide support for the dependence assumption. If not, the BRM provides a systematic framework for re-examining how to impose alternative forms of dependence.

The flexibility of the PR approach allows for alternative specifications of the BRM that can accommodate a myriad of alternative empirical modelling situations. Furthermore, the tractability of the logistic regression model (given the wide array of existing software packages for estimating such model types) combined with the PR approach provides an accessible and theoretically sound approach to specifying and estimating Bernoulli Regression Models using observational data. The probabilistic reduction approach extends the work of Kay and Little (1987) by rationalizing the dependence of the specification on the inverse conditional distribution and formalizing the path from the joint distribution to an estimable model.

4 Model Specification, Estimation and Simulation

4.1 Model Specification

Functional specification of the BRM amounts to determining the functional form for the index function $\eta(\mathbf{X}_i;\beta)$ via proper specification of the inverse conditional distribution. This is examined in detail for both univariate and multivariable cases.

4.1.1 Univariate Models

While most discrete choice models will have multiple explanatory variables, it will be of interest to focus first on the univariate case. Kay and Little (1987) and Scrucca and Weisberg (2004) derived a number of univariate cases that satisfy the linearity property when the inverse conditional distribution is conditionally distributed beta, binomial, gamma, geometric, log-normal, normal or Poisson. Each of these distributions belongs to the simple exponential family given by equation (2). This paper contributes an additional two models to this group, the logarithmic and Pareto. Table 2 provides the functional forms for $h(x_i;\beta)$ and $\eta(x_i;\beta(\theta_j,\, j=0,1))$ needed to obtain a properly specified univariate BRM for each of these inverse conditional distribution assumptions. The models presented in Table 2 emphasize the relationship between the parameters of the BRM, β, and the inverse conditional distributions, $\theta_j,\, j=0,1$. In addition, the last column in Table 2 provides the needed transformation(s) of the explanatory variable to include in the index function, which will be linear in the parameters.

Table 2. Specification of $\eta(x_i;\beta)$ with one explanatory variable and conditional distribution, $f_{X|Y}(X_i;\theta_j)$, for $j=0,1$.

Distribution of X_i given Y_i	$f_{X\|Y}(X_i;\theta_j)=$ [3]	$h(x_i;\beta)=$ [4]	**Terms required in $\eta(x_i;\beta)$** [5]
Beta[1,2]	$\frac{X_i^{\alpha_j-1}(1-X_i)^{\gamma_j-1}}{\mathbf{B}[\alpha_j,\gamma_j]}$, 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 $\beta_2=(\gamma_1-\gamma_0)$.	$(1,\ln(x_i),\ln(1-x_i))$ If $\alpha_1=\alpha_0$: $(\mathbf{1},\ln(1-x_i))$ If $\gamma_1=\gamma_0$: $(\mathbf{1},\ln(x_i))$
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 $n=1,2,3,\ldots$	$[1+\exp\{-\beta_0-\beta_1x_i\}]^{-1}$, where $\beta_0=\left[\kappa+n\ln\left(\frac{1-\theta_1}{1-\theta_0}\right)\right]$ and $\beta_1=\ln\left(\frac{\theta_1}{\theta_0}\right)-\ln\left(\frac{1-\theta_1}{1-\theta_0}\right)$.	$(\mathbf{1},x_i)$
Gamma[1,2]	$\frac{1}{\gamma_j\Gamma[\alpha_j]}\left(\frac{X_i}{\gamma_j}\right)^{\alpha_j-1}\exp\left\{-\frac{X_i}{\gamma_j}\right\}$, where $(\alpha_j,\gamma_j)\in\mathbf{R}_+^2$ and $X_i\in\mathbf{R}_+$.	$[1+\exp\{\beta_0+\beta_1x_i+\beta_2\ln(x_i)\}]^{-1}$, 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 $\beta_2=(\alpha_1-\alpha_0)$.	$(1,x_i,\ln(x_i))$ If $\alpha_1=\alpha_0$: $(\mathbf{1},x_i)$ If $\gamma_1=\gamma_0$: $(\mathbf{1},\ln(x_i))$
Geometric[2]	$\theta_j(1-\theta_j)^{X_i-1}$, where $0\le\theta_j\le 1$ and $X_i=1,2,3,\ldots$	$[1+\exp\{-\beta_0-\beta_1x_i\}]^{-1}$, 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 $\beta_1=\ln\left(\frac{1-\theta_1}{1-\theta_0}\right)$.	$(\mathbf{1},x_i)$
Logarithmic	$\alpha_j\left(\frac{\theta_j^{X_i}}{X_i}\right)$, where $\alpha_j=-[\ln(1-\theta_j)]^{-1}$, $0<\theta_j<1$ and $X_i=1,2,3,\ldots$	$[1+\exp\{-\beta_0-\beta_1x_i\}]^{-1}$, where $\beta_0=\left[\kappa+\ln\left(\frac{\alpha_1}{\alpha_0}\right)\right]$ and $\beta_1=\ln\left(\frac{\theta_1}{\theta_0}\right)$.	$(\mathbf{1},x_i)$

Table 2. Continued

Distribution of X_i given Y_i	$f_{X\|Y}(X_i;\theta_j)=$ [3]	$h(x_i;\beta)=$ [4]	**Terms required in $\eta(x_i;\beta)$** [5]
Log-Normal[2]	$\frac{1}{X_i}\cdot\frac{1}{\sigma_j\sqrt{2\pi}}\exp\left\{-\frac{(\ln(X_i)-\mu_j)^2}{2\sigma_j^2}\right\}$, where $\mu_j\in\mathbf{R}$, $\sigma_j^2\in\mathbf{R}_+$ and $X_i\in\mathbf{R}$.	$\left[1+\exp\left\{\beta_0+\beta_1\ln(x_i)+\beta_2(\ln(x_i))^2\right\}\right]^{-1}$, 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 $\beta_2=\left(\frac{1}{2\sigma_0^2}-\frac{1}{2\sigma_1^2}\right)$.	$(\mathbf{1},\ln(x_i),\ln(x_i)^2)$ If $\sigma_1^2=\sigma_0^2$: $(\mathbf{1},\ln(x_i))$
Normal[1,2]	$\frac{1}{\sigma_j\sqrt{2\pi}}\exp\left\{-\frac{1}{2\sigma_j^2}(X_i-\mu_j)^2\right\}$, where $\mu_j\in\mathbf{R}$, $\sigma_j^2\in\mathbf{R}_+$ and $X_i\in\mathbf{R}$.	$\left[1+\exp\left\{\beta_0+\beta_1x_i+\beta_2x_i^2\right\}\right]^{-1}$, 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 $\beta_2=\left(\frac{1}{2\sigma_0^2}-\frac{1}{2\sigma_1^2}\right)$.	$(\mathbf{1},x_i,x_i^2)$ If $\sigma_1^2=\sigma_0^2$: $(\mathbf{1},x_i)$
Pareto		$[1+\exp\{-\beta_0-\beta_1\ln(x_i)\}]^{-1}$, where $\beta_0=\left[\kappa+\ln\left(\frac{\theta_1}{\theta_0}\right)+(\theta_1-\theta_0)\ln(x_0)\right]$ and $\beta_1=(\theta_0-\theta_{1.})$	$(\mathbf{1},\ln(x_i))$
Poisson[1,2]	$\frac{e^{-\theta_j}\theta_j^{X_i}}{X_i!}$, where $\theta_j>0$ and $X_i=1,2,3,...$	$[1+\exp\{-\beta_0-\beta_1x_i\}]^{-1}$, where $\beta_0=[\kappa+\theta_0-\theta_{\cdot}]$ and $\beta_1=\ln\left(\frac{\theta_1}{\theta_0}\right)$.	$(\mathbf{1},x_i)$

