A Small-Sample Estimator for the Sample-Selection Model

by

Amos Golan, Enrico Moretti, and Jeffrey M. Perloff

February 2000

ABSTRACT

A semiparametric estimator for evaluating the parameters of data generated under a sample selection process is developed. This estimator is based on the generalized maximum entropy estimator and performs well for small and ill-posed samples. Theoretical and sampling comparisons with parametric and semiparametric estimators are given. This method and standard ones are applied to three small-sample empirical applications of the wage-participation model for female teenage heads of households, immigrants, and Native Americans.

Key Words: Maximum Entropy, Sample Selection, Monte Carlo Experiments.

Contact:
Amos Golan
Department of Economics
American University
Roper 200
4400 Massachusetts Avenue, NW
Washington, DC 20016-8029

(202) 885-3788
agolan@american.edu

We thank James Powell for very helpful suggestions. Moretti and Perloff thank the Institute for Industrial Relations at Berkeley and the Giannini Foundation for support.

Table of Contents

A Small-Sample Estimator for the Sample-Selection Model

1. INTRODUCTION

The problem of sample selection arises frequently in econometric studies of individuals' wages or labor supply and other topics. When sample sizes are small, existing parametric (full and limited information maximum likelihood, Heckman 1976, 1979) and semiparametric estimators (Manski, 1975, 1985; Cosslett, 1981; Han, 1987; Ahn and Powell, 1993) have difficulties.

We have three objectives. First, we develop a semiparametric estimator for the sample-selection problem that performs well when the sample is small. This estimator has its roots in information theory and is based on the generalized maximum entropy (GME) approach of Golan, Judge, and Miller (1996) and Golan, Judge, and Perloff (1997). Second, we use Monte Carlo experiments to compare and contrast the small-sample behavior of our GME estimator with other parametric and semiparametric estimators. Third, we apply this method to examine the wage-participation of several groups of females where the data sets are small.

Section 2 discusses the sample-selection model. Section 3 reviews GME and develops a GME sample-selection estimator with the relevant inferential statistics. Section 4 lays out the experimental design and discusses the sampling results. Section 5 applies the various methods to the wage-participation model for female teenage heads of households, immigrants, and Native Americans, all of which involve relatively small samples. Section 6 briefly summarizes the results.

2. THE MODEL

Many sample-selection models exist. For specificity, we consider a common labor model (see, e.g., Maddala, 1983). Suppose the *i*th person values staying home (working in the home) at y_{1i}^* and can earn y_{2i}^* in the marketplace. If $y_{2i}^* > y_{1i}^*$, the individual chooses to work in the marketplace, $y_{1i} = 1$, and we observe the market value, $y_{2i} = y_{2i}^*$. Otherwise, $y_{1i} = 0$ and $y_{2i} = 0$.

The individual's value at home and in the marketplace depends on demographic characteristics:

$$y_{1i}^* = \underline{x}_1' \underline{\beta}_1 + e_{1i} \tag{2.1}$$

$$y_{2i}^* = \underline{x}_2' \underline{\beta}_2 + e_{2i}, \tag{2.2}$$

where $\underline{x}_1 = (1, x_{12},..., x_{1L})'$, $\underline{x}_2 = (1, x_{22},..., x_{2K})'$, $\underline{\beta}_1$ and $\underline{\beta}_2$ are L and K-dimensional vectors of unknowns. We observe

$$y_{1i} = \begin{cases} 1 & \text{if } y_{2i}^* > y_{1i}^* \\ 0 & \text{if } y_{2i}^* \le y_{1i}^*. \end{cases} \tag{2.3}$$

$$y_{2i} = \begin{cases} \underline{x}_2' \underline{\beta}_2 + e_{2i} & \text{if } y_{2i}^* > y_{1i}^* \\ 0 & \text{if } y_{2i}^* \le y_{1i}^*. \end{cases} \tag{2.4}$$

Our objective is to estimate $\underline{\beta}_1$ and $\underline{\beta}_2$. Typically in these types of studies, the researcher is interested primarily in $\underline{\beta}_2$.

3. Estimation Approach

We use a GME approach to estimate the sample-selection model. We start by providing some background as to how the generalized maximum entropy approach works, and then develop the GME sample-selection estimator.

3.1 Review of Maximum Entropy and Generalized Maximum Entropy Estimators

The GME estimator is based on the classic maximum entropy (ME) approach of Jaynes (1957a, 1957b, 1984), which uses the entropy-information measure of Shannon (1948) to recover the unknown probability distribution of underdetermined problems. Shannon's (1948) entropy measure reflects the uncertainty (state of knowledge) we have about the occurrence of a collection of events. Letting x be a random variable with possible outcomes x_s, $s = 1, 2, \ldots, n$, with probabilities p_s such that $\Sigma_s\, p_s = 1$, Shannon (1948) defined the *entropy* of the distribution $\underline{p} = (p_1, p_2, ..., p_n)'$, as

$$H \equiv -\sum_s p_s \ln p_s , \tag{3.1}$$

where $0 \ln 0 \equiv 0$. The function H, reaches a maximum of $\ln(n)$ when $p_1 = p_2 = \ldots = p_n = 1/n$. It is zero when $p_s = 1$ for one value of s. To recover the unknown probabilities $\underline{p}$ that characterize a given data set, Jaynes (1957a, 1957b) proposed maximizing entropy, subject to available sample-moment information and adding up constraints on the probabilities. For an axiomatic development of ME and a review of its properties, see Kullback (1959), Levine (1980), Shore and Johnson (1980), Skilling (1989), and Csiszár (1991).

The GME approach generalizes the maximum entropy problem to noisy problems where each observation is taken into account. This approach uses a dual-objective (precision and prediction) function. To illustrate the GME approach, we examine the linear, noisy-inverse model

$$\underline{y} = X\underline{\beta} + \underline{e}, \tag{3.2}$$

where $\underline{\beta}$ is a K-dimensional vector of unobserved parameters, $\underline{y} = (y_1, y_2, ..., y_T)'$ is a T-dimensional vector of observed data, and X is a $(T \times K)$ design matrix. Our objective is to recover the unknown vector $\underline{\beta}$ using as few assumption as possible. Consistent with our goal, we impose no distributional assumptions and no assumptions regarding the exact relationship between sample and population moments. That is, our objective is to simultaneously recover the signal $\underline{\beta}$ and the noise (unknown error distribution) $\underline{e}$ where both are unknown.

