A MONTE CARLO STUDY OF A GENERALIZED MAXIMUM ENTROPY ESTIMATOR OF THE BINARY CHOICE MODEL

Lee Adkins

ABSTRACT

Monte Carlo simulation is used to study the finite sample statistical properties of a maximum entropy estimator of the binary choice model. The estimator is similar in spirit to generalized maximum entropy estimators (GME) proposed by Golan et al. (1996) and is nonparametric in the sense that no assumption is made about the underlying density function generating the data. The GME is compared to the probit and logit MLEs as well as the OLS estimator of the linear probability model. The results indicate that the GME has squared error risk lower than that of the probit and logit for all experimental designs. The GME has lower squared error risk than OLS when the inherent variability of $x'\beta$ is large and when the number of independent variables is small. Increasing the number of support points improves performance. Unfortunately, the in-sample and out-of-sample predictive abilities of the GME estimator leaves something to be desired. Only when the signal-to-noise ratio is very small is the GME a good choice.

Advances in Econometrics, Volume 12, pages 183–197.

ISBN: 0-7623-0187-2

I. INTRODUCTION

In binary choice models the probability of observing an event is related to a set of explanatory variables. Members of the class of binary choice models differ in the particular way the probabilities are modeled. In the probit model they are modeled using the normal cumulative distribution function (cdf), in the logit model the logistic cdf, in the linear probability model they are simply linear functions of the independent variables. The choice of model is often made based on common practice within the researcher's discipline and without regard to the actual data generation process associated with the sample.

Maximum likelihood estimation of the probit and logit models require numerical optimization since the first derivatives of the log-likelihood function cannot be explicitly solved for the unknown parameters. The MLE is attractive since it is known to have good asymptotic properties.

In this chapter, a maximum entropy estimator of the binary choice model is proposed that requires a minimum number of assumptions about the stochastic nature of the equation's error. The estimator is similar in spirit to generalized maximum entropy estimators proposed by Golan, Judge and Miller (1996a) and Golan, Judge and Miller (1996b). The required assumptions that are:

1. the underlying probability of an event must be positive, but not greater than 1
2. the underlying probability of an event is a function of the independent variables
3. (optional) feasible support for the model's errors is the interval $[-1,1]$

Probability measures require that probabilities meet the first assumption. The second assumption is reasonable if data are being used to ascertain the underlying probabilities. The third assumption is optional, but when used it enables one to improve inference in the model. The proposed estimator is nonparametric in the sense that no assumption is made about the underlying density function generating the data. All that we need to know is that the process has a binary outcome and that the independent variables contain information about the probabilities of observing the event.

II. MODEL AND ITS ESTIMATORS

In the binary choice model the dependent variable takes the value of one or zero, indicating whether or not an event occurs. Let y_t be a binary random variable taking the value of 1 or 0 and x_t a known $K \times 1$ vector of explanatory variables associated with the t^{th} observation. The probability of observing the event for individual t is

$$p_t = Pr(y_t = 1 \mid x_t'\beta) = G(x_t'\beta) > 0 \tag{1}$$

for $t = 1, 2, \ldots, T$, β is $(K \times 1)$ vector of unknown parameters and $G(\cdot)$ is a function linking the probabilities to the linear index $(x_t'\beta)$. The function $G(\cdot)$ maps the real line into the [0,1] interval.

Following Golan et al. (1996a, 1996b), the model can be written

$$y_t = G(x_t'\beta) + e_t = p_t + e_t \tag{2}$$

where the p_t are the unknown probabilities and the e_t are unobservable errors contained in the $[-1,1]$ interval. In vector notation equation (2) can be written

$$y = p + e \tag{3}$$

where y, p, and e are $(T \times 1)$ vectors.

Explanatory variables can be used in conjunction with (3) to form the following moment condition

$$X'y = X'p + X'e \tag{4}$$

At this point, there are more unknown parameters than observations. As will be discussed in the next section, Jaynes (1957a) has proposed a practical solution to this problem using Shannon's (1948) entropy measure.

The traditional maximum likelihood method maximizes

$$L(G, \beta) = \sum_{t=1}^{T} y_t \ln G(x_t'\beta) + (1 - y_t)\ln(1 - G(x_t'\beta)) \tag{5}$$

The logit MLE is obtained if $G(\cdot)$ is a logistic cdf and the probit MLE is obtained if it is a standard normal (i.e., $N(0,1)$) cdf.

