Rochester Center for Economic Research

On the Choice Between Sample Selection and Two-Part Models

Leung, Siu Fai and Shihti Yu

Working Paper No. 337
December 1992

University of
Rochester

On the Choice Between Sample Selection and Two-Part Models

Siu Fai Leung
and
Shihti Yu

Rochester Center for Economic Research
Working Paper No. 337

ON THE CHOICE BETWEEN SAMPLE SELECTION AND TWO-PART MODELS

Siu Fai Leung

University of Rochester

and

Shihti Yu

University of Rochester

Abstract: This paper contributes to resolve the vigorous debates between advocates of the sample selection model and the two-part model. Recent Monte Carlo studies by Hay, Leu, and Rohrer (1987) and Manning, Duan, and Rogers (1987) find that the two-part model performs better than the sample selection model even when the latter is the true model. We show that Manning, Duan, and Rogers' negative results regarding the sample selection model are caused by a critical design problem. We demonstrate that their data generating process produces serious collinearity problems that bias against the sample selection model. Once the design problem is rectified, the poor performance of the sample selection model evaporates. Our Monte Carlo results offer a more balanced view on the relative merits of the two models as each model performs well under different conditions. Among our results are that the sample selection model is susceptible to collinearity problems and that a t-test can be used to distinguish between the two models as long as there are no collinearity problems.

First Draft: October 1992
This Version: December 1992

Department of Economics
Harkness Hall
University of Rochester
Rochester, NY 14627

We thank Bruce Hansen and Charles Phelps for comments and suggestions, as well as Shao-Chung Chang for research assistance. This research was supported by the NIAAA.

1. Introduction

Sample selection models have dominated much of the literature in microeconometrics. Heckman's (1976, 1979) two-step estimation procedure and its variants have routinely been adopted in almost every empirical study that involves selection bias (Maddala (1983), Amemiya (1985), Greene (1990)). In a series of articles beginning with an empirical investigation on the demand for medical care from the Rand Health Insurance Study, Duan et al. (1983, 1984, 1985) and Manning, Duan, and Rogers (1987) offer by far the strongest criticisms against the sample selection model. They contend that "the selection models are intrinsically flawed because they have to rely on untestable assumptions and have poor statistical and numerical properties" and therefore they "may be inappropriate for any applications involving either actual or potential outcomes" (Duan et al. 1984). They propose a two-part non-selection model as an alternative and argue that it is much better than the sample selection model. Their ardent stand against the sample selection model has inevitably and predictably provoked some intense debates. In an attempt to defend the sample selection model, Hay and Olsen (1984) argue that the two-part model is built on some unusual assumptions and that it can be nested in the sample selection model. By constructing a counter-example, Duan et al. (1984) show that Hay and Olsen's arguments are erroneous. Perplexed by these exchanges, Maddala (1985a, 1985b) tries to sort out the different issues raised in Hay and Olsen (1984) and Duan et al. (1984). However, Duan et al. (1985) reject much of Maddala's analysis and respond that from a policy perspective, their exchanges with Hay and Olsen and Maddala "are much ado about nothing." Clearly, the debates have yet to be settled.

The origin of the two-part model can at least be traced back to Goldberger (1964) who labels it the twin linear probability approach. Cragg (1971), who appears to be the first to use the term "two-part model," discusses several of its variants.[1] Researchers associated with the Rand Corporation in the

[1] Neither Goldberger (1964) nor Cragg (1971) is mentioned in Duan et al. (1983, 1984, 1985) and Manning, Duan, and Rogers (1987).

seventies and the early eighties make extensive uses of the two-part model in their empirical work (e.g., Manning et al. (1981, 1984, 1985), Newhouse et al. (1981)). Although the term "two-part model" is never mentioned, the model has actually been frequently used in numerous applied works.[2] In these studies, ordinary least squares estimates obtained from regressions that omit the inverse Mills' ratio, which are usually reported along with Heckman's two-step estimates for comparison purposes, can be interpreted as the estimates of the second part of the two-part model. Given the widespread uses of the two-part and the sample selection models in empirical work, and the strong claims by Duan et al. (1984) that the sample selection model is intrinsically flawed, the debates between advocates of the two models should not be overlooked.

While earlier comparisons between the sample selection and the two-part models focus primarily on theoretical issues, recent investigations have turned to Monte Carlo simulation experiments. Hay, Leu, and Rohrer (1987) simulate a data set from the 1981 Population of Switzerland Survey and use a variety of sample sizes, true model parameter values, and error term distributions to compare the performance of the two-part and the sample selection models. They find that the two-part model performs at least as well as the sample selection model in terms of mean prediction bias and mean squared prediction error, and significantly outperforms the sample selection model in terms of parameter squared error. Although Hay has criticized the two-part model in an earlier theoretical study (Hay and Olsen 1984), he and his associates (Leu and Rohrer) have to admit that their Monte Carlo evidence lends some support to the claims in Duan et al. (1982) and that the two-part model appears to be a more robust estimator than the sample selection model. In a different Monte Carlo investigation, Manning, Duan, and Rogers (1987) put the two-part model to a worst-case test by assuming that the true model is a selection model. When there are no exclusion restrictions (i.e., the same regressor appears in the choice and the level equations), they

[2] See, e.g., Dudley and Montmarquette (1976), Blau and Robins (1990), Grossman and Joyce (1990), Kostiuk (1990), McLaughlin (1991).

find that the two-part model is much better than the sample selection model in terms of mean squared prediction error and mean prediction bias, despite the fact that the selection model is the true one. The sample selection model performs better than the two-part model only when there are exclusion restrictions. Manning, Duan, and Rogers (1987, p.80) conclude that their "results are convincing for the use of the data-analytic two-part model, because we stacked the comparisons against the two-part models, and in favor of the selection models" and they maintain that "Given the uncertainty about the true specification, these [sample selection] models will perform poorly in practice" (p. 81).

Although their simulation designs are different, both Hay, Leu, and Rohrer (1987) and Manning, Duan, and Rogers (1987) reach the same conclusion that the two-part model appears to dominate the sample selection model. The main and most intriguing finding is that even when the sample selection model is the true model, the two-part model still considerably outperforms the selection model in most of their simulation experiments. If this striking result is robust, then it would cast doubts on the reliability of all the empirical findings that are based on the sample selection model in the literature in the past two decades. This is undoubtedly an important issue that deserves further investigations.

In this paper, we conduct a different set of Monte Carlo experiments to compare the performance of the sample selection and the two-part models. In contrast to the overwhelming rejection of the sample selection model found in previous Monte Carlo studies, we offer a more balanced account on the merits of the sample selection and the two-part models. We demonstrate that the failure of the sample selection model in Manning, Duan, and Rogers can be traced back to a subtle design problem in their simulation experiments. The underlying shortcoming lies in the way the regressors are generated. In all of their experiments, Manning, Duan, and Rogers draw the regressors from a uniform distribution with a range of [0,3]. When there are no exclusion restrictions, the level equation contains the same U(0,3) regressor as the choice equation.[3] The inverse Mills' ratio, which is a function of the regressor, turns out to be

[3] Throughout the paper, we will use U(a,b) to denote a uniform distribution with range [a,b].

highly correlated with the regressor because the range [0,3] is far too narrow. As a result of high collinearity between the regressor and the inverse Mills' ratio, the Heckman two-step estimators have large standard errors and behave badly. We prove this point by using different measures of collinearity (such as correlation coefficient and condition number) and by widening the range of the uniform distribution for the regressors. We show that when the regressors are drawn from U(0,10), the collinearity problems vanish and the sample selection model behaves much better than the two-part model. To further substantiate our claims, we conduct several t-tests to check whether the two-part model will be rejected when the data are generated from the sample selection model. We find that the t-tests fail to reject the two-part model when there are collinearity problems. When collinearity is lessened, the t-tests strongly reject the two-part model. Based on these results, we can therefore explain Manning, Duan, and Rogers' striking result that the two-part model outperforms the sample selection model even when the latter is the true model. Consequently, the favorable results on the merits of the two-part model reported in Manning, Duan, and Rogers are not reliable because their simulation setups are biased against the sample selection model.[4]

Without burdening the sample selection model with collinearity problems, we generate the regressors from U(0,10) and thereby put the two-part and the selection models to a fair competition. We conduct a series of experiments with and without exclusion restrictions, using various true models and different degrees of censoring. Six criteria (mean prediction bias, mean squared prediction error, parameter bias, parameter squared error, elasticity bias, and elasticity squared error) are employed to evaluate the estimators. Our results stand in sharp contrast to those of Hay, Leu, and Rohrer and Manning, Duan, and Rogers. We find that when the sample selection model is the true model, it performs

[4] To our surprise, we find that the design problem on the data generating process in Manning, Duan, and Rogers is actually very common in the sample selection literature. We will show below that a number of widely cited Monte Carlo studies on the sample selection model are also marred by the same design problem.

substantially better than the two-part model as long as there are no collinearity problems. When the two-part model is the true model, the sample selection model is inferior, but is still reasonably close to the two-part model. Hence, our results do not support the contention that the two-part model dominates the sample selection. Nor do we find that the selection model is superior to the two-part model. We believe that a balanced view is more appropriate because each model performs well under different conditions.

In addition to resolving the debates between the sample selection and the two-part models, another contribution of the paper is the finding that Heckman's two-step estimator is susceptible to collinearity problems. Although several researchers have noted that collinearity is a potential problem in the two-step estimation method, none have investigated it systematically.[5] We show that little exclusion restrictions, a high degree of censoring, a low variability among the regressors, or a large error variance in the choice equation can all contribute to near collinearity between the regressors and the inverse Mills' ratio, rendering the two-step estimator ineffective. In view of this, we suggest that applied researchers should examine whether there exists high collinearity in the level equation whenever they implement the two-step procedure. After investigating the performance of several different measures of collinearity, we believe that the condition number is more accurate and dependable than the other measures.

The plan of the paper is as follows. Section 2 briefly reviews the models and the estimation methods. The designs of the Monte Carlo experiments, the criteria used to assess the estimators, and our measures of collinearity are all described in section 3. The simulation results are reported in section 4. Section 5 discusses the results and section 6 concludes the paper.

---

[5] See, e.g., Heckman (1979) and Manning, Duan, and Rogers. Nelson (1984) reports some simulation results on the adverse effects of collinearity on Heckman's two-step estimators; however, he only considers the efficiency of the estimators. Our focus is different from Nelson's and our study is also broader in that we use several measures of collinearity and employ six criteria to assess the collinearity problems. As will be shown below, our results are notably different from Nelson's.