To achieve this goal, Golan, Judge, and Miller (1996, Chapter 6) developed the generalized maximum entropy estimator. As a first step, the parameters of Equation 3.2 are reparameterized as

$$\underline{y} = X\underline{\beta} + \underline{e} = XZ\underline{p} + V\underline{w}. \tag{3.3}$$

In this reformulation, the coefficient β_k on the k^{th} variable in X is defined as

$$\beta_k \equiv \sum_m z_{km} p_{km} \tag{3.4}$$

where

$$\sum_m p_{km} = 1 \tag{3.5}$$

and $\underline{z}_k = (z_{k1}, z_{k2}, ..., z_{kM})'$ is an $M \geq 2$ dimensional vector. This vector serves as a discrete support space for each one of the K unknowns and is specified to span the possible (unknown) values of β_k. This reformulation converts the unknown vector $\underline{\beta}$ from the real line to the [0, 1] interval with the properties of K proper probability distributions $\underline{p}_k$ defined over the K support spaces $\underline{z}_k$.

How should we specify Z? If we possess no knowledge as to the possible values of β_k, we specify $\underline{z}_k$ to be symmetric around zero, with large negative and positive boundaries. For example, $z_{k1} = -z_{kM} = -10^6$. Often, however, we have some knowledge regarding the possible values of $\underline{\beta}$ and use that information to specify Z.

Similarly, we transform each e_t into T proper probability distributions. Define a discrete support space $\underline{v}$ of dimension $J \geq 2$, *equally spaced and symmetric around zero* and associate with it a set of weights w_t such that

$$e_t \equiv \sum_j v_j w_{tj} \tag{3.6}$$

and

$$\sum_j w_{tj} = 1, \tag{3.7}$$

and V is a $T*J$ matrix of the T identical vectors v. The end points, v_1 and v_J, are chosen to be $-3\sigma_y$ and $3\sigma_y$ where σ_y is the empirical standard deviation of $\underline{y}$.

Having converted the two sets of unknowns into probability distributions, the estimation problem is to maximize a dual-loss function where emphasis is placed on both prediction and precision (smoothness) of the estimates:

$$\max_{\underline{p},\underline{w}} H(\underline{p},\underline{w}) = -\sum_k \sum_m p_{km} \ln p_{km} - \sum_t \sum_j w_{tj} \ln w_{tj} \tag{3.8}$$

subject to

$$y_t = \sum_k \sum_m x_{tj} z_{km} p_{km} + \sum_j v_j w_{tj} \tag{3.9}$$

$$\sum_m p_{km} = 1 \tag{3.10}$$

$$\sum_j w_{tj} = 1. \tag{3.11}$$

This estimator shrinks *all* unknown parameters to the center of the support given the data. The ME estimator is a special case of the GME, in which no weight is placed on the noise component and the T observations (3.4) are represented as K zero moments.

Letting $\hat{\lambda}_t$ be the estimate of the relevant Lagrange multiplier, the optimal solution is

$$\hat{p}_{km} = \frac{\exp\left(-\sum_t \hat{\lambda}_t x_{tk} z_{km}\right)}{\sum_m \exp\left(-\sum_t \hat{\lambda}_t x_{tk} z_{km}\right)} \equiv \frac{\exp\left(-\sum_t \hat{\lambda} x_{tk} z_{km}\right)}{\Omega_k(\hat{\lambda})} \tag{3.12}$$

and

$$\hat{w}_{tj} = \frac{\exp\left(-\hat{\lambda}_t v_j\right)}{\sum_j \exp\left(-\hat{\lambda}_t v_j\right)} \equiv \frac{\exp\left(-\hat{\lambda}_t v_j\right)}{\Psi_t(\hat{\lambda})}. \tag{3.13}$$

The resulting point estimates are $\hat{\beta} \equiv \Sigma_m z_{km} \hat{p}_{km}$ and $\underline{\hat{e}}_t \equiv \Sigma_j v_j \hat{w}_{tj}$.

Golan, Judge, and Miller (1996) show that a dual, concentrated model can be constructed. Substituting into the first element of the right-hand side of the Lagrangean corresponding to Equations 3.8 - 3.9, the post-data $\underline{\hat{p}}$ and $\underline{\hat{w}}$, yields

$$L(\underline{\lambda}) = \sum_t y_t \lambda_t + \sum_k \ln \Omega_k(\lambda) + \sum_t \ln \Psi_t(\lambda). \tag{3.14}$$

Setting the derivative of $L(\underline{\lambda})$ with respect to $\underline{\lambda}$ equal to zero yields $\underline{\hat{\lambda}}$, from which we can derive $\underline{\hat{\beta}}$ and $\underline{\hat{e}}$.

3.2 A GME Sample-Selection Estimator

We now apply the same approach to the sample-selection problem. We start by reparameterizing the signals $\underline{\beta}_1$ and $\underline{\beta}_2$ to be proper probability distributions that are defined over some support. We start by choosing a support space with $M \geq 2$ of discrete points $\underline{z}_{1l} = [z_{1l1}, z_{1l2}, ..., z_{1lM}]'$ and $\underline{z}_{2k} = [z_{2k1}, z_{2k2}, ..., z_{2kM}]'$ that span the possible range of the unknowns $\underline{\beta}_1$ and $\underline{\beta}_2$.

Then we let

$$\beta_{1l} = \sum_m p_{1lm} z_{1lm}, \qquad l = 1, \ldots, L, \tag{3.15a}$$

and

$$\beta_{2k} = \sum_m p_{2km} z_{2km}, \qquad k = 1, \ldots, K, \tag{3.15b}$$

where $\underline{p}_{1k}$ and $\underline{p}_{2k}$ are proper probability vectors that correspond to the M-dimensional support vectors of weights. As the dimension of M increases, we recover more moments for each

point estimate $\underline{\beta}_1$ and $\underline{\beta}_2$. However, for all practical purposes, the results are not sensitive to *M* as long as *M* is at least of dimension 3.

Similarly, we treat the errors as unknowns and use the following parametrization. Let each $\underline{e}_1$ and $\underline{e}_2$ be specified as

$$e_{1i} = \sum_j w_{1ij} v_j \quad \text{and} \quad e_{2i} = \sum_j w_{2ij} v_j \tag{3.16}$$

where $\underline{w}_1$ and $\underline{w}_2$ are proper probability vectors and $\underline{v}$ is a support space of dimension greater than or equal to two that is symmetric about zero. The lower and upper bounds of the support space $\underline{v}$ are $-3\sigma_{y_2}$ and $3\sigma_{y_2}$ respectively where σ_{y_2} is the empirical standard deviation.

