Computational Economics **15:** 273–289, 2000.

On the Use of Enumeration for Investigating the Performance of Hypothesis Tests for Economic Models with a Discrete Response Variable

SIMON PETERS [1] and ANDREW CHESHER [2]

[1] *School of Economic Studies, Dover St., Manchester University, Manchester, U.K.;*
[2] *Department of Economics, University College London, London, U.K.*

Abstract. This article notes that it is now practical to use the method of enumeration to analyse the performance of estimators and hypothesis tests of fully parametric binary data models. The general method is presented and then employed to investigate the power performance of a common misspecification test for the Probit model. The advantages, disadvantages and limitations of enumeration compared with standard Monte Carlo simulation are then discussed. Finally, an example from experimental economics is used to demonstrate that the methodology can also be used in small empirical studies.

Key words: enumeration, hypothesis test, Probit model, size, power

1. Introduction

In practice, the standard inferential methods for econometric models proceed under the assumption that the asymptotic approximation to the distribution of a test or estimator is reasonable. This may not be the case in finite samples, and the performance of a test statistic or estimator is usually evaluated by standard simulation techniques, commonly known as Monte Carlo experimentation. However, one can obtain the exact distribution of any test or estimator for any discrete data model, although it is not always computationally feasible to do so. This article presents a general technique for enumerating such a distribution, and demonstrates it for two examples: obtaining the power curves for test statistics obtained from the classical Probit binary data model, and examining the risk behaviour of individuals.

The idea of enumeration has been around for some time, and its main use has been for exact inference in logit and related biometric models. Cox (1970) showed that it is possible to evaluate the exact distribution for a binary data model by enumerating the sample space of the minimal sufficient statistic for a logit model. Hirji, Mehta and Patel (1987) present an algorithm for obtaining the distribution of sufficient statistics of a logistic regression model that can enumerate the probabilities for a larger model than that previously envisaged at the time; a sample

of size 40 with 3 binary covariates and a constant. This type of data is common in biometric applications, and versions of the method of enumeration have been used to evaluate, for example, the performance of the likelihood based tests (Hirji, 1991) and Markov chain Monte Carlo methods (McDonald, Smith and Forster, 1999). This form of enumeration can be extended to any model as long as it has, in the terminology of McCullagh and Nelder (1989), a canonical link function. Hirji (1992) presents an extension of the earlier algorithm to polytomous response models.

The technique presented in this paper can be applied to a greater variety of estimators and tests than those referred to above, although it was first used by Berkson (1955) for evaluating the performance of maximum likelihood and minimum χ^2 estimators of logit models for case-control studies. This work has been re-examined recently by Hughes and Savin (1994). Its first application to an economic problem appears to be in Hausman and McFadden (1984), where general enumeration is used to evaluate the performance of a misspecification test for a simple discrete choice model.

In economics, however, one might consider an average *small* cross section data set to consist of 1000 observations and 10 explanatory variables, which is much larger than those found in the field of biometrics. This makes enumeration impractical for everyday inference in standard applied economic models. It does not preclude its use, however, in distributional experimentation where the design of an experiment can be chosen by the researcher. Enumeration can, therefore, be used instead of simulation to examine the behaviour of estimators and test statistics, and is more efficient computationally in small samples when used to study the power properties of statistical hypothesis tests, or the robustness of estimators to model misspecification. The power properties of a hypothesis test can also be analysed by standard asymptotics, however there is strong evidence that these theoretical results can be misleading (Nelson and Savin, 1990; Davidson and MacKinnon, 1984). Further, it is also possible to use this methodology to analyse the responses of individuals to experimentally designed studies or surveys. There are at least three areas of modern economics where such a study or survey might arise in practice: experimental economics (examining an individual's perception of uncertainty), market research (analysing the responses of individual focus group members), and human resource management (testing job candidates for suitability).

The remainder of this article is laid out as follows: Section 2 presents the details of general enumeration for a binary response model, Section 3 uses it to evaluate the power performance of a misspecification test for the Probit model, Section 4 discusses its advantages over standard simulation, and Section 5 presents a small empirical example that uses experimental economics data. Section 6 concludes.

2. A General Approach to Enumeration

The discrete nature of binary response variables allows one to generate all the possible outcomes (0s or 1s) for a data set of size N. For example, if $N = 3$ then one has 8 possible configurations of 3 responses: (0, 0, 0), (0, 0, 1), (0, 1, 0), (1, 0, 0), (0, 1, 1), (1, 0, 1), (1, 1, 0) and (1, 1, 1). These are fixed no matter what binary data model is used to obtain an estimate (or test statistic) for a configuration. The estimates from a given model make up a sample space, and the probability distribution associated with this sample space is given by the probability of each of these configurations occurring. So, the probability assigned to a configuration is also the probability of any statistic calculated at that configuration. The technique of enumeration calculates all these statistics and their associated probabilities.

This raw enumeration can be split into two parts by noticing that the probability of a configuration occurring does not depend upon the model chosen to calculate the estimates at that configuration. The sample space, therefore, has two models associated with it; the true model that generates the probability of an outcome and the maintained model that is used to compute estimates and related statistics. Such a split enables the points of support of the sample space, which are determined by the maintained model, to be obtained separately from their probabilities, which are determined by the true model.

This method only appears to be computationally tractable for small sample sizes. However, computational efficiencies can be obtained by restricting the enumeration to the observations that might be generated by a replicated design matrix. The underlying methodology is that of a classical simulation study, but without the Monte Carlo error.

The basis for this is the following binary data specification. Given a column vector of coefficients, $\mathbf{b}$, and a column vector of covariates, $\mathbf{x}$, then for the response, y, $\Pr(y = 1|\mathbf{x}) = F(\mathbf{b}'\mathbf{x})$ where $F(.)$ is a cumulative distribution function. The responses are assumed to be independent across observations.