[1] *Source*: Kay and Little (1987).
[2] *Source*: Scrucca and Weisberg (2004) and Cook and Weisberg (1999).
[3] *Source*: Spanos (1999). $\mathbf{B}[\ ]$ represents the beta function and $\Gamma[\ ]$ represents the gamma function.
[4] $\kappa=\ln(\pi_1)-\ln(\pi_0)$
[5] '**1**' refers to the inclusion of an intercept term in the model, which is needed to represent κ

Parameterizations can differ depending on whether or not the location and/or shape parameters of the distribution vary with Y_i. For example, if $f_{\mathbf{X}|Y}(X_i;\theta_j)$ is conditional normal with both the mean (μ_j) and variance (σ_j^2) dependent on Y_i then the transformations in the index function would include x_i and x_i^2, as well as an intercept term. If the mean varies with Y_i, but the variance does not (i.e. $\sigma_1^2 = \sigma_0^2$), then the x^2 term can be dropped.[8] Two other interesting cases include when $f_{\mathbf{X}|Y}(X_i;\theta_j)$ is distributed conditional chi-square or exponential. Both are special cases of the conditional gamma distribution in Table 2. If $f_{\mathbf{X}|Y}(X_i;\theta_j)$ is conditional exponential then the index function would include an intercept and the term x_i. If $f_{\mathbf{X}|Y}(X_i;\theta_j)$ is conditional chi-square then the index function would include an intercept and the term $\ln(x_i)$.

While the inverse conditional distributions in Table 2 result in index functions linear in the parameters, this may not always be the case. Examples of inverse conditional distributions that give rise to index functions nonlinear in the parameters include the F, extreme value[9] and logistic distributions. In such cases, one option is to explicitly specify the inverse conditional distribution and estimate the model using equation (10). This may be difficult numerically if the parameterization of the model, given the mapping $\beta = \beta(\theta_j, j = 0,1)$, is not well-defined. Another option is to transform X_i so that it has one of the inverse conditional distributions specified in Table 2. Consider the following case, which is explored in more detail later in the paper. If the inverse conditional distribution follows a Weibull distribution, $W(\alpha_j, \gamma)$, then an appropriate transformation of the explanatory variable X_i would be X_i^γ, given $X_i^\gamma \mid Y_i = j$ is exponentially distributed, $Exp(\alpha_j)$. In this case, the index function is:

$$\eta(\mathbf{x}_i;\beta,\gamma) = \beta_0 + \beta_1 x_i^\gamma, \tag{14}$$

where $\beta_0 = \left[\kappa + \gamma \ln\left(\frac{\alpha_0}{\alpha_1}\right)\right]$ and $\beta_1 = \left(\frac{1}{\alpha_0}\right)^\gamma - \left(\frac{1}{\alpha_1}\right)^\gamma$.[10] This example is the first known attempt to derive a BRM that is nonlinear in the parameters using a conditional Weibull distribution.

[8] Scrucca and Weisberg (2004) provide other conditions under which x^2 can be dropped when $\eta(\mathbf{x}_i;\beta)$ is monotone.

[9] The extreme value distribution being referred to here is the Gumbel type extreme value distribution (see Spanos 1999; p. 139).

[10] The estimable parameters of the model are β_0, β_1, and γ. Furthermore, the range of γ does not restrict the range of β_0 or β_1.

4.1.2 Multivariable Models using Multivariate Inverse Conditionals

While the case with one explanatory variable is readily manageable, as the number of explanatory variables increase, specification of the model becomes more complex. There are a number of different approaches for model specification in this instance. The first approach for the multivariable case is to explicitly specify the multivariate distribution $f_{\mathbf{X}|Y}\left(\mathbf{X}_i;\theta_j\right)$ and derive the model. This approach is particularly useful if all the explanatory variables follow the same distribution. For example, if $f_{\mathbf{X}|Y}\left(\mathbf{X}_i;\theta_j\right)$ is multivariate normal with the same covariance matrix for $j=0,1$, then:

$$\eta(\mathbf{x}_i;\beta)=\beta_0+\sum_{k=1}^{K}\beta_k x_{k,i}$$

On the other hand, if the covariance matrix exhibits heterogeneity and is not equal for $j=0,1$, then:

$$\eta(\mathbf{x}_i;\beta)=\beta_0+\sum_{k=1}^{K}\beta_k x_{k,i}+\sum_{j=1}^{K}\sum_{l\geq j}^{K}\beta_{j,l}x_{j,i}x_{l,i} \tag{15}$$

(Kay and Little 1987). If $f_{\mathbf{X}|Y}\left(\mathbf{X}_i;\theta_j\right)$ is multivariate Bernoulli made up of K explanatory variables, then the index function would include an intercept, as well as first order , second order, and so on up to order K interaction terms (e.g. Kay and Little 1987).

A more general multivariate distributional assumption can be utilized following the logistic discriminant model proposed by Day and Kerridge (1967). In this case, the inverse conditional distribution is:

$$f_{\mathbf{X}|Y}\left(\mathbf{X}_i;\theta_j\right)=\alpha_j\exp\left\{-\frac{1}{2}(\mathbf{X}_i-\Lambda_j)'\mathrm{A}_j^{-1}(\mathbf{X}_i-\Lambda_j)\right\}\delta(\mathbf{X}_i), \tag{16}$$

where Λ_j are mean vectors for $j=0,1$, A_j is a covariance matrix for $j=0,1$, and $\delta(\mathbf{X}_i)$ is a non-negative scalar function of $\mathbf{X}_i$. When $\delta(\mathbf{X}_i)=1$, the density given by equation (16) is the multivariate normal distribution. When $\delta(\mathbf{X}_i)\neq 1$, the density function can represent a wide range of alternatives, including skewed distributions (see Byth and McLachlan 1980). The advantage of this distributional assumption is that $\delta(\mathbf{X}_i)$ does not have to be specified explicitly to arrive at a estimable model (Day and Kerridge 1967). If the distributional assumption given by equation (16) holds (which should be tested *a posteriori*) then the functional form of the model takes that given by equation (15).