Having parameterized the unknowns, we now wish to maximize the dual loss (objective) function, which is the sum of the joint entropies of the signal and noise in the system, subject to the sample-selection model, Equations 2.3 and 2.4:

$$\begin{aligned} \max_{\underline{p}_1,\underline{p}_2,\underline{w}_1,\underline{w}_2} H(\underline{p}_1,\underline{p}_2,\underline{w}_1,\underline{w}_2) = & -\sum_l \sum_m p_{1lm} \ln p_{1lm} - \sum_k \sum_m p_{2km} \ln p_{2km} \\ & - \sum_i \sum_j w_{1ij} \ln w_{1ij} - \sum_i \sum_j w_{2ij} \ln w_{2ij} \end{aligned} \tag{3.17}$$

subject to

$$\sum_k \sum_m x_{2ik} z_{2km} p_{2km} + \sum_j v_j w_{2ij} = y_{2i}, \quad \text{if } y_{2i}^* > 0, \tag{3.18}$$

$$\sum_k \sum_m x_{2ik} z_{2km} p_{2km} + \sum_j v_j w_{2ij} > \sum_l \sum_m x_{1ik} z_{1lm} p_{1lm} + \sum_j v_j w_{1ij}, \quad \text{if } y_{2i}^* > 0, \tag{3.19}$$

$$\sum_k \sum_m x_{2ik} z_{2km} p_{2km} + \sum_j v_j w_{2ij} \leq \sum_l \sum_m x_{1il} z_{1lm} p_{1lm} + \sum_j v_j w_{1ij}, \quad \text{if } y_{2i}^* = 0, \tag{3.20}$$

$$\sum_m p_{1lm} = 1 \; ; \quad \sum_m p_{2km} = 1 \; ; \tag{3.21}$$

$$\sum_j w_{1ij} = 1 \; ; \quad \sum_j w_{2ij} = 1. \tag{3.22}$$

The optimization yields estimates of $\underline{\hat{p}}_1$, $\underline{\hat{p}}_2$, $\underline{\hat{w}}_1$, and $\underline{\hat{w}}_2$, from which we obtain estimates $\underline{\hat{\beta}}_1$, $\underline{\hat{\beta}}_2$, $\underline{\hat{e}}_1$, and $\underline{\hat{e}}_2$ using Equations 3.15 and 3.16.

In the Appendix, we establish the asymptotic properties of these estimates and then derive some useful statistics. We show that the GME is consistent and asymptotically normal under some mild regularity conditions.

3.3 Discussion

Because the GME estimator is more stable than the ML or LS, the GME-sample selection variances are smaller than the corresponding ML or LS variances. The greater stability from sample to sample is demonstrated in the Monte Carlo experiments in the next section and analytically (for the linear model) in Golan, Judge, and Miller (1996). This greater stability is due to the GME's relatively relaxed data specification, whereby the restriction that $E(X'\underline{e}) = \underline{0}$ is not imposed. Further, no assumptions are made about the distribution or the covariance structure between the two equations. The only disadvantage of note is that the computation time increases markedly as the number of observations increases.

However, if one is analyzing a single data set (rather than running simulations), the increase in time is not a major consideration.

The GME differs from other models in how it handles identification. Most previous approaches use an exclusion restriction to identify the outcome equation in the sample selection model. On the other hand, the GME approach achieves identification from the inequality structure of Equations 3.19-3.20, which allows the covariance elements to be nonzero.

4. Sampling Experiments

In recent years, there have been several Monte Carlo studies of sample selection estimators for relatively large data sets. These studies include Hay, Leu, and Rohrer (1987), Manning, Duan, and Rogers (1987), Hartman (1991), and Leung and Yu (1996). Their results differ because of differences in their experimental designs.

4.1 Experimental Designs

Leung and Yu (1996) argue that several of the earlier studies that found superior performance of ordinary least squares (OLS) over maximum likelihood (ME) sample-selection estimators was due to unusual experimental designs. In particular, they argue that studies such as Manning, Duan, and Rogers (1987) got their results because they drew regressors from a uniform distribution with a range of [0, 3]. Because this range is narrow, the covariates are highly collinear and the Mills' ratio term used in two-stage estimators is highly correlated with the regressor. Leung and Yu find that the ME sample-selection estimators perform better than OLS when they draw regressors from a larger range, [0, 10]. In order to

give maximum likelihood estimators the greatest possible advantage, we use Leung and Yu's larger range for the right-hand-side variables.

Following Leung and Yu, most of our experiments involve only a single regressor (in addition to the intercept) in both the choice and level equations, so $L = K = 2$. In all designs, $\beta_{12} = \beta_{22}$. We vary β_{11}, β_{21}, and the intercepts to control the level of censoring. The support spaces for $\underline{z}_1$ and $\underline{z}_2$ are all specified to be symmetric about zero with large negative and positive bands, $\underline{z}_{1m} = \underline{z}_{2m} = (-100, -50, 0, 50, 100)'$ for all the unknown $\underline{\beta}_1$ and $\underline{\beta}_2$. These supports reflect our state of ignorance relative to the unknown $\underline{\beta}$'s in the range [-100, 100]. The support spaces for the errors $\underline{e}_1$ and $\underline{e}_2$ are based on the *empirical* standard deviations of the observed y_{2i}, σ_2^*, such that $\underline{v}_1 = \underline{v}_2 = (-3\sigma_2^*, 0, 3\sigma_2^*)'$ for all $i = 1,2, ..., T$.

We used the computer program GAMS to generate the data. We repeated each experiment 1,000 times. To show the robustness of the GME estimator, we repeated the experiments for different right-hand-side variables, different number of observations, different number of regressors, normal and non-normal distributions, and for correlations between $\underline{e}_1$ and $\underline{e}_2$ of $\rho = 0$ and $\rho = 0.5$. Table 1 describes the various designs.

We use the performance criteria of Leung and Yu (1996) to summarize the performance of each experiment. The first measure is the mean square error (MSE) for the wage equation. We also use the slope parameter bias and its mean square error, where $\text{Bias}(\hat{\beta}_{22}) \equiv \hat{\beta}_{22} - \beta_{22}$ (and the subscript "22" indicates the second coefficient in the $\underline{\beta}_2$ vector from the wage equation). The final criterion is the mean square prediction error (MSPE) for the second equation where

$$MSPE \equiv \frac{1}{1000} \sum_{i=1}^{1000} \left[\hat{E}(y_{2i}) - E(y_{2i}) \right]^2. \tag{4.1}$$

4.2 Alternative Estimators

We compare our new estimator to alternative parametric and semiparametric estimators. The alternative estimators include OLS, Heckman's two-step approach (2-Step) method, full-information maximum likelihood (FIML), and a semiparametric estimator with a nonparametric selection mechanism (AP) due to Ahn and Powell (1993).

The simplest alternative is to estimate the second equation using ordinary least squares, ignoring the sample-selection problem. We used GAMS to estimate both the GME and OLS models.