In order to allow for replicated designs one assumes that the covariate vector, $\mathbf{x}$, takes values at M distinct points, and the number of responses (dependent variables) at each point is known. At each point i, therefore, there are N_i responses. So, there will be $J = \prod_{i=1}^{M}(N_i + 1)$ configurations in total and $N = \sum_{i=1}^{M} N_i$ observations per configuration. For the case where there is no replication ($N = M$), this gives 2^N configurations for a sample size of N, resulting in just over a million configurations for sample size 20 and over 10^{12} for sample size 40. At the present time, the former is computationally feasible but the latter is not. Larger sample sizes have to be obtained by replication. For example, assuming 2^{20} configurations is an operational upper limit, one can obtain larger sample sizes of $N = 75$ when $M = 5$ ($N_i = 15$), and $N = 301$ when $M = 3$ ($\{N_i\} = \{100, 100, 101\}$).

2.1. GENERATING CONFIGURATIONS

This section describes how one generates a sample in a repeatable manner. In this way the parameter estimates from the maintained model can be matched up with the associated probability from the true model. One must be able to generate all the possible response configurations for the maintained design. This is done by noting that the total number of configurations (J) is equivalent to the number of elements in an M dimensional array with indices running from 0 to N_i. Knuth (1973, p. 296) shows how elements of such an array can be sequentially allocated to locations in a computer's memory. The opposite of this allocation returns the array's index values from a given memory location.[1]

An index value gives the number of unit responses at point i, and is obtained for configuration number j as:

$$S_{i,j} = \left\lfloor \frac{j-1}{D_i} \right\rfloor \text{ modulo } (N_i + 1), \quad i = 1 \text{ to } M, \quad j = 1 \text{ to } J\,, \tag{1}$$

where $D_i = \prod_{r=0}^{i-1}(N_r + 1)$, $N_0 = 0$, and $\lfloor . \rfloor$ represents the *floor* function ($\lfloor x \rfloor$ is the greatest integer not exceeding x). This can then be used to generate a standard binary regression sample by creating $S_{i,j}$ responses of $y = 1$ and $N_i - S_{i,j}$ responses of $y = 0$ at each point i.

2.2. THE MAINTAINED MODEL

The maintained probability model is used to obtain the estimator. It consists of:

(a) a fixed N by p design matrix X,
(b) a specified probability model, F_m, that links a covariate point to the outcome, and
(c) a method for computing the parameter estimates.

The application of the maintained model will produce a vector of p parameter estimates, $\hat{\mathbf{b}}$, at each configuration.

The probability model is decoupled from the estimation method, allowing one to distinguish between different estimation methods for the same F_m and X. An example of this could occur in the Probit model ($F_m \equiv \Phi$, the standard Normal cumulative distribution function) if one needed to compare maximum likelihood and minimum χ^2 estimators.

A discussion of the computational details of obtaining parameter estimates is beyond the scope of this article, however, no matter what model and method are chosen there will always be the problem of coping with data configurations that admit infinite parameter estimates. This is discussed by Silvapulle (1981) among others, and an algorithm to identify these *improper* data configurations is presented in Burridge and Silvapulle (1986) for the general case. Configurations of responses and covariates are categorised as completely separated, quasicompletely separated

and overlapped. If there exists a hyperplane in the covariate space separating design points with unit responses from design points with zero responses then there is complete separation. If there is not complete separation only because one or more design points lie on what would otherwise be a separating hyperplane then there is quasicomplete separation, otherwise there is overlap. Only in overlapped configurations do unique finite valued maximum likelihood estimates exist. The detection of non-overlapped configurations is always possible using linear programming methods but straightforward inspection is all that is required in simple designs.

Given X, F_m and an estimation technique, all the maintained model's unknown parameter estimates can be obtained by following these steps:

Algorithm M (obtaining all the finite parameter estimates).

M0. [Initialize.] Set j to 0.

M1. [Loop.] Increment j. While $j < J$ perform the following:

M2. [Obtain a configuration.] Generate the $\{S_{ij}\}$ using formula 1, and use them to create the responses, $\mathbf{y}_j$ say, associated with X.

M3. [Check for an *improper* configuration.] Examine the $\mathbf{y}_j$ and X using the methods discussed above. if the data configuration is overlapping, continue. If the configuration is non-overlapping note the configuration number in a monitoring file and skip to M1.

M4. [Obtain the estimates.] Use the maintained model's estimation technique to obtain the parameter estimates for the responses, $\mathbf{y}_j$ and model specification X and F_m.

M5. [Save.] Write the estimates and their configuration number, j, to a file.

M6. [Continue.] Skip to M1.

Functions of the parameter estimates, such as test statistics, are best obtained in a second pass. This avoids repeating the computationally intensive step M4 if one decides to calculate another test statistic. In this case, the binary file containing the estimates should also be opened in step M0. M4 is now replaced by:

M4*. [Obtain test statistics.] Read in the estimates associated with configuration j from the file they were saved to. Calculate the desired test statistic from these estimates for the responses, $\mathbf{y}_j$ and model specification X and F_m.

2.3. THE TRUE MODEL

The true probability model consists of:

(a) the true N by q design matrix Z, where q is the length of a covariate vector $\mathbf{z}$,

(b) a probability model, F_t, that generates the true probability of an outcome, and

(c) the true parameter values, $\mathbf{b}$.

A correctly specified model has $F_m \equiv F_t$ and $X = Z$ $(p = q)$. If $p < q$ there is omitted variable misspecification present, and distributional misspecification occurs when F_m is different from F_t.

The probability for each configuration under the true model is:

$$\pi_j = \Pr(\text{configuration } j) = \prod_{i=1}^{M} \binom{N_i}{S_{i,j}} F_t(\mathbf{b}'\mathbf{z}_i)^{S_{i,j}} (1 - F_t(\mathbf{b}'\mathbf{z}_i))^{N_i - S_{i,j}}, \quad (2)$$

where $\mathbf{z}_i$ is a distinct point from the design matrix $\mathbf{Z}$ and $S_{i,j}$ is the number of unit responses at point i for configuration j.