Another common choice for $G(x_t'\beta)$ is the simple linear function $(x_t'\beta)$. This function is not a proper cdf since it does not map into the [0,1] interval. Nevertheless, it is often used because β can be estimated consistently using ordinary least squares.

In this chapter a generalized maximum entropy estimator is presented and its properties are compared to those of the MLE logit (LMLE), MLE probit (PMLE), and the ordinary least squares estimator (OLSE) of the linear probability model in a Monte Carlo study.

III. MAXIMUM ENTROPY ESTIMATORS

Jaynes (1957a), (1957b) proposed a way to estimate the unknown probabilities of a discrete probability distribution when there are fewer observations than parameters. Given a set of distributions that are consistent with the way the data are generated, he proposed using the one that selects the most "uncertain" one. Jaynes measured uncertainty using Shannon's (1948) entropy measure

$$H(p) = -\sum_{t=1}^{T} p_t \ln(p_t). \tag{6}$$

The entropy function (6) is maximized subject to any known restrictions on the probabilities, p_t.

The maximum entropy (ME) approach assumes that our prior knowledge is limited and assumes that each outcome is equally likely *a priori*. Thus, according to Jaynes (1985) the ME solution "agrees with what is known, but expresses 'maximum uncertainty' with respect to all other matters."

The approach has been used to estimate the parameters of multinomial discrete choice problems by Denzau, Gibbons and Greenburg (1989) and Soofi (1992) where it is shown to be equivalent to MLE multinomial logit estimator. Golan, Judge, and Perloff (1996c) generalized this approach by introducing error terms into the formulation of the unknown probabilities as in equation (2). The error terms are modeled as the mean of a finite set of known points defined on that support; the distribution of the errors is otherwise unspecified and the probabilities that the errors take specified values on the support are treated as unknown parameters.

Following Judge (1991) and Judge, Golan, and Miller (1993) the errors are reparameterized in the following way. The errors are assigned probabilities of taking on various values on the interval [−1,1]. The hypothesized values for the errors are denoted $v_t' = [v_{t1}, v_{t2}, \ldots, v_{tM}]$ where $M \geq 2$ and have corresponding unknown probabilities (or weights), $w_t = [w_{t1}, w_{t2}, \ldots, w_{tM}]'$. The transformed errors become

$$e = Vw = \begin{bmatrix} v_1' & & & \\ & v_2' & & \\ & & \ddots & \\ & & & v_T' \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_T \end{bmatrix} \tag{7}$$

with $w = (w_1, w_2, \ldots, w_T)'$ which is $MT \times 1$. In the binary choice model, the generalized entropy function is

$$\max_{p,w} -p' \ln p - w' \ln w \tag{8}$$

subject to the data consistency constraints

$$X'y = X'p + X'Vw \tag{9}$$

and a restriction that ensures the weights on the errors sum to one

$$\sum_{i=1}^{M} w_{ti} = 1 \quad t = 1, 2, \ldots, T \tag{10}$$

The corresponding Lagrangian is

$$L = -p' \ln p - w' \ln w + \lambda'(X'y - X'p - X'Vw) + \delta' \left(1 - \sum_{i=1}^{M} w_i \right) \tag{11}$$

which is maximized with respect to p, w, λ, and δ. In addition, the Lagrange multiplier associated with the data consistency constraints, $-\lambda$, is the GME estimate of β.

The path taken in this paper is similar to that the multinomial approach suggested by Golan, Judge, and Perloff (1996c) and is completely consistent with the entropy approach and analogous to that taken in binary choice models estimated based on the maximum likelihood principle. In the multinomial case each choice is parameterized by a unique set of parameters *after* normalization and the maximum entropy estimator (ME) yields a set of coefficients that correspond to those of the multinomial logit model's MLE.

In the setup considered in this paper a single set of probabilities are estimated for the event which implicitly yields a single set of parameter estimates for the underlying βs; no separate normalization is required. The probability of the alternative choice, p_A, is treated as it is in the usual binary choice model where it is implicitly assumed to be $1 - p$; a formal constraint is not imposed. One disadvantage of formulating the GME estimator in this way is that the estimated probabilities and coefficients are not invariant to how the choice is coded. The major advantages of this approach is the dramatic improvement in terms of squared error loss over the ME (MLE logit) specification.

