ELSEVIER

Journal of Econometrics 72 (1996) 197–229

JOURNAL OF Econometrics

On the choice between sample selection and two-part models

Siu Fai Leung*[a,b], Shihti Yu[c]

[a]*Department of Economics, University of Rochester, Rochester, NY 14627, USA*
[b]*Department of Economics, Hong Kong University of Science & Technology, Kowloon, Hong Kong*
[c]*National Chung Cheng University, Chia-Yi 621, Taiwan*

(Received August 1993; final version received November 1994)

Abstract

This paper resolves 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. In particular, the sample selection model is susceptible to collinearity problems and a t-test can be used to distinguish between the two models as long as there are no collinearity problems. As an example, we employ Mroz's (1987) labor supply data to illustrate how his tests for selectivity bias might have been affected by collinearity problems.

Key words: Sample selection model; Two-part model; Censored regression model; Collinearity; Health economics; Monte Carlo simulation
JEL classification: C15; C24; I10

* Corresponding author.

We thank Bruce Hansen, Adrian Pagan, Charles Phelps, William Rogers, an Editor, and an Associate Editor for comments and suggestions, as well as Shao-Chung Chang for research assistance. This research was supported by National Institute on Alcohol Abuse and Alcoholism grant AA08393. An earlier version of the paper appeared as a Rochester Center for Economic Research Working Paper No. 337, December 1992.

SSDI 0304076940172O K

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 empirical studies that involve censoring and selection bias. In a series of articles, 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 nonselection 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 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 flawed. Perplexed by these exchanges, Maddala (1985a, b) tries to sort out the various 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 vigorous debates have yet to be settled.

The origin of the two-part model can be traced back at least 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. Researchers associated with the Rand Corporation in the seventies and the early eighties made extensive use of the two-part model in their empirical studies in health economics (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 studies (e.g., Dudley and Montmarquette, 1976; Grossman and Joyce, 1990; McLaughlin, 1991). 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 use of two-part and 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 of sample selection and two-part models focus primarily on theoretical issues, recent investigations have turned to Monte Carlo simulation experiments. Using a simulated data set from the 1981 Population of Switzerland Survey, Hay, Leu, and Rohrer (1987) 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 it in terms of parameter squared error. Although Hay strongly criticizes 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. (1983) 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 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 sample selection model is the true one. The sample selection model performs better than the two-part model only when there are exclusion restrictions. They 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' (p. 80) and that '[g]iven 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: 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 will cast doubts on the reliability of all empirical findings based on the sample selection model in the literature of the past two decades. This is undoubtedly an important issue that deserves further investigation.

In this paper, we conduct a different set of Monte Carlo experiments to compare the performance of sample selection and 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 of the merits of the two 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. 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.[1] The inverse Mills' ratio,

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

which is a function of the regressor, turns out to be highly correlated with the regressor because the range [0, 3] is far too narrow. As a result of this high collinearity, the Heckman two-step estimates 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 employ a t-test 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-test fails to reject the two-part model when there are collinearity problems. When collinearity is lessened, the t-test strongly rejects the two-part model. Based on these results, we can therefore explain Manning, Duan, and Rogers' striking finding that the two-part model outperforms the sample selection model even when the latter is the true model. Consequently, their favorable results on the merits of the two-part model are not robust because their simulation setups are biased against the sample selection model. In the process of our inquiry, we discover, to our surprise, that such inadvertent bias in the data generating process is actually very prevalent in the literature. We will show that a number of widely cited Monte Carlo studies of the sample selection model are also marred by the same design problem.

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 comparison. We conduct a series of experiments with and without exclusion restrictions, using different true models and various 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 of Manning, Duan, and Rogers. When the sample selection model is the true model, it performs 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 it 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 model. 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.[2] We show that models with few 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 problem, we suggest that applied researchers should check for high collinearity in the level equation whenever they implement the two-step procedure. After investigating several different measures of collinearity, we believe that the condition number is more accurate and dependable than the other measures. As an example, we employ Mroz's (1987) data on female labor supply to illustrate the importance of checking for collinearity problems.

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

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,b) adjusted Tobit model [Type 2 Tobit in Amemiya's (1985) classification system]:

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

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

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

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

where $\underset{\sim}{x}_1 = (1, x_{12}, \ldots, x_{1J})$, $\underset{\sim}{x}_2 = (1, x_{22}, \ldots, x_{2K})$, $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_J)'$, $\beta = (\beta_1, \beta_2, \ldots, \beta_K)'$, and individual subscripts are suppressed for simplicity. The error terms u_1 and u_2 are assumed to be independent of the regressors $\underset{\sim}{x}_1$ and $\underset{\sim}{x}_2$, and

[2] See, e.g., Heckman (1976), Nelson (1984), Manning, Duan, and Rogers (1987), and Nawata (1993, 1994). Our investigation is more thorough and our results are also notably different.

follow a bivariate normal distribution:

$$(u_1, u_2) \sim \mathrm{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\sigma\\ \rho\sigma & \sigma^2\end{pmatrix}\right),$$

where the covariance matrix has been normalized for identification reasons.