Given Z, F_t and $\mathbf{b}$, the probability of an estimate (or a function of an estimate) can be obtained by following these steps:

Algorithm T (obtaining all the probabilities).

T0. [Initialize.] Open the file containing the saved values and their configuration code.

T1. [Loop.] Until the end of the file, perform the following:

T2. [Input.] Read in the saved values and their configuration code j.

T3. [Obtain a configuration.] Generate the $\{S_{ij}\}$ using formula 1.

T4. [Obtain the probability.] Calculate the probability, π_j, of this configuration using formula 2, given the chosen Z, F_t and $\mathbf{b}$.

Given the maintained probability model estimate, $\widehat{\mathbf{b}}_j$, for configuration j, our true model generates its probability π_j. However, because not all configurations admit finite estimates it may be necessary to work with the probability conditional upon a configuration being estimable (admitting finite parameter estimates). This probability is simply calculated as $\pi_e = \sum_{j=1}^{J} \delta_j \pi_j$, where J is the total number of configurations for the maintained model, and $\delta_j = 1$ if configuration j is estimable, 0 otherwise. The probability of obtaining finite statistics calculated at configuration j, such as $\widehat{\mathbf{b}}_j$, is then just $\pi_j^* = \frac{\pi_j}{\pi_e}$.

2.4. THE EXACT DISTRIBUTION

Once one has obtained the estimate, $\widehat{\mathbf{b}}_j$, for configuration j, and its probability π_j (or possibly π_j^*), the sampling distribution of any statistic, t, calculated from $\widehat{\mathbf{b}}_j$, $\widehat{t}_j = g(\widehat{\mathbf{b}}_j)$ say, can be obtained in the following way:[2]

Algorithm E (obtaining the distribution).

E0. [Order the points of support and probabilities.] Sort the $\widehat{t}_j$ in ascending order, and impose the same ordering on the probabilities. All duplicates of a statistic $\widehat{t}_j$ are ignored at this stage, but the probability associated with this configuration is multiplied by the number of replicates of $\widehat{t}_j$.

E1. [Obtain the cumulative distribution function.] Let $j*$ represent the new index for the ordered and unique set of statistics, $\widehat{t}_{j*}$, and their associated probabilities π_{j*}. The sets of all these values give the points of support

and probability mass function respectively. The cumulative distribution is obtained simply by applying a cumulative sum operation to the vector of probabilities to give $P_{j*} = \sum_{l=1}^{j*} \pi_l$.

The exact distribution of t then consists of:

(a) the sample space $\{\widehat{t}_{j*}\}$,
(b) the probability space $\{\pi_{j*}\}$,
(c) and the cumulative distribution function $\{P_{j*}\}$.

2.5. SOME COMMENTS ON IMPLEMENTATION

Since there may be very large numbers of probabilities to compute, each very small in value, care has to be taken to compute them precisely. In practice this is achieved using high precision arithmetic, by computing $\log(F_t(\mathbf{b}'\mathbf{z}_i))$ utilising accurate routines for computation of the log-gamma function, only reverting to the probability metric when probabilities have to be summed during calculation of distribution functions and moments, and by summing sorted probabilities so that smallest probabilities are accumulated first. Any precision problems associated with these calculations, and the uniqueness of the tests or estimates, did not appear to cause any trouble for the examples studied in Section 3 and 5.

In order to economise on storage space, the estimates and test statistics obtained using algorithm M should be saved in binary files. It is also good practice to monitor the process and record the indicators of the improper configurations and any others that might appear to unsettle the optimisor at step M4. Binary files will also have to be used if algorithms E and T are unable to process the configurations using available computer memory.

However, it must be emphasised that the method is only feasible for small samples. A comfortable upper limit to the number of configurations is 2^{24} using standard computing technology, a sample size of 24 without replication.[3]

Further, the probabilities can only be obtained by specifying the true model. This is analogous to the data generating process in a Monte Carlo study. In the following section, the true model is changed in order to investigate the size performance of a distributional misspecification test. Algorithm T is run with $F_t \equiv F_m$ and $Z = X$ for each choice of $\mathbf{b}$. The power of the test is investigated by running algorithm T with different F_t. Algorithm M is run twice, once to obtain the estimates and once to obtain the test statistics.

3. An Application to the Probit Model[4]

One of the original motivations behind this work was to examine the performance of tests for model inadequacy. Chesher and Peters (1995) used this technique to investigate the bias of Probit model estimates, and associated tests for omitted variables, while Peters (1995) explored the size and power properties of a com-

mon test for error misspecification in the standard Probit model, the score test for non-Normality. This later work complements several studies of related tests in the literature, see for example Davidson and MacKinnon (1984, 1998), Horowitz (1994), and Skeels and Vella (1998).

As test performance is not the focus of this article, only one set of size and power results will be reported for brevity. This is enough to demonstrate the utility of general enumeration for investigating the finite sample behaviour of statistics obtained from discrete response models.

The score test is obtained from the locally augmented log-likelihood for a Probit model, based on a single observation $(y_n, \mathbf{x}_n')$ with a parameter vector of $\theta = (\mathbf{b}', \mathbf{w}')'$:

$$\begin{aligned} L^*(\theta|\mathbf{x}_n, y_n) &= y_n \log(\Phi(\mathbf{b}'\mathbf{x}_n)) + (1 - y_n)\log(1 - \Phi(\mathbf{b}'\mathbf{x}_n)) \\ &\quad + \log\left(1 + \frac{w_1}{2}(e_n^{(3)} - 3e_n^{(1)}) + \frac{w_2}{4}(e_n^{(4)} - 3e_n^{(2)})\right), \end{aligned} \tag{3}$$

where the *test* parameters w_1 (associated with skewness) and w_2 (associated with kurtosis) make up the vector $\mathbf{w} = (w_1, w_2)'$. This takes the value $\mathbf{w} = (0, 0)'$ when the score test statistic is evaluated. The response, y, takes the value 1 or 0, $n = 1$ to N, and $\Phi(.)$ is the standard Normal cumulative distribution function with an associated density $\phi(.)$.