2. Review of Models and Estimation Methods

2.1 Sample Selection Model

There are many variants of the sample selection model. Following Hay, Leu, and Rohrer and Manning, Duan, and Rogers, we focus on van de Ven and van Praag's (1981a, 1981b) Adjusted Tobit Model (a Type 2 Tobit model in Amemiya's (1985) classification system):

$$I = \underline{x}_1\alpha + u_1, \tag{1}$$

$$m = \underline{x}_2\beta + u_2, \tag{2}$$

$$\ln(y) = m \quad \text{if } I > 0,$$

$$= -\infty \quad \text{if } I \leq 0,$$

where $\underline{x}_1 = (1,x_{12},\ldots,x_{1J})$, $\underline{x}_2 = (1,x_{22},\ldots,x_{2K})$, $\alpha = (\alpha_1,\alpha_2,\ldots,\alpha_J)'$, $\beta = (\beta_1,\beta_2,\ldots,\beta_K)'$, and individual subscripts are suppressed for simplicity. The error terms u_1 and u_2 are assumed to be independent and identically distributed bivariate normal random variables:

$$(u_1,u_2)' \sim N(0,\Sigma),$$

$$\Sigma = \begin{bmatrix} 1 & \rho\sigma \\ \rho\sigma & \sigma^2 \end{bmatrix}.$$

Eq. (1) is the choice equation that governs whether $y = 0$ or $y > 0$ is observed. If $I > 0$, then $\ln(y)$ follows the level equation given by (2). The logarithmic transformation serves to reduce the skewness of the y variable. There are many examples in economics in which the dependent variable is highly skewed, e.g., individual medical expenditures in van de Ven and van Praag (1981a, 1981b) and Duan et al. (1983).

The two most popular estimation strategies for the sample selection model are Heckman's two-step procedure (limited-information maximum-likelihood, LIML) and maximum likelihood (full-information maximum-likelihood, FIML). For the LIML, the first step is to obtain an estimate of α from (1) by means of maximum likelihood on the probit model. The estimates of β and $\rho\sigma$ are then obtained

by running a simple ordinary least squares regression on the model

$$\ln(y) = \underline{x}_2\beta + \rho\sigma\hat{\lambda} + \epsilon, \qquad (3)$$

using the sample of positive y's, where $\hat{\lambda} = \phi(\underline{x}_1\hat{\alpha})/\Phi(\underline{x}_1\hat{\alpha})$ is the estimated inverse Mills' ratio. The estimates for ρ and σ are then calculated as in Manning, Duan, and Rogers. For the FIML, the likelihood function is simply

$$L = \Pi_0[1-\Phi(\underline{x}_1\alpha)]\bullet\Pi_1\Phi((\underline{x}_1\alpha+\rho(m-\underline{x}_2\beta)/\sigma)(1-\rho^2)^{-1/2})\phi((m-\underline{x}_2\beta)/\sigma)/\sigma,$$

where Π_0 and Π_1 denote the products over the censored and the uncensored samples, respectively. Although the FIML estimator is more efficient than the LIML estimator, the Heckman two-step procedure is more popular because computationally it is faster and much easier to use. Furthermore, the LIML estimator is more robust than the FIML estimator in some circumstances. Stapleton and Young (1984) show that the FIML estimator will no longer be consistent when there are measurement errors in the dependent variable of the uncensored observations. The LIML estimator, however, will remain consistent because the measurements errors are fully absorbed into the error term (u_2) of the level equation and thus will not affect the consistency of the estimators.

2.2 Two-part Model

The two-part model separates the dependent variable into two parts: $y > 0$ and $y|y > 0$. For the first part, it is assumed to be a standard probit model

$$I = \underline{x}_1\alpha + u_3, \qquad (4)$$

$$u_3 \sim N(0,1),$$

where $y > 0$ if $I > 0$, and $y = 0$ otherwise. For the second part, it is a linear model with

$$\ln(y|I > 0) = \underline{x}_2\beta + u_4, \qquad (5)$$

where $E(u_4|I > 0) = 0$ and u_4 is not necessarily normally distributed. The two-part model does not involve any sample selection or selectivity bias. Eq. (5) implies that $E[\ln(y)|I > 0] = \underline{x}_2\beta$, as opposed

to $E[\ln(y)|I > 0] = \underline{x}_2\beta + \rho\sigma\lambda$ in the sample selection model, where $\lambda = \phi(\underline{x}_1\alpha)/\Phi(\underline{x}_1\alpha)$ is the inverse Mills' ratio. The two-part model maintains that the level of use, given any, is conditionally independent of the decision to use.

The fundamental distinction between the sample selection and the two-part models lies in the assumptions on the error terms u_2 and u_4. The sample selection model presumes that $E(u_2) = 0$ (and hence $E(u_2|I > 0) = \rho\sigma\lambda$), whereas $E(u_4|I > 0) = 0$ is assumed in the two-part model. This is the core of the debate between Hay and Olsen (1984) and Duan et al. (1984). Hay and Olsen (1984) argues that for $E(u_4|I > 0) = 0$ to be consistent with eqs. (4) and (5), the two-part model imposes some unusual assumptions on the distribution of the error terms u_3 and u_4. Duan et al. (1984) counter their criticism by constructing an example in which both the joint and the marginal distributions of u_3 and u_4 are consistent with (4) and (5). In particular, Duan et al. (1984) show that u_3 and u_4 can be correlated, but the correlation coefficient need not be estimated and is irrelevant for the purpose of estimating the two-part model.[6]

In Maddala's (1985a, 1985b) adjudication of the disputes between Hay and Olsen (1984) and Duan et al. (1984), he points out that the two sides are talking about two different types of models. He maintains that the sample selection model deals with joint decision problems such as the decision to buy a car and the expenditures on the car. In contrast, the two-part model is designed for sequential decision problems such as the decision to visit a doctor and the medical expenditures. If an individual decides to see a doctor, he puts matters in the doctor's hands for he has little control over the costs and there may also be unforeseen expenditures. Hence, Maddala argues that the decision process (joint or sequential) largely determines which model is pertinent. Nevertheless, Duan et al. (1985, p.22) dismiss Maddala's

[6] On this part of the debate, Duan et al. (1984) are indisputably correct. In fact, one of our simulation experiments below, which is based on the example in Duan et al. (1984), illustrates that it is possible to generate the error terms u_3 and u_4 in a way that is consistent with (4) and (5).

argument as "more semantic than substantive."[7]

Because (4) and (5) are conditionally independent, the estimation procedure for this model is straightforward. For the probit model (4), α is estimated by maximum likelihood. For the linear model (5), β is estimated by regressing ln(y) on $\underline{x}_2$ using the sample with positive observations. Following Manning, Duan, and Rogers, we call this the naive two-part (N2P) model.

In addition to the naive two-part model, Manning, Duan, and Rogers propose a data-analytic two-part (DA2P) model that tries to find the best specification for (5) by adding higher-order terms to the right-hand side of (5) and by looking for heteroskedasticity in the observed residuals. They use Mallow's (1973) C_p rule to determine whether higher-order terms should be included into the model. For simplicity, they only consider the second-order (squared) terms of $\underline{x}_2$. If the t statistic of a second-order term is greater than 1.414, then the term will be included in the specification. In other words, the data-analytic two-part model will coincide with the naive two-part model when each of the t-statistics of the second-order terms is less than 1.414.

3. Experimental Designs, Performance Criteria, and Collinearity Measures

3.1 Experimental Designs

In our Monte Carlo experiments, we will focus on the case where there is only one regressor in the choice and the level equations, i.e., $J = K = 2$. For brevity, let $x_1 = x_{12}$ and $x_2 = x_{22}$, then $\underline{x}_1 = (1,x_1)$ and $\underline{x}_2 = (1,x_2)$. Let $\theta(x_1,x_2)$ denote the theoretical correlation coefficient of x_1 and x_2. We conduct five experiments based on the following designs:

Design [1]: The sample selection model is the true model. The regressors are identical: $x_1 = x_2 = x$; and x is drawn randomly from U(0,3). The error terms u_1 and u_2 are drawn randomly from a bivariate

[7] In his rejoinder, Maddala (1985) insists that the issue is substantive and not semantic.

normal distribution with zero means, $var(u_1) = 1$, $var(u_2) = 1$, and $\rho = 0.5$.

Design [2]: The same as design [1] except that x is drawn from U(0,10).

Design [3]: The sample selection model is the true model. The regressors x_1 and x_2 have zero correlation ($\theta(x_1,x_2) = 0$), and they are drawn independently from U(0,10). The error terms u_1 and u_2 are drawn randomly from a bivariate normal distribution with zero means, $var(u_1) = 1$, $var(u_2) = 1$, and $\rho = 0.5$.

Design [4]: The same as design [3] except that $\theta(x_1,x_2) = 0.5$.[8]

Design [5]: The two-part model is the true model. The regressors are identical: $x_1 = x_2 = x$; and x is drawn randomly from U(0,10). The error term u_3 is drawn randomly from a standard normal distribution and u_4, given $I > 0$, is drawn from U(-1.5,1.5).[9]

We create 1,000 observations for each sample and perform 100 repetitions for each experiment. In all of the experiments, we set $\alpha_2 = \beta_2 = 1$, and $\alpha_1 = \beta_1$. Three different probabilities of a positive outcome ($I > 0$) are examined: 0.75, 0.50, and 0.25. These probabilities, denoted by P_+, are obtained by varying the intercept α_1. It is clear from (1) and (3) that the larger the value of α_1, the higher will be P_+ (the probability that $I > 0$).[10] For each experiment, four estimators are considered: LIML and FIML

---

[8] The regressors for the second case ($\theta(x_1,x_2) = 0.5$) are created in the following way. First, we generate two random variables w_1 and w_2 from a bivariate normal distribution with zero means, unit variances, and correlation coefficient 0.526. Then we obtain x_1 and x_2 by setting $\Psi(x_1) = \Phi(w_1)$ and $\Psi(x_2) = \Phi(w_2)$, where $\Phi(.)$ denote the c.d.f. of N(0,1) and $\Psi(z) = z/10$ is the c.d.f. of U(0,10). Then x_1 and x_2 are uniformly distributed with a sample correlation coefficient of 0.5. Hence, accurately speaking, 0.5 is the sample, not the theoretical, correlation coefficient of x_1 and x_2.

[9] We adopt the example in Duan et al. (1984) and generate u_3 and u_4 in the following way. First we draw (u_{3i},z_i) from a bivariate normal distribution with zero means, unit variances, and correlation coefficient 0.5, $i = 1,2,\ldots,1000$. For each observation i we check whether $I_i = \underline{x}_1\alpha + u_{3i} > 0$. If $I_i > 0$, we set $F(u_{4i}) = \Phi(z_i)$, where F(.) denote the c.d.f. of U(-1.5,1.5). If $I_i \leq 0$, we set $u_{4i} = -\infty$. In this way, u_{3i} and u_{4i} satisfy the assumptions of the two-part model.