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, Eq. (2). The two most popular estimation methods for the model are maximum likelihood (full-information maximum-likelihood, FIML) and Heckman's (1976, 1979) two-step procedure (limited-information maximum-likelihood, LIML). For the LIML, the first step is to obtain $\hat{\alpha}$ from (1) by means of maximum likelihood. The estimates of β and $\rho\sigma$ are then obtained by running an ordinary least squares regression on the model $\ln(y) = \underset{\sim}{x}_2\beta + \rho\sigma\hat{\lambda} + \varepsilon$, using all the observations in which $I > 0$, where $\hat{\lambda} = \lambda(\underset{\sim}{x}_1\hat{\alpha}) = \phi(\underset{\sim}{x}_1\hat{\alpha})/\Phi(\underset{\sim}{x}_1\hat{\alpha})$ is the estimated inverse Mills' ratio, $\phi(\cdot)$ and $\Phi(\cdot)$ are the p.d.f. and c.d.f. of the standard normal distribution, and $E(\varepsilon | I > 0) = 0$. The estimates for ρ and σ are then calculated as in Manning, Duan, and Rogers. For the FIML, the likelihood function is given by $L = \Pi_0 \times [1 - \Phi(\underset{\sim}{x}_1\alpha)] \cdot \Pi_1 \Phi((\underset{\sim}{x}_1\alpha + \rho(m - \underset{\sim}{x}_2\beta)/\sigma)(1 - \rho^2)^{-1/2})\phi((m - \underset{\sim}{x}_2\beta)/\sigma)/\sigma$, where Π_0 and Π_1 denote the products over the censored and the uncensored samples, respectively.

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:

$$I = \underset{\sim}{x}_1\alpha + u_3, \qquad u_3 \sim \mathrm{N}(0, 1), \tag{3}$$

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

$$\ln(y | I > 0) = \underset{\sim}{x}_2\beta + u_4, \tag{4}$$

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. (4) implies that $E[\ln(y) | I > 0] = \underset{\sim}{x}_2\beta$, as opposed to $E[\ln(y) | I > 0] = \underset{\sim}{x}_2\beta + \rho\sigma\lambda(\underset{\sim}{x}_1\alpha)$ in the sample selection model, where $\lambda(\underset{\sim}{x}_1\alpha) = \phi(\underset{\sim}{x}_1\alpha)/\Phi(\underset{\sim}{x}_1\alpha)$ 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. Hence, Eq. (2) is an unconditional equation while Eq. (4) is a conditional one. As a result, the interpretation of β is different between the two models.

As emphasized in Duan et al. (1983, 1984, 1985), the two-part model was designed to predict the actual outcome y. Based on the observation that

$E(y) = P(y > 0)E(y \mid y > 0)$ (assuming that $y \geqslant 0$), Duan et al. (1983, 1984) employ Eq. (3) to model $P(y > 0)$ and Eq. (4) to model $E(y \mid y > 0)$. Clearly, one can also employ the sample selection model to predict y by using (1) to model $P(y > 0)$ and (2) to obtain $E(y \mid y > 0)$. However, Duan et al. (1984, pp. 286–287) maintain that the specification of the two-part model is more direct and parsimonious than the sample selection model because the latter obtains the conditional mean $E(y \mid y > 0)$ via the unconditional equation, Eq. (2), which only introduces unnecessary complications (such as distributional assumptions on u_2, numerical and statistical problems) into the analysis. In addition, they contend that there is no direct interest in the potential outcome and the parameters of the unconditional equation in their analysis of the medical expense data. Their disinterest in these parameters contrasts sharply with the views of Heckman (1990) who argues that identifying the potential outcome is economically more interesting.

A fundamental distinction between the two models lies in the assumptions on the error terms u_2 and u_4. The sample selection model assumes that $E(u_2) = 0$ [hence $E(u_2 \mid I > 0) = \rho\sigma\lambda(\underset{\sim}{x}_1\alpha)$], whereas $E(u_4 \mid I > 0) = 0$ is assumed in the two-part model. This difference is the core of the debate between Hay and Olsen (1984) and Duan et al. (1984). Hay and Olsen (1984) argue that for $E(u_4 \mid I > 0) = 0$ to be consistent with Eqs. (3) and (4), the two-part model must impose some unusual assumptions on the distributions of the error terms u_3 and u_4. They also contend that the two-part model can be nested in the sample selection model. Duan et al. (1984) counter their criticisms by constructing an example in which both the joint and the marginal distributions of u_3 and u_4 are consistent with (3) and (4). In particular, they 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. They prove convincingly that the two-part model is not nested in the sample selection model.

As (3) and (4) are conditionally independent, the estimation procedure for this model is straightforward. For the probit model (3), α is estimated by maximum likelihood. For the linear model (4), β is estimated by regressing $\ln(y)$ on $\underset{\sim}{x}_2$ using all the observations in which $I > 0$. Following Manning, Duan, and Rogers, we call it 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 (4) by adding higher-order terms to the right-hand side of (4) and by looking for heteroskedasticity in the observed residuals. They use Mallows' (1973) C_p rule to determine whether higher-order terms should be included into the model. For simplicity, they consider only the second-order (squared) terms of $\underset{\sim}{x}_2$. 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 every t-statistic of the second-order terms is less than 1.414.