The two most commonly used maximum likelihood, parametric approaches are the Heckman two-step and maximum likelihood estimators. We estimated these models using the computer program Limdep. Because of the relatively small sample sizes, these ME models often failed to produce plausible estimates. Indeed, as Nawata and Nagase (1996) show, the FIML estimator may not converge or may converge to a local optimum. They use Monte Carlo experiments to show that FIML may not be a proper estimator when there is a high degree of multicollinearity between the estimated indicator value of the first equation and the right-hand-side variable in the second equation. For these two estimators, we reject an

estimate if Limdep reports a failure to converge or if the estimated correlation coefficient between the two errors does not lie within the range (-1, 1).[1]

As the GME estimator can be viewed as an estimator from the class of semiparametric estimators we also compare the sampling experiments with the Ahn and Powell (1993) estimator. We now briefly discuss the motivation for their estimator and its characteristics.

The AP approach is designed to deal with a well-known problem of the parametric likelihood estimators, which assume that the errors in the two equations are jointly normally distributed. If the joint distribution of the error terms is misspecified, these parametric estimators are inconsistent. Some recent proposals for semiparametric estimation of selection models have relaxed this strong assumption but have kept the additional "single-index" restriction on the form of the selection equation (Cosslett, 1981; Han, 1987). In most instances this restriction, as the one on the joint distribution of errors, does not have any theoretical justification besides convenience.

Ahn and Powell (1993) show that these additional restrictions are not needed to attain a root-*n*-consistent estimator. They propose a two-step approach where both the joint distribution of the error term and the functional form of the selection equation is unknown. In the first step, a multivariate kernel estimator is used to obtain the conditional mean of the selection variable y_1 given a vector of conditioning exogenous variables. This step is analogous to the first step in the Heckman 2-Step procedure, where the selectivity term is

---

[1] Nawata and Nagase (1996) suggest an alternative method to that used by Limdep that may be more likely to converge; however, we did not learn about their approach until after we completed our simulations. In our experiments, these failures are due primarily to small sample sizes. This failure virtually disappears with samples of 500 or 1,000 observations.

estimated. In the second step, parameters of the outcome equation (the wage, y_2, equation) are estimated by a weighted instrumental variables regression of pairwise differences in dependent variables on corresponding differences in explanatory variables. Decreasing weight is put on a pair of observations which have larger differences in the selectivity correction term.

Intuitively, the second stage is based on a comparison of pairs of observations (i, j) for which the estimated ($\hat{y}_{1i}$) and ($\hat{y}_{1j}$) are "close". The selection bias is eliminated through differencing: Differences in the dependent variable is regressed on corresponding differences in explanatory variables. Each of the $\binom{T}{2}$ distinct pairs of observation is assigned a weight that falls (toward zero) as the magnitude of the difference $\hat{y}_{1i}$ - $\hat{y}_{1j}$ increases. Therefore, greater weights is put on pairs with difference in error terms that are approximately unbiased. The weights depend on the sample size T as well, with larger values of T corresponding to lower weight on pairs with a constant value of y_{1i} - $\hat{y}_{1j}$.[2]

The AP estimator is robust to misspecification of the distribution of residuals and the form of the selection equation. When the distribution of residuals is not normal and the sample size is large, we expect the AP estimator to perform better than FIML and 2-Step

[2] When selection depends on one variable only, the first step is not needed. The second step weights can be obtained by simply conditioning on the selection equation regressor. The one-step estimator is asymptotically identical to the two-step estimator. For those experiment designs where selection depends on only one variable, we calculated both one- and two-step estimators and got similar results. To be consistent across designs, we report the two-step estimates in all tables.

estimators. When the sample size is small, however, it is not clear whether AP would dominate FIML and 2-Step estimators, as the large sample size requirement for the AP estimator is not met. So far as we know, no previous study has examined the small-sample performance of the AP estimator.

In our experiments, we use Matlab for the AP estimator, where the kernel functions are taken to be normal density functions. Following Ahn and Powell (1993), in the first-step kernel regression, the data were first linearly transformed so that the components of the vector of exogenous variables are orthogonal in the sample, with variances that are equal one for each component. The first-step bandwidth parameter h_1 was selected in each iteration by least-square cross validation over a crude grid of possible values.

The choice of the second-step kernel bandwidths, h_2, is less straightforward. Cross validation does not necessarily produces the best bandwidths (Powell and Stoker, 1996). We set $h_2 = 0.7$. We experimented with different values of h_2 between 0.0005 and 1. For each of these values between 0.0005 to 1, the point estimates were equal to the ones presented here up to the third decimal point.

Although it is possible to estimate the structural form of the selection equation using Heckman's method, it is not possible using the AP semiparametric estimator. Only the outcome equation is identified. Moreover, since the outcome equation is estimated by regressing differences in dependent variables on corresponding differences in explanatory variables, the intercept term is not identified.

In contrast, in the GME model the structural form of the selection model and $\underline{\beta}_1$ are estimated and there is no need for any weighing procedure. The unknown Lagrangean

multipliers (one for each observation) in the GME are the implicit and natural weight of each observation (Golan and Judge, 1996).

4.3 Discussion of Results

In all experiments, the 2-Step and FIML approaches failed to estimate a large proportion of the samples while the OLS, AP, and GME models *always* produced estimates. When we compare various measures of fit in the following discussion, we discuss only the "plausible" 2-Step and FIML repetitions, which favors these two approaches substantially. The summary statistics for the other approaches include the difficult samples that the 2-Step and FIML approaches could not handle.

Tables 2-5 report results for the five experimental designs. The first column shows the technique. The number in the parentheses following the FIML or 2-Step label is the percent of the 1,000 repetition where that estimator converged and produced plausible values (the estimated correlation was between -1 and 1). The next two columns show the number of observations and the exact proportion of censored observations. The next two columns report the mean square error for all the coefficients in the second equation including the constant, MSE($\hat{\underline{\beta}}_2$), and for just the second coefficient in the second equation, MSE($\hat{\beta}_{22}$). The following column shows the bias for this coefficient. The last column shows the mean squared prediction error, MSPE. We cannot report the MSE($\hat{\underline{\beta}}_2$) and MSPE for the AP approach because it does not produce an estimate of the constant term.

In all the sampling experiments reported in Tables 2-5, the GME strictly dominates the OLS (hence we do not discuss this comparison further). In virtually all of the sampling results, the GME has smaller MSE($\tilde{\underline{\beta}}_2$) and MSE($\tilde{\beta}_{22}$), indicating the stability of the GME

estimator relative to the other estimators for both the whole vector of estimated parameters and of the slope parameter β_{22}. Further, the bias of the GME estimator is smaller than for the other estimators in many cases.

In general, the GME dominates the AP. The AP method is designed to provide robust estimates with large samples and has been shown to perform well with large samples. However, it performs relatively poorly in our small-sample experiments, presumably because it imposes very little structure on the data. The AP bias is lower than OLS bias in all designs, but does not otherwise provide significantly better results than OLS.