The generalised errors, $e_n^{(j)}$, $j = 1$ to 4, introduced by Chesher and Irish (1987) are defined as follows for the Probit model:

$$e_n^{(1)} = y_n \frac{\phi(\mathbf{b}'\mathbf{x}_n)}{\Phi(\mathbf{b}'\mathbf{x}_n)} - (1 - y_n)\frac{\phi(\mathbf{b}'\mathbf{x}_n)}{1 - \Phi(\mathbf{b}'\mathbf{x}_n)}, \tag{4}$$

$$e_n^{(2)} = -\mathbf{b}'\mathbf{x}_n e_n^{(1)}, \tag{5}$$

$$e_n^{(3)} = (2 + (\mathbf{b}'\mathbf{x}_n)^2)e_n^{(1)}, \tag{6}$$

$$e_n^{(4)} = -\mathbf{b}'\mathbf{x}_n(3 + (\mathbf{b}'\mathbf{x}_n)^2)e_n^{(1)}. \tag{7}$$

The corresponding residuals, $\hat{e}_n^{(j)}$, are evaluated at $\hat{\mathbf{b}}$, and used to calculate the test statistic. The score vector under the null is obtained as:

$$\frac{\partial L^*(\theta|\mathbf{x}_n, y_n)}{\partial \theta} = \left(e_n^{(1)}\mathbf{x}_n', \frac{e_n^{(3)} - 3e_n^{(1)}}{2}, \frac{e_n^{(4)} - 3e_n^{(2)}}{4}\right)'. \tag{8}$$

Suppose $\hat{\theta} = (\widehat{\mathbf{b}}', \mathbf{0}')'$, then define the score quantities as $\mathbf{s}_n = \frac{\partial L^*(\theta|\mathbf{x}_n, y_n)}{\partial \theta}$, $\widehat{\mathbf{s}}_n = \mathbf{s}_n|_{\theta=\hat{\theta}}$ and $\hat{\mathbf{s}} = \sum_{n=1}^N \widehat{\mathbf{s}}_n$. A fully efficient (FE) score test statistic for the null hypothesis of correct distributional specification ($F_m \equiv \Phi$ for a Probit model) can now be calculated as $\hat{\mathbf{s}}'(\hat{\mathbf{V}}_{FE})^{-1}\hat{\mathbf{s}}$, where $\hat{\mathbf{V}}_{FE}$ is the estimated sample information matrix for (3) under the null. Given the augmented covariate vector $\hat{\mathbf{x}}_n = (\mathbf{x}_n', \frac{\hat{\mathbf{b}}'\mathbf{x}_n{}^2 - 1}{2}, \frac{-\hat{\mathbf{b}}'\mathbf{x}_n{}^3}{4})'$, this matrix is defined as:

$$\hat{\mathbf{V}}_{FE} = \sum_{n=1}^N \frac{\phi(\hat{\mathbf{b}}'\mathbf{x}_n)^2}{\Phi(\hat{\mathbf{b}}'\mathbf{x}_n)(1 - \Phi(\hat{\mathbf{b}}'\mathbf{x}_n))}\hat{\mathbf{x}}_n\hat{\mathbf{x}}_n'. \tag{9}$$

The score test statistic has an asymptotic reference distribution of χ^2_2 under the null.

The first problem that needs addressing is how to handle improper configurations. If one is testing for an omitted variable, as in Chesher and Peters (1995), then there is some scope for obtaining test statistics when the full configuration is improper. One is able to use extended maximum likelihood estimates (Clarkson and Jennrich, 1991) when calculating a likelihood ratio statistic, while the score test only requires that the null configuration be estimable. Neither method is applicable here because the observed test statistic has components that are powers of the estimated linear predictor ($\widehat{\mathbf{b}}'\mathbf{x}$). If this is not finite, then the test statistic is undefined. The size and power of the score tests are, therefore, calculated conditionally on a configuration admitting estimable parameters.

The enumeration experiments are designed to try and allow for a reasonable sample size, N, to be obtained, while retaining enough distinct design points, M, to be meaningful. In this way one could get some idea of what might occur in a larger sample and avoid any of the small sample problems that occur because of the presence of improper configurations.

The study uses the standard Probit specification for the maintained model, with the design matrix $\mathbf{X}$ containing M distinct design points of the type $\mathbf{x}_m = (1, \frac{m}{M-1})'$ where $m = 0$ to $M - 1$. The parameter vector $\mathbf{b}$ has length $p = 2$. The results reported in this article use a design with $N = 50$ with $M = 5$. The number of responses at each point is set as $N_i = \frac{N}{M}$. Size and power are investigated by changing the true model. Peters (1995) reports a larger selection of experiments with $N = 20$, other choices of parameter coefficients and the outer product of gradient (OPG) form of the test statistic.[5]

3.1. TEST SIZE

Orme (1990) and other authors have examined the size properties of related tests in a limited dependent variable context. The perceived opinion is that both the FE and OPG forms of the test have asymptotic approximations of their sampling distributions that are over-sized, and that this is much worse for the OPG case (which is confirmed in Peters, 1995). However, the simulations in Skeels and Vella (1998) suggest that the FE form of the test could also be under-sized.