3. Experimental designs, performance criteria, and collinearity measures

3.1. Experimental designs

As in Manning, Duan, and Rogers, we focus on the case in which there is only one regressor in the choice and the level equations, i.e., $J = K = 2$. We set $\alpha_2 = \beta_2 = 1$ and $\alpha_1 = \beta_1$. For brevity, let $x_1 = x_{12}$, $x_2 = x_{22}$, and $\text{corr}(x_1, x_2)$ be the population correlation coefficient of x_1 and x_2, then $\underset{\sim}{x}_1 = (1, x_1)$ and $\underset{\sim}{x}_2 = (1, x_2)$. Our simulations are based on five different designs, which are listed in Table 1.[3]

We use the GAUSS program to generate the random numbers, create 1,000 observations for each sample, and perform 100 repetitions for each experiment. 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.[4] For each experiment, four estimators are considered: LIML and FIML estimators for the sample selection model and N2P and DA2P estimators for the two-part model.[5]

3.2. Performance criteria

We assess the performance of the estimators in six different ways. The first two performance criteria are mean prediction bias (*MPB*) and mean squared

[3] The regressors in design [4] are generated in the following way. First, we generate two random variables ω_1 and ω_2 from a bivariate normal distribution with zero means, unit variances, and a correlation coefficient of 0.526. Then we obtain x_1 and x_2 by setting $\Psi(x_1) = \Phi(\omega_1)$ and $\Psi(x_2) = \Phi(\omega_2)$, 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 (range [0, 10]) with a sample correlation coefficient of 0.5. Hence, precisely speaking, 0.5 is the sample, not the population, correlation coefficient of x_1 and x_2. In design [5], 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 $I_i = \underset{\sim}{x}_{1i}\alpha + u_{3i} > 0$. If $I_i > 0$, we set $G(u_{4i}) = \Phi(z_i)$, where $G(.)$ denotes 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.

[4] 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 percent 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 percent censoring, respectively.

[5] 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 '[l]ess 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 solely for the two-part model.

Table 1
Design of the experiments; $\alpha_1 = \beta_1$, $\alpha_2 = \beta_2 = 1$

Design	True model	Regressors	Error terms
[1]	Sample selection model	$x_1 = x_2 = x$, $x \sim U(0, 3)$	
[2]	Sample selection model	$x_1 = x_2 = x$, $x \sim U(0, 10)$	$(u_1, u_2) \sim$
[3]	Sample selection model	$x_1 \sim U(0, 10)$, $x_2 \sim U(0, 10)$, $\text{corr}(x_1, x_2) = 0$	$\text{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0.5\\0.5 & 1\end{pmatrix}\right)$
[4]	Sample selection model	$x_1 \sim U(0, 10)$, $x_2 \sim U(0, 10)$, $\text{corr}(x_1, x_2) = 0.5$	
[5]	Two-part model	$x_1 = x_2 = x$, $x \sim U(0, 10)$	$u_3 \sim \text{N}(0, 1)$, $u_4 \mid I > 0 \sim U(-1.5, 1.5)$

prediction error (*MSPE*):

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

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

where $n = 1{,}000$. Both criteria are used in Hay, Leu, and Rohrer and Manning, Duan, and Rogers. As emphasized in van de Ven and van Praag (1981a, b) and Duan et al. (1983, 1984, 1985), the quantity $E(y)$ plays an important role in their empirical analyses. For the sample selection model, it is easy to verify with some calculations that

$$E(y) = [\Phi(\underset{\sim}{x}_1\alpha + \rho\sigma)]\exp(\underset{\sim}{x}_2\beta + \sigma^2/2). \tag{5}$$

The LIML and FIML estimates of $E(y)$ are obtained by plugging the LIML and FIML estimates of α, β, ρ, and σ into (5), respectively. For the two-part model, the expression for $E(y)$ is simply

$$E(y) = [\Phi(\underset{\sim}{x}_1\alpha)]\,[\exp(\underset{\sim}{x}_2\beta)]\,E[\exp(u_4)]. \tag{6}$$

If u_4, given $I > 0$, is assumed to be $U(a, b)$ (as in design [5]), then $E[\exp(u_4)] = [\exp(b) - \exp(a)]/(b - a)$. Instead of assuming a specific probability distribution for u_4, Duan et al. (1983, 1984, 1985) and Manning, Duan, and Rogers advocate Duan's (1983) distribution-free smearing estimator for the retransformation factor $E[\exp(u_4)]$ in (6). Let q be the number of observations in which $I > 0$ and $\hat{u}_{4i}$ denote the ordinary least squares residual from (4), then Duan's (1983) smearing estimate is given by $s = \sum_{i=1}^{q} [\exp(\hat{u}_{4i})]/q$. Duan (1983)

demonstrates that s is a consistent nonparametric estimator of $E[\exp(u_4)]$. Hence, the N2P and DA2P estimates of $E(y)$ are obtained by substituting the N2P and DA2P estimates of α, β, and the smearing estimate of $E[\exp(u_4)]$ into (6), respectively.