[10] When x is drawn from U(0,3), we choose $\alpha_1 = -0.57$, -1.5, and -2.43 to obtain 25%, 50%, and 75% censoring, respectively. When x is drawn from U(0,10), we pick $\alpha_1 = -2.5$, -5, and -7.5 to achieve 25%, 50%, and 75% censoring, respectively.

estimators for the sample selection model, and N2P and DA2P estimators for the two-part model.[11]

3.2 Performance Criteria

We assess the performance of the estimators in six different ways. The first pair of performance criteria is mean prediction bias (MPB) and mean squared prediction error (MSPE):

$$\text{MPB} = (1/n)\,\Sigma_{i=1}^{n}\,[\hat{E}(y_i)-E(y_i)],$$

$$\text{MSPE} = (1/n)\,\Sigma_{i=1}^{n}\,[\hat{E}(y_i)-E(y_i)]^2,$$

where n = 1,000. Both criteria are used in Hay, Leu, and Rohrer and Manning, Duan, and Rogers. Notice that we measure the prediction error of the expected untransformed unconditional outcome E(y).[12] As emphasized in van de Ven and van Praag's (1981a, 1981b) and Duan et al. (1983, 1984, 1985), the quantity E(y) plays an important role in empirical work. For the sample selection model, it is easy to verify with some calculations that

$$\begin{aligned} E(y) &= [\text{Prob}(I > 0)]E[\exp(m)\,|\,I > 0] \\ &= [\Phi(\underline{x}_1\alpha+\rho\sigma)]\exp(\underline{x}_2\beta+\sigma^2/2). \end{aligned} \qquad (6)$$

The LIML and FIML estimators for E(y) are obtained by plugging the LIML and FIML estimators of α, β, ρ, and σ into (6), respectively. For the two-part model, the expression for E(y) is simply

$$\begin{aligned} E(y) &= [\text{Prob}(I > 0)]E[\exp(m)\,|\,I > 0] \\ &= [\Phi(\underline{x}_1\alpha)][\exp(\underline{x}_2\beta)]E[\exp(u_4)]. \end{aligned} \qquad (7)$$

---

[11] In Manning, Duan, and Rogers' version of the data-analytic two-part model, they also test and adjust for heteroskedasticity in the level equation. We do not follow this procedure for several reasons. First, they do not indicate what test and adjustment they used. Second, they find that "Less than 10 percent of the time did the data-analytic two-part model use a heteroscedastic retransformation" (p. 79). Third, unless one also tests and adjusts for heteroskedasticity for the LIML estimators, it does not seem to be fair to do the procedures merely for the two-part model.

[12] We have also calculated the MPB and MSPE using the conditional outcome $E(y\,|\,y>0)$ instead of the unconditional outcome E(y). The simulation results are found to be very similar and therefore will not be reported below.

Hence, E(y) depends on u_4 through the retransformation factor $E[\exp(u_4)]$. If u_4 is assumed to be $N(0,\tau^2)$, then $E[\exp(u_4)] = \tau^2/2$. Hence,

$$E(y) = [\Phi(\underline{x}_1\alpha)][\exp(\underline{x}_2\beta+\tau^2/2)]. \qquad (8)$$

If u_4, given $I > 0$, is assumed to be U(a,b) (as in design [5]), then

$$E(y) = [\Phi(\underline{x}_1\alpha)][\exp(\underline{x}_2\beta)][\exp(b)-\exp(a)]/(b-a).$$

Instead of specifying a particular distributional form for u_4, Duan et al. (1983, 1984, 1985) and Manning, Duan, and Rogers advocate Duan's (1983) distribution-free smearing estimator for $E[\exp(u_4)]$. Let m be the number of observations with positive outcomes and $\hat{u}_{4i}$ denote the ordinary least squares residual from (5), then the smearing estimate is given by $s = \Sigma_{i=1}^{m} [\exp(\hat{u}_{4i})]/m$. Duan (1983) demonstrates that s is a consistent nonparametric estimator of $E[\exp(u_4)]$. Thus, the N2P and DA2P estimates of E(y) are obtained by substituting the N2P and DA2P estimates of α, β, ρ, σ, and $E[\exp(u_4)]$ into (7), respectively.

The second pair of performance criteria is the parameter bias (PB) and the parameter squared error (PSE):

$$PB = \hat{\beta}_2 - \beta_2,$$

$$PSE = (\hat{\beta}_2 - \beta_2)^2.$$

The parameter squared error is used in Hay, Leu, and Rohrer's Monte Carlo study. We focus on β_2 because it is usually the parameter of interest in the model. Since Manning, Duan, and Rogers argue that the parameter β of the sample selection and the two-part models are not comparable,[13] we also examine the elasticity bias (EB) and the elasticity squared error (ESE):

$$EB = \hat{\eta} - \eta,$$

$$ESE = (\hat{\eta} - \eta)^2,$$

---

[13] Manning, Duan, and Rogers (1987, p.60) maintain that "the coefficients in the level-of-use equations for the two models are incomparable. In the two-part model, the level-of-use equation models the conditional distribution of the actual outcome (e.g., for those with any use). For the selection model, the same equation models the unconditional distribution of the potential outcome."

where $\eta = [\partial E(y)/\partial z][z/E(y)]$ for some variable z in $\underline{x}_2$. All elasticities are evaluated at the mean values of the regressors. Apart from the quantity E(y), the elasticity η has also played a major role in empirical work involving the two-part model; see, e.g., Manning, Blumberg, and Moulton's (1991) recent study on the demand for alcoholic beverages. The expression for η depends on the assumptions on $\underline{x}_1$ and $\underline{x}_2$.

For the sample selection model, if $z = x_1 = x_2$, then (6) implies that

$$\eta = [\partial E(y)/\partial z][z/E(y)] = [\alpha_2\lambda(\underline{x}_1\alpha+\rho\sigma) + \beta_2]z. \qquad (9)$$

If $z = x_2$ but $z \neq x_1$, then

$$\eta = \beta_2 z. \qquad (10)$$

For the two-part model, if $z = x_1 = x_2$, then (7) implies that

$$\eta = [\partial E(y)/\partial z][z/E(y)] = [\alpha_2\lambda(\underline{x}_1\alpha) + \beta_2]z, \qquad (11)$$

regardless of the distributional assumption on u_4. The expression for η is the same as (10) when $z = x_2$ but $z \neq x_1$. Since we iterate each experiment 100 times, the numbers for MPB, MSPE, PB, PSE, EB, and ESE reported below are the mean values out of 100 replications.

3.3 Measures of collinearity

One of the main objectives of the paper is to show that the LIML estimator for the sample selection model is particularly vulnerable to collinearity problems. To that end we need to utilize some measures of collinearity. It is well known that collinearity is a data problem and there is hardly a consensus among econometricians on the best measure of collinearity. While we recognize that there may not be a perfect diagnostic procedure for collinearity, we employ several measures to detect the problem. The first one is simply $r(x_2,\hat{\lambda})$, the sample correlation coefficient of x_2 and $\hat{\lambda}$. Because by design there are only two regressors (x_2 and $\hat{\lambda}$) in the second step of the LIML in our experiments, $r(x_2,\hat{\lambda})$ is obviously an appropriate diagnostic measure for collinearity. Furthermore, the R^2 of the auxiliary regression (regressing $\hat{\lambda}$ against x_2), which is a widely used measure of collinearity, is just $[r(x_2,\hat{\lambda})]^2$ (as there is only

one regressor). It also follows that the variance inflation factor (VIF = $(1-R^2)^{-1}$), which is another yardstick of collinearity suggested in the literature, has a one-to-one correspondence with $[r(x_2,\hat{\lambda})]^2$.

The second diagnostic tool we employed is the condition number advocated by Belsley, Kuh, and Welsch (1980). The condition number is defined as the square root of the ratio of the largest to the smallest eigenvalue of the moment matrix X'X. As the eigenvalues are dependent on the scale of the data, we follow Belsley, Kuh, and Welsch's suggestion by normalizing the data matrix X to have unit column length. Clearly, the condition number of any matrix is bounded below by one. It will attain unity when the matrix has orthonormal columns (the perfect case). In general, the higher the condition number, the more likely will there be collinearity problems. Based on a series of Monte Carlo experiments, Belsley, Kuh, and Welsch suggest that a condition number beyond 30 is indicative of collinearity problems.

In addition to using sample correlation coefficient and condition number, we also follow Belsley, Kuh, and Welsch's (1980, p.113) two-step diagnostic procedure. They suggest that "an appropriate means for diagnosing degrading collinearity is the following double condition: 1* A singular value judged to have a high condition index, and which is associated with 2* High variance-decomposition proportions for <u>two or more</u> estimated regression coefficient variances."[14] The number of high condition indices identifies the number of near dependencies among the columns of the X matrix. The high variance-decomposition proportions associated with each high condition index identify those regressors that are involved in the corresponding near dependency.

---

[14] Let e_j (j = 1,2,...,K) denote the jth eigenvalue of the KxK matrix X'X (K = 3 in our regression model) and e_{max} = Max $\{e_1,...,e_K\}$. The condition index associated with the jth eigenvalue is defined as $(e_{max}/e_j)^{1/2}$. By definition, the largest condition index is the condition number. Let M be a KxK matrix that diagonalizes X'X, then $M^{-1}(X'X)M = D$, where D is a KxK diagonal matrix with e_j on the diagonal. Consider a linear model $Y = X\theta + Z$, with i.i.d. error terms z_i and $Var(z_i) = s^2$. Then $V(\theta_{OLS}) = s^2(X'X)^{-1} = s^2MD^{-1}M'$, or $Var(\theta_{OLS,i}) = s^2\sum_{j=1}^{K}[(m_{ij})^2/e_j]$, where $M = (m_{ij})$. Then the variance-decomposition proportion of $\theta_{OLS,i}$ associated with e_j is defined as $[(m_{ij})^2/e_j]/\sum_{j=1}^{K}[(m_{ij})^2/e_j]$. See Belsley, Kuh, and Welsch (1980) for details.