IV. EXPERIMENTAL DESIGN

The goal is to examine how the probit maximum likelihood and OLS estimators compare to the generalized maximum entropy estimator (GME) in terms of squared error loss. Towards this end, three economic examples are chosen, each representing a distinct experimental design. The models vary by sample size, number of parameters to estimate, and in the inherent variability of $x_i'\beta$. In the probit model, if the variability of $x_i'\beta$ is too small, then values of $F(x_i'\beta)$ will not show much variation and the probabilities will be clustered around .5. On the other hand, if the variability is too great, then probabilities are clustered around the extreme values of 0 or 1 (see Griffiths, Hill and Pope 1987 for details).

$Var(X\beta)$ is computed using $(T-K)^{-1}\beta'(X-\bar{X})'(X-\bar{X})\beta$. In practice this quantity is unknown since it depends on the unknown parameters, β. Researchers are more familiar with the likelihood based pseudo-R^2 which is also reported. Although the correspondence is not perfect, higher values of R^2 tend to be associated with larger values of $Var(X\beta)$. To achieve different amounts of variability, the data are generated using 3 values of the parameter vector. One set is estimated from the data set using

the probit MLE, another using the logit MLE, and the third using OLS. As Maddala (1983) points out, the probit MLE, logit MLE and the OLS estimator can be made comparable by simple linear transformations. In general, $|\beta_{logit}| \geq |\beta_{probit}| \geq |\beta_{OLS}|$. The data generation procedure is as follows:

1. The probit MLE, the logit MLE, and the OLS estimates are obtained for β using the original T observations on y_t and the K explanatory variables.
2. Letting X be the $T \times K$ matrix of explanatory variables, 500 samples of the latent vector $y_j = X\beta_i + e_j$, i = probit, logit, OLS, that are generated using (e_j) which is a $T \times 1$ vector of random deviates from the standard normal $(j = 1)$ or logistic $(j = 2)$ distributions and $y_{jt} = 1$ assigned if $y_{jt} \geq 0$. Hence in one set of experiments, the probit model is the true DGP (i.e., $j = 1$) and in the other the logit model is the appropriate one to use.
3. For each sample and each value of β the probit, logit, OLS, and GME estimators are obtained.
4. For each estimator and each sample the squared error loss, $L(\hat{\beta}_s - \beta_s)'(\hat{\beta}_s - \beta_s)$, is computed where $\beta_s = (\beta_2, \beta_3, \ldots, \beta_K)'$ is the vector of coefficients excluding the constant. Empirical risk is the arithmetic average of the loss function. Of course, small estimation risk does not necessarily imply that a specific parameter will be estimated with small mean square error.
5. For each estimator and each sample the in-sample prediction loss is computed based on $L(\hat{p} - p)'(\hat{p} - p)$ where $\hat{p}$ is the $T \times 1$ vector of in-sample predicted probabilities associated with each model and p is vector containing the actual probabilities. The true probabilities for the probit model $(j = 1)$ are computed $F(X\beta)$ with $F()$ being the normal CDF; for the logit model $(j = 2)$ they are computed using $L_*(X\beta)$ with $L_*()$ being the logistic CDF. $F(GME)$ and $L_*(GME)$ are also computed by plugging in the GME estimates of β into the normal and logistic cdfs, respectively. These are offered as semi-parametric alternatives to the usual probabilities estimated directly by the GME estimator. The rationale for $F(GME)$ and $L_*(GME)$ is that better estimates of β provided by the GME should lead to improved in- and out-of-sample predictions. The obvious problem with this in practice is that $F()$ or $L_*()$ must be chosen without knowledge of the actual DGP. Again, the arithmetic average loss and its Monte Carlo standard error are reported.[1]
6. For each estimator and each sample the out-of-sample prediction loss is computed using $L(\hat{p}_o - p_o)'(\hat{p}_o - p_o)$ where $\hat{p}_o$ is the $T \times 1$ vector containing the out-of-sample predicted probabilities associated with each model and p_o contains the actual probabilities. The true probabilities for the probit model are $F(X_o\beta)$ with $F()$ being the normal CDF and $L_*(X\beta)$ with $L_*()$ being the logistic CDF. Out-of-sample predicted probabilities based on the GME are computed using $F(GME)$ and $L_*(GME)$ by plugging in the GME estimates into the normal and logistic cdfs, respectively. The out-of-sample