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

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

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 (1987, p. 60) maintain that the parameter β in the sample selection and the two-part models is not comparable, we also examine the elasticity bias (*EB*) and the elasticity squared error (*ESE*):

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

where $\eta = [\partial E(y)/\partial z][z/E(y)]$ for some variable z in $\underset{\sim}{x}_2$. 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 (1995) recent study on the demand for alcoholic beverages. The expression for η depends on the assumptions on x_1 and x_2. For the sample selection model, if $z = x_1 = x_2$, then (5) implies that

$$\eta = [\alpha_2 \lambda(\underset{\sim}{x}_1 \alpha + \rho\sigma) + \beta_2] z. \tag{7}$$

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

$$\eta = \beta_2 z. \tag{8}$$

For the two-part model, if $z = x_1 = x_2$, then (6) implies that

$$\eta = [\alpha_2 \lambda(\underset{\sim}{x}_1 \alpha) + \beta_2] z, \tag{9}$$

regardless of the distributional assumption on u_4. The expression for η is the same as (8) 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 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 employ 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. As we recognize that there may not be a perfect diagnostic procedure for collinearity, we employ several measures to detect the problem. The first 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 widely used as a 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 $[VIF = (1 - R^2)^{-1}]$, another suggested measure of collinearity, has a one-to-one correspondence with $[r(x_2, \hat{\lambda})]^2$.

The second diagnostic tool we employ 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 $X'X$. Since the eigenvalues depend 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. In general, the higher the condition number, the more likely that there are 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.[6]

4. Simulation results

4.1. *True model = Sample selection model with* $x_1 = x_2$

As a basis of comparison, we first ran 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 given in Table 2. 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.

The second experiment is based on design [2] and the results are reported in Table 3. Compared to Table 2, Table 3 gives a completely different picture. The

[6] 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 (variance-decomposition proportions). Because of space limitations, we do not report the results here; see Leung and Yu (1992) for details.

Table 2
Simulation results based on design [1] (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	1.7030	− 0.0082	0.0326	0.0464	0.0203
		(0.062)	(0.568)	(0.018)	(0.005)	(0.014)	(0.005)
	FIML	0.0979	0.4570	− 0.0419	0.0209	0.0132	0.0079
		(0.037)	(0.068)	(0.014)	(0.004)	(0.009)	(0.0014)
	N2P	− 0.0169	0.4220	− 0.1714	0.0313	− 0.0214	0.0059
		(0.034)	(0.040)	(0.004)	(0.002)	(0.007)	(0.0007)
	DA2P	0.0103	0.4773	− 0.3083	0.1406	− 0.0363	0.0066
		(0.035)	(0.058)	(0.021)	(0.017)	(0.007)	(0.0008)
0.5	LIML	0.1420	0.8286	− 0.0099	0.1451	0.0961	0.0497
		(0.040)	(0.529)	(0.038)	(0.024)	(0.020)	(0.015)
	FIML	0.0564	0.2151	− 0.1125	0.0805	0.0167	0.0511
		(0.037)	(0.118)	(0.026)	(0.014)	(0.023)	(0.029)
	N2P	0.0065	0.0652	− 0.2784	0.0806	0.0351	0.0164
		(0.014)	(0.007)	(0.006)	(0.003)	(0.012)	(0.002)
	DA2P	0.0113	0.0750	− 0.4283	0.2582	− 0.0004	0.0138
		(0.014)	(0.010)	(0.028)	(0.031)	(0.012)	(0.002)
0.25	LIML	30.025	6.9×10^5	− 0.0112	1.3618	0.2907	0.5351
		(29.65)	(6.9×10^5)	(0.117)	(0.249)	(0.068)	(0.249)
	FIML	0.0116	0.0569	− 0.1910	0.4022	0.0784	0.3204
		(0.015)	(0.020)	(0.061)	(0.188)	(0.056)	(0.235)
	N2P	0.0082	0.0112	− 0.3539	0.1352	0.0371	0.0606
		(0.006)	(0.002)	(0.010)	(0.008)	(0.025)	(0.01)
	DA2P	0.0084	0.0136	− 0.4269	0.3592	0.0112	0.0599
		(0.006)	(0.002)	(0.042)	(0.079)	(0.025)	(0.01)

For $P_+ = 0.25$, the FIML failed 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 3
Simulation results based on design [2] (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	4202.4	0.00057	0.00057	0.0055	0.0130
		(2.528)	(670.99)	(0.002)	(0.0001)	(0.011)	(0.0018)
	FIML	6.4979	3999.2	0.00013	0.00048	0.0033	0.01128
		(2.484)	(606.46)	(0.002)	(0.0001)	(0.011)	(0.0015)
	N2P	− 14.73	5174.5	− 0.0444	0.00225	− 0.1562	0.0332
		(2.242)	(557.51)	(0.002)	(0.0002)	(0.009)	(0.0034)
	DA2P	11.376	8985.8	− 0.2759	0.08605	− 0.3576	0.1428
		(2.833)	(1577.3)	(0.010)	(0.0052)	(0.012)	(0.0084)
0.5	LIML	0.6207	44.031	0.00167	0.00240	0.0872	0.2772
		(0.238)	(9.747)	(0.005)	(0.0004)	(0.052)	(0.042)
	FIML	0.4999	39.933	− 0.0024	0.00199	0.1114	0.2167
		(0.233)	(7.182)	(0.005)	(0.0003)	(0.045)	(0.034)
	N2P	− 1.153	50.986	− 0.0910	0.00918	1.0639	1.3266
		(0.204)	(4.978)	(0.003)	(0.0005)	(0.044)	(0.111)
	DA2P	0.2696	52.055	− 0.6280	0.45681	0.2496	0.4174
		(0.236)	(9.610)	(0.025)	(0.0276)	(0.06)	(0.064)
0.25	LIML	0.0793	0.6736	− 0.0034	0.0293	0.3277	5.5860
		(0.025)	(0.137)	(0.017)	(0.0039)	(0.235)	(1.447)
	FIML	0.0415	0.3762	− 0.0353	0.0208	0.4789	5.2457
		(0.021)	(0.078)	(0.014)	(0.0031)	(0.225)	(1.305)
	N2P	− 0.0269	0.3731	− 0.2143	0.0493	1.4662	6.4335
		(0.02)	(0.040)	(0.006)	(0.0026)	(0.208)	(1.251)
	DA2P	− 0.0070	0.3581	− 0.9020	1.6406	0.1147	7.5091
		(0.020)	(0.041)	(0.091)	(0.235)	(0.275)	(1.287)