4. Simulation Results

4.1 True Model = Sample Selection Model with $x_1 = x_2$

As a basis of comparison, we first run an experiment to replicate the key findings in Manning, Duan, and Rogers. The experiment is based on design [1], which is essentially the same as the first experiment in Manning, Duan, and Rogers, and the results are contained in Table 1. It verifies their claims that the two-part model performs better than the selection model even though the latter is the true model. The LIML estimator is worse than the other three estimators: the mean squared prediction errors are substantially larger than those of the others regardless of the degree of censoring. In general, the LIML estimators are poorer the higher the degree of censoring (the smaller the proportions of uncensored observations). Although the LIML estimator has the smallest parameter bias, the parameter squared error, the elasticity bias, and the elasticity squared error are all greater than those of the N2P estimator, which indicates that the LIML estimator is less stable and therefore less reliable than the two-part model.[15]

The second experiment is based on design [2] and the results are reported in Table 2. Compared to Table 1, Table 2 gives a totally different picture. The sample selection model outperforms the two-part model in all but one case. When $P_+ = 0.25$, the two-part model performs better in terms of MSPE, a finding which will be explained below.

The striking differences between Table 1 and Table 2 can be explained by collinearity. Table 3 reports various measures of collinearity between the regressors in the second step of the LIML in the first two experiments. When x is drawn from U(0,3), the absolute values of the sample correlation coefficients are all greater than 0.95 and the condition numbers are exceedingly high for $P_+ = 0.5$ and 0.25.[16]

---

[15] Notice that even though the true model is the selection model, the Heckman two-step estimator (LIML) is not unbiased. Hence one should not expect that the bias of the LIML estimator to be close to zero.

[16] One may wonder why the condition number in Table 3 (and also Table 8) is not equal to the square root of the ratio of the largest to the smallest eigenvalue (e.g., $(2.3574/0.0046)^{1/2} = 22.63 \neq 23.036$).

Using Belsley, Kuh, and Welsch's two-step diagnostic procedure, we find that the high condition indices (53.708 and 155.68) are associated with the smallest eigenvalue (0.0009 and 0.0001) for $P_+ = 0.5$ and 0.25, and the corresponding high variance-decomposition proportions, all provide clear evidence that there are collinearity problems. With such a high collinearity, the LIML estimators are unstable and hence perform much worse than the two-part model, even though the selection model is the true model. However, when x is drawn from U(0,10), the absolute values of the sample correlation coefficients and the condition numbers for both $P_+ = 0.5$ and 0.75 are substantially lowered. There being no collinearity problems, the LIML estimators behave much better than those of the two-part models. For the case $P_+ = 0.25$, where the condition number reaches 64.45, collinearity problems reappear again. This explains why the LIML estimator has a larger MSPE than the two-part model in this case.

Tables 1-3 indicate that collinearity problems become serious when the condition number is higher than 20, which is lower than the threshold condition number (30) that Belsley, Kuh, and Welsch find in their Monte Carlo studies. Table 4 shows that the degree of censoring has a dramatic impact on the degree of collinearity. Even when the regressor is drawn from U(0,10), the model can still suffer from near collinearity. Taking 20 as the threshold condition number, the sample must contain at least 80 percent uncensored data in order to avoid high collinearity in the U(0,3) case. For the U(0,10) case, at least 50 percent uncensored observations are required.

Censoring increases collinearity in two ways. First, censoring reduces the number of positive observations and hence lowers the variability of the inverse Mills' ratio. Second, the range of the inverse Mills' ratio $\lambda(\alpha_1 + \alpha_2 x_1)$ diminishes as censoring increases. To see this, notice that given $\alpha_2 = 1$ and $x_1 \sim U(0,\xi)$ ($\xi = 3$ or 10), the range of λ is given by $[\lambda(\alpha_1 + \xi),\lambda(0)]$. The upper bound is $\lambda(0)$

The reason is that each condition number reported there is the mean of 100 condition numbers (each of which is the square root of the ratio of the largest to the smallest eigenvalue) as there are 100 iterations. Similarly, each eigenvalue in Table 3 is the mean of 100 eigenvalues. Hence, the mean of the condition numbers does not exactly coincide with the square root of the ratio of the mean of the largest eigenvalues to the mean of the smallest eigenvalues.

because m is observed only if $I = \alpha_1 + \alpha_2 x_1 = \alpha_1 + \xi > 0$ [see eqs. (1) and (2)] and $\lambda(.)$ is a decreasing function. Since $\alpha_1 < 0$, a higher degree of censoring is achieved by a larger value of $|\alpha_1|$. As a result, given ξ, $\alpha_1 + \xi$ falls as censoring ($|\alpha_1|$) increases. Hence, more censoring decreases the range, and hence the variability, of λ.

The poor performance of the LIML estimator stems from the highly conditioned moment matrix in the second step of the two-step procedure. Although the FIML does not appear to depend on this moment matrix, Tables 1 and 2 reveal that high collinearity also impairs the FIML estimator.[17] In most cases, the N2P estimator dominates the FIML estimator when collinearity is high. Hence our results do not support Nelson's (1984) finding that collinearity has relatively little effect on the FIML estimator and his conclusion that the FIML should be used when collinearity is high. We believe that the N2P estimator may be better than the FIML in these circumstances.

To compare the MSPEs across the three values of P_+ in Table 1, we normalize (scale) the MSPE for each P_+ by taking the square root of MSPE and then dividing it by E(y). Figure 1 plots the normalized root mean squared prediction error against the condition number. There are three distinct clusters of data points because the ranges of the condition numbers of the three values of P_+ do not overlap. From the left to the right, the clusters refer to $P_+ = 0.75$, 0.5, and 0.25, respectively. The figure reveals that the normalized root MSPEs are roughly the same for $P_+ = 0.75$ and 0.5, but the errors increase considerably when $P_+ = 0.25$.[18] By normalizing the MSPEs from Table 2, a similar

---

[17] As the likelihood equation for the FIML might not have a unique root, we tried many different starting values to ensure that the weak performance of the FIML estimator was not due to the particular starting values that we chose. Among the starting values we used were the LIML estimates and the true parameter values, and the FIML estimates always remained virtually the same. We also estimated a Type 1 Tobit model (of which the root of the likelihood equation is unique) and still found that high collinearity impairs the FIML estimator. These results suggest that the choice of the starting values is not the cause of the problem.

[18] Notice that a hugh outlier (156,15155) was deleted from the figure because it would dramatically change the scale of the diagram.

diagram is obtained as shown in Figure 2. It again indicates that the normalized root MSPEs are roughly the same for $P_+ = 0.75$ and 0.5. The errors are in general higher when $P_+ = 0.25$. From these two figures, one can see that the normalized root MSPEs increase appreciably only when the condition numbers get very high.

Table 5 reports the effects of collinearity on hypothesis testing. Under the null hypothesis $\rho = 0$, a simple "asymptotic t-test" on the coefficient of $\hat{\lambda}$ can be used to test whether the sample selection model is the true specification.[19] As there are 100 replications in each experiment, the frequency of not rejecting the null hypothesis should be close to the prescribed level of significance, given that the sample selection is the true model. Using the 5 percent significance level, the t-ratios in Table 5 show that, when x is drawn from U(0,3) and $P_+ = 0.75$, the rejection frequency is only 24 percent. The rejection frequency diminishes with the degree of censoring. When $P_+ = 0.25$, the t-test fails completely because the null hypothesis is never rejected. The rejection frequencies are clearly substantially below the expected 95 percent. When x is drawn from U(0,10), the rejection frequencies are higher, although they are still considerably below 95 percent (especially for $P_+ = 0.5$ and 0.25). Thus, there is a considerable divergence between the actual and the nominal sizes, although the discrepancy falls as collinearity decreases. Only two of the six mean values of the t-ratios (column 3) are greater than 1.96 (the critical t-value at the 5 percent significance level), and it is no coincidence that the four cases in which the mean t-ratios are less than 1.96 are exactly the ones with high collinearity. Therefore, high collinearity renders the t-tests ineffective because they fail to reject the two-part model even when the true model is the sample selection model. The lack of power of these tests manifest the harmful effects of collinearity.[20]

---

[19] Heckman (1979) suggests this t-test and Melino (1982) proves that the test is equivalent to the Lagrange multiplier test and therefore has desirable asymptotic properties.

[20] Using Belsley, Kuh, and Welsch's terminology, the collinearity is harmful because "it is first degrading and then ... important tests based on the degraded estimates are considered inconclusive, for these tests could be refined and made more trustworthy (even if the outcome is the same) when based on better conditioned data" (p.172).

4.2 True Model = Sample Selection Model with $x_1 \neq x_2$

There are two different designs on the regressors in Manning, Duan, and Rogers' experiments. In the first design (the no-exclusion-restrictions case), x_1 and x_2 are identical and therefore perfectly correlated. In this case, they find that the LIML estimator performs poorly. In the second design (the exclusion-restrictions case), x_1 and x_2 are not correlated, and they find that the LIML estimator behaves much better than the N2P and DA2P estimators. Based on these contrasting results from the two limiting cases ($\theta(x_1,x_2) = 0$ and 1), they "conjecture that the LIML estimator will be less well behaved if x_1 and x_2 are correlated. ... If this conjecture is correct, then the performance of LIML estimator may depend on how correlated the measures are, not just the presence of exclusions" (p.74). The next two experiments are designed to evaluate their conjecture.

Table 6 reports the simulation results based on design [3]. Similar to Manning, Duan, and Rogers, the LIML estimator performs very well. The MSPEs of the N2P estimator are at least twice as large as those of the LIML and FIML in all three cases. The N2P estimator is also inferior to the LIML and FIML estimators in almost all of the other criteria, except in the case $P_+ = 0.75$ where the PB and EB are notably smaller. In general, the FIML is slightly better than the LIML.

Table 7 describes the results based on design [4]. When x_1 and x_2 are imperfectly correlated, the LIML and FIML estimators continue to outperform the N2P and DA2P estimators. Both Tables 6 and 7 indicate that the DA2P estimator is almost always the worst in terms of mean squared prediction error, parameter squared error, and elasticity squared error. These two sets of experiments therefore clearly disprove Manning, Duan, and Rogers' conjecture that the LIML estimator will not behave well if x_1 and x_2 are correlated. Regardless of the degree of correlation between x_1 and x_2, the LIML estimator will perform well as long as $\hat{\lambda}$ and x_2 are not highly correlated.

Table 8 shows that for both $\theta(x_1,x_2) = 0$ and 0.5, the absolute values of $r(x_2,\hat{\lambda})$, as well as the condition numbers, are very low. There are no signs of collinearity problems. This again verifies our

claims that the sample selection estimators will perform well when there are no collinearity problems. The t-tests also perform very well. Unlike the results in Table 5, Table 9 demonstrates that when there are no collinearity problems, the t-tests on the coefficient of $\hat{\lambda}$ are very effective. In all of the six cases, the rejection frequencies are very close to 95 percent and the mean t-ratios are well above 1.96. The two-part model is properly rejected when the true model is the sample selection model.