A number of bivariate distributions give rise to tractable models, which are derived by the authors below. If $f_{\mathbf{X}|Y}\left(\mathbf{X}_i;\theta_j\right)$ is a conditional bivariate gamma distribution of the form:

$$f_{\mathbf{X}|Y}\left(X_{1,i}, X_{2,i}; \theta_j\right) = \frac{\alpha_j \theta_{1,j} \theta_{2,j}}{\Gamma\left[\theta_{1,j}\right]\Gamma\left[\theta_{2,j}\right]} e^{-\alpha_j X_{2,i}} X_{1,i}^{\theta_{1,j}-1}\left(X_{2,i} - X_{1,i}\right)^{\theta_{2,j}-1},$$

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:

$$\eta\left(x_{1,i}, x_{2,i}; \beta\right) = \beta_0 + \beta_1 x_{2,i} + \beta_2 \ln\left(x_{1,i}\right) + \beta_3 \ln\left(x_{2,i} - x_{1,i}\right).$$

If $f_{\mathbf{X}|Y}\left(\mathbf{X}_i; \theta_j\right)$ is a conditional bivariate beta distribution of the form:

$$f_{\mathbf{X}|Y}\left(X_{1,i}, X_{2,i}; \theta_j\right) = \left(\frac{\Gamma\left(\alpha_j + \delta_j + \gamma_j\right)}{\Gamma\left(\alpha_j\right)\Gamma\left(\delta_j\right)\Gamma\left(\gamma_j\right)}\right)\left[X_{1,i}^{\alpha_j - 1} \cdot X_{2,i}^{\delta_j - 1} \cdot \left(1 - X_{1,i} - X_{2,i}\right)^{\gamma_j - 1}\right] \quad (17)$$

where $X_{1,i} \geq 0$, $X_{2,i} \geq 0$ and $X_{1,i} + X_{2,i} \leq 1$ for $i = 1,\ldots,N$; $\left(\alpha_j, \delta_j, \gamma_j\right) > 0$ for $j = 0,1$; and $\Gamma(.)$ is the gamma function (Spanos 1999); then:

$$\eta\left(x_{1,i}, x_{2,i}; \beta\right) = \beta_0 + \beta_1 \ln\left(x_{1,i}\right) + \beta_2 \ln\left(x_{2,i}\right) + \beta_3 \ln\left(1 - x_{1,i} - x_{2,i}\right).$$

4.1.3 Multivariable Models Assuming Independence of Explanatory Variables

The second approach for the multivariate case involves examining and testing the conditional dependence structure (on Y_i) between the explanatory variables. While this may be problematic when comparing discrete and continuous variables, traditional measures of correlation and association may still be useful in gauging independence (see Tate 1954; Oklin and Tate 1961).

The simplest scenario is when all the explanatory variables are conditionally independent. In this case, $f_{\mathbf{X}|Y}\left(\mathbf{X}_i; \theta_j\right) = \prod_{k=1}^{K} f_{X_k|Y}\left(X_{k,i}; \theta_{k,j}\right)$, making the index function $\eta\left(\mathbf{x}_i; \beta\right) = \sum_{k=1}^{K} \ln\left(\frac{f_{X_k|Y=1}\left(X_{k,i}; \theta_{k,1}\right)}{f_{X_k|Y=0}\left(X_{k,i}; \theta_{k,0}\right)}\right) + \kappa$ (Kay and Little 1987). In such a case, the results in Table 2 can be used to specify the model. For example, consider the case of two independent explanatory variables X_1 and X_2, where $f_{X_1|Y}\left(X_1; \theta_{1,j}\right)$ is normally distributed and $f_{X_2|Y}\left(X_2; \theta_{2,j}\right)$ is gamma distributed with all parameters dependent on Y_i in both distributions. Then according to Table 2, the index function would be linear in the parameters and include an intercept term, as well as the terms x_1, x_1^2, x_2 and $\ln\left(x_2\right)$.

4.1.4 Multivariable Models using Sequential Conditioning

If some or none of the explanatory variables are conditionally independent, then another approach for decomposing $f_{\mathbf{X}|Y}\left(\mathbf{X}_i;\theta_j\right)$ is sequential conditioning, i.e.:

$$f_{\mathbf{X}|Y}\left(\mathbf{X}_i;\theta_j\right)=f_{X_1|Y}\left(X_{1,i};\theta_{1,j}\right)\prod_{k=2}^{K}f_{X_k|Y}\left(X_{k,i}\mid X_{k-1,i},\ldots,X_{1,i};\theta_{k,j}\right).$$

Given the potential complexity of this approach, it can be combined with the previous approaches to reduce the dimensionality and increase the tractability of the problem. Consider the following original derivation by the authors where $f_{\mathbf{X}|Y}\left(\mathbf{X}_i;\theta_j\right)$ is the product of conditional binomial, exponential and bivariate beta distributions. Let

$$f_{\mathbf{X}|Y}\left(X_{1,i},X_{2,i},X_{3,i};X_{4,i};\theta_j\right)=f_{\mathbf{X}_1|Y}\left(X_{1,i},X_{2,i};\boldsymbol{\theta}_{\mathbf{1},j}\right)\cdot f_{\mathbf{X}_2|Y}\left(X_{3,i},X_{4,i};\boldsymbol{\theta}_{\mathbf{2},j}\right),$$

where $X_{1,i}$ and $X_{2,i}$ are conditionally independent of $X_{3,i}$ and $X_{4,i}$. Now assume that (i) $X_{1,i}$ given $Y_i=j$ is distributed $Bin(1,q_j)$; (ii) $X_{2,i}$ given $X_{1,i}=l$ $(l=0,1)$ and $Y_i=j$ is distributed exponential, i.e.