Table 4
Measures of collinearity

Design	Proportion of uncensored observations	$r(x_2, \hat{\lambda})$	Condition number
[1]	0.75	−0.9573	23.036
	0.5	−0.9844	53.708
	0.25	−0.9926	155.68
[2]	0.75	−0.6564	8.2399
	0.5	−0.7963	17.337
	0.25	−0.9434	64.449
[3]	0.75	−0.0155	3.9068
	0.5	−0.0299	4.0645
	0.25	−0.0701	4.5570
[4]	0.75	−0.0155	3.9356
	0.5	−0.0394	4.0510
	0.25	−0.0211	4.3765

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 that will be explained below.

The striking differences between Tables 2 and 3 can be explained by collinearity. Table 4 reports two 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. 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. When x is drawn from $U(0, 10)$, however, the absolute values of the sample correlation coefficients and the condition numbers for both $P_+ = 0.5$ and 0.75 are substantially lowered, and the LIML estimators behave much better than those of the two-part models. For the case $P_+ = 0.25$, the condition number reaches 64.45 and collinearity problems reappear again. This explains why the LIML estimator has a larger *MSPE* than the two-part model in this case.

Tables 2–4 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 5 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

Table 5
Effect of censoring on collinearity

Design	P_+	$r(x_2, \hat{\lambda})$	Condition number
[1]	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
[2]	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

high collinearity in the $U(0, 3)$ case. For the $U(0, 10)$ case, at least 50 percent uncensored observations are required.

A graph of the inverse Mills' ratio would help to explain the source of the collinearity problems. Fig. 1 plots the inverse Mills' ratio $\lambda(z)$ against z. The curve $\lambda(z)$ is virtually a straight line for $z \leqslant 0$,[7] takes a sharp turn at around $z = 2$, and is essentially flat for $z \geqslant 3$. Hence, if the range of z is small and z is less than 2, then $\lambda(z)$ and z will be highly correlated. Taking 100 points evenly spaced over each range of z, Table 6 gives the sample correlation coefficient of $\lambda(z)$ and z for ten different ranges of z. When the range of z is less than or equal to $[0, 3]$, the correlation between $\lambda(z)$ and z is indeed very high.

Censoring further raises the correlation because the argument of the inverse Mills' ratio decreases with censoring. To see this, notice from the Monte Carlo design that λ is a function of $(\alpha_1 + \alpha_2 x_1)$ and a higher degree of censoring is

[7] This can be seen by fitting a straight line to $\lambda(z)$ for $z \in [-10, 0]$. If one regresses $\lambda(z)$ against an intercept and z using 101 observations equally spaced on $[-10, 0]$, one obtains $\hat{\lambda}(z) = 0.5176 - 0.9476z$. The t-ratios for the intercept and the slope coefficients are 32.58 and − 345.23, respectively. The adjusted R-squared of the regression is 0.9992, which indicates that it is almost a perfect fit.

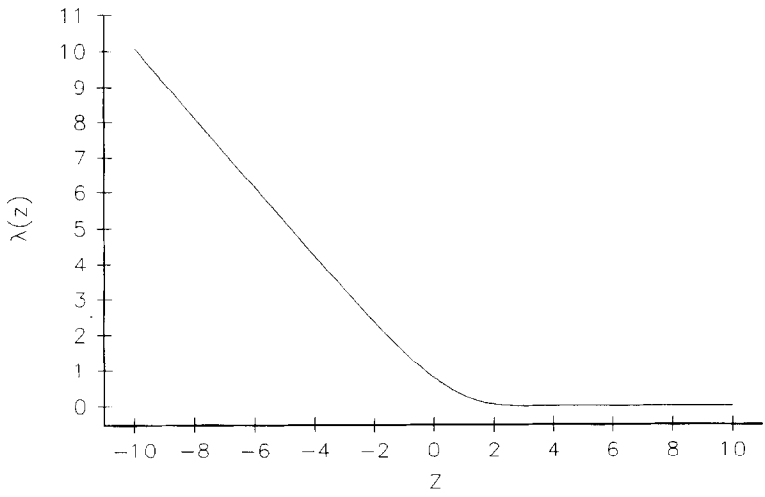

Fig. 1. The inverse Mills' ratio: $\lambda(z) = \phi(z)/\Phi(z)$.