The behaviour of the FE test is compared with its first order asymptotic reference distribution by the use of quantile-quantile (Q-Q) plots. These types of graphs are now widely used in the statistics literature, and give an overall picture of distributional behaviour, especially in the tails, and are, therefore, of greater utility than tabulations. (Alternatives have been proposed, see Davidson and MacKinnon (1998) for example). In these enumeration experiments, the plots are constructed by plotting $\widehat{t}_{j*}$ against $F^{-1}(P^c_{j*})$ where P^c_{j*} is a continuity corrected version ($P^c_{j*} = \frac{P_{j*} - P_{j*-1}}{2}$ with $P_0 = 0$) of the distribution function at $\widehat{t}_{j*}$ and F^{-1} is the inverse χ^2_2 cumulative distribution function. If these distributions are the same, the

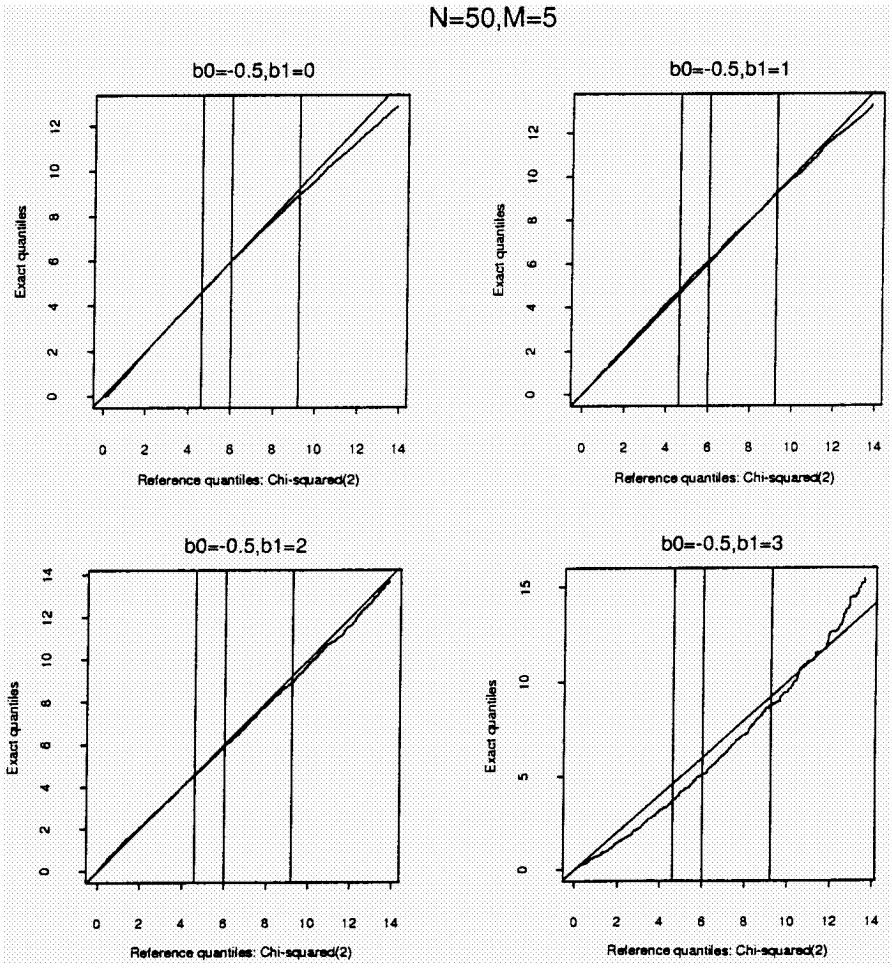

Figure 1. Q-Q plots for different b1 values.

plot will follow the 45° line on the graph. Three further lines have been drawn on the graphs, corresponding to the 10%, 5% and 1% critical values for the target χ^2_2 distribution.

The FE test appears to be remarkably well behaved, in that it follows the line 45° on the graphs, for most of the plots in Figure 1. However for $b_1 = 3$ the test starts to become over-sized in the tail. This poorer performance occurs when the expected ratio of unit to zero responses moves away from 1 : 1. For the designs used in Figure 1 these ratios are approximately 1 : 2, 1 : 1, 2 : 1, and 3 : 1 for $b_1 = 0$, 1, 2, and 3 respectively.

3.2. TEST POWER

One power study in the literature that deals with Probit models and the type of tests under scrutiny here is that of Skeels and Vella (1998). They examined versions of the conditional moment test for omitted variable, heteroscedastic and distributional misspecification. Their tests are not algebraically equivalent to those used here, but they are related. Their findings suggest that non-Normality and heteroscedasticity tests have very poor power. Given that their design had 610 observations and 7 coefficients, this is not very encouraging. This is partly contradicted by the study of Horowitz (1994), who shows that the OPG full Information Matrix test has good power properties for his heteroscedastic alternative, even with a sample of size 50. Davidson and MacKinnon's (1984) results tend to support Skeels and Vella's (1999) results for heteroscedasticity, though they are slightly more optimistic. The enumerations in this section are an attempt to try to examine and explain this behaviour by obtaining the exact power curves for a wider range of alternatives than commonly used in Monte Carlo experimentation.

Four alternatives were chosen. The first two used the BurrII density as the distributional misspecification. This density is discussed in Fry (1993) and is used here because it has both non-Normal skewness and kurtosis. The standardised version was used to enforce zero mean and unit variance, so that the misspecification entered through the distributional shape. This was then relaxed to allow for mean and variance distributional misspecification. The BurrII distribution is $F(x) = (1 + e^{-x})^{-k}$ where $k > 0$ and $x \in (-\infty, \infty)$. This alternative was controlled by its nuisance parameter k, with $k = 1$ giving the logit specification, a distribution is skewed to the left when $k < 1$, and to the right when $k > 1$.

The third alternative was a mixture of the Normal, $N(0, 1)$, and Cauchy, $C(0, 0.674)$ distributions. Here the nuisance parameter controlled the amount of mixing, with 0 being the null and 1, Cauchy. The Cauchy distribution was standardised to have the same inter-quartile range as the Normal. This alternative has kurtotic misspecification only.