$$f_{X_2|X_1,Y}\left(X_{2,i};\theta_{j,l}\right)=\frac{1}{\theta_{j,l}}\exp\left\{-\frac{X_{2,i}}{\theta_{j,l}}\right\};$$

making:

$$\begin{aligned}f_{\mathbf{X}_1|Y}\left(X_{1,j},X_{2,j};\boldsymbol{\theta}_{\mathbf{1},j}\right)&=f_{X_2|X_1,Y}\left(X_{2,i};\theta_{j,l}\right)\cdot f_{X_1|Y}\left(X_{1,i};q_j\right)\\&=\left[\frac{q_j}{\theta_{j,1}}\exp\left\{-\frac{X_{2,i}}{\theta_{j,1}}\right\}\right]^{X_{1,i}}\left[\frac{\left(1-q_j\right)}{\theta_{j,0}}\exp\left\{-\frac{X_{2,i}}{\theta_{j,0}}\right\}\right]^{1-X_{1,i}}.\end{aligned}$$

Let $X_{3,i}$ and $X_{4,i}$ given $Y_i=j$ be jointly distributed bivariate beta following equation (17). Based on these assumptions:

$$\eta\left(\mathbf{x}_i;\beta\right)=\beta_0+\beta_1x_{1.i}+\beta_2x_{2,i}+\beta_3x_{1,i}x_{2,i}+\beta_4\ln\left(x_{3,i}\right)+\beta_5\ln\left(x_{4,i}\right)+\beta_6\ln\left(1-x_{3,i}-x_{4,i}\right).$$

Kay and Little (1987) provide some other bivariate examples involving discrete and continuous variables.

4.1.5 A General Approach for Multivariable Models

If the decomposition of $f_{\mathbf{X}|Y}\left(\mathbf{X}_i;\theta_j\right)$ involved an unknown multivariate conditional distribution of continuous variables, then it becomes considerably more difficult to derive the specification of $h\left(\mathbf{x}_i;\beta\right)$. A strategy may be to transform the variables to achieve a more estimable form of the model. Consider the more general case of a

mixture of continuous $(\mathbf{X}_{1,i})$ and discrete variables $(\mathbf{X}_{2,i})$. Now use sequential conditioning of the inverse conditional distribution to obtain:

$$f_{\mathbf{X}|Y}(\mathbf{X}_{1,i},\mathbf{X}_{2,i};\eta_{1,i}) = f_{\mathbf{X}_1|\mathbf{X}_2,Y}(\mathbf{X}_{1,i}|\,\mathbf{X}_{2,i};\boldsymbol{\theta}_{\mathbf{1},j})\cdot f_{\mathbf{X}_2|Y}(\mathbf{X}_{2,i};\boldsymbol{\theta}_{\mathbf{2},j}),$$

where the parameters $\boldsymbol{\theta}_{\mathbf{1},j}$ are dependent upon $\mathbf{X}_{2,i}$ and Y_i. To obtain an operational model, appropriate transformations of the continuous variables $\mathbf{X}_{1,i}$ could be found to try and make $f_{\mathbf{X}_1|\mathbf{X}_2,Y}(\mathbf{X}_{1,i}|\,\mathbf{X}_{2,i};\boldsymbol{\theta}_{\mathbf{1},j})$ close to multivariate normal or the more flexible distribution following equation (16) (Scrucca and Weisberg 2004).[11] To transform the discrete variables, recode these variables as binary so that $f_{\mathbf{X}_2|Y}(\mathbf{X}_{2,i};\boldsymbol{\theta}_{\mathbf{2},j})$ is multivariate Bernoulli. In this case, $f_{\mathbf{X}_2|Y}(\mathbf{X}_{2,i};\boldsymbol{\theta}_{\mathbf{2},j})$ can be represented in log-linear form. That is:

$$f_{\mathbf{X}_2|Y}(\mathbf{X}_{2,i};\boldsymbol{\theta}_{\mathbf{2},j}) = \exp\left(\begin{array}{l} u_{0,j} + \sum_{k=1}^{K_2} u_{k,j} X_{2,k,i} + \sum_{k=1}^{K_2}\sum_{r>k}^{K_2} u_{k,r,j} X_{2,k,i} X_{2,r,i} \\ \qquad + \ldots + u_{1,2,\ldots,K_2\,j} X_{2,1,i} \cdots X_{2,K_2,i} \end{array}\right),$$

where $\mathbf{X}_{2,i}$ is a $(1\times K_2)$ vector of binary variables (Liang *et al*. 1992). Based on these distributional assumptions and assuming $\mathbf{X}_{1,i}$ is a $(1\times K_1)$ vector of normally distributed variables (or following the multivariate distribution given by equation (16)), the index function for the associated BRM would take the form:

$$\begin{aligned}\eta(\mathbf{x}_i;\beta) = &\left(\alpha_0 + \sum_{s=1}^{K_1}\alpha_s x_{1,s,i} + \sum_{s=1}^{K_1}\sum_{t\geq s}^{K_1}\alpha_{s,t} x_{1,s,i} x_{1,t,i}\right)\times \\ &\left(u_{0,j} + \sum_{k=1}^{K_2} u_k x_{2,k,i} + \sum_{k=1}^{K_2}\sum_{r>k}^{K_2} u_{k,r} x_{2,k,i} x_{2,r,i} + \ldots + u_{1,2,\ldots,K_2} x_{2,1,i} \cdots x_{2,K_2,i}\right).\end{aligned} \tag{18}$$

The index function given by equation (18) can be made linear in the parameters via reparameterization of the model by letting $\beta = \beta(\alpha,u)$. To further improve the tractability of the model, the number of interaction terms in $f_{\mathbf{X}_2|Y}(\mathbf{X}_{2,i};\boldsymbol{\theta}_{\mathbf{2},j})$ may need to be restricted to order $\overline{K} \leq K_2$, which can be tested *a posteriori*.

4.2 Estimation

The method of maximum likelihood can be used to estimate the parameters of the Bernoulli Regression Model. Gourieroux (2000) and Train (2003) provide background for estimation of logistic regression models, which is applicable for the BRM, as well. When $\eta(\mathbf{x}_i;\beta)$ is nonlinear in the parameters estimation becomes more difficult, because the likelihood function may no longer be globally concave and many

[11] For examples of appropriate transformations see Box and Cox (1964); Draper and Cox (1969); and Yeo and Johnson (2000).

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 (e.g. Newton-Raphson, quasi-Newton and conjugate gradient methods) 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).

4.3 Simulation

A significant benefit of using the probabilistic reduction approach for developing the BRM is that it provides a mechanism for randomly generating binary choice data using the relationship given by equation (5). 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 $f_{\mathbf{X}|Y}(\mathbf{X}_i;\theta_j)$ generate a realization of the vector stochastic process, $\{\mathbf{X}_i, i = 1,..., N\}$ using appropriate random number generators with the parameters given by $\theta_j = \theta_0$ when $Y_i = 0$ and $\theta_j = \theta_1$ when $Y_i = 1$.