achieved by lowering α_1. Since $\alpha_2 = 1$ and $x_1 \sim U(0, \xi)$ ($\xi = 3$ or 10), the range of $\alpha_1 + \alpha_2 x_1$ is $[\alpha_1, \alpha_1 + \xi]$. As $\alpha_1 < 0$ (see Footnote 5), a higher degree of censoring is achieved by a larger value of $|\alpha_1|$. Given ξ, increases in censoring will move the interval $[\alpha_1, \alpha_1 + \xi]$ further to the left and closer to the region where $\lambda(z)$ is approximately linear in z (see Fig. 1). For example, when $\xi = 3$, the intervals for $[\alpha_1, \alpha_1 + \xi]$ are $[-0.57, 2.43]$, $[-1.5, 1.5]$, and $[-2.43, 0.57]$ for 25, 50, and 75 percent of censoring, respectively. The correlation coefficients between $\lambda(z)$ and z over these three ranges are -0.9603, -0.9872, and -0.9975, respectively.[8] When $\xi = 10$, the intervals are $[-2.5, 7.5]$, $[-5, 5]$, and $[-7.5, 2.5]$, and the corresponding correlation coefficients between $\lambda(z)$ and z are -0.6766, -0.8090, and -0.9523, for 25, 50, and 75 percent of censoring, respectively. Hence, more censoring increases the correlation between $\lambda(\alpha_1 + \alpha_2 x_1)$ and $(\alpha_1 + \alpha_2 x_1)$ because it lowers the values of $(\alpha_1 + \alpha_2 x_1)$. Thus, the small range of x_1 explains why there are serious collinearity problems for LIML when x_1 is drawn from $U(0, 3)$, and the high degree of censoring explains why collinearity problems for LIML appear in the 75 percent censoring case even when x_1 is drawn from $U(0, 10)$. These results are all consistent with those discussed earlier.

The poor performance of the LIML estimator stems from the highly conditioned moment matrix in the second step of the two-step procedure. Although

[8] In calculating these coefficients, we have taken into account that lower values of $(\alpha_1 + \alpha_2 x_1)$ are less likely to be included into the sample because of the selection rule that $\alpha_1 + \alpha_2 x_i + u_i > 0$ [see Eq. (1)].

Table 6
Correlation between $\lambda(z)$ and z

Range of z	[0, 1]	[0, 2]	[0, 3]	[0, 4]	[0, 5]	[0, 6]	[0, 7]	[0, 8]	[0, 9]	[0, 10]
$r(\lambda(z), z)$	−0.998	−0.981	−0.938	−0.879	−0.821	−0.770	−0.727	−0.690	−0.657	−0.629

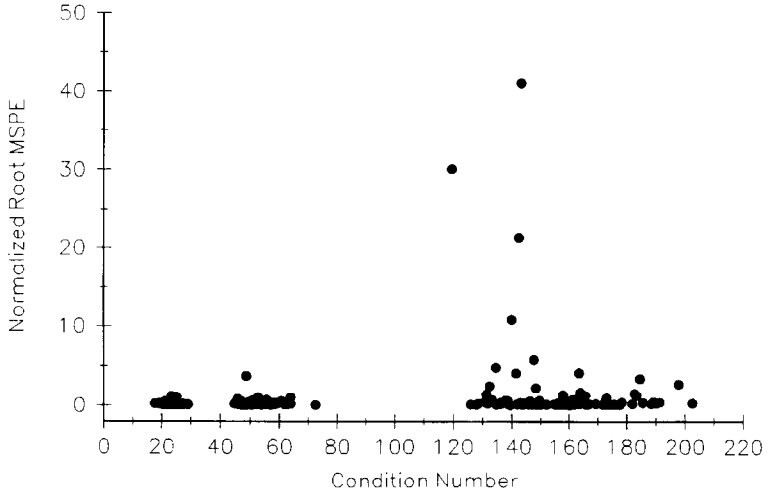

Fig. 2. Normalized root *MSPE* versus condition number: Design [1].

the FIML does not appear to depend on this moment matrix, Tables 2 and 3 reveal that high collinearity also impairs the FIML estimator. In most cases, the N2P estimator dominates the FIML estimator when collinearity is high. Our results thus do not support Nelson's (1984) finding that collinearity has relatively little effect on the FIML estimator and his conclusion that the FIML estimator should be used when collinearity is high. We believe that the N2P estimator may be better than the FIML estimator in these circumstances.

To compare the *MSPE*s across the three values of P_+ in Table 2, we normalize the *MSPE* for each P_+ by dividing the square root of *MSPE* by $E(y)$. Fig. 2 plots the normalized root mean squared prediction error against the condition number. There are three distinct clusters of data points. From left to right, the clusters refer to $P_+ = 0.75$, 0.5, and 0.25, respectively. The figure reveals that the normalized root *MSPE*s are roughly the same for $P_+ = 0.75$ and 0.5, but the errors increase considerably when $P_+ = 0.25$.[9] By normalizing

[9] A huge outlier (156, 15155) was deleted from the figure because it would dramatically change the scale of the diagram.

