COWLES FOUNDATION FOR RESEARCH IN ECONOMICS

AT YALE UNIVERSITY

Box 2125, Yale Station
New Haven, Connecticut 06520

COWLES FOUNDATION DISCUSSION PAPER NO. 636

A BOOKMAKER OR MARKET TYPE TEST FOR SPECIFICATION IN DISCRETE CHOICE MODELS

John J. Beggs

June 29, 1982

A BOOKMAKER OR MARKET TYPE TEST FOR SPECIFICATION IN DISCRETE CHOICE MODELS

by

John J. Beggs

ABSTRACT

This paper suggests that the predicted probabilities of outcomes given by an estimated discrete choice model be thought of as prices (or bookmaker odds) associated with those outcomes. By buying or selling contracts (gambling) at those prices (odds) it should not be possible to, on average, make a profit if the model is well specified and is generating "correct" prices. This notion then forms the basis of a model specification test.

INTRODUCTION

This paper suggests a simple test for the specification of discrete choice models which is motivated by conventional economic notions about the behavior of markets. Faced with evaluating the probabilities of discrete outcomes an efficient (better still, unbiased) market will set prices, in the form of odds, in such a way that neither a buyer nor a seller can expect on average to make profits. The estimated form of a discrete choice model should also have this simple desirable property. The predicted probabilities of each possible outcome should be such that if an individual offered to buy or sell contracts at the prices (odds) predicted by the model there would be an expected zero net gain; otherwise the model could be thought of as biasing the predictions in one direction or the other. The analogy to the case of a continuous dependent variable model is immediate, namely that the expected value of the residuals from the regression should be zero. This has been of limited interest in the regression context since least-squares regression ensures that either the sample mean, or some suitably weighted sum, of the regression residuals will equal zero, and most other estimation techniques are close relatives of these two cases.

The particular appeal of the test suggested below is that it poses the question of "goodness-of-fit" in a discrete choice model in a metric which is easily understood, and the test itself is particularly simple to mount.

TEST

Let P_i be a k-dimensional vector containing the probabilities associated with each of the k-possible outcomes for the i^{th} observation in the data set. P_i will be modelled quite generally as

$$(1) \qquad P_i = G(X_i, \theta_0)$$

where $G(\cdot)$ is a vector valued function on the unit simplex, X_i is a vector of explanatory variables pertaining to the i^{th} observation in space X, and θ_0 is an m-dimensional vector of model parameters known to lie in a compact set Θ. The null hypothesis is that the model is correctly specified, so that G is known, and that θ can be estimated consistently. Denote the estimate of the parameters from N observations as $\hat{\theta}$ where $(\hat{\theta} - \theta) = O_p(a_N)$ and $\lim_{N\to\infty} a_N = 0$. The predicted probabilities associated with the i^{th} observation are written

(2) $\hat{P}_i = G(X_i, \hat{\theta})$

$G(\cdot)$ will be assumed to have continuous second derivatives for all θ in Θ and that the mean values of the first and second derivatives of the elements of G and of the reciprocals of the elements of $G(\cdot)$ each converge uniformly in Θ to continuous functions of θ. This assumption augments other similar assumptions which might be placed on the $\{X_i\}$ sequence in the event that $G(\cdot)$ was given an explicit functional form. Assumptions of this type are typically required to ensure estimability of models and correspond to the "excitation conditions" in the familiar regression problem. Their purpose is to rule out certain pathological cases. To illustrate, consider the case where the $\{X_i\}$ sequence was such that on successive draws the resulting P vector converged towards a vertex of the simplex. The odds associated with one of the outcomes would then be increasing unboundedly. It is intuitively clear that there will be special difficulties in estimating and testing such a model specification since there is progressively less information available about one of the possible outcomes.

The test being considered here envisages a hypothetical gamble. It may be useful to first motivate the test with an example. The unit of observation could be thought of as a horse race. The model can be thought of as estimating the probability of a win for each of the k horses running the race. These estimates are stored in the vector $\hat{P}$. One may think of the model as being "unbiased" if by acting as a bookmaker and offering the pay-offs implied by $\hat{P}$ one would not on average either win or lose money.

Consider taking each observation and placing a \$1 bet on each of the k possible outcomes to yield a pay-off of $\$(1/\hat{P}_{ij})$ if the j^{th}

outcome is realized. The net profitability of following this strategy for each observation can be written,

$$\Pi_i(\hat{\theta}) = \sum_{j=1}^{k} \left(\frac{1}{P_{ij}(\hat{\theta})}\right) \cdot \Delta_{ij} - k \tag{3}$$

where,

$$\Delta_{ij} = \begin{cases} 1 & \text{if } j^{th} \text{ outcome realized on } i^{th} \text{ observation} \\ 0 & \text{otherwise} \end{cases} \qquad i = 1, 2, \ldots, N .$$