The fourth alternative introduces heteroscedasticity by specifying the Normal as $N(0, e^{c\beta'\mathbf{z}})$ where $c = 0$ gives the null. This alternative was used by Horowitz (1994) in his study, and assumes that heteroscedasticity is a function of the linear predictor. Given that this factor enters directly into $e^{(j)}$, $j = 1$ to 4, the Normality test should respond to this alternative.

The power for these alternatives is summarised by plotting the curves for the different values of b_1, given $b_0 = -0.5$, and the alternative distribution. The mixing parameter for the Cauchy/Normal alternative, and for the heteroscedastic alternative range from 0.0, the null, to 1.0 with increments of 0.1. The BurrII alternatives take the same values of k. It ranges from 0.1 to 5.0 with increments of 0.1 for the left skewed alternatives, and 1.0 for the right skewed alternatives. The logit misspecification is emphasised in Figure 2 by the vertical line at $k = 1$. The power is calculated at the value obtained for the exact sampling distribution obtained under the null for a size of 5%.

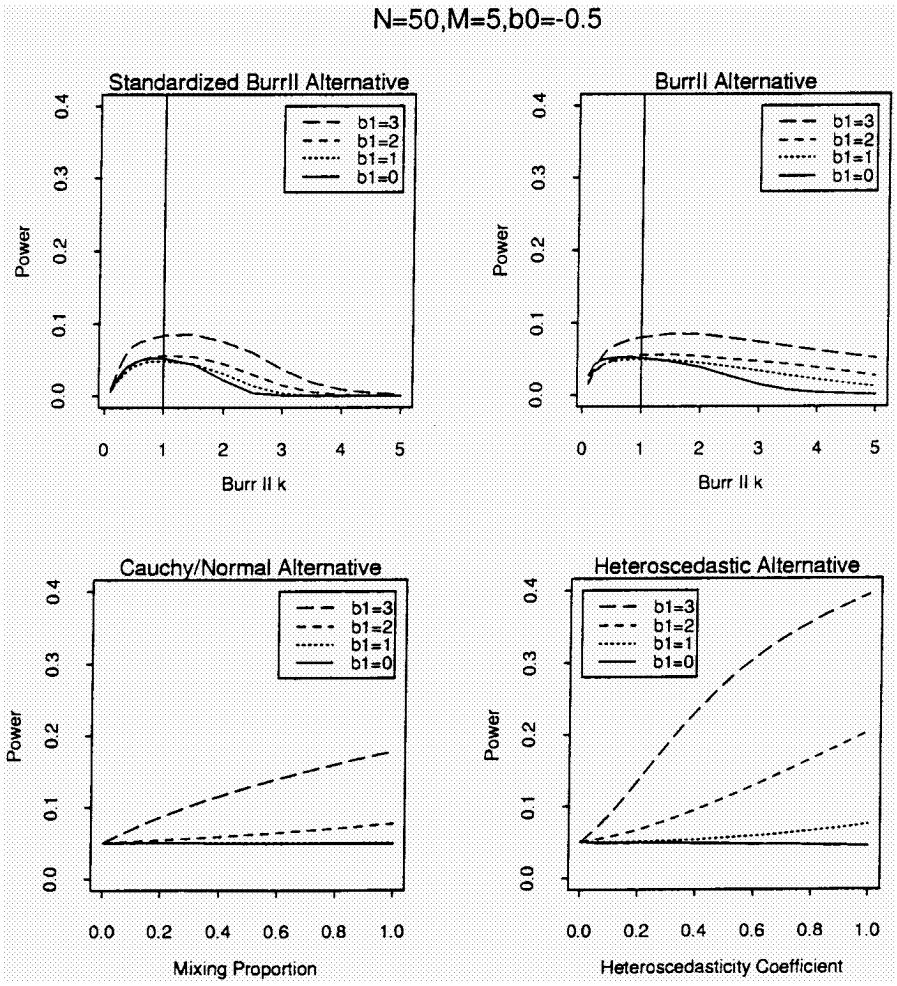

Figure 2. Power curves.

The overall result superficially confirms the observations of Skeels and Vella (1998), that these forms of distributional misspecification test have poor power. Their comment that they have *little useful power except when testing for certain asymmetric distributions* is too pessimistic however, as certain of the alternatives presented here do exhibit power, and for a much smaller sample size than their experiments. This is most marked for the heteroscedastic alternative in Figure 2.

The main weakness of this method, in general, is that the small sample sizes and restricted covariate designs can admit improper configurations with high probability. This can affect the size by making the discrete nature of the exact sampling distribution more apparent, and so rendering it difficult to evaluate power over a variety of designs. This did not, however, cause any serious problems for the configurations examined here. On the other hand, when this design effect occurs

via the alternative specification, one's ability to evaluate the power performance of a test is obscured and it becomes difficult to extrapolate the results in a qualitative fashion to larger sample sizes. This occurred for the more extreme of the skewed BurrII alternatives, and can also occur through the choice of **b**. As the coefficients become large in size, the probability of having all unit or zero responses also increases. A study of this problem in binary data models using large parameter asymptotics is presented in Savin and Würtz (1999).

4. Enumeration Versus Simulation

Exact enumeration avoids the sampling error associated with simulation studies. It also generates all of the extreme data configurations, an advantage noted by Diaconis and Holmes (1994). These may not be generated within a simulation study, either because of their rarity with respect to the total number of Monte Carlo replications or because of computational limitations associated with the random number generator. Once a maintained model has been enumerated, these estimates are fixed no matter how the true model is changed. This is not the case in a Monte Carlo procedure, when the estimates have to be re-calculated if the true data generating process is altered.

These efficiency gains can be quantified using the maintained work ratio, which compares the computations done to obtain the estimates and tests for the maintained model with those needed for a similar simulation study. This is given as:

$$\frac{\{1+(\frac{M(N_A+1)}{10.N})\}\prod_{i=1}^{M}(N_i+1)}{R_{MC}N_D(N_A+C_0)}, \tag{10}$$

where R_{MC} is the number of replications used for the misspecified simulations, N_D is the number of design parameter choices, N_A is the number of alternatives investigated and C_0 is the multiplier for the number of replications used for the null specification. The configurations require more processing time than a simulation, because of the probabilities associated with the true model. This is measured by the inflator for the processing time: $1+(\frac{M(N_A+1)}{10N})$, which is a crude measure of the relative number of operations needed to calculate a point's probability (Equation (2)) compared with that from an estimation (optimising (3) using Newton's method with $p=2$).