values X_o are generated by randomly drawing (with replacement) T observations from each of the independent variables. For example, suppose $X = \{x_1, x_2, \ldots, x_K\}$ where x_k is the $T \times 1$ vector of observations on the k^{th} independent variable. The elements of x_k are sampled randomly with replacement to obtain x_{ko} the $T \times 1$ vector of out-of-sample values. The process is repeated for each of the independent variables and becomes $X_o = \{x_{1o}, x_{2o}, \ldots, x_{Ko}\}$.

7. Several choices of the support for each of the GME errors are used. They are $v_1 = T^{1/2}(-1, 0, 1)$, $v_2 = T^{1/2}(-1, -.5, 0, .5, 1)$, and $v_3 = T^{1/2}(-1, -.75, -.5, -.25, 0, .25, .5, .75, 1)$. Their respective starting values are $w_1 = (.33, .34, .33)'$, $w_2 = (.2, .2, .2, .2, .2)'$, and $w_3 = (.1, .1, .1, .1, .2, .1, .1, .1, .1)'$,
8. The starting values for the probabilities, p, in the GME are obtained using the relation $\hat{p} = \exp(X\beta_{Logit})/(1 + \exp(X\beta_{Logit}))$ which initially satisfies the data consistency constraint.
9. The empirical probability that the squared error risk of the GME is greater than that of the probit, logit, and OLS estimator are computed. Each of the data generation processes is described in detail below and the results from the Monte Carlo experiments are summarized.

A. School Vote Data

The probit model has been used by Rubinfeld (1977) to study the voting decisions of individuals in a local school milage referendum. The abridged data set consists of 95 observations and can be found in Pindyck and Rubinfeld (1981). The binary dependent variable takes the value 1 if an individual votes yes and 0 if no. Of the eight independent variables, five are binary and three are continuous. A constant term is added for a total of nine variables.

B. The Mortgage Data

Dhillion, Shilling, and Sirmans (1987) investigate the financial and personal characteristics that influence home-buyers to select either a fixed rate or variable (adjustable) rate mortgage. Their probit model and data is discussed in Lott and Ray (1992). The binary dependent variable, y, takes the value 1 if a variable rate mortgage is chosen and 0 if a fixed rate is chosen. There are $K = 16$ explanatory variables, including a constant, that may be characterized as financial variables (5) or personal characteristics (10). There are $T = 78$ observations on household choices and the explanatory variables.

C. Voting Data

The data for this example appear in Greene (1990). The purpose is to explore factors which affect the probability that the Democratic candidate wins a state in the 1976 presidential election. The dependent variable is 1 if the state was won by

the Democrat, Jimmy Carter, and 0 if won by the Republican, Gerald Ford. The independent variables are: 1975 median income by state, the percentage of the population living in a Standard Metropolitan Statistical Area as defined by the Census Bureau, average years schooling in the state, and a regional dummy variable which takes the value of 1 for southern states and 0 otherwise.

V. RESULTS

The results under squared error loss for the simulation appear in Tables 1 and 2 below. For each data set, the true parameters were generated using β_{Probit}, β_{Logit}, and β_{OLS}. The support is v_3 which resulted in slight risk improvements over v_2 and v_1. Empirical risk, the Monte Carlo estimate of risk's standard error (S.E.), and empirical probability of the GME having greater risk than OLS, logit, and probit are given for each parameter vector and data set.

The GME has squared error risk lower than that of the probit and logit for all samples. The GME has lower squared error risk than OLS for larger values of $Var(X\beta)$ and for smaller values of K.