Now consider the average net profit taken over N observations, $\overline{\Pi}(\hat{\theta})$, and examine its Taylor's series expansion about θ_0 . This may be written,

$$\overline{\Pi}(\hat{\theta}) = \sum_{i=1}^{N} \frac{\Pi_i(\hat{\theta})}{N} = \sum_{i=1}^{N} \frac{\Pi_i(\theta_0)}{N} + (\hat{\theta}-\theta_0)' \left[\sum_{i=1}^{N} \frac{\Pi_i'(\theta_0)}{N}\right] - \frac{(\hat{\theta}-\theta_0)'}{2} \left[\sum_{i=1}^{N} \frac{\Pi_i''(\theta^+)}{N}\right] (\hat{\theta} - \theta_0) \tag{4}$$

where $\Pi_i'(\theta_0)$ is an m-dimensional vector of first derivatives of $\Pi_i(\cdot)$ evaluated at θ_0 , Π_i'' is an $m \times m$ matrix of second derivatives of $\Pi_i(\cdot)$ evaluated at θ^+ , a point on the line segment joining θ^0 and $\hat{\theta}$. It follows immediately from our assumptions that (4) may be rewritten as

$$\overline{\Pi}(\hat{\theta}) = \sum_{i=1}^{N} \frac{\Pi_i(\theta_0)}{N} + O_p(a_N) . \tag{5}$$

Attention then focuses on the first term on the right-hand side of equation (5). If the model is correctly specified then the $E[\Pi_i(\theta_0)] = 0$

and $\Pi_i(\theta_0)$ has some variance which was assured sumable by earlier assumption. It follows then that $\overline{\Pi}(\theta_0)$ will be $O_p(N^{-1/2})$. Hence average profits $\overline{\Pi}(\hat{\theta})$ will converge to zero with $O_p(\max(N^{-1/2}, a_N)]$.

The limiting distribution of $\overline{\Pi}(\hat{\theta})$ now is seen to depend on the order of magnitude of a_N . Each of the three cases can be considered. The first is that $a_N = o(N^{-1/2})$, so that the second and third terms on the right-hand side of equation (4) become asymptotically irrelevant. Then by the central limit theorem

$$(6) \qquad \sqrt{N}\cdot\overline{\Pi}(\hat{\theta}) \overset{\mathcal{L}}{\sim} N(0, \sigma_1^2)$$

where

$$(7) \qquad \sigma_1^2 = \lim_{N\to\infty} \frac{\sum_{i=1}^{N}}{N}\left(\sum_{j=1}^{k} \frac{1}{P_{ij}(\theta_0)} - k\right)^2 .$$

By arguments analogous to those above it can be shown that σ_1^2 can be estimated consistently by replacing $P_{ij}(\theta_0)$ by $P_{ij}(\hat{\theta})$ in (7).

The second case to consider is where $N^{-1/2} = o(a_N)$, so that the first and third terms on the right-hand side of (4) become asymptotically irrelevant. Consider the typical case of $\hat{\theta}$ having a limiting normal distribution denoted,

$$(8) \qquad a_N(\hat{\theta} - \theta_0) \overset{\mathcal{L}}{\sim} N(0,S)$$

and denote $\lim_{N\to\infty} \sum_i [\Pi_i'(\theta_0)/N]$ by the vector s . Then it follows that the limiting distribution of $\overline{\Pi}(\hat{\theta})$ is given

$$(9) \qquad a_N\, \overline{\Pi}(\hat{\theta}) \overset{\mathcal{L}}{\sim} N(0, \sigma_2^2)$$

where

(10) $$\sigma_2^2 = s'Ss$$

and s is estimated consistently by evaluation at $\hat{\theta}$.

Finally consider the case where $a_N = N^{-1/2}$, so that the first two terms on the right-hand side of (4) are of the same order in probability. The limiting distribution will then depend on the covariance between these two terms and will in general be unknown. However, in the most common cases encountered in practice (e.g., in logit, probit and log-linear models) $\hat{\theta}$ will be a maximum likelihood estimator and, given that the appropriate regularity conditions are satisfied for the model, it will be fully efficient. Consequently $\sqrt{N}(\hat{\theta}_{ML} - \theta_0)$ must be asymptotically uncorrelated with any other statistic of the data which has zero mean; otherwise there would be a possibility of improving the efficiency of the estimator by taking a linear combination of the two (a similar argument is used in Hausman (1978, 1981). It follows then that

(11) $$\sqrt{N}\ \bar{\Pi}(\hat{\theta}_{ML}) \overset{\mathcal{L}}{\sim} N(0,\ \sigma_1^2 + \sigma_2^2) \ .$$

This discussion then suggests that a test statistic could be created for each of these cases. These are shown below in equations (12a, b, c). They are denoted b to indicate bookmaker-type test. Each of the b's will have a limiting standard normal distribution.