The following table shows the efficiencies based on the power investigations used above and in Peters (1995), which have $N_D=8$ and $N_A=76$, compared to a comparable simulation assuming $R_{MC}=1000$ and $C_0=10$ (implying 10000 replications are used to obtain the size distribution).

The entries in Table I are conservative in that they are more likely to be biased in favour of simulation. The work ratio has been calculated assuming the generation of the simulated samples is equivalent to an enumeration sample and negligible

Table I. Design work ratios: enumeration to simulation.

$N = 20, M = 10$	$N = 20, M = 5$	$N = 50, M = 5$
0.416	0.013	0.414

compared to an estimation that required a numerical optimisation. Taking this into account, Table I shows that the experiments reported in Section 3 above required, at most, 41.4% of the computational effort for a simulation study with a sample size of 50.

5. An Empirical Example[6]

Hey and Orme (1994) test expected utility theory (EU) using data on choices between lotteries. Their experimental subjects were confronted with a set of choice problems that involved two lotteries with the money amounts £0, £10, £20 and £30. A single lottery is represented by a set of four probabilities, corresponding to the four money amounts. Hey and Orme (1994) distinguish between the two lotteries by referring to one as the *left-hand* lottery and the other as the *right-hand* lottery. They define the left-hand lottery by the vector of probabilities $\mathbf{p} = (p_1, p_2, p_3, p_4)'$, and the right-hand lottery by $\mathbf{q} = (q_1, q_2, q_3, q_4)'$. The experimental design is that implied if one assumes that subjects have a risk-neutral utility function. If one measures wealth in units of £10, then the risk-neutral expected utility of the left hand lottery is simply $x = d_2 + 2 * d_3 + 3 * d_4$ where $d_2 = p_2 - q_2$, $d_3 = p_3 - q_3$ and $d_4 = p_4 - q_4$.

Let the binary variable under analysis be y, taking the value 1 if the left-hand lottery is chosen by the subject, and zero if the right hand lottery is chosen. Following the same strategy as Hey and Orme (1994), assume that the left hand lottery is chosen if the difference between the expected utilities of the two lotteries, plus a random term, exceeds zero.[7] So, under risk neutral EU, $\Pr(y = 1) = \Pr(x + \epsilon > 0)$ where $\epsilon \sim N(0, \sigma^2)$. This can be specified as a standard Probit model where $\Pr(y = 1) = \Phi(kx)$. Under the assumption of risk-neutrality, if EU holds, $k > 0$ and can be interpreted as $1/\sigma$. If $k <= 0$, EU is violated.

Hey and Orme (1994)'s original study involved 80 subjects and 100 questions. The subjects were also allowed to return an indifferent response. The example here uses 8 subjects who never reported indifference, and is a subset of the original experiment that uses the first two responses of a subject to problems that have x values ranging from –0.5 to 0.5, with increments of 0.125.

Enumeration can be used to find the exact probability that an individual exhibits EU by testing $H0 : k > 0$ vs. $HA : k <= 0$. Algorithm M is run just once to obtain all the possible values of k for the Probit model given above. This design has 2 observations on 9 unique points, giving a sample size of 18 and 19683 configurations.

Table II. The probability that a subject exhibits EU.

Subject	$\hat{k}$	Exact	Asymptotic
1	0.104	0.855	0.864
2	0.162	0.954	0.947
3	–0.143	0.048	0.072
4	–0.346	0.000	0.005
5	0.255	0.996	0.988
6	–0.017	0.392	0.428
7	0.122	0.894	0.897
8	–0.104	0.107	0.136

The exact distribution for an individual subject is obtained by running algorithm T using the same design but with **b** set to the value of $\hat{k}$ observed for that subject. Algorithm E is then run to obtain the unique points of support, along with their probabilities. The exact probability that $k > 0$ for a subject is obtained by simply adding up all the probabilities for the positive points of support. The results for the eight subjects are reported in the following table, along with the probability one would have obtained using standard asymptotics.[8]

The probabilities above are the p-values for the test that k is positive for a given subject. This suggests that one would reject the null at the 5% level for subjects 3 and 4 using the exact values. The asymptotic results suggest that only the null for subject 4 would be rejected.

This example has been presented to demonstrate that enumeration can be used in empirical work. The further examination of all the designs presented in the Hey and Orme (1994) study, including the possibility of enumerating the 801900000 configurations required for the full risk neutral EU design, is left for future work.

6. Concluding Comments

This article has presented a method for the enumeration of the exact sampling distributions of statistics and estimators of binary data models. The technique was used to examine the finite sample behaviour of a misspecification test for the classical Probit model. The method has definite advantages over the alternative strategy of standard simulation. The efficiency gain of this method over simulation lets the researcher tackle a larger number of experimental designs than commonly used in econometric simulation experiments and is of greater applicability than competing enumeration algorithms.

The disadvantage of enumeration is that the method is restricted in the experimental designs it can cover by the need to have a small number of distinct sets of explanatory variables, **x**, making up the covariate matrix. This makes the method

unattractive for the data sets commonly used in applied work today. However, there are areas of modern economics where small samples and experimental designs can be found, and it is in these situations where asymtptotic approximations are likely to be poor. This was demonstrated in Section 5.