The following regression models are used to further condense the results. The risk difference between OLS and GME is regressed on $Var(X\beta)$ and K. The results are:

$$\underset{}{\text{Risk}_{OLS} - \text{Risk}_{GME}} = \underset{(2.19)}{15.91} + \underset{(4.39)}{5.32\, Var(X\beta)} - \underset{(-2.56)}{1.84 K}$$

T-statistics appear in parentheses. Increasing $Var(X\beta)$, holding K constant, increases the riskiness of OLS relative to the GME. Increasing K, holding $Var(X\beta)$ constant, decreases the riskiness of OLS relative to the GME. Thus, equations with large numbers of explanatory variables would require a large $Var(X\beta)$ in order for the GME to have lower risk than the OLS estimator. On the other hand, for small models like that associated with the Voting data, the GME is more likely to produce lower risk than OLS. Since $Var(X\beta)$ is unknown in practice, caution is advised. The clear result is that the GME will work better the smaller K, other things equal.

A similar approach is taken to describe the effect of K and R^2 on the probability of GME having greater squared error risk than the PMLE and OLS estimators. In this case $Var(X\beta)$ is a less reliable guide than the population pseudo-R^2. The regression for the PMLE is:

$$\underset{}{\text{Prob}(L_{GME} > L_{Probit})} = \underset{(-0.99)}{-0.231} + \underset{(2.12)}{1.09 R^2} + \underset{(1.17)}{0.017 K}$$

Increasing K actually increases the probability that the loss of the GME is greater than that of the PMLE. Although the risk of using PMLE is generally higher than that of the GME, increasing the number of independent variables increases the probability that GME loss is greater than that of PMLE. This suggests that when

Table 1. Empirical Risks and Estimated Probabilities, Normal Errors

		Estimator			
	Data Set	*Probit*	*Logit*	*OLS*	*GME*
		$\beta = \beta_{OLS}$			
Var($X\beta$) = .044	School Vote				
K = 9	Risk	2.17	6.52	.399	1.08
R^2 = .064	S.E.	.073	.235	.009	.044
	Prob	.03	0	.942	—
Var($X\beta$) = .132	Mortgage				
K = 16	Risk	14.33	44.26	.714	2.55
R^2 = .202	S.E.	1.91	5.99	.023	.123
	Prob	.01	0	.962	—
Var($X\beta$) = .136	Voting				
K = 5	Risk	6.09	20.29	1.89	2.40
R^2 = .223	S.E.	.511	1.90	.073	.138
	Prob	.370	.198	.410	—
		$\beta = \beta_{PROBIT}$			
Var($X\beta$) = .664	School Vote				
K = 9	Risk	3.46	16.46	4.03	2.00
R^2 = .311	S.E.	.148	.678	.024	.068
	Prob	.118	.014	.100	—
Var($X\beta$) = 2.759	Mortgage				
K = 16	Risk	59.33	115.75	8.89	9.10
R^2=.430	S.E.	12.34	5.40	.065	.521
	Prob	.102	.006	.296	—
Var($X\beta$) = 2.14	Voting				
K = 5	Risk	14.21	84.09	21.31	9.96
R^2 = .498	S.E.	1.49	6.84	.183	.716
	Prob	.630	.306	.042	—
		$\beta = \beta_{LOGIT}$			
Var($X\beta$) = 1.81	School Vote				
K = 9	Risk	5.60	23.74	12.84	3.49
R^2 = .270	S.E.	.316	1.35	.033	.104
	Prob	.408	.030	.008	—
Var($X\beta$) = 8.04	Mortgage				
K = 16	Risk	103.84	140.93	33.49	13.55
R^2 = .368	S.E.	12.19	5.06	.089	.397
	Prob	.226	.016	.032	—
Var($X\beta$) = 6.12	Voting				
K = 5	Risk	48.68	266.40	76.92	22.40
R^2 = .419	S.E.	7.66	19.93	.235	1.29
	Prob	.608	.198	.022	—