The objective of the 2-Step and FIML estimators is to maximize prediction within the sample. It is, therefore, not surprising that the likelihood methods produce the best results in terms of the MSPE in most experiments (where we compare just the successful likelihood estimates to all the estimates for the alternative approaches). The fraction of repetitions for which the 2-Step and FIML estimators fail to provide plausible estimates is very large, ranging from 11% to 99%. In samples of 20 and 50 observations, both the 2-Step and FIML estimators fail to provide plausible estimates in more than half of the repetitions. As the sample size increases, the percentage of implausible estimates decreases.

We now discuss each of the different experiments in more detail. Table 2 reports the effect of sample size on the estimates in Experimental Design 1. The GME approach dominates the AP method on all criteria except bias in the single case where $T = 20$. For the smallest sample size, $T = 20$, the GME estimator is superior to the likelihood methods based on all criteria. For $T = 50$, GME is superior on all criteria except MSPE. For sample sizes

of $T = 75$ and larger, the GME is superior in terms of MSE while the likelihood estimators (where they work at all) have smaller bias and better MSPE.

Table 3 reports the effect of the level of censoring on the estimates. In general, the performance of the estimators, as measured by MSPE, gets worse as the proportion of censored observations increases. The results are similar to Table 1 where the GME always has the lowest MSE, thus exhibiting the highest level of stability from sample to sample. We view this result as a strong one because the GME is superior even for a small proportion of censored observations.

In Table 4, we investigate the robustness of the estimators to various distribution or ill-posed specifications. The first row reports results based on the $\chi^2_{(4)}$ distribution normalized to have a unit of variance. Again the GME is the most stable estimator while the two likelihood estimators have the smallest MSPE. Surprisingly, where they work, the likelihood approaches often perform better than the AP method in terms of bias (though not in terms of the MSE) even when the distribution is misspecified. This result may be due to the small sample size. In small samples, the first-step, AP nonparametric regression estimator is likely to be imprecise. Because they impose "less structure" on data, nonparametric estimators typically need many observations to achieve good precision levels (Silverman, 1986). Hartman (1991), using a different experiment with a sample of a thousand, found that maximum likelihood performed badly with a misspecified error distribution, particularly with respect to the MSE.

The second row of Table 4 shows results based on $\underline{x}_1 = \underline{x}_2$. In this case the GME is superior to all other estimators under all the different statistics reported. Further, both the

FIML and 2-Step estimators "work" only for a very small proportion of the samples. This last result is not surprising because the likelihood estimators are not identified (or are identified only by the nonlinearity of the inverse Mills ratio). As the AP estimator does not impose any restriction on the form of the selection equation, it is not identified in the limit. We report results for AP for completeness.

In Table 5, we compare estimators where $K = 3$, $\rho = 0$ or 0.5, and $T = 50$ or 100. The GME dominates the other methods in terms of mean square error (except in the $\rho = 0$, $T = 100$ case), while the maximum likelihood (in the relatively few repetition where it produces plausible estimates) is superior in terms of prediction.

5. Empirical Applications

We used each approach on three empirical applications with small data sets drawn from the March 1996 Current Population Survey (CPS). In each application, we estimated the wage-participation model (Equation 2.3 and 2.4) for the subset of respondents in the labor market. In all three application, we exclude from the sample workers who are self-employed. In no case did the Heckman's full-information maximum likelihood model converge, so we report results for only the OLS, Heckman two-step, AP, and GME models.

5.1 Teenage Heads of Households

We first examine the labor market behavior of female teenagers who are heads of households. Recently, political debates concern the relationship between the increasing number of teen-age pregnancies and the generosity of welfare payments. Knowing which teenagers choose to work and how much they earn may help inform such a debate.

The wage equation covariates include years of education, potential experience (age - education - 6) and potential experience squared, a dummy for Black, and a dummy for rural location. The covariates in the selection equation include all the variables in the wage equation and the amount of welfare payments received in the previous year, a dummy equal one for married teenagers, and the number of children. The March 1996 CPS has 43 female teenagers who are head of an household for whom all the relevant variables are available. Of these, 29 are employees.

Table 6 shows the estimated wage equation coefficients and asymptotic standard errors and an selection outcome table representing the probability of correct prediction. Except for the GME, in all models, an individual is assigned to a category if the fitted probability of being in that category is greater than 0.5. With the GME, we determine the category directly from the inequalities. GME predicts selection better than the probit model and AP. As was discussed above, the intercept is not identified for AP (consequently, the MSPE is not identified either). Heckman's two-step estimator failed to yield an estimated correlation coefficient, ρ, between -1 and 1, so the table reports a ρ that is censored at 1.

None of the estimators finds a positive return to education that is statistically significantly larger than zero using the 0.05 criterion. Indeed, some of the estimates returns are negative, possibly because there is little variation in years of education among these teenagers. The coefficient on Black is positive and surprisingly large. The remaining coefficients are as expected.

5.2 Recent Immigrants

Next we analyze a sample of 107 female immigrants who entered the United States in the five years preceding the interview (27 of whom are in the labor force). Although there is now a significant literature on labor market performance of recent immigrants, most of the research has been conducted on men rather than women. Table 7 reports estimates for the same model as for the teenagers. Return to education are now positive, although smaller than the 8 to 10% returns usually reported in the literature for U.S.-born workers. Once again, the GME methods predicts selection better than do the Heckman two-step and AP models.

5.3 Native Americans

Finally, we analyze a sample of 151 Native American females, of whom 65 are in the labor force. We are unaware of any previous study of wages and participation by Native American females. The model in the two preceding applications is modified by dropping the (irrelevant) race dummy. The estimated return to education is around 6% across estimation methods. Surprisingly, native American women who live in rural areas earn more than similar women who live in urban areas. Again, the GME does a superior job of predicting selection.

5.4 Summary

In all three of these applications based on small samples, we obtain fairly similar coefficients estimates (though the GME estimates tend to be slightly closer to the OLS estimates than to the other two sets), the GME does a better job of predicting labor force participation, and we cannot estimate Heckman's full-information maximum likelihood model.

In one of these cases, Heckman's two-step model fails to produce a plausible estimate of the correlation coefficient, which brings the entire estimate into question. In each case, the GME's estimated asymptotic standard errors are much smaller than those of the other methods (followed by those of the OLS and AP).

6. CONCLUSION

In a large number of empirical economic analysis, the data sets are relatively small and there is a need for a stable and consistent estimator that converges to an optimal solution and performs well under these circumstances. Our new generalized maximum entropy (GME) estimator meets this objective. For small samples, the GME approach has smaller mean square error measures than other well-known estimators such as ordinary least squares, Heckman's 2-step method, full-information maximum likelihood, and Ahn and Powell's method.

We compared GME to these alternative estimators in small sample experiments. All but one of our experimental designs uses a normal distribution, which favors the likelihood approaches. In these small samples, the OLS, Ahn and Powell, and GME methods always work, but the 2-Step and FIML methods frequently fail to converge or provide estimates of the correlation coefficient that do not lie within the plausible range.