There are three different strands one might pursue when trying to increase the applicability of enumeration. The first *brute force* extension would just rely on computational power to tackle larger sample sizes. However, it may be better to move away from serial computation, as enumeration will lend itself to parallel or distributed operations. A second *theoretical* approach could be to employ some form of approximation or smoothing, as suggested in Diaconis and Holmes (1994), to reduce the overheads of this procedure while obtaining most of its advantages. A third area of extension is *design related* and involves conditioning the model on a fixed response set. Instead of generating all possible configuration, one would condition on all samples that admitted a fixed number of unit responses.

Acknowledgements

This article was based on chapter 3 of Peters (1995), which was supported by the University of Bristol postgraduate scholarship fund. The material presented here was supported by ESRC grant R000237386, and has benefitted from discussions with Ken Clark, John McDonald, Grant Hillier, Chris Orme, Gary Phillips, Peter Smith, and from the comments of the two anonymous referees. Any errors are the sole responsibility of the first author.

Notes

[1] Formula 1 decodes Knuth's formula 5 assuming the base location is 0 and the word length, Knuth's c, is 1. Knuth's array indices, dimensions and memory location correspond to S_{ij}, N_i and j respectively.

[2] The function $g(.)$ is a mapping from $\Re^p \rightarrow \Re$ and could be anything from a selection operation to obtain a specific estimate from the vector $\widehat{\mathbf{b}}_j$, to the calculation of a test statistic such as those used in Section 3 below.

[3] In October 1999, personal computers with 600 MHz processors, 37.5 GB hard disks, 768 MB RAM and writable CDs (650 MB disks) are readily available.

[4] All calculations were performed in double precision arithmetic on a Sun SPARCstation 2 running SUN-OS 4.1.2 or on a Hewlett-Packard HP9000/827s running HP-UX Release 9.0. Results' summaries were performed using Statistical Science Inc.'s S-Plus version 3.1. Estimation and service routines were written in Fortran77.

[5] The OPG test is calculated as $\hat{\mathbf{s}}'(\hat{\mathbf{V}}_{OPG})^{-1}\hat{\mathbf{s}}$ where $\hat{\mathbf{V}}_{OPG} = \sum_{n=1}^{N} \hat{s}_n \hat{s}_n'$.

[6] This example used a personal computer with a 400 MHz processor running Red Hat Linux version 5.1. Gentleman and Ihaka's R was used instead of Splus.

[7] The random term represents the numerical error made in the comparison of the two expected utilities.

[8] Results are given to 3 decimal places.

References

Berkson, J. (1955). Maximum likelihood and minimum χ^2 estimates of the logistic function, *Journal of the American Statistical Association*, **39**, 130–162.

Chesher, A.D. and Irish, M. (1987). Residual analysis in the grouped and censored normal linear model, *Journal of Econometrics*, **34**, 33–61.

Chesher, A.D. and Peters, S.A. (1995). Exact sampling distributions in binary data models. Bristol University Department of Economics, D.P. 95/392.

Clarkson, D.B. and Jennrich, R.I. (1991). Computing extended maximum likelihood estimates for linear parameter models. *Journal of the Royal Statistical Society Series B*, **53**, 417–426.

Cox, D.R. (1970). *The Analysis of Binary Data*. Chapman and Hall, London.

Davidson, R. and MacKinnon, J.G. (1984). Convenient specification tests for logit and probit models. *Journal of Econometrics*, **25**, 241–262.

Davidson, R. and MacKinnon, J.G. (1998). Graphical methods for investigating the size and power of hypothesis tests. *The Manchester School*, **66**, 1–26.

Diaconis, P. and Holmes, S. (1994). Gray codes for randomization procedures. *Statistics and Computing*, **4**, 287–302.

Fry, T.R.L. (1993). Univariate and multivariate burr distributions: a survey. *Pakistan Journal of Statistics*, **9A**, 1–24.

Hausman, J. and McFadden, D. (1984). Specification tests for the multinomial logit model. *Econometrica*, **52**, 1219–1240.

Hey, J.D. and Orme, C. (1994). Investigating generalisations of expected utility theory using experimental data. *Econometrica*, **62**, 1291–1326.

Hirji, K.F., Mehta, C.R. and Patel, N.R. (1987). Computing distributions for exact logistic regression. *Journal of the American Statistical Association*, **47**, 1110–1117.

Hirji, K.F. (1992). Computing exact distributions for polytomous response data. *Journal of the American Statistical Association*, **87**, 487–492.

Horowitz, J.L. (1994). Bootstrap-based critical values for the information matrix test, *Journal of Econometrics*, **61**, 395–411.

Hughes, G.A. and Savin, N.E. (1994). Is the minimum chi-square estimator the winner in logit regression? *Journal of Econometrics*, **61**, 345–366.

Knuth, D.E. (1973). *The Art of Computer Programming Volume 1: Fundamental Algorithms*. Addison-Wesley, Reading, Massachusetts.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. Chapman and Hall, London.

McDonald, J.W., Smith, P.W.F. and Forster, J.J. (1999). Exact tests of goodness of fit of log-linear models of rates. *Biometrics*, **55**, 620–624.

Nelson, F.D. and Savin, N.E. (1990). The danger of extrapolating asymptotic local-power. *Econometrica*, **58**, 977–981.

Orme, C. (1990). The small-sample performance of the information matrix test. *Journal of Econometrics*, **46**, 309–331.

Peters, S.A. (1995). Computational research tools for evaluating microeconometric misspecification tests. Ph.D. Thesis, Department of Economics, Bristol University, U.K.

Savin, N.E. and Würtz, A.H. (1999). The power of tests in binary response models. *Econometrica*, **67**, 413–421.

Silvapulle, M.J. (1981). On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society Series B*, **43**, 310–313.

Silvapulle, M.J. and Burridge, J. (1986). Existence of maximum likelihood estimates in regression models for grouped and ungrouped data. *Journal of the Royal Statistical Society Series B*, **48**, 100–106.

Skeels, C.L. and Vella, F. (1999). A Monte Carlo investigation of the sampling behaviour of conditional moments tests in tobit and probit models. *Journal of Econometrics*, **92**, 275–294.