No *a priori* theoretical interpretation is imposed on the generation process; it is purely statistical in nature.[12] The parameters β can be determined via the relationship $\beta(\theta_j, j = 0,1)$ as seen in examples presented above. Thus, this modelling framework provides a straightforward set-up to simulate binary choice processes.

To illustrate, consider the BRM given by equation (14) specified assuming a univariate inverse conditional Weibull distribution. Let $Y_i \sim Bin(1,0.6)$ and $X_i \mid Y_i = j$ be Weibull distributed with $\alpha_0 = 1, \alpha_1 = 1.4$ and $\gamma = 3$. In this case, $\beta_0 = -0.6040$, $\beta_1 = 0.6356$ and $\gamma = 3$ for the parameters in equation (14). To examine the asymptotic properties of the parameters β_0, β_1 and γ a Monte Carlo simulation was conducted with 1000 runs for sample sizes of N = 50, 100, 250, 500, 1000, 2500 and 5000. For each run, the regression equation given by equation (14) was estimated using a derivative-free algorithm developed by Nelder and Mead (1965).[13] The results of the simulation are reported in Table 3.

Two desirable asymptotic properties of estimators are consistency and asymptotic normality. For BRM or logistic regression models with nonlinear index functions, ensuring these properties hold is important for statistical reliability. To check for consistency, two verifiable conditions are: (i) $\lim_{N\to\infty} E(\hat{\beta}_N) = \beta$ and (ii) $\lim_{N\to\infty} \mathrm{var}(\hat{\beta}_N) = 0$, where $\hat{\beta}_N$ is the estimator of interest and N is the sample size (Spanos 1999). As the sample size (N) increases, the mean estimate for each parameter converges to its true value and the standard error converges to zero. Thus,

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

[13] It was found that this algorithm provided the best convergence properties for the given problem.

there is evidence that the estimators for each of the parameters are consistent. To examine asymptotic normality, we can examine the asymptotic skewness and kurtosis.[14] As *N* increases, the skewness and kurtosis for each coefficient converge to 0 and 3, respectively, providing evidence the estimators are asymptotically normal. In addition, the model was compared for each *N* to a more traditional specification with index function linear in the variables using a Hosmer-Lemeshow test (Hosmer and Lemeshow 2000). For each *N*, the traditional specification was rejected over 95 percent of the time in favour of the BRM specification.[15]

5 Empirical Application

The primary purpose of the empirical example is to illustrate the specification of a Bernoulli Regression Model in an applied setting and compare it to the traditional

Table 3. Monte Carlo Simulation Results for Univariate Bernoulli Regression Model with Inverse Conditional Weibull Distribution

Parameter	Number of Observations (*N*)	Mean	Standard Deviation	Skewness	Kurtosis	Min	Max
			True Value = -0.60				
β_0	N = 50	-1.38	2.84	-6.26	51.09	-30.58	0.86
	N = 100	-1.12	2.26	-7.79	78.15	-29.51	0.50
	N = 250	-0.66	0.45	-2.00	11.96	-4.35	0.28
	N = 500	-0.62	0.25	-0.56	3.88	-1.67	0.10
	N = 1000	-0.62	0.18	-0.55	3.76	-1.42	-0.20
	N = 2500	-0.61	0.11	-0.27	2.94	-0.95	-0.28
	N = 5000	-0.61	0.08	-0.03	3.15	-0.91	-0.31
			True Value = 0.64				
β_1	N = 50	1.36	2.92	6.07	48.36	0.00	31.13
	N = 100	1.14	2.34	7.47	72.76	0.00	29.91
	N = 250	0.68	0.50	2.24	13.03	0.00	4.82
	N = 500	0.64	0.28	0.84	4.54	0.05	2.04
	N = 1000	0.65	0.200	0.74	4.15	0.16	1.60
	N = 2500	0.64	0.13	0.34	2.95	0.27	1.07
	N = 5000	0.64	0.09	0.07	3.00	0.32	0.98
			True Value = 3.0				
γ	N = 50	4.67	4.35	2.42	11.64	-6.62	36.22
	N = 100	4.15	3.51	2.63	13.00	0.08	28.01
	N = 250	3.53	1.70	2.58	16.42	0.45	17.76
	N = 500	3.24	0.92	1.26	6.85	1.13	9.15
	N = 1000	3.08	0.58	0.65	4.32	1.63	6.12
	N = 2500	3.04	0.37	0.29	2.95	2.11	4.23
	N = 5000	3.03	0.26	0.37	3.28	2.29	4.15

[14] The skewness and kurtosis take a value of 0 and 3 for the normal distribution. Within the Pearson family of distributions, the normal distribution is characterized by these two moments (Spanos 1999). If the skewness and kurtosis of the estimator of interest converge to 0 and 3 respectively, then this should provide evidence of asymptotic normality.

[15] Results not shown, but are available from the authors upon request.

linear specification of the logistic regression model. The empirical example uses data from Al-Hmoud and Edwards (2004) who examined private sector participation in the water and sanitation sectors of developing countries. A simplified model was constructed examining 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 continuous variables were *per capita GDP* ($000s) ($X_1$) and *percent urban population growth* (X_2). The two binary variables were *low renewable water resources* (X_3) and *government effectiveness* (X_4). X_3 takes a value of '1' when renewable water resources are below 2000 cubic meters per capita per year, and X_4 takes a value of '1' when the World Bank indicator for government effectiveness indicates the presence of effective government operations (see Al-Hmoud and Edwards 2004). The dataset contained 149 observations for 39 countries from 1996 to 2001, but data were not available for all countries for all years, resulting in an unbalanced panel (Al-Hmoud and Edwards 2004).