Under all scenarios, the GME proved to be the most stable estimator (had the lowest variance and mean square errors), while the likelihood approaches predicted within the sample better when it worked at all (except for small sample sizes where the GME out-performed the other estimators under all criteria). The GME approach performed better than the OLS in all cases and better than the AP estimator in most cases. Finally, the GME works where the

right-hand-side variables are identical in both equations, a situation where the likelihood methods cannot work at all and the AP method does not perform as well. Thus, if precision and stability of the estimates of a sample-selection model based on a relatively small data set are the objective, the GME estimator appears to be the appropriate choice.

References

Ahn, H. and J. L. Powell, "Semiparametric estimation of censored selection models with a nonparametric selection mechanism," *Journal of Econometrics*, 58, 1993:3-29.

Cosslett, S. R., "Distribution-free maximum likelihood estimator of the binary choice model," *Econometrica*, 51, 1981:765-782.

Csiszár, I., "Why Least Squares and Maximum Entropy? An Axiomatic Approach to Inference for Linear Inverse Problems," *The Annals of Statistics*, 19, 1991:2032-2066.

Efron, B., "Bootstrapping Methods: Another Look at the Jackknife," *Annals of Statistics*, 7, 1979:1-26.

Golan, A., and G. Judge, "A Maximum Entropy Approach to Empirical Likelihood Estimation and Inference," Working paper, University of California, Berkeley, 1996.

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

Golan, A,, G. Judge, J. M. Perloff, "Recovering Information from Censored and Ordered Multinomial Response Data," *Journal of Econometrics*, 79, 1997:23-51.

Han, A. K., "Non-parametric analysis of a generalized regression model: The maximum rank correlation estimator," *Journal of Econometrics*, 35, 1987:303-316

Hartman, R. S., "A Monte Carlo Analysis of Alternative Estimators in Models Involving Selectivity," *Journal of Business & Economic Statistics*, 9, 1991:41-9.

Hay, J., R. Leu, and P. Rohrer, "Ordinary least squares and sample-selection models of health-care demand," *Journal of Business & Economic Statistics*, 5, 1987:499-506.

Hinkley, D., "Discussion of 'Jackknife, Bootstrap and other Resampling Methods in Regression Analysis,' by C. F. J. Wu, *Annals of Statistics*, 14, 1312-6.

Jaynes, E. T., "Information Theory and Statistical Mechanics," *Physics Review*, 106, 1957a:620-630.

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

Jaynes, E. T., "Prior Information and Ambiguity in Inverse Problems," *Inverse Problems*, D. W. McLaughin, ed., Providence, Rhode Island: American Mathematical Society, 1984.

Kullback, J., *Information Theory and Statistics*, New York: John Wiley & Sons, 1959.

Leung, S. F., and S. Yu, "On the choice between sample selection and two-part models," *Journal of Econometrics*, 72, 1996:197-229.

Levine, R. D., "An Information Theoretical Approach to Inversion Problems," *Journal of Physics*, A, 13, 1980:91-108.

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

Manning, W., N. Duan, and W. Rogers, "Monte Carlo evidence on the choice between sample selection and two-part models," *Journal of Econometrics*," 35, 1987:59-82.

Nawata, K., and N. Nagase, "Estimation of Sample Selection Bias Models," *Econometric Review*, 15, 1996:387-400.

Owen, A.B., "Empirical Likelihood Ratio Confidence Regions," *Annals of Statistics*, 18, 1990:90-120.

Powell, J. L., and T. M. Stoker, "Optimal bandwidth choice for Choice for Density-Weighted Averages," *Journal of Econometrics*, 75, 291-316, 1996.

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

Shore, J. E., and R. W. Johnson, "Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy," *IEEE Transactions on Information Theory*, IT-26, 1980:26-37.

Silverman, B. W., *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, 1986.

Skilling, J., "The Axioms of Maximum Entropy," in J. Skilling, ed., *Maximum Entropy and Bayesian Methods in Science and Engineering*, Dordrecht: Kluwer Academic, 1989:173-87.

Wu, C.F.J., "Jackknife, Bootstrap and other Resampling Methods in Regression Analysis," *Annals of Statistics*, 14, 1986:1261-1350.

Table 1
Experimental Design

Design	*Regressors*	*T*	*K*	*Approximate Percent Censored*	*Error Distribution*	ρ
1	$\underline{x}_1 \sim U(0,10)$, $\underline{x}_2 \sim U(0,10)$	20, 50, 75, 100, 125	2	50	bivariate *N*	0.5
2	$\underline{x}_1 \sim U(0,10)$, $\underline{x}_2 \sim U(0,10)$	100	2	25, 50, 75	bivariate *N*	0.5
3	$\underline{x}_1 \sim U(0,10)$, $\underline{x}_2 \sim U(0,10)$	100	2	50	bivariate χ^2	0.5
4	$\underline{x}_1 = \underline{x}_2 \sim U(0,10)$	100	2	50	bivariate *N*	0.5
5	$\underline{x}_1 \sim U(0,10)$, $\underline{x}_2 \sim U(0,10)$	50, 100	3	50	bivariate *N*	0, 0.5

Table 2
Sample Results for Experimental Design 1

Estimation Method	*Number of Observations*	*Proportion of Censored Observation*	*MSE*($\hat{\beta}_2$)	*MSE*($\hat{\beta}_{22}$)	*Bias*($\hat{\beta}_{22}$)	*MSPE*
FIML (0.044)*	20	44.86	1.6238	0.05627	–0.0750	1.5177
2-Step (0.040)*			3.5148	0.71140	–0.2115	1.5616
OLS			0.5937	0.11994	–0.0257	1.0092
AP				0.01201	–0.0110	
GME			0.5576	0.01141	–0.0131	1.0014
FIML (0.462)*	50	42.20	0.3590	0.00720	–0.0136	0.6383
2-Step (0.695)*			0.3605	0.00723	–0.0129	0.6420
OLS			0.3390	0.00686	–0.0257	1.0206
AP				0.00692	–0.0250	
GME			0.2966	0.00631	–0.0131	1.0079
FIML (0.589)*	75	51.41	0.4621	0.00847	–0.0106	0.5425
2-Step (0.776)*			0.4525	0.00831	–0.0116	0.5453
OLS			0.4541	0.00819	–0.0364	1.0756
AP				0.00802	–0.0341	
GME			0.3734	0.00712	–0.0201	1.0539
FIML (0.792)*	100	51.37	0.2939	0.00472	–0.0052	0.3662
2-Step (0.883)*			0.2663	0.00445	–0.0065	0.3670
OLS			0.3091	0.00487	–0.0361	1.0550
AP				0.00485	–0.0348	
GME			0.2301	0.00397	–0.0215	1.0300
FIML (0.836)*	125	51.41	0.2045	0.00349	–0.0050	0.3068
2-Step (0.908)*			0.1956	0.00343	–0.0037	0.3063
OLS			0.2283	0.00379	–0.0232	1.0363
AP				0.00370	–0.0222	
GME			0.1925	0.00345	–0.0153	1.0234

* The fraction in parentheses indicates the share of repetitions for which this estimator converged and produced "plausible" results.