Although Nelson (1984) finds that the FIML estimator dominates the LIML estimator in terms of efficiency, we observe that the LIML is not always inferior to the FIML when other criteria are considered. Tables 2, 6, and 7 illustrate that the FIML is in general better than the LIML in terms of squared errors (MPSE, PSE, ESE), but the LIML can be better than the FIML in terms of biases (MPB, PB, and EB).

4.3 True Model = Two-part Model with $x_1 = x_2$

Table 10 contains the simulation results based on design [5]. Given that the true model is the two-part model, it is clear from Table 10 that the N2P model dominates the sample selection model in every aspect. Nevertheless, the performance (in terms of order of magnitude) of the LIML and FIML estimators is comparable to the N2P estimator especially for $P_+ = 0.75$ and 0.5. When the true model is the two-part model, the LIML estimator is expected to have larger squared errors because an irrelevant variable, $\hat{\lambda}$, has been admitted into the regression. The LIML estimator is not expected to behave well when $P_+ = 0.25$, for there exists high collinearity in the selection model (as we have seen from Table 3). Table 10 confirms these conjectures. A somewhat surprising finding is that the DA2P estimator behaves worse than the LIML and FIML estimators in terms of PSE when $P_+ = 0.5$ and 0.25, and in terms of ESE when $P_+ = 0.25$.

Table 11 indicates that when the two-part model is the true model, the frequency of rejecting the null hypothesis $\rho = 0$ is very close to the prescribed 5 percent level of significance. For the three

different values of P_+, the t-tests incorrectly reject the two-part model only 4, 5, and 11 times out of 100 replications. This again confirms that the t-tests are very effective when there are no collinearity problems in the model.

5. Discussions

The results in the previous section clearly demonstrate that the merits of the two-part model have been grossly exaggerated in the literature. We have proved that a deficient design in Manning, Duan, and Rogers' data generating process causes serious collinearity problems that lead to the poor performance of the sample selection model. When the deficient design is corrected, the sample selection model clearly dominates the two-part model when the former is the true model. In fact, Manning, Duan, and Rogers' own findings corroborate with our argument that there is a design problem in the regressors in their experiments. They chose a range of [0,3] for the regressors because they "tried a narrower range of 1, but had severe numerical problems with the LIML version of the selection model" (p.65, footnote 10). This suggests that the sample selection model experiences serious collinearity problems when the range of the regressor is [0,1]. Had they gone further to pick a broader range than [0,3], they would have found that their negative findings regarding the sample selection model are not robust.[21] Hence, the extreme and negative remarks against the sample selection model made by Duan et al. (1983, 1984, 1985) are obviously unwarranted and misleading. Although the sample selection model is susceptible to collinearity problems, one cannot reject it in favor of the two-part model because of its numerical weakness. A model is rejected if its implications are contradicted by data. Numerical problems encountered with a particular data set does not invalidate the model.

---

[21] We believe that Hay, Leu, and Rohrer's (1987) negative results regarding the sample selection model can also be explained by collinearity problems. As we do not have access to their data, we cannot replicate their results and verify our conjecture.

From a practical point of view, the distinction between the two-part and the sample selection models is whether a selection-bias adjustment (adding the inverse Mills' ratio in the normality case) should be made. In a study using some simple truncation models, Goldberger (1983) finds that the normal selection-bias adjustment procedure is quite sensitive to modest departures from the normality assumption on the error term. While the bias of the normal selection-bias adjustment can be quite substantial when the true distribution of the error term is not normal, the normal selection-bias adjustment appears to be better than no adjustment at all. Partly based on Goldberger's findings, Duncan (1983) extends the case to censored (Tobit) models and recommends that given unknown error distributions, it will still be better to include the normal selection-bias term than to drop it entirely. Our results, however, suggest that Duncan's recommendation is questionable because the normal selection-bias adjustment procedure may do more harm than good. There is no selection-bias adjustment in the two-part model and we have seen from Table 10 that when the two-part model is the true model, the LIML estimator is significantly inferior to the N2P estimator. Hence, one should not blindly adopt the normal selection-bias adjustment procedure in all circumstances.

In contrast to Duan et al. (1985) who argue that Maddala's distinction between the sample selection and the two-part models is semantic and not testable, we have shown that the two models are testable in principle. With the null hypothesis that the two-part model is the true model, a t-test can be used to test against the alternative hypothesis that the true model is the sample selection model. However, the power of the test will be limited by the presence of collinearity problems, as we have seen from the results in Table 5. Another problem with the t-test is that it is possible to find that the coefficient of $\hat{\lambda}$ to be significant (say, t-ratio > 2) and yet the data matrix has a high condition number. Figures 3 and 4 plot the t-ratio against the condition number for all three values of P_+. In Figure 3, although the middle cluster of data (i.e., when $P_+ = 0.5$) all have condition numbers higher than 40, some of the t-ratios are larger than 2. Similarly, in Figure 4, some of the data in the right cluster (i.e., when $P_+ = 0.25$) have

t-ratios larger than 2, even though the condition numbers are greater than 50. Hence, a t-ratio above 2 does not guarantee that the data are free of collinearity problems. The high condition numbers indicate the presence of high collinearity and that the estimates may be very sensitive and unstable.

A high collinearity between x_2 and $\lambda(\underline{x}_1\hat{\alpha})$ can arise in a number of ways. We have seen the case that if $x_2 = x_1$, and x_1 has little variation, then x_2 and $\lambda(\underline{x}_1\hat{\alpha})$ will be highly collinear. A high degree of censoring can also generate near collinearity because more censoring reduces the sample size and the variation of $\hat{\lambda}$. This has been shown in Table 4, where the absolute value of the sample correlation coefficient and the condition number decrease with the proportion of uncensored observations. A higher variance of u_1 can also cause near collinearity because the variation of the inverse Mills' ratio decreases with the standard error of u_1.[22] In view of this drawback, we suggest that one should examine whether there are collinearity problems whenever Heckman's two-step procedure is applied in empirical work. Our experience is that a condition number above 20 is indicative of collinearity problems. This is lower than the threshold condition number (30) suggested by Belsley, Kuh, and Welsch.

While Nelson (1984) recommends using the R^2 (from the regression of $\hat{\lambda}$ against x_2) to detect collinearity, our results suggest that the condition number is a better measure of collinearity. To see this, consider the following two cases in Table 4: (i) $x \sim U(0,3)$ and $P_+ = 0.9$, and (ii) $x \sim U(0,10)$ and $P_+ = 0.3$. Although the condition number in case (i) is considerably smaller than that in case (ii) (13.63 versus 44.87), the sample correlation coefficients (the square of which are the R^2s since there is only one regressor in the model) are approximately the same (-0.9123 and -0.9157). The condition number in case (i) does not signify collinearity problems whereas the condition number in case (ii) indicates serious collinearity problems. In contrast, given the same sample correlation coefficient (about -0.91) in cases (i) and (ii), one cannot tell whether there are collinearity problems. This example illustrates that the

---

[22] When σ_1 (standard error of u_1) is not necessarily unity, the inverse Mill's ratio is given by $\lambda(\underline{x}_1\alpha/\sigma_1)$. Other things being equal, the variation of $\underline{x}_1\alpha/\sigma_1$, and hence λ, decreases with σ_1.

condition number is superior to the sample correlation coefficient in detecting collinearity.

Because the condition numbers reported in the tables are the averages of 100 replications, it may be useful to examine their sample distributions. Table 12 describes some basic summary statistics of the condition numbers. They show that the distribution of the condition numbers is fairly symmetric and concentrated around the mean. Because the standard errors are small relative to the means, the condition numbers are therefore reliable indicators of collinearity. In view of this and the previous results, and the fact that condition numbers (or eigenvalues) can readily be obtained from most statistics and econometrics softwares (such as GAUSS, SAS, SPSS), we recommend that they be used to detect collinearity.[23]

We believe that collinearity problems provide an additional (or alternative) explanation for the anomalous results that many have found in their applications of the two-step procedure. For example, a large number of empirical studies find (somewhat surprisingly) that the coefficient estimates on the inverse Mills' ratio are generally insignificant, contrary to what one would expect from economic theory. Non-normality of the error terms and heteroskedasticity have often been employed to explain the anomalies (e.g., Duncan (1983)). However, based on our Monte Carlo results, we believe that collinearity may also be responsible for the large standard errors that give rise to the insignificance of the coefficient estimates of the inverse Mills' ratio.

6. Conclusion

Our results demonstrate that a biased experimental design can seriously distort the results of a Monte Carlo study. Extensive Monte Carlo experiments are required before any reliable conclusions can be drawn from the simulations. While we argue that there is a major shortcoming in Manning, Duan, and Rogers' experimental design that severely hampers their conclusions, we also recognize that our own

[23] Of course, one should not indiscriminately rely on just one measure of collinearity. The condition number is not perfect, nor are any other collinearity indices (see the discussions in Stewart (1987)).

Monte Carlo experiments are not complete because we have not investigated other design issues. For instance, we have not studied how the various biases and squared errors may change with the parameters (e.g., α_2, β_2) of the true model and the distributional assumptions on the error terms.

As the sample selection and the two-part models perform well under different simulation setups, we believe that a balanced view on the merits of the two models is more appropriate. The two-part model does provide an alternative approach to the sample selection model that has dominated the literature. Each model describes a different mode of decision making process. If the choice and the level of use decisions are made jointly, which theoretically seems to be the case in most economic problems, the sample selection model is the proper one to use. If the decisions are made sequentially, then the two-part model is more appropriate.

Finally, we remark that the design problem in Manning, Duan, and Rogers is actually a pervasive one in the sample selection literature. In most Monte Carlo studies on the sample selection model, the designs are usually only focused on exploring the impact of different error distributions, sample sizes, or degrees of censoring on the performance of the estimators. The role of the regressors and the data generating process have often been ignored. We demonstrate in a separate paper (Leung and Yu, 1992) that the ways the regressors are generated in these studies also produce collinearity problems: either the ranges of the regressors are too narrow or the variances of the error terms are too high. For example, Powell (1986) and Peters and Smith (1991) generate the regressors from U(-1.7,1.7), and the error terms have unit variance. This design is clearly similar to the U(0,3) regressors in Manning, Duan, and Rogers. Although Paarsch (1984) generates the regressors from U(0,20), the variance of the error term is 100. With such a high standard deviation (10), the effective range of the regressor becomes [0,2]. Consequently, Heckman's two-step estimator does not behave well in Paarsch's simulations. Paarsch's results are particularly influential and have led many to believe that Heckman's two-step procedure is an inferior one. Many researchers have since then adopted Paarsch's designs and excluded Heckman's two-

step estimator in their Monte Carlo studies (e.g., Powell (1986), Duncan (1986), Fernandez (1986), Moon (1989), Nawata (1990)). Hence, there are some notable misleading results in the sample selection literature because of the inadvertent bias against Heckman's two-step procedure.