Given that Y is distributed Bernoulli, the BRM provides a natural statistical modelling framework for this problem. The general multivariable modelling approach using sequential conditioning was used to specify the functional form of the model. In this case:

$$f_{\mathbf{X}|Y}\left(X_{1,i}, X_{2,i}, X_{3,i}, X_{4,i}; \theta_j\right) = f_{\mathbf{X}_1|\mathbf{X}_2,Y}\left(X_{1,i}, X_{2,i} \mid X_{3,i}, X_{4,i}; \boldsymbol{\theta}_{\mathbf{1},j}\right) \cdot f_{\mathbf{X}_2|Y}\left(X_{3,i}, X_{4,i}; \boldsymbol{\theta}_{\mathbf{2},j}\right)$$

where $f_{\mathbf{X}_1|\mathbf{X}_2,Y}\left(X_{1,i}, X_{2,i} \mid X_{3,i}, X_{4,i}; \boldsymbol{\theta}_{\mathbf{1},j}\right)$ was assumed to be multivariate normal and $f_{\mathbf{X}_2|Y}\left(X_{3,i}, X_{4,i}; \boldsymbol{\theta}_{\mathbf{2},j}\right)$ bivariate Bernoulli. To obtain multivariate normality, $X_{1,i}$ was transformed using the natural log, so that $\overline{X}_{1,i} = \ln\left(X_{1,i}\right)$. An Omnibus Test for multivariate Normality was conducted following Doornik and Hansen (1994) to test the distributional assumption for $f_{\mathbf{X}_1|\mathbf{X}_2,Y}\left(\overline{X}_{1,i}, X_{2,i} \mid X_{3,i}, X_{4,i}; \boldsymbol{\theta}_{\mathbf{1},j}\right)$, which gave a test statistic of 1.39 and associated *p*-value of 0.85, indicating support for the normality assumption.[16]

Using equation (18) and reparameterizing to make the index function linear in the parameters, gave:

[16] Two additional multivariate tests using the skewness and kurtosis coefficients were conducted as well following procedures in Spanos (1986). The skewness test gave a test statistic of 3.65 with associated p-value of 0.46; and the kurtosis test gave a test statistic of 1.41 with an associated *p*-value of 0.24. Both tests provide support for the multivariate normality assumption. All tests were conducted using the residuals of OLS regressions of $\overline{X}_{1,i}$ and X_{2i} on $X_{3,i}$, $X_{4,i}$ and Y_i.

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

Given the linearity property is satisfied, a standard computer statistical software package for estimating logistic regression models was used to estimate the corresponding BRM. Estimation results for the BRM using equation (19) and a more common specification of the logistic regression model found in the applied literature:

$$\eta(\mathbf{x}_i;\beta) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \beta_4 x_{4,i}, \tag{20}$$

are presented in Table 4. It should be noted that $x_{1,i}$ in the traditional logistic formulation was not transformed using the logarithmic transformation, as was done in the BRM. Misspecification testing results for the BRM using equation (19) indicated the presence of heterogeneity across time and space, so fixed effects for years and regions were incorporated into both models.[17]

The BRM and more traditional model specifications were compared using a likelihood ratio test, with the null hypothesis being that the more common specification of the logit model using equation (20) with fixed effects across time and space was correct. The computed likelihood ratio test statistic was 80.22 with an associated *p*-value of less than 0.001, indicating the more common specification of the logistic regression model was misspecified. In addition, the BRM formulation was tested for additional functional misspecification using RESET-type tests, which indicated strong support for the functional specification used.[18]

If the traditional formulation was used for statistical inference to make substantive claims about what variables significantly influence private investment in the water and sanitation sectors of developing countries, then such inferences would have been statistically invalid. The significance of many of the nonlinear and interaction terms provides evidence of the nonlinear nature of the index function, in contrast to the linear in variables functional form commonly used in applied

[17] There is only one observation for 2001 in the dataset, which made it difficult to include a fixed effect (dummy variable) without encountering convergence problems during estimation. A likelihood ratio test was conducted in a Fisherian testing framework to examine the BRM without fixed effects across time and regions (see Spanos 1999). The null hypothesis was no fixed effects and the likelihood test statistic was 88.10 with an associated p-value less than 0.001, indicating a strong lack of support for the null hypothesis. The regions tested included (1) Central America and the Caribbean; (2) South America; (3) Europe; (4) Africa; (5) Middle East; (6) Asia Major; and (7) South East Asia and Australia. The last region was dropped in the model for estimation purposes.

[18] The RESET-type test used tested for the significance of a squared and cubed term of the fitted values of the index function (i.e. $\hat{\eta}^2$ and $\hat{\eta}^3$) as additional covariates. The null hypothesis was that the coefficients on these terms were equal to zero. A likelihood ratio test was used to conduct the test, giving a test statistic equal to 0.91 and associated *p*-value of 0.63. Test results indicate strong support for the null hypothesis.

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

Variable[1]	**BRM using Equation (19)**	**Traditional Logit (LRM) using Equation (20)**
	Coefficient Estimate (Standard Error)[2]	**Coefficient Estimate (Standard Error)**[2]
	Fixed Effects	
1996	-4. 99	-4.22***
	(5.91)	(1.65)
1997	-0.58	-3.41**
	(5.88)	(1.55)
1998	2.66	-2.74*
	(6.18)	(1.55)
1999	7.37	-1.98
	(7.12)	(1.53)
2000	13.53*	-0.73
	(8.61)	(1.56)
Central America and Caribbean	-33.94***	-2.48***
	(12.32)	(0.97)
South America	6.84**	0.06
	(3.32)	(0.87)
Europe	-36.78***	-3.08**
	(13.60)	(1.28)
Africa	-24.73***	-4.27***
	(9.04)	(1.09)
Middle East	-35.27***	-5.70***
	(13.44)	(1.68)
Asia Major	-35.98***	-2.98**
	(12.99)	(1.47)
	Covariates	
$\overline{X}_{1,i}$ ($X_{1,i}$ for LRM)	-88.33***	0.28
	(34.57)	(0.21)
$X_{2,i}$	-8.64	0.76**
	(6.09)	(0.31)
$X_{3,i}$	21.60**	3.80***
	(9.33)	(0.94)
$X_{4,i}$	974.24**	0.37
	(409.78)	(0.63)
$\overline{X}_{1,i}^2$	27.44***	---
	(10.49)	
$\overline{X}_{1,i}X_{2,i}$	23.97**	---
	(9.48)	
$X_{2,i}^2$	2.07*	---
	(1.22)	
$\overline{X}_{1,i}X_{3,i}$	89.64**	---
	(36.28)	
$X_{2,i}X_{3,i}$	8.63	---
	(8.43)	

Table 4. continued.