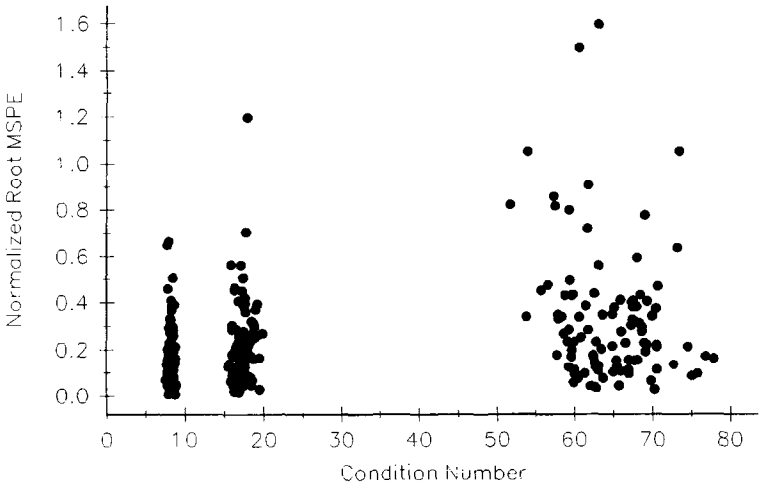

Fig. 3. Normalized root MSPE versus condition number: Design [2].

the *MSPE*s from Table 3, a similar diagram is obtained, as shown in Fig. 3. It again indicates that the normalized root *MSPE*s are roughly the same for $P_+ = 0.75$ and 0.5, and that the errors are in general higher when $P_+ = 0.25$. From these two figures, one can see that the normalized root *MSPE* increases appreciably only when the condition number gets large.

Table 7 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 (Heckman, 1979; Melino, 1982). Using the 5 percent significance level, the t-ratios in Table 7 show that, when x is drawn from $U(0, 3)$ and $P_+ = 0.75$, the rejection frequency is only 24 percent. The rejection frequency decreases with the degree of censoring. When $P_+ = 0.25$, the t-test fails completely because the null hypothesis is never rejected. When x is drawn from $U(0, 10)$, however, the rejection frequencies are considerably higher. Only two of the six mean values of the t-ratios in designs [1] and [2] (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 collinearity problems. High collinearity therefore 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 manifests the harmful effects of collinearity.

4.2. *True model = Sample selection model with* $x_1 \neq x_2$

There are two different designs of the regressors in Manning, Duan, and Rogers' experiments. In the first design (the no-exclusion-restrictions case), x_1

Table 7
Summary statistics of t-ratio and rejection frequency

Design	P_+	Mean	Standard deviation	Skewness	Kurtosis	Minimum	Maximum	Rejection frequency
[1]	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
[2]	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
[3]	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
[4]	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
[5]	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	-6×10^{-5}	1.179	0.047	3.299	−2.993	3.200	11

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 [$\text{corr}(x_1, x_2) = 0$ 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 8 reports the simulation results based on design [3]. Similar to Manning, Duan, and Rogers, the LIML estimator performs very well. The *MSPE*s 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 9 gives 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 8 and 9 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 experiments therefore 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.[10]

Table 4 shows that the condition numbers and the absolute values of $r(x_2, \hat{\lambda})$ for both $\text{corr}(x_1, x_2) = 0$ and 0.5 (designs [3] and [4]) are very low. There are no signs of collinearity problems. This again verifies our claims that the sample

[10] In design [3], we follow Manning, Duan, and Rogers' setup and generate x_1 and x_2 independently. Hence x_1 and x_2 are not only uncorrelated but also independent. This independence assumption may be too strong for the purpose of studying the effect of the degree of correlation between x_1 and x_2 on the performance of the models. For example, it favors the two-part model because, when x_1 and x_2 are independent, the omitted variable λ (which is a function of x_1) is uncorrelated with x_2, so that the OLS estimate of β_2 remains unbiased. On the other hand, the independence assumption also favors the sample selection model because λ and x_2 are uncorrelated so that there are no collinearity problems (see Table 4). We keep the independence assumption in design [3] in order that our results (Table 8) can be made comparable to those in Manning, Duan, and Rogers (Table 3). Nevertheless, we have conducted a similar experiment in which x_1 and x_2 are uncorrelated but not independent [x_1 is drawn from $U(-5, 5)$ and $x_2 = x_1^2/5$, so $\text{corr}(x_1, x_2) = 0$]. As expected, the two-part model has a larger parameter bias because the omitted variable λ is now correlated with x_2. Similar to the results in Table 3, the LIML performs better than the N2P as long as $\hat{\lambda}$ and x_2 are not highly correlated.