Table 2. Empirical Risks and Estimated Probabilities, Logistic Errors

		Estimator			
	Data Set	*Probit*	*Logit*	*OLS*	*GME*
		$\beta = \beta_{OLS}$			
Var($X\beta$) = .044	School Vote				
K = 9	Risk	1.99	5.33	.493	1.26
R^2 = .046	S.E.	.062	.181	.010	.042
	Prob	.08	0	.964	—
Var($X\beta$) = .132	Mortgage				
K = 16	Risk	8.63	24.87	.907	3.42
R^2 = .136	S.E.	.710	1.96	.027	.242
	Prob	.024	0	.98	—
Var($X\beta$) = .136	Voting				
K = 5	Risk	5.10	13.52	2.33	3.11
R^2 = .153	S.E.	.282	.842	.083	.161
	Prob	.320	.208	.476	—
		$\beta = \beta_{PROBIT}$			
Var($X\beta$) = .664	School Vote				
K = 9	Risk	2.97	6.34	4.81	2.65
R^2 = .311	S.E.	.076	.253	.030	.065
	Prob	.294	.144	.034	—
Var($X\beta$) = 2.759	Mortgage				
K = 16	Risk	19.49	58.66	9.81	8.91
R^2 = .377	S.E.	1.57	4.19	.078	.366
	Prob	.246	.052	.226	—
Var($X\beta$) = 2.14	Voting				
K = 5	Risk	10.76	20.23	25.63	15.62
R^2 = .415	S.E.	.512	1.64	.252	.488
	Prob	.886	.680	.046	—
		$\beta = \beta_{LOGIT}$			
Var($X\beta$) = 1.81	School Vote				
K = 9	Risk	5.25	9.70	13.96	5.48
R^2 = .250	S.E.	.12	.44	.046	.113
	Prob	.702	.426	.002	—
Var($X\beta$) = 8.04	Mortgage				
K = 16	Risk	960	95.41	34.56	18.14
R^2 = .420	S.E.	913	6.17	.120	.479
	Prob	.428	.162	.040	—
Var($X\beta$) = 6.12	Voting				
K = 5	Risk	25.07	36.19	83.42	45.16
R^2 = .460	S.E.	1.17	4.79	.348	1.02
	Prob	.952	.768	.006	—

the PMLE is bad, it is really bad. The GME is apparently less prone to large losses. The large S.E. associated with PMLE is consistent with this. The probability regression for OLS is:

$$\text{Prob}(L_{GME} > L_{OLS}) = 1.277 - 2.264R^2 - 0.026K$$
$$(4.40) \quad (-3.51) \quad (-1.45)$$

Increasing K and R^2 decreases the probability that the loss of the GME is greater than that of the OLS.

Table 3. In-Sample Prediction Risks

		Estimator				
Var($X\beta$)	Data Set	*Probit*	*Logit*	*OLS*	*GME*	*F(GME)*
			$\beta = \beta_{OLS}$			
.044	School Vote					
K = 9	Risk	1.99	2.00	1.93	2.69	1.14
	S.E.	.044	.044	.042	.054	.032
.132	Mortgage					
K = 16	Risk	3.49	3.50	2.98	3.89	1.98
	S.E.	.052	.052	.043	.062	.045
.136	Voting					
K = 5	Risk	1.03	1.03	.960	1.52	.670
	S.E.	.030	.030	.028	.037	.022
			$\beta = \beta_{PROBIT}$			
.664	School Vote					
K = 9	Risk	1.84	1.86	1.81	1.97	2.47
	S.E.	.038	.038	.033	.038	.033
2.759	Mortgage					
K = 16	Risk	2.96	2.94	2.66	3.02	2.72
	S.E.	.045	.040	.033	.041	.038
2.14	Voting					
K = 5	Risk	.676	.695	.669	.723	.735
	S.E.	.022	.023	.016	.021	.021
			$\beta = \beta_{LOGIT}$			
1.81	School Vote					
K = 9	Risk	1.64	1.66	2.04	2.07	2.79
	S.E.	.033	.034	.025	.037	.035
8.04	Mortgage					
K = 16	Risk	2.33	2.28	2.89	2.60	2.48
	S.E.	.034	.032	.023	.030	.029
6.12	Voting					
K = 5	Risk	.706	.716	.983	.739	.689
	S.E.	.019	.019	.010	.019	.019

The squared error risks for the models with logistic errors are presented in Table 2. Some differences should be noted. First, the PMLE has lower risk than the GME in several cases. It appears that as Var($X\beta$) (or R^2) increases the PMLE tends to perform better both in terms of risk and in the probability of having lower risk than the GME estimator. In addition, increasing K holding variability constant now increases the risk difference between OLS and the GME estimators. This is apparent in the following regression based on the results in Table 2.