Variable[1]	BRM using Equation (19)	Traditional Logit (LRM) using Equation (20)
	Coefficient Estimate (Standard Error)[2]	**Coefficient Estimate (Standard Error)[2]**
$\overline{X}_{1,i}X_{4,i}$	-942.74** (399.63)	---
$X_{2,i}X_{4,i}$	-515.14** (208.38)	---
$X_{3,i}X_{4,i}$	-1038.46** (431.22)	---
$\overline{X}_{1,i}^2X_{3,i}$	-6.96 (9.84)	---
$\overline{X}_{1,i}X_{2,i}X_{3,i}$	-24.94** (10.05)	---
$X_{2,i}^2X_{3,i}$	-1.97 (1.95)	---
$\overline{X}_{1,i}^2X_{4,i}$	226.91** (97.00)	---
$\overline{X}_{1,i}X_{2,i}X_{4,i}$	251.44** (102.26)	---
$X_{2,i}^2X_{4,i}$	68.02*** (26.71)	---
$\overline{X}_{1,i}X_{3,i}X_{4,i}$	1042.99** (432.76)	---
$X_{2,i}X_{3,i}X_{4,i}$	551.43** (220.56)	---
$\overline{X}_{1,i}^2X_{3,i}X_{4,i}$	-279.48** (113.48)	---
$\overline{X}_{1,i}X_{2,i}X_{3,i}X_{4,i}$	-274.65** (110.11)	---
$X_{2,i}^2X_{3,i}X_{4,i}$	-72.43*** (28.17)	---
Other Statistics		
Log-Likelihood	-22.57	-62.68
McFadden's Pseudo R^2	0.77	0.37
Estrella's R^2	0.86	0.47
Fraction Correctly Predicted	94%	81%

[1] $\overline{X}_{1,i}$ is the log of per capita GDP for the BRM and $X_{1,i}$ is per capita GDP; X_2 is percent urban population growth; X_3 is binary variable indicating low renewable water resources; and X_4 is a binary variable indicating government effectiveness.

[2] The standard errors are calculated using the estimate of the asymptotic information matrix. An * indicates the coefficient was significantly different from zero at the 0.10 level of significance, ** at the 0.05 level of significance, and *** at the 0.01 level of significance.

modelling. Further evidence that the BRM using equation (19) was superior to the more common specification of the logistic regression model is given by the higher pseudo R^2 values, higher within-sample prediction and lower mean square error.[19]

The purpose of the BRM estimated was to examine the correlations between private investment in the water and sanitation sector and the explanatory factors included in the model, in order to determine factors that create the proper environment for private sector participation to exist (following Al-Hmoud and Edwards 2004). No underlying theoretical or latent variable model was posited by Al-Hmoud and Edwards (2004). Thus, the estimated model becomes the structural model, once substantive questions are asked of the estimated model. A particularly useful substantive interference that can be made is the examination of marginal effects.

While the BRM provides a more statistically adequate representation of the underlying probabilistic process giving rise to the data, the explanatory power of the model could be seen as questionable given the highly nonlinear nature of the index function. While the coefficients of the model may or may not be readily interpretable, the marginal effects can be used for substantive inferences. The marginal effects are the partial derivatives of the conditional mean with respect to the explanatory variables of interest. For the logistic formulation of the BRM the vector of marginal effects is:

$$\frac{\partial h(\mathbf{x}_i;\beta)}{\partial \mathbf{x}_i} = h(\mathbf{x}_i;\beta)\cdot(1-h(\mathbf{x}_i;\beta))\cdot\frac{\partial \eta(\mathbf{x}_i;\beta)}{\partial \mathbf{x}_i}. \tag{21}$$

For binary variables the marginal effect is the change in probability from changing from a value of ‘0’ to a value of ‘1’ holding all other explanatory variables at their present values (or mean). The standard errors can be calculated using the delta method (see Greene 2003), to test for statistical significance.

The marginal effects with standard errors for the estimated BRM and the traditional logistic regression model are presented in Table 5. The substantive inferences that can be made from each model are significantly different. For the traditional logistic regression model, the marginal effects tell us that an increase in percent urban population growth or having low water renewable resources will have a positive and statistically significant impact on the probability of the private sector in a developing country investing in the water and sanitation sector. Given the misspecified nature of the traditional logistic regression model in this case, such inferences are not statistically valid, rendering the substantive inferences unreliable. The properly specified BRM indicates that an increase in percent urban population growth will have a positive and statistically significant impact on private sector investment in the water and sanitation sector of developing countries, but the impact of having low water renewable resources is not statistically significant from zero. Furthermore, effective governance has a negative and statistically significant impact

[19] Given the unbalanced panel and potential of serial correlation the BRM model was re-estimated using a generalized estimating equations (GEE) estimator in SAS (PROC GENMOD) assuming observations over time from each country followed an AR(1) process. To test model fit of the re-estimated model, a Hosmer-Lemeshow test was conducted for the original model and the GEE model. The tests indicated a poor fit for the BRM with AR(1) covariance structure (8.58, *p-value* = 0.035) and relatively better fit for the BRM with independent covariance structure (5.21, *p-value* = 0.157).

Table 5. Marginal Effects for the Empirical BRM and Traditional Logit Models

Variable[1]	Marginal Effect[2]	
	BRM using Equation (19)	**Traditional Logit (LRM) using Equation (20)**
X_1	-0.04	0.04
	(1.52)	(0.03)
X_2	1.04	0.10*
	(1.53)	(0.06)
X_3	0.62***	0.40***
	(0.05)	(0.14)
X_4	-0.42***	0.05
	(0.06)	(0.09)

[1] X_1 is per capita GDP, X_2 is percent urban population growth, X_3 is binary variable indicating low renewable water resources, and X_4 is a binary variable indicating government effectiveness.
[2] The standard errors are in parentheses and are calculated using the delta method (Greene 2003). Marginal effects and standard errors are calculated for each individual and then averaged following recommendations in Greene (2003). Marginal effects for X_1 in the BRM model are calculated by taking the derivative in equation (21) with respect to X_1 instead of $\overline{X}_1 = \ln(X_1)$. An * indicates the coefficient was significantly different from zero at the 0.10 level of significance, ** at the 0.05 level of significance, and *** at the 0.01 level of significance.

on private sector investment, opposite in sign from the marginal effect for the traditional logistic specification (which was not statistically significant).

6 Conclusion

Rooted in the analysis of observational data, the probabilistic reduction approach provides a systematic way to specify regression models with binary dependent variables. More traditional approaches to model specification (e.g. the latent variable and GLM) provide alternatives that may be less suited for modelling observational data. The BRM imposes no *a priori* theoretical or ad hoc restrictions (or assumptions) upon the model, thereby providing a theory-neutral statistical model of the conditional binary choice process being examined. In addition, the Bernoulli Regression Model (BRM) can provide a parsimonious description of the probabilistic structure of the conditional binary choice process being modelled. By understanding the probabilistic assumptions underlying the BRM, a modeller can test the statistical adequacy of their model *a posteriori*. Statistical adequacy ensures that any substantive or theoretical inferences obtained from the model are statistically valid, providing the needed evidence for supporting or rejecting hypotheses concerning the underlying theoretical model of interest.