Table 1
Simulation Results based on Design [1]
True Model = Sample Selection Model
$x_1 = x_2 \sim U(0,3)$
(Standard Errors in Parentheses)

Proportion of Uncensored Observations (P_+)	Estimation Method	Mean Prediction Bias (MPB)	Mean Squared Prediction Error (MSPE)	Parameter Bias (PB)	Parameter Squared Error (PSE)	Elasticity Bias (EB)	Elasticity Squared Error (ESE)
0.75	LIML	0.2383 (0.624)	1.7030 (5.677)	-0.0082 (0.181)	0.0326 (0.049)	0.0464 (0.135)	0.0203 (0.050)
	FIML	0.0979 (0.366)	0.4570 (0.675)	-0.0419 (0.139)	0.0209 (0.041)	0.0132 (0.088)	0.0079 (0.0135)
	N2P	-0.0169 (0.340)	0.4220 (0.400)	-0.1714 (0.044)	0.0313 (0.015)	-0.0214 (0.074)	0.0059 (0.007)
	DA2P	0.0103 (0.348)	0.4773 (0.579)	-0.3083 (0.214)	0.1406 (0.171)	-0.0363 (0.073)	0.0066 (0.008)
0.5	LIML	0.1420 (0.403)	0.8286 (5.289)	-0.0099 (0.383)	0.1451 (0.238)	0.0961 (0.202)	0.0497 (0.150)
	FIML	0.0564 (0.372)	0.2151 (1.179)	-0.1125 (0.262)	0.0805 (0.141)	0.0167 (0.227)	0.0511 (0.292)
	N2P	0.0065 (0.141)	0.0652 (0.071)	-0.2784 (0.055)	0.0806 (0.031)	0.0351 (0.124)	0.0164 (0.022)
	DA2P	0.0113 (0.143)	0.0750 (0.104)	-0.4283 (0.275)	0.2582 (0.311)	-0.0004 (0.118)	0.0138 (0.020)
0.25	LIML	30.025 (296.5)	6.9×10^5 (6.9×10^6)	-0.0112 (1.173)	1.3618 (2.494)	0.2907 (0.675)	0.5351 (2.491)
	FIML	0.0116 (0.154)	0.0569 (0.203)	-0.1910 (0.608)	0.4022 (1.879)	0.0784 (0.563)	0.3204 (2.352)
	N2P	0.0082 (0.057)	0.0112 (0.016)	-0.3539 (0.100)	0.1352 (0.075)	0.0371 (0.245)	0.0606 (0.098)
	DA2P	0.0084 (0.057)	0.0136 (0.023)	-0.4269 (0.423)	0.3592 (0.788)	0.0112 (0.246)	0.0599 (0.098)

Note: For $P_+ = 0.25$, the FIML fails to locate the maximum of the likelihood function in one of the iterations. Hence the FIML estimates reported in the table are based on 99 iterations of the model.

Table 2
Simulation Results based on Design [2]
True Model = Sample Selection Model
$x_1 = x_2 \sim U(0,10)$
(Standard Errors in Parentheses)

Proportion of Uncensored Observations (P_+)	Estimation Method	Mean Prediction Bias (MPB)	Mean Squared Prediction Error (MSPE)	Parameter Bias (PB)	Parameter Squared Error (PSE)	Elasticity Bias (EB)	Elasticity Squared Error (ESE)
0.75	LIML	6.5810 (25.28)	4202.4 (6709.9)	0.00057 (0.024)	0.00057 (0.0008)	0.0055 (0.114)	0.0130 (0.0179)
	FIML	6.4979 (24.84)	3999.2 (6064.6)	0.00013 (0.022)	0.00048 (0.0006)	0.0033 (0.107)	0.01128 (0.0148)
	N2P	-14.73 (22.42)	5174.5 (5575.1)	-0.0444 (0.017)	0.00225 (0.0016)	-0.1562 (0.094)	0.0332 (0.0337)
	DA2P	11.376 (28.33)	8985.8 (15773)	-0.2759 (0.100)	0.08605 (0.0516)	-0.3576 (0.123)	0.1428 (0.0835)
0.5	LIML	0.6207 (2.377)	44.031 (97.47)	0.00167 (0.049)	0.00240 (0.0042)	0.0872 (0.522)	0.2772 (0.421)
	FIML	0.4999 (2.329)	39.933 (71.82)	-0.0024 (0.045)	0.00199 (0.0031)	0.1114 (0.454)	0.2167 (0.344)
	N2P	-1.153 (2.040)	50.986 (49.78)	-0.0910 (0.030)	0.00918 (0.0052)	1.0639 (0.443)	1.3266 (1.106)
	DA2P	0.2696 (2.362)	52.055 (96.10)	-0.6280 (0.251)	0.45681 (0.2761)	0.2496 (0.599)	0.4174 (0.638)
0.25	LIML	0.0793 (0.248)	0.6736 (1.374)	-0.0034 (0.172)	0.0293 (0.0386)	0.3277 (2.352)	5.5860 (14.47)
	FIML	0.0415 (0.208)	0.3762 (0.778)	-0.0353 (0.141)	0.0208 (0.0309)	0.4789 (2.251)	5.2457 (13.05)
	N2P	-0.0269 (0.195)	0.3731 (0.402)	-0.2143 (0.059)	0.0493 (0.0259)	1.4662 (2.080)	6.4335 (12.51)
	DA2P	-0.0070 (0.200)	0.3581 (0.414)	-0.9020 (0.914)	1.6406 (2.347)	0.1147 (2.752)	7.5091 (12.87)

Table 3
Measures of Collinearity: Designs [1] and [2]
True Model = Sample Selection Model
$x_1 = x_2 = x$

Distribution of x	Proportion of Uncensored Observations	$r(x,\hat{\lambda})$	Eigenvalue	Condition Index	Variance-decomposition Proportions		
					Intercept	x_2	$\hat{\lambda}$
U(0,3)	0.75	-0.9573	2.3574	1	0.0014	0.002	0.004
			0.6381	1.9236	0.0001	0.008	0.040
			0.0046	23.036*	0.9985	0.990	0.956
	0.5	-0.9844	2.5665	1	0.00023	0.00037	0.001
			0.4326	2.4382	0.00005	0.00260	0.013
			0.0009	53.708*	0.99972	0.99703	0.987
	0.25	-0.9926	2.7742	1	0.00002	0.00005	0.0001
			0.2257	3.5141	0.00001	0.00079	0.0030
			0.0001	155.68*	0.99997	0.99916	0.9969
U(0,10)	0.75	-0.6564	2.0638	1	0.013	0.0135	0.02484
			0.9057	1.5097	0.001	0.0092	0.46675
			0.0305	8.2399*	0.986	0.9773	0.50841
	0.5	-0.7963	2.2272	1	0.0030	0.0029	0.02146
			0.7653	1.7063	0.0007	0.0032	0.30015
			0.0075	17.337*	0.9963	0.9939	0.67839
	0.25	-0.9434	2.4964	1	0.00018	0.0002	0.006
			0.5030	2.2310	0.00017	0.0005	0.086
			0.0006	64.449*	0.99965	0.9993	0.908

Note: A condition index with an asterisk indicates that it is the condition number.

Table 4
Effect of Censoring on Collinearity
$x_1 = x_2 = x$

Distribution of x	Proportion of Uncensored Observations	$r(x,\hat{\lambda})$	Condition Number
U(0,3)	0.9	-0.9123	13.63
	0.8	-0.9466	19.46
	0.7	-0.9657	27.36
	0.6	-0.9768	37.60
	0.5	-0.9844	53.71
	0.4	-0.9892	78.74
	0.3	-0.9919	120.82
	0.2	-0.9926	209.72
	0.1	-0.9989	447.86
U(0,10)	0.9	-0.6241	6.04
	0.8	-0.6409	7.25
	0.7	-0.6791	9.33
	0.6	-0.7370	12.43
	0.5	-0.7963	17.34
	0.4	-0.8577	26.19
	0.3	-0.9157	44.87
	0.2	-0.9648	100.02
	0.1	-0.9843	327.64

Table 5
Summary Statistics of the T-ratios and Rejection Frequency: Designs [1] and [2]

Design on x_1 and x_2	P_+	Mean	Standard Deviation	Skewness	Kurtosis	Minimum	Maximum	Rejection Frequency
$x_1 = x_2 \sim$ U(0,3)	0.75	1.073	1.087	-0.064	2.769	-2.023	3.188	24
	0.5	0.750	1.010	-0.265	3.007	-2.473	2.895	13
	0.25	0.251	0.749	-0.307	2.103	-1.243	1.766	0
$x_1 = x_2 \sim$ U(0,10)	0.75	3.378	1.137	0.054	3.259	-0.002	6.274	89
	0.5	2.669	1.044	0.147	2.332	0.292	4.970	68
	0.25	1.229	0.992	0.005	2.372	-1.200	3.388	27

Table 6
Simulation Results based on Design [3]
True Model = Sample Selection Model
$x_1, x_2 \sim U(0,10)$, $\theta(x_1,x_2) = 0$
(Standard Errors in Parentheses)

Proportion of Uncensored Observations (P_+)	Estimation Method	Mean Prediction Bias (MPB)	Mean Squared Prediction Error (MSPE)	Parameter Bias (PB)	Parameter Squared Error (PSE)	Elasticity Bias (EB)	Elasticity Squared Error (ESE)
0.75	LIML	3.5985 (17.84)	2427.49 (3130.5)	-0.00049 (0.0136)	0.000183 (0.000257)	-0.0025 (0.069)	0.00471 (0.0066)
	FIML	3.5123 (17.71)	2373.34 (3060.2)	-0.00058 (0.0135)	0.000179 (0.000256)	-0.0029 (0.068)	0.00462 (0.0066)
	N2P	2.5682 (18.45)	6167.52 (4409.6)	0.00027 (0.0138)	0.000190 (0.000261)	0.0014 (0.070)	0.00487 (0.0067)
	DA2P	2.5722 (19.23)	6697.10 (5188.1)	0.00038 (0.0406)	0.001634 (0.004285)	0.00147 (0.070)	0.0049 (0.0068)
0.5	LIML	0.2794 (1.311)	16.480 (22.29)	0.00039 (0.0167)	0.0002763 (0.000384)	0.00198 (0.0847)	0.00711 (0.0099)
	FIML	0.2579 (1.305)	16.002 (22.17)	0.00024 (0.0167)	0.0002759 (0.000384)	0.00120 (0.0847)	0.00710 (0.0099)
	N2P	0.2260 (1.347)	40.090 (33.15)	-0.00147 (0.0166)	0.0002759 (0.000398)	-0.00744 (0.0843)	0.00709 (0.0102)
	DA2P	0.2236 (1.451)	44.995 (41.26)	-0.00011 (0.0484)	0.002320 (0.006378)	-0.00740 (0.0841)	0.00706 (0.0102)
0.25	LIML	0.0179 (0.093)	0.1186 (0.154)	0.000566 (0.0220)	0.000479 (0.00074)	0.0029 (0.112)	0.0123 (0.019)
	FIML	0.0166 (0.092)	0.1169 (0.152)	0.000127 (0.0221)	0.000482 (0.00077)	0.00064 (0.112)	0.0124 (0.020)
	N2P	0.0224 (0.097)	0.2551 (0.254)	-0.00478 (0.0226)	0.000530 (0.00083)	-0.0242 (0.115)	0.0136 (0.021)
	DA2P	0.0207 (0.099)	0.2685 (0.274)	-0.00061 (0.0684)	0.004627 (0.0156)	-0.0238 (0.116)	0.0138 (0.021)