Table 4. Out-of-Sample Prediction Risks

		Estimator			
Var($X\beta$)	Data Set	*Probit*	*Logit*	*OLS*	*F(GME)*
		$\beta = \beta_{OLS}$			
.044	School Vote				
K = 9	Risk	2.92	2.93	3.05	1.57
	S.E.	.074	.074	.084	.044
.132	Mortgage				
K = 16	Risk	7.03	7.05	6.97	3.70
	S.E.	.123	.121	.171	.096
.136	Voting				
K = 5	Risk	1.81	1.80	1.85	1.17
	S.E.	.056	.055	.070	.045
		$\beta = \beta_{PROBIT}$			
.664	School Vote				
K = 9	Risk	2.46	2.45	3.48	2.85
	S.E.	.062	.062	.083	.050
2.759	Mortgage				
K = 16	Risk	6.96	6.92	5.96	5.74
	S.E.	.139	.135	.120	.118
2.14	Voting				
K = 5	Risk	1.21	1.23	1.22	1.91
	S.E.	.050	.051	.042	.063
		$\beta = \beta_{LOGIT}$			
1.81	School Vote				
K = 9	Risk	2.13	2.14	4.48	3.14
	S.E.	.054	.055	.068	.049
8.04	Mortgage				
K = 16	Risk	7.16	7.04	6.67	6.40
	S.E.	.138	.133	.081	.126
6.12	Voting				
K = 5	Risk	1.10	1.11	1.76	1.71
	S.E.	.042	.043	.033	.057

$$\underset{\text{(-3.3)}}{\text{Risk}_{OLS} - \text{Risk}_{GME} = -13.26} + \underset{\text{(5.35)}}{3.24 Var(X\beta)} + \underset{\text{(3.74)}}{1.34 K}$$

The probability that the GME estimator has a higher loss than the LMLE the results increases as the data become more variable.

The results for in-sample prediction for the models having standard normal errors appear in Table 3. When Var($X\beta$) is very small, F(GME) has lower prediction risk than the other estimators. Interestingly enough, the usual GME never outperforms probit, logit, or OLS estimators. In-sample predictions from logit and probit are very similar to one another and perform fairly well even as Var($X\beta$) increases. The results for the models having logistic errors are very similar and are not reported.

The results for out-of-sample prediction appear in Table 4. When Var($X\beta$) is very small, F(GME) performs better than the other estimators. The predictions based on the GME estimates of β has higher risk than probit or logit except for in the Mortgage data with Var($X\beta$) = 2.759. Again, it is difficult to come to any general conclusions about the choice of estimator for out-of-sample prediction. The results for the models having logistic errors are completely similar and are not reported.

As for choosing the number of support points, using more improves performance. Support v_3 performs results in very modest risk improvements over v_2 which in turn performs slightly better than v_1. Even though the feasible support is [−1,1], substantial gains occurred by making the endpoints equal to $-/+T^{1/2}$.

VI. CONCLUSION

If the goal is to obtain estimates of the underlying parameters of a binary choice model, then the GME estimator is a good choice under certain circumstances. In particular, if the number of independent variables is relatively small (e.g., less than 7 or 8) the likelihood of squared error risk improvements over the next best alternative (OLS) is fairly high. This is especially true if the inherent variability of the data are high. Unfortunately, this last feature cannot be known with certainty since the variability depends on the unknown parameters.

Improved specification of the error's support also improves inference using the GME. The GME appears to be more precise than either the PMLE or the LMLE under most circumstances and may yet be used to increase the power of hypothesis tests about the parameters of the model.

Unfortunately, the in-sample and out-of-sample predictive abilities of the GME estimator as specified in this paper leaves something to be desired. As the signal-to-noise ratio gets small (i.e., when $Var(X\beta)$ small), the GME is a good choice. Unfortunately, as this value increases, its performance relative to the other estimators diminishes. A clear cut recommendation is not possible, although the performance of the GME is better the larger the number of βs.

In addition, the computational burden of estimating the GME using the constrained optimization routine increases rapidly with sample size. One sample was chosen that has 753 observations. The CO algorithm (Schoenberg 1995) used to obtain the results under GAUSS 3.2.13 was unable to process this data set under 32 Megabytes of system RAM. The algorithm is fairly quick however when samples are small (<100) and when there are relatively few explanatory variables. One possible solution to the computational problems is to reformulate the entropy using the multinomial approach followed by Golan et al. (1996a). They derive a dual unconstrained generalized entropy function which is used to estimate the parameters of interest.