The paper provides a modelling framework for specifying Bernoulli Regression Models that allows the statistical information to play a crucial role in ensuring a statistically adequate model is obtained. Furthermore, specification and application issues concerning the BRM are addressed along with the presentation of an empirical example. This example provides evidence that when an underlying statistically adequate model is not obtained prior to making substantive inferences; those inferences are likely to be statistically unreliable and potentially erroneous. The

Bernoulli Regression Model provides a tractable and statistically sound framework upon which to construct statistically adequate models for applied problem solving. In addition, the work in this paper provides a framework upon which to gain further insight into the specification and estimation of other discrete choice models, such as the multinomial regression and ordered logistic regression models.

Acknowledgements

The authors would like to acknowledge two blind reviewers for their helpful comments and suggestions.

References

Aldrich, J. and Nelson, F., 1984. *Linear Probability, Logit, and Probit Models*, Sage Publications, Beverly Hills, CA.

Al-Hmoud, R. B. and Edwards, J., 2004. A means to an end: studying the existing environment for private sector participation in the water and sanitation sector, *International Journal of Water Resources Development*, 20(4), 507-522.

Anderson, J. A., 1972, Separate sample logistic discrimination, *Biometrika*, 59(1), 19-35.

Aranda-Ordaz, F. J., 1981. On two families of transformations to additivity for binary response data. *Biometrika*, 68(2), 357-363.

Arnold, B. C., Castillo, E. and Sarabia, J. M., 1999. *Conditional Specification of Statistical Models*, Springer Verlag, New York, NY.

Box, G. E. P. and Cox, D. R., 1964. An analysis of transformations, *Journal of the Royal Statistical Society*, Series B Methodological, 26(2), 211-252.

Byth, K. and McLachlan, G. J., 1980. Logistic regression compared to normal discrimination for non-normal populations, *Australian and New Zealand Journal of Statistics*, 22(2), 188-196.

Cook, D. R. and Weisberg, S., 1999. *Applied Regression Including Computing and Graphics*, John Wiley and Sons, Inc., New York, NY.

Coslett, S. R., 1983. Distribution-free maximum likelihood estimator of the binary choice model, *Econometrica*, 51(3), 765-782.

Cox, D. R. and Snell, E. J., 1989. *Analysis of Binary Data*, 2nd ed., Chapman and Hall, New York, NY.

Day, N. E. and Kerridge, D. F., 1967. A general maximum likelihood discriminant, *Biometrics*, 23(2), 313–323.

Doornik, J. A. and Hansen, H., 1994. An Omnibus test for univariate and multivariate normality, http://www.doornik.com/research/normal2.pdf.

Draper, N. R. and Cox, D. R., 1969. On distributions and their transformations to normality, *Journal of the Royal Statistical Society*, Series B Methodological, 31(3), 472-476.

Fahrmeir, L. and Tutz, G., 2001. *Multivariate Statistical Modeling Based on Generalized Linear Models*, Springer-Verlag, New York, NY.

Gourieroux, C., 2000, *Econometrics of Qualitative Dependent Variables*, Cambridge University Press, Cambridge, UK.

Greene, W. H., 2003. *Econometric Analysis*, 5th ed., Prentice Hall, Upper Saddle River, NJ.

Hardin, J. W. and Hilbe, J. M., 2007. *Generalized Linear Models and Extensions*, 2nd ed., StataCorp, LP, College Station, TX.

Hogg, R. W. and Craig, A. T., 1978. *Introduction to Mathematical Statistics*, 4th ed., MacMillan Publishing Co., Inc., New York, NY.

Hosmer, D. W. and Lemeshow, S., 2000. *Applied Logistic Regression*, 2nd ed., John Wiley & Sons, Inc., Hoboken, NJ.

Kay, R. and Little, S., 1986. Assessing the fit of the logistic model: a case study of children with the Haemolytic Uraemic Syndrome, *Applied Statistics*, 35(1), 16-30.

Kay, R. and Little, S., 1987. Transformations of the explanatory variables in the logistic regression model for binary data, *Biometrika*, 74(3), 495-501.

Lauritzen, S. L. and Wermuth, N., 1989. Graphical models for association between variables, some which are qualitative and some quantitative, *Annals of Statistics*, 17(1), 31-57.

Liang, K. Y., Zeger, S. L., and Qaqish, B., 1992. Multivariate regression analyses for categorical data, *Journal of the Royal Statistical Society*, Series B Methodological, 54(1), 3-40.

Maddala, G. S., 1983. *Limited Dependent and Qualitative Variables in Econometrics*, Cambridge University Press, Cambridge, UK.

Marschak, J., 1960. Binary choice constraints on random utility indications. In:, *Stanford Symposium on Mathematical Methods in the Social Sciences*, K. Arrow eds. Stanford University Press, Stanford, CA, 312-329.

McFadden, D. L., 1984. Econometric analysis of qualitative response models. In: *Handbook of econometrics*, vol. 2, Z. Griliches and M. D. Intriligator eds., North Holland, New York, NY.

Nelder, J. A. and Wedderburn, R. W. M., 1972. Generalized linear models, *Journal of the Royal Statistical Society*, Series A General, 135(3), 370-384.

Nelder, J. A. and Mead, R., 1965. A simplex method for function minimization, *Computer Journal*, 7(4), 308-313.

Oklin, I. and Tate, R. F., 1961, Multivariate correlation models with mixed discrete and continuous variables, *The Annals of Mathematical Statistics*, 32(1), 448-465.

Powers, D. A. and Xie, Y., 2000, *Statistical Methods for Categorical Data Analysis*, Academic Press, San Deigo, CA.

Scrucca, L. and Weisberg, S., 2004. A simulation study to investigate the behavior of the log-density ratio under normality, *Communications in Statistics: Simulation and Computation*, 33(1), 159-178.

Spanos, A., 1986. *Statistical Foundations of Econometric Modeling*, Cambridge University Press, Cambridge, UK.

Spanos, A., 1999. *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*, Cambridge University Press, Cambridge, UK.

Tate, R. F., 1954, Correlation between a discrete and a continuous variable: point-biserial correlation, *The Annals of Mathematical Statistics*, 25(3), 603-607.

Train, K. E., 2003. *Discrete Choice Methods with Simulation*, Cambridge University Press, Cambridge, UK.

Warner, S. L., 1963. Multivariate regression of dummy variables under normality assumptions, *Journal of the American Statistical Association*, 58, 1054-1063.

Williams, D. A., 1982, Extra-binomial variation in logistic linear models, *Applied Statistics*, 31, 144-148.

Yeo, I. K. and Johnson, R. A., 2000. A new family of power transformations to improve normality or symmetry, *Biometrika*, 87(4), 954-995.