Table 7
Simulation Results based on Design [4]
True Model = Sample Selection Model
$x_1, x_2 \sim U(0,10)$, $\theta(x_1,x_2) = 0.5$
(Standard Errors in Parentheses)

Proportion of Uncensored Observations (P_+)	Estimation Method	Mean Prediction Bias (MPB)	Mean Squared Prediction Error (MSPE)	Parameter Bias (PB)	Parameter Squared Error (PSE)	Elasticity Bias (EB)	Elasticity Squared Error (ESE)
0.75	LIML	4.5801 (21.49)	2762.82 (3535.0)	-0.00061 (0.0132)	0.000173 (0.00027)	-0.0031 (0.067)	0.0044 (0.007)
	FIML	4.3635 (21.57)	2753.13 (3578.4)	-0.00086 (0.0132)	0.000174 (0.00027)	-0.0043 (0.067)	0.0045 (0.007)
	N2P	3.7209 (21.74)	3747.82 (4107.1)	-0.01427 (0.0131)	0.000373 (0.00041)	-0.0722 (0.066)	0.0096 (0.011)
	DA2P	5.8444 (22.79)	4648.41 (4884.0)	-0.02476 (0.0471)	0.002808 (0.00516)	-0.0749 (0.068)	0.0102 (0.011)
0.5	LIML	0.2247 (1.733)	20.3768 (23.479)	-0.00187 (0.0158)	0.000252 (0.00043)	-0.0095 (0.080)	0.0065 (0.011)
	FIML	0.2308 (1.747)	20.5593 (24.426)	-0.00172 (0.0163)	0.000265 (0.00048)	-0.0087 (0.082)	0.0068 (0.012)
	N2P	0.2617 (1.772)	31.6954 (31.512)	-0.0202 (0.0157)	0.000655 (0.00086)	-0.1023 (0.080)	0.0168 (0.022)
	DA2P	0.3126 (1.908)	39.5396 (41.768)	-0.0249 (0.0669)	0.00505 (0.0100)	-0.1047 (0.087)	0.0184 (0.022)
0.25	LIML	0.0135 (0.138)	0.1677 (0.183)	-0.0031 (0.024)	0.00058 (0.00070)	-0.0157 (0.121)	0.0147 (0.018)
	FIML	0.0136 (0.137)	0.1647 (0.174)	-0.0024 (0.024)	0.00056 (0.00066)	-0.0123 (0.120)	0.0145 (0.017)
	N2P	0.0309 (0.141)	0.3654 (0.274)	-0.0251 (0.023)	0.00117 (0.00141)	-0.1268 (0.118)	0.0299 (0.036)
	DA2P	0.0262 (0.152)	0.4155 (0.353)	-0.0143 (0.114)	0.01307 (0.02691)	-0.1177 (0.152)	0.0367 (0.042)

Table 8
Measures of Collinearity: Designs 3 and 4
True Model = Sample Selection Model
$x_1, x_2 \sim U(0,10)$

$\theta(x_1,x_2)$	Proportion of Uncensored Observations	$r(x,\hat{\lambda})$	Eigenvalue	Condition Index	Variance-decomposition Proportions		
					Intercept	x_2	$\hat{\lambda}$
0	0.75	-0.0155	2.0857	1	0.05006	0.05035	0.0704
			0.7777	1.6382	0.02182	0.03458	0.9168
			2.0857	3.9068*	0.92812	0.91507	0.0128
	0.5	-0.0299	2.1797	1	0.04355	0.04473	0.0753
			0.6883	1.7805	0.02361	0.05545	0.88153
			0.1320	4.0645*	0.93284	0.89982	0.04317
	0.25	-0.0701	2.3917	1	0.03072	0.03481	0.06423
			0.4930	2.2060	0.02250	0.11822	0.80370
			0.1153	4.5570*	0.94678	0.84697	0.13207
0.5	0.75	-0.0155	2.0932	1	0.04917	0.04949	0.07100
			0.7717	1.6473	0.02207	0.03515	0.91554
			0.1352	3.9356*	0.92876	0.91536	0.01346
	0.5	-0.0394	2.1574	1	0.04453	0.04555	0.07429
			0.7112	1.7427	0.02179	0.05129	0.88232
			0.1315	4.0510*	0.93368	0.90316	0.04339
	0.25	-0.0211	2.3796	1	0.03329	0.03687	0.06671
			0.4961	2.1943	0.02793	0.11490	0.83688
			0.1243	4.3765*	0.93878	0.84823	0.09641

Note: A condition index with an asterisk indicates that it is the condition number.

Table 9
Summary Statistics of the T-ratios and Rejection Frequency: Designs [3] and [4]

Design on x_1 and x_2	P_+	Mean	Standard Deviation	Skewness	Kurtosis	Minimum	Maximum	Rejection Frequency
$x_1 \sim U(0,10)$ $x_2 \sim U(0,10)$ $\theta(x_1,x_2) = 0$	0.75	4.474	1.080	0.047	3.027	1.257	6.909	99
	0.5	4.409	1.152	-0.180	3.771	1.140	8.045	98
	0.25	3.812	1.089	-0.003	3.375	0.441	6.440	97
$x_1 \sim U(0,10)$ $x_2 \sim U(0,10)$ $\theta(x_1,x_2) = 0.5$	0.75	4.574	1.145	0.586	3.726	1.902	8.519	99
	0.5	4.372	1.097	0.289	2.812	1.976	7.650	100
	0.25	3.694	1.016	0.353	3.297	1.063	6.668	97

Table 10
Simulation Results based on Design 5
True Model = Two-part Model
$x_1 = x_2 \sim U(0,10)$, $u_4 \sim U(-1.5,1.5)$
(Standard Errors in Parentheses)

Proportion of Uncensored Observations (P_+)	Estimation Method	Mean Prediction Bias (MPB)	Mean Squared Prediction Error (MSPE)	Parameter Bias (PB)	Parameter Squared Error (PSE)	Elasticity Bias (EB)	Elasticity Squared Error (ESE)
0.75	LIML	20.3178 (13.84)	3302.34 (3346.6)	-0.00109 (0.0186)	0.00034 (0.0005)	-0.0025 (0.0862)	0.00735 (0.0099)
	FIML	20.2447 (13.83)	3278.64 (3220.1)	-0.00127 (0.0186)	0.00034 (0.0004)	-0.0030 (0.0860)	0.00733 (0.0100)
	N2P	11.9245 (12.14)	1631.90 (1776.6)	-0.00074 (0.0135)	0.00018 (0.0003)	-0.0057 (0.0774)	0.0060 (0.0090)
	DA2P	11.2889 (13.80)	2033.24 (2963.5)	0.00645 (0.0509)	0.00260 (0.0073)	0.0008 (0.0856)	0.0073 (0.0103)
0.5	LIML	2.3069 (1.391)	41.490 (46.65)	-0.00338 (0.03716)	0.00138 (0.0020)	0.08899 (0.574)	0.3343 (0.507)
	FIML	2.2981 (1.383)	41.073 (45.33)	-0.00385 (0.03741)	0.00140 (0.0020)	0.1118 (0.595)	0.3624 (0.550)
	N2P	1.5183 (1.219)	22.2057 (28.77)	-0.002359 (0.02536)	0.00064 (0.0010)	0.0596 (0.384)	0.1498 (0.245)
	DA2P	1.5243 (1.250)	24.4951 (31.81)	-0.004188 (0.14654)	0.02128 (0.0529)	0.0570 (0.465)	0.2177 (0.385)
0.25	LIML	0.3176 (0.134)	0.8760 (0.940)	-0.0115 (0.198)	0.0390 (0.059)	0.2621 (2.286)	5.2400 (8.719)
	FIML	0.2825 (0.163)	0.7055 (0.648)	-0.0576 (0.192)	0.0398 (0.053)	0.3340 (2.131)	4.6069 (7.332)
	N2P	0.2201 (0.103)	0.3678 (0.298)	-0.0103 (0.051)	0.0027 (0.004)	0.2397 (2.038)	4.1699 (8.208)
	DA2P	0.2211 (0.108)	0.4287 (0.365)	-0.0178 (0.683)	0.4624 (1.074)	0.2253 (2.426)	5.8789 (10.815)

Table 11
Summary Statistics of the T-ratios and Rejection Frequency: Design [5]

Design on x_1 and x_2	P_+	Mean	Standard Deviation	Skewness	Kurtosis	Minimum	Maximum	Rejection Frequency
$x_1 = x_2 \sim$ U(0,10)	0.75	-0.029	1.004	0.165	2.792	-2.063	2.676	4
	0.5	-0.033	0.999	0.276	2.759	-1.924	2.580	5
	0.25	$-6x10^{-5}$	1.179	0.047	3.299	-2.993	3.200	11

Table 12
Summary Statistics of the Condition Numbers

Design on x_1 and x_2	P_+	Mean	Standard Deviation	Skewness	Kurtosis	Minimum	Maximum
$x_1 = x_2 \sim$ U(0,3)	0.75	23.036	2.2704	0.385	3.013	17.63	29.07
	0.5	53.708	4.9941	0.631	3.831	44.80	72.46
	0.25	155.68	17.597	0.453	2.648	119.4	202.5
$x_1 = x_2 \sim$ U(0,10)	0.75	8.2399	0.2920	0.043	2.264	7.542	8.927
	0.5	17.337	0.8825	0.432	2.991	15.59	19.81
	0.25	64.449	5.326	0.203	2.661	51.67	77.92
$x_1 \sim$ U(0,10) $x_2 \sim$ U(0,10) $\theta(x_1,x_2) = 0$	0.75	3.9068	0.0284	0.327	2.903	3.851	3.991
	0.5	4.0645	0.0405	0.341	2.923	3.988	4.200
	0.25	4.5570	0.0886	0.297	2.850	4.360	4.784
$x_1 \sim$ U(0,10) $x_2 \sim$ U(0,10) $\theta(x_1,x_2) = 0.5$	0.75	3.9356	0.0271	0.246	3.784	3.862	4.015
	0.5	4.051	0.0398	0.103	2.756	3.959	4.157
	0.25	4.3765	0.0824	0.581	3.885	4.205	4.685

References

Amemiya, Takeshi. Advanced Econometrics. Cambridge: Harvard University Press, 1985.