Table 3
Sample Results for Experimental Design 2

Estimation Method	*Number of Observations*	*Proportion of Censored Observation*	$MSE(\hat{\beta}_2)$	$MSE(\hat{\beta}_{22})$	$Bias(\hat{\beta}_{22})$	*MSPE*
FIML (0.685)*	100	26.06	0.0932	0.00194	–0.0071	0.2848
2-Step (0.830)*			0.0883	0.00186	–0.0059	0.2860
OLS			0.0939	0.00197	–0.0179	0.9877
AP				0.00190	–0.0172	
GME			0.0808	0.00178	–0.0112	0.9839
FIML (0.792)*	100	51.37	0.2939	0.00472	–0.0052	0.3662
2-Step (0.883)*			0.2663	0.00445	–0.0065	0.3670
OLS			0.3091	0.00487	–0.0361	1.0550
AP				0.00487	–0.0348	
GME			0.2301	0.00397	–0.0215	1.0300
FIML (0.661)*	100	75.11	1.1321	0.01679	–0.0073	0.3312
2-Step (0.823)*			1.1422	0.01669	–0.0062	0.3309
OLS			1.1177	0.01611	–0.0356	1.3045
AP				0.01613	–0.0351	
GME			0.9500	0.01440	–0.0177	1.2480

* The fraction in parentheses indicates the share of repetitions for which this estimator converged and produced "plausible" results.

Table 4
Sample Results for Experimental Designs 3 and 4

Design	*Estimation Method*	*Number of Observations*	*Proportion of Censored Observation*	$MSE(\hat{\beta}_2)$	$MSE(\hat{\beta}_{22})$	$Bias(\hat{\beta}_{22})$	*MSPE*
3 $\chi^2_{(4)}$	FIML (0.759)*	100	51.29%	0.3766	0.00511	–0.0077	0.3698
	2-Step (0.898)*			0.2833	0.00416	–0.0088	0.3699
	OLS			0.2515	0.00394	–0.0273	1.0476
	AP				0.00391	–0.0266	
	GME			0.1918	0.00333	–0.0161	1.0275
4	FIML (0.123)*	100	49.92	9217.5900	9193.80	30.9090	116.370
	2-Step (0.015)*			0.7771	0.00401	0.0303	0.6106
	OLS			0.2339	0.00227	–0.0019	1.1448
	AP				0.00233	–0.0019	
	GME			0.1267	0.00234	0.0003	1.0419

* The fraction in parentheses indicates the share of repetitions for which this estimator converged and produced "plausible" results.

Table 5
Sample Results for Experimental Design 5

Estimation Method	*Correlation*	*Number of Observations*	*Proportion of Censored Observation*	$MSE(\hat{\beta}_2)$	$MSE(\hat{\beta}_{22})$	$Bias(\hat{\beta}_{22})$	*MSPE*
FIML (0.313)*	$\rho = 0$	50	49.79	0.43298	0.00626	–0.0022	0.19302
2-Step (0.538)*				0.46207	0.00727	–0.0108	0.19694
OLS				0.44895	0.00716	–0.0156	1.11194
AP					0.00710	–0.0149	
GME				0.41531	0.00680	–0.0128	1.07460
FIML (0.252)*	$\rho = .5$	50	49.52	0.45369	0.01187	–0.0072	0.37473
2-Step (0.411)*				0.42840	0.00649	–0.0005	0.18853
OLS				0.43692	0.00671	–0.0023	1.05764
AP					0.00673	–0.0018	
GME				0.40531	0.00635	0.0004	1.04195
FIML (0.583)*	$\rho = 0$	100	49.40	0.19493	0.00333	–0.0093	0.09365
2-Step (0.789)*				0.19040	0.00322	–0.0124	0.09459
OLS				0.24461	0.00442	–0.0383	1.09071
AP					0.00434	–0.0370	
GME				0.19992	0.00371	–0.0300	1.05963
FIML (0.569)*	$\rho = .5$	100	49.40	0.20378	0.00348	–0.0077	0.09351
2-Step (0.793)*				0.20240	0.00339	–0.0097	0.09379
OLS				0.20345	0.00357	–0.0218	1.03300
AP					0.00357	–0.0210	
GME				0.17967	0.00319	–0.0155	1.02113

* The fraction in parentheses indicates the share of repetitions for which this estimator converged and produced "plausible" results.

Table 6
Wage Equation for Female Teen Heads of Households
(N = 43; 29 in the labor force)

	OLS	*2-Step*	*AP*	*GME*
Constant	1.923	0.782	NA	1.892
	(1.012)	(2.683)	NA	(0.460)
Education	-0.019	0.070	0.021	-0.012
	(0.083)	(0.215)	(0.080)	(0.038)
Black	0.372	0.405	0.441	0.347
	(0.189)	(0.241)	(0.148)	(0.083)
Experience	0.140	0.091	0.137	0.064
	(0.148)	(0.207)	(0.178)	(0.085)
Experience squared	-0.079	-0.070	-0.091	-0.063
	(0.044)	(0.059)	(0.055)	(0.020)
Rural	-0.004	-0.017	-0.004	-0.027
	(0.116)	(0.146)	(0.102)	(0.057)
λ		0.351		
		(0.723)		
ρ		1		
R^2	0.204	0.215		0.188
MSPE	0.058	0.043		0.058

Predicted

	2-Step's Probit		*AP*		*GME*	
Actual	0	1	0	1	0	1
0	6	8	3	11	11	3
1	3	26	0	29	2	27

Table 7
Wage Equation for Female Immigrants
(*N* = 107; 27 in the labor force)

	OLS	*2-Step*	*AP*	*GME*
Constant	1.103	0.755	-	1.174
	(0.489)	(1.216)	-	(0.031)
Education	0.052	0.062	0.057	0.046
	(0.031)	(0.043)	(0.054)	(0.010)
Black	0.015	0.172	0.040	0.012
	(0.401)	(0.626)	(0.185)	(0.173)
Experience	0.019	0.026	0.024	0.024
	(0.040)	(0.043)	(0.057)	(0.016)
Experience squared	-0.0005	-0.0007	-0.0008	-0.0009
	(0.001)	(0.001)	(0.002)	(0.0005)
Rural	0.331	0.415	0.365	0.328
	(0.652)	(0.649)	(0.238)	(0.105)
λ		0.164		
		(0.535)		
ρ		0.287		
R^2	0.162	0.165		0.153
MSPE	0.310	0.228		0.314