ACKNOWLEDGMENT

I would like to thanks Professor's George Judge, Amos Golan, Carter Hill, and Tom Fomby for their comments. Any errors in the manuscript are mine. Send correspondence to Department of Economics, Oklahoma State University, Stillwater, OK 74075, internet: Ladkins@okway.okstate.edu

NOTE

1. The results were quite similar when prediction loss was computed based on the average number of misses. A miss is measured whenever the $\hat{p}_t > .5$ and $y_t = 0$ or $\hat{p}_t \leq .5$ and $y_t = 1$.

REFERENCES

Denzau, A.T., P.C. Gibbons, and E. Greenburg (1989). "Bayesian Estimation of Proportions with a Cross-Entropy Prior." *Communications in Statistics Theory and Methods* 18, 1843–1861.

Dhillion, U.S., J.D. Shilling, and C.F. Sirmans (1987). "Choosing Between Fixed and Adjustable Rate Mortgages." *Journal of Money Credit and Banking* 19, 260–267.

Golan, A., G. Judge, and D. Miller (1997). "The Maximum Entropy Approach to Estimation and Inference: An Overview." In *Advances In Econometrics*, Vol. 12, edited by T. Fomby and R.C. Hill. Greenwich, CT: JAI Press.

Golan, A., G. Judge, and D. Miller (1996a). *Maximum Entropy Econometrics: Robust Estimation with Limited Data.* New York: John Wiley and Sons.

Golan, A., G. Judge, and J.M. Perloff (1996b). "A Maximum Entropy Approach to Recovering Information from Multinomial Response Data." *Journal of the American Statistical Association* 87, 841–853.

Greene, W.H. (1990). *Econometric Analysis.* New York: Macmillan.

Griffiths, W.E., R.C. Hill, and P. Pope (1987). "Small Sample Properties of Probit Model Estimators." *Journal of the American Statistical Association* 82, 929–937.

Jaynes, E.T. (1957a). "Information Theory and Statistical Mechanics." *Physics Review* 106, 620–630.

Jaynes, E.T. (1957b). "Information Theory and Statistical Mechanics II." *Physics Review* 108, 171–190.

Jaynes, E.T. (1984a). *Inverse Problems*, edited by D.W. McLaughlin. Providence, RI: American Mathematical Society, 151–166.

Jaynes, E.T. (1984b). "Prior Information and Ambiguity in Inverse Problems." *Physics Review* 108, 620–630.

Jaynes, E.T. (1985). "Where Do We Go From Here?" In *Maximum Entropy and Bayesian Methods in Inverse Problems*, edited by C.R. Smith and W.T. Grandy. New York: D. Reidel Publishing Company.

Judge, G.G. (1991). *A Reformulation of Ill-posed Inverse Problems with Noise.* Berkeley CA: University of California Press.

Judge, G.G., A. Golan, and D. Miller (1993). *Recovering Information in the Case of Ill-posed Inverse Problems with Noise.* Berkeley, CA: University of California Press.

Lott, W.F. and S.C. Ray (1992). *Applied Econometrics: Problems with Data Sets.* Fort Worth, TX: Dryden Press,

Maddala, G.S. (1983). *Limited Dependent and Qualitative Variables in Econometrics.* Cambridge, MA: Cambridge University Press.

Pindyck, R.S. and D. Rubinfeld (1981). *Econometric Models and Economic Forecasts*, 2nd Ed. New York: McGraw-Hill.

Rubinfeld, D.L. (1977). "Voting in a Local School Election: A Micro Analysis." *Review of Economics and Statistics* 59, 30–42.

Schoenberg, R. (1995). *Constrained Optimization.* Maple Valley, WA: Aptech Systems, Inc.

Shannon, C.E. (1948). "A Mathematical Theory of Communication." *Bell System Technical Journal* 27, 379–423.

Soofi, E.S. (1992). "A Generalizable Formulation of Conditional Logit with Diagnostics." *Journal of the American Statistical Association* 87, 812–816.