Belsley, David A.; Kuh, Edwin; and Welsch, Roy E. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: John Wiley, 1980.

Blau, David, and Robins, Philip. "Job Search Outcomes for the Employed and Unemployed," Journal of Political Economy 98 (1990): 637-655.

Cragg, John G. "Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods," Econometrica 39 (1971): 829-844.

Duan, Naihua. "Smearing Estimate: A Nonparametric Retransformation Method," Journal of the American Statistical Association 78 (1983): 605-610.

Duan, Naihua; Manning, Willard G.; Morris, Carl N.; and Newhouse, Joseph P. "A Comparison of Alternative Models for the Demand for Medical Care," Journal of Business and Economics Statistics 1 (1983): 115-126.

Duan, Naihua; Manning, Willard G.; Morris, Carl N.; and Newhouse, Joseph P. "Choosing Between the Sample-Selection Model and the Multi-Part Model," Journal of Business and Economics Statistics 2 (1984): 283-289.

Duan, Naihua; Manning, Willard G.; Morris, Carl N.; and Newhouse, Joseph P. "Comments on Selectivity Bias," Advances in Health Economics and Health Services Research 6 (1985): 19-24.

Dudley, Leonard. and Montmarquette, Claude. "A Model of the Supply of Bilateral Foreign Aid." American Economic Review 66 (1976): 132-142.

Duncan, Gregory M. "Sample Selectivity as a Proxy Variable Problem: On the Use and Misuse of Gaussian Selectivity Corrections," Research in Labor Economics 2 (1983): 333-345.

Fernandez, Luis. "Non-parametric Maximum Likelihood Estimation of Censored Regression Models," Journal of Econometrics 32 (1986): 35-57.

Goldberger, Arthur. Econometric Theory. New York: John Wiley, 1964.

Goldberger, Arthur. "Abnormal Selection Bias," in: S. Karlin, T. Amemiya, and L. Goodman, eds., Studies in Econometrics, Time Series, and Multivariate Statistics. New York: Academic Press, 1983.

Greene, William. Econometric Analysis. New York: Macmillan, 1990.

Grossman, Michael, and Joyce, Theodore. "Unobservables, Pregnancy Resolutions, and Birth Weight Production Functions in New York City," Journal of Political Economy 98 (1990): 983-1007.

Hay, Joel W., and Olsen, Randall J. "Let Them Eat Cake: A Note on Comparing Alternative Models of the Demand for Medical Care," Journal of Business and Economic Statistics 2 (1984): 279-282.

Hay, Joel W.; Leu, Robert; and Rohrer, Paul. "Ordinary Least Squares and Sample-Selection Models of Health-care Demand" Journal of Business and Economics Statistics 5 (1987): 499-506.

Heckman, James. "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models," Annals of Economic and Social Measurement 5 (1976): 475-592.

Heckman, James. "Sample Selection Bias as a Specification Error," Econometrica 47 (1979), 153-161.

Kostiuk, Peter. "Compensating Differentials for Shift Work," Journal of Political Economy 98 (1990): 1054-1075.

Leung, Siu Fai, and Yu, Shihti. "Collinearity and Heckman's Two-step Estimators: A Critical Review of Paarsch's Monte Carlo Study," working paper, University of Rochester, October 1992.

Maddala, G. S. Limited-dependent and Qualitative Variables in Econometrics. New York: Cambridge University Press, 1983.

Maddala, G. S. (1985a) "A Survey of the Literature on Selectivity Bias as it Pertains to Health Care Markets," Advances in Health Economics and Health Services Research 6 (1985): 3-18.

Maddala, G.S. (1985b) "Further Comments on Selectivity Bias," Advances in Health Economics and Health Services Research 6 (1985): 25-26.

Manning, Willard G.; Bailit, H. L.; Benjamin, B.; and Newhouse, J. "The Demand for Dental Care: Evidence from a Randomized Trial in Health Insurance," Journal of the American Dental Association 110 (1985): 895-902.

Manning, Willard G.; Blumberg, Linda; and Moulton, Lawrence H. "The Demand for Alcohol: The Differential Response to Price," unpublished manuscript, University of Minnesota, 1991.

Manning, Willard G.; Duan, Naihua; and Rogers, W. H. "Monte Carlo Evidence on the Choice Between Sample Selection and Two-Part Models," Journal of Econometrics 35 (1987): 59-82.

Manning, Willard G.; Leibowitz, A.; Goldberg, G. A.; Rogers, W. H.; and Newhouse, J. "A Controlled Trial of the Effect of a Prepaid Group Practice on the Use of Services," New England Journal of Medicine 310 (1984): 1505-1510.

Manning, Willard G.; Morris, C. N.; and Newhouse, J. P. et al. "A Two-part Model of the Demand for Medical Care: Preliminary Results from the Health Insurance Experiment," in: J. van der Gaag and M. Perlman, eds., Health, Economics, and Health Economics. Amsterdam: North Holland, 1981.

Melino, Angelo. "Testing for Sample Selection Bias," Review of Economic Studies 49 (1982): 151-153.

McLaughlin, Kenneth J. "A Theory of Quits and Layoffs with Efficient Turnover," Journal of Political Economy 99 (1991): 1-29.

Moon, Choon-Geol. "A Monte Carlo Comparison of Semiparametric Tobit Estimators," Journal of Applied Econometrics 4 (1989): 361-382.

Nawata, Kazumitsu. "Robust Estimation Based on Grouped-adjusted Data in Censored Regression Models," Journal of Econometrics 43 (1990):337-362.

Nelson, Forrest D. "Efficiency of the Two-Step Estimator for Models with Endogenous Sample Selection," Journal of Econometrics 24 (1984): 181-196.

Newhouse, J. P.; Manning, W. G.; and Morris, C. N. et al. "Some Interim Results from a Controlled Trial of Cost Sharing in Health Insurance," New England Journal of Medicine 305 (1981): 1501 -1507.

Paarsch, Harry. "A Monte Carlo Comparison of Estimators for Censored Regression Models," Journal of Econometrics 24 (1984): 197-213.

Peters, Simon and Smith, Richard J. "Distributional Specification Tests Against Semiparametric Alternatives," Journal of Econometrics 47 (1991): 175-194.

Powell, James. "Symmetrically Trimmed Least Squares Estimation for Tobit Models," Econometrica 54 (1986): 1435-1460.

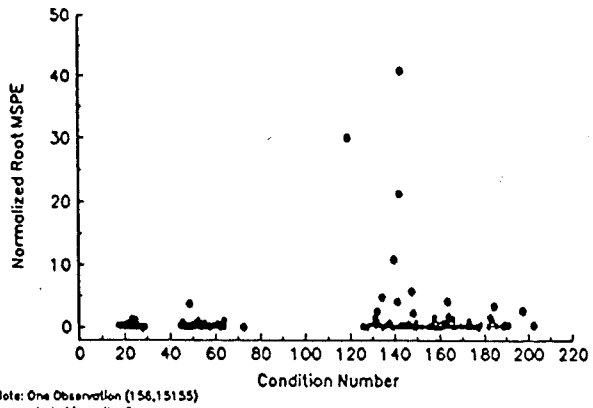
Stapleton, David C., and Young, Douglas J. "Censored Normal Regression with Measurement Error on the Dependent Variable," Econometrica 52 (1984): 737-760.

Stewart, G. W. "Collinearity and Least Squares Regression," (with discussions) Statistical Science 2 (1987): 68-100.

van de Ven, Wynand P. M. M., and van Praag, Bernard M. S. (1981a) "Risk Aversion and Deductibles in Private Health Insurance: Application of an Adjusted Tobit Model to Family Health Care Expenditures," in: J. van der Gaag and M. Perlman, eds., Health, Economics, and Health Economics. Amsterdam: North Holland, 1981.

van de Ven, Wynand P. M. M., and van Praag, Bernard M. S. (1981b) "The Demand for Deductibles in Private Health Insurance: A Probit Model with Sample Selection," Journal of Econometrics 17 (1981): 229-252.

Figure 1
Normolized Root MSPE versus Condition Number: Design (1)



Note: One Observation (156,15155) was excluded from the figure

Figure 2
Normolized Root MSPE versus Condition Number: Design (2)

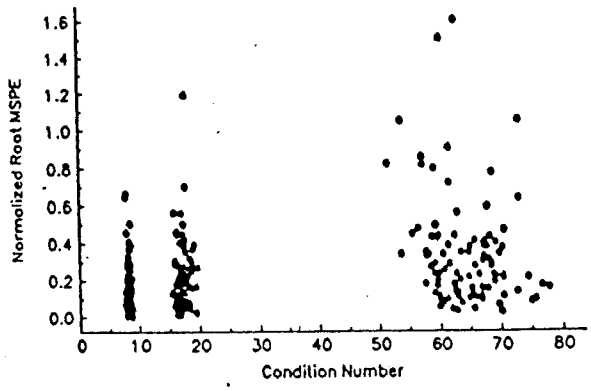


Figure 3
T-ratio versus Condition Number: Design (1)

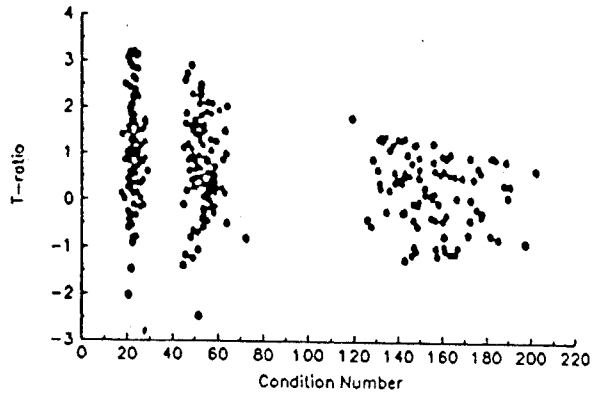


Figure 4
T-ratio versus Condition Number: Design (2)