	Predicted					
	2-Step's Probit		*AP*		*GME*	
Actual	0	1	0	1	0	1
0	75	5	80	0	80	0
1	18	9	27	0	0	27

Table 8
Wage Equation for Native Americans Females
(*N* = 151; 65 in the labor force)

	OLS	*2-Step*	*AP*	*GME*
Constant	0.763	1.113	-	0.807
	(0.347)	(0.596)	-	(0.027)
Education	0.063	0.058	0.055	0.061
	(0.026)	(0.030)	(0.040)	(0.007)
Experience	0.047	0.042	0.049	0.047
	(0.013)	(0.016)	(0.020)	(0.008)
Experience squared	-0.001	-0.0008	-0.001	-0.001
	(0.0003)	(0.0004)	(0.0005)	(0.009)
Rural	0.299	0.182	0.457	0.0001
	(0.113)	(0.197)	(0.223)	(0.002)
λ		-0.264		
		(0.349)		
ρ		-0.606		
R^2	0.324	0.332		0.324
MSPE	0.150	0.135		0.151

	Predicted					
	2-Step's Probit		*AP*		*GME*	
Actual	0	1	0	1	0	1
0	67	19	68	18	86	0
1	30	35	23	42	0	65

Appendix: Asymptotic Results

Given some mild conditions, our sample-selection GME estimator is consistent and asymptotically normal. These conditions are that (i) the errors' supports $\underline{v}$ for each equation are symmetric around zero, (ii) the support spans the true values for each one of the unknown parameters $\underline{\beta} = (\underline{\beta}_1', \underline{\beta}_2')'$ and has finite lower and upper bounds (z_{1l1} and z_{1lM} for $\underline{\beta}_1$ and z_{2k1} and z_{2kM} for $\underline{\beta}_2$), (iii) the errors are independently and identically distributed, and (iv) plim $(1/T)X'X$ exists and is nonsingular, where X is a block diagonal matrix consisting of X_1 and X_2. [We note that assumption (iii) *does not* restrict the errors to be uncorrelated across equation.]

The proofs of consistency and asymptotic normality follow immediately from those in Golan, Judge, and Miller (1996), Golan, Judge, and Perloff (1997), and Mittelhammer and Cardell (1996). These asymptotic properties can also be established via the empirical likelihood approach (Owen 1990, 1991, Qin and Lawless 1994, and Golan and Judge 1996).

To estimate the variances of $\underline{\beta}_1$ and $\underline{\beta}_2$, we can use a resampling inference approach, such as the jackknife (Hinkley 1977, Wu 1986) or the bootstrap (Efron 1979). For example, using the bootstrap approach, we start by estimating our system of equations and then use the estimated errors to generate data sets for resampling from the fitted system plus independent errors from the estimated distribution.

The asymptotic variances can be computed in a number of ways. We discuss the simplest approach here. We calculate

$$\hat{\sigma}_{\delta}^2 = \frac{1}{T}\sum_t \hat{e}_{\delta i}^2,$$

for $\delta = 1, 2$, where $\hat{e}_{\delta i} \equiv \sum_j v_j \hat{w}_{\delta ij}$. The elements of the asymptotic variance-covariance matrix, Σ, for the error terms of the entire system are

$$\hat{\sigma}_{12} = \frac{1}{T}\sum_t \hat{e}_{1i}\hat{e}_{2i}.$$

Given these estimates, the asymptotic covariance matrix for the GME sample-selection (logically similar to a seemingly unrelated regression) system is

$$\Omega_{\text{GME}} = \text{plim } T^{-1}\left[X'(\hat{\Sigma}^{-1} \otimes I_T)X\right]^{-1}.$$

We can now establish:

Theorem 1: Under the four assumptions (i) - (iv), the restricted GME sample-selection estimate $\hat{\underline{\beta}}$ is consistent and asymptotically normally distributed

$$\sqrt{T}(\hat{\underline{\beta}} - \underline{\beta}) \xrightarrow{d} \text{N}(\underline{0}, \Omega_{\text{GME}}).$$

The proof of this theorem is an immediate extension of Lemma 1 and Theorem 1 of Qin and Lawless (1994), since $\underline{\beta}$ is a continuous function of $\underline{\lambda}$, the Lagrangean multipliers, and is bounded within $\underline{z}_\beta$ by assumption (ii).

Next, let $\underline{\lambda}$ be the vector of Lagrange multipliers associated with Equations 3.18 - 3.20, $\underline{\beta} \equiv (\underline{\beta}_1', \underline{\beta}_2')'$, and $H_c(\underline{\lambda})$ be the Shannon entropy measure of the (data) constrained GME model where $\hat{\underline{\beta}} \neq 0$ or, equivalently, $\underline{\lambda} \neq 0$, where the value of the maximum joint entropies in Equation 3.17 is maximized subject to the data as represented by Equations 3.18-3.21. Now let $H_u(\underline{\lambda})$ be the Shannon entropy measure of the unconstrained problem where $\underline{\lambda} = 0$ (or, equivalently, $\hat{\underline{\beta}} = 0$ or equals the center of the supports). Thus, $H_u(\underline{\lambda})$ is the maximum value of the joint entropies where no data restrictions are imposed and the only restrictions are the four probability distributions are proper, which is the entropy values of the four discrete, uniform distributions:

$$H_u(\underline{\lambda}) = 2T\ln J + K\ln M + L\ln M.$$

Therefore, the entropy-ratio statistic is

$$R_E(\underline{\lambda}) \equiv 2[H_u(\underline{\lambda}) - H_c(\underline{\lambda})].$$

Following Owen (1990) and Qin and Lawless (1994), if $\underline{\beta} = \beta_0$, then $R_E(\underline{\lambda})$ converges in distribution to $\chi^2_{(K+L-2)}$ as $T \to \infty$. Finally, to construct an approximate α-level confidence region for $\underline{\beta} = (\underline{\beta}_k', \underline{\beta}_l')'$, we note that $R_E(\underline{\lambda}) \leq C_\alpha$, where C_α is the $\Pr(\chi^2_{(K+L-2)} \leq C_\alpha) = \alpha$.

One measure of the "goodness of fit" is the pseudo-R^2:

$$PR^2 = 1 - S^*(\underline{\hat{p}}),$$

where $S^*(\cdot) = H_c(\cdot)/H_u(\cdot)$ is a *normalized* version of $H(\cdot)$ with respect to $\underline{p}$. This normalized measures lies in the interval [0, 1], where 0 reflects perfect knowledge about the estimates and 1 reflects a state of complete ignorance or uncertainty regarding $\underline{\hat{\beta}}$.