(12a) $$b_1 = \frac{\sqrt{N}\ \bar{\Pi}(\hat{\theta})}{\sigma_1(\hat{\theta})}$$

(12b) $$b_2 = \frac{\sqrt{a_N}\ \bar{\Pi}(\hat{\theta})}{\sigma_2(\hat{\theta})}$$

$$b_3 = \frac{\sqrt{N}\,\overline{\Pi}(\hat{\theta}_{ML})}{\sqrt{\sigma_1^2(\hat{\theta}_{ML}) + \sigma_2^2(\hat{\theta}_{ML})}} \quad (12c)$$

This test can be considered as a form of specification test in the absence of a postulated alternative hypothesis. As with all specification tests it is necessary to identify those alternative hypotheses against which the test will be powerful. Suppose that the predicted outcome probabilities associated with the estimated misspecified model are denoted $\tilde{P}(\tilde{\theta})$, where $\tilde{\theta}$ are the estimates of the parameters of the misspecified model. The net outcome of writing contracts (gambling) at the prices (probabilities) predicted by the misspecified model may be written

$$\tilde{\overline{\Pi}}_N(\tilde{\theta}) = \frac{\sum_{i=1}^{N}}{N}\left(\sum_{j=1}^{k} \frac{\Delta_{ij}}{\tilde{P}_{ij}(\tilde{\theta})} - k\right) . \quad (13)$$

The above test will then be powerful for those cases where $\lim_{N\to\infty}\tilde{\overline{\Pi}}_N(\tilde{\theta}) \neq 0$ That is, for those cases of misspecification where, in the limit, the errors associated with offering incorrect prices do not manage to cancel one another out. In general such canceling out will occur only under very special conditions regarding the $\{X_i\}$ sequence. However, one pathological case can be noted, namely the $\tilde{P}_{ij}$ always equal $(1/k)$, so that $\tilde{\overline{\Pi}}$ is always equal to zero. That is, this test has no power against the alternative hypothesis that all outcomes are equally likely. Fortunately there are other simple tests that can be constructed to test the plausibility of equi-probable outcomes.

ILLUSTRATIVE EXAMPLE

Interfirm tender offers have become, in recent years, a major means of corporate growth. A substantial amount has been written regarding the economic efficiency aspects of takeovers but only recently has more attention been focused on the actual process involved in making and receiving a tender offer. This example is directed at the question of why a particular tender offer is or is not successful. The simple model proposed is "descriptive" rather than "structural" in the economic sense, but it does suggest ways in which the "bookmaker" test can be applied to a data set.

The data used for this study are a U.S. sample of 124 cash tender offers made in the years 1978-80. This represents approximately 40 percent of the cash tender offers made in those three years. A logit procedure was used to model the success or failure of the cash tender offers. Seven explanatory variables were considered. These are defined in Table 1, which also shows the results of two logit models fitted to the data. Individual t-statistics on the final three variables included in the model indicate that these are not individually statistically significant terms. The likelihood-ratio test against the null hypothesis that these three coefficients are equal to zero leads us to not reject the null hypothesis. The "bookmaker" test suggests, however, that we should not accept the specification which restricts the final three coefficients to be zero. In doing so the test statistic suggests that the second specification is a "correct" specification of the model in that the b_3 statistic is not significantly different from zero at a 95 percent confidence level.

The interpretation of the results is rather self evident though it seems a good deal can be said about the likely success of a merger

TABLE 1

LOGIT MODEL RESULTS
SUCCESS OF CASH TENDER OFFER

Variable		Model 1	Model 2
Success	1 if approx. number of shares requested were acquired, 0 otherwise.		
Foreign	1 if bidder a foreign firm, 0 otherwise	0.20 (0.04)	0.15 (0.03)
Industrial	1 if target a mining/manufacturing firm, 0 otherwise	-0.21 (0.08)	-0.32 (0.12)
Previous Holding	1 if any shares in target previously held by acquiror, 0 otherwise	-0.16 (0.02)	-0.09 (0.02)
Premium	percentage offer price over market price 2 weeks before offer	0.42 (0.18)	0.71 (1.31)
Total Offer Price	1 if $100 m, 0 otherwise		0.12 (0.08)
P-E Ratio	Difference between P-E ratio of target and average of all targets in sample		-1.32 (0.90)
Return on Equity	Ratio of rate of return on equity of target firm to its industry average		1.21 (0.89)
		$b_3 = 2.84$	$b_3 = 1.62$

Log Likelihood Ration Statistic on H_0 that last three variables are zero is 2.39 which leads us to accept the null hypothesis as it is well with the 95% confidence interval.

Standard errors are given in parentheses.

from the characteristics of the bidder and target firms. The estimation and testing techniques are asymptotic in nature and the usual caution must be taken regarding their interpretation in the current sample of 124 observations.

CONCLUSION

The test suggested here is inspired by the idea that if the discrete choice model were predicting prices in the form of odds it should not be possible on average to make profits by selling contracts (gambling) at those prices. From the point of view of the user of statistics the test is conceptually more transparent than the familiar tests of these models and has the added advantage that it is not necessary to formally specify the alternative hypotheses. In offering a novel perspective on the discrete choice model the paper is quite suggestive of areas for future research.

REFERENCES

Hausman, J. (1978), "Specification Tests in Econometrics," Econometrica, V. 46.

Hausman, J., D. McFadden (1981), "Specification Tests for the Multinormal Logit Model," Massachusetts Institute of Technology Working Paper No. 292.