Table 8
Simulation results based on design [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	3.5985	2427.49	− 0.00049	0.000183	− 0.0025	0.00471
		(1.784)	(313.05)	(0.0014)	(0.000026)	(0.007)	(0.0007)
	FIML	3.5123	2373.34	− 0.00058	0.000179	− 0.0029	0.00462
		(1.771)	(306.02)	(0.0014)	(0.000026)	(0.007)	(0.0007)
	N2P	2.5682	6167.52	0.00027	0.000190	0.0014	0.00487
		(1.845)	(440.96)	(0.0014)	(0.000026)	(0.007)	(0.0007)
	DA2P	2.5722	6697.10	0.00038	0.001634	0.00147	0.0049
		(1.923)	(518.81)	(0.0041)	(0.000429)	(0.007)	(0.0007)
0.5	LIML	0.2794	16.480	0.00039	0.0002763	0.00198	0.00711
		(0.131)	(2.227)	(0.0017)	(0.000038)	(0.0085)	(0.001)
	FIML	0.2579	16.002	0.00024	0.0002759	0.00120	0.00710
		(0.131)	(2.217)	(0.0017)	(0.000038)	(0.0085)	(0.001)
	N2P	0.2260	40.090	− 0.00147	0.0002759	− 0.00744	0.00709
		(0.135)	(3.315)	(0.0017)	(0.00004)	(0.0084)	(0.001)
	DA2P	0.2236	44.995	− 0.00011	0.002320	− 0.00740	0.00706
		(0.145)	(4.126)	(0.0048)	(0.000638)	(0.0084)	(0.001)
0.25	LIML	0.0179	0.1186	0.000566	0.000479	0.0029	0.0123
		(0.009)	(0.015)	(0.0022)	(0.00007)	(0.011)	(0.002)
	FIML	0.0166	0.1169	0.000127	0.000482	0.00064	0.0124
		(0.009)	(0.015)	(0.0022)	(0.00008)	(0.011)	(0.002)
	N2P	0.0224	0.2551	− 0.00478	0.000530	− 0.0242	0.0136
		(0.01)	(0.025)	(0.0023)	(0.00008)	(0.012)	(0.002)
	DA2P	0.0207	0.2685	− 0.00061	0.004627	− 0.0238	0.0138
		(0.01)	(0.027)	(0.0068)	(0.0016)	(0.012)	(0.002)

Table 9
Simulation results based on design [4] (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	2762.82	− 0.00061	0.000173	− 0.0031	0.0044
		(2.149)	(353.5)	(0.0013)	(0.00003)	(0.007)	(0.001)
	FIML	4.3635	2753.13	− 0.00086	0.000174	− 0.0043	0.0045
		(2.157)	(357.84)	(0.0013)	(0.00003)	(0.007)	(0.001)
	N2P	3.7209	3747.82	− 0.01427	0.000373	− 0.0722	0.0096
		(2.174)	(410.71)	(0.0013)	(0.00004)	(0.007)	(0.001)
	DA2P	5.8444	4648.41	− 0.02476	0.002808	− 0.0749	0.0102
		(2.279)	(488.4)	(0.0047)	(0.00052)	(0.007)	(0.001)
0.5	LIML	0.2247	20.3768	− 0.00187	0.000252	− 0.0095	0.0065
		(0.173)	(2.348)	(0.0016)	(0.00004)	(0.008)	(0.001)
	FIML	0.2308	20.5593	− 0.00172	0.000265	− 0.0087	0.0068
		(0.175)	(2.443)	(0.0016)	(0.00005)	(0.008)	(0.001)
	N2P	0.2617	31.6954	− 0.0202	0.000655	− 0.1023	0.0168
		(0.177)	(3.151)	(0.0016)	(0.00009)	(0.008)	(0.002)
	DA2P	0.3126	39.5396	− 0.0249	0.00505	− 0.1047	0.0184
		(0.191)	(4.177)	(0.0067)	(0.001)	(0.009)	(0.002)
0.25	LIML	0.0135	0.1677	− 0.0031	0.00058	− 0.0157	0.0147
		(0.014)	(0.018)	(0.002)	(0.00007)	(0.012)	(0.002)
	FIML	0.0136	0.1647	− 0.0024	0.00056	− 0.0123	0.0145
		(0.014)	(0.017)	(0.002)	(0.00007)	(0.012)	(0.002)
	N2P	0.0309	0.3654	− 0.0251	0.00117	− 0.1268	0.0299
		(0.014)	(0.027)	(0.002)	(0.00014)	(0.012)	(0.004)
	DA2P	0.0262	0.4155	− 0.0143	0.01307	− 0.1177	0.0367
		(0.015)	(0.035)	(0.011)	(0.00269)	(0.015)	(0.004)

selection estimators perform well when there are no collinearity problems. Unlike the results in design [1], Table 7 demonstrates that when there are no collinearity problems, the t-tests on the coefficient of $\hat{\lambda}$ are very effective. In all six cases in designs [3] and [4], the rejection frequencies are close to 100 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 3, 8, and 9 show that the FIML is in general better than the LIML in terms of squared errors (*MSPE*, *PSE*, *ESE*), but the LIML can be better than the FIML in terms of biases (*MPB*, *PB*, *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 of the LIML and FIML estimators is still 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 4). 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 7 indicates that the frequency of rejecting the null hypothesis $\rho = 0$ is very close to the prescribed 5 percent level of significance when the two-part model is the true model (design [5]). 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. These results again confirm that the t-tests are very effective when there are no collinearity problems in the model.

5. Discussions

The previous section clearly demonstrates that the merits of the two-part model have been grossly exaggerated in the literature. We prove 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. After the deficient design is corrected, the sample selection model clearly dominates the two-part model when the former is the true model. Hence, the

Table 10
Simulation results based on design [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	3302.34	− 0.00109	0.00034	− 0.0025	0.00735
		(1.384)	(334.66)	(0.0019)	(0.00005)	(0.0086)	(0.001)
	FIML	20.2447	3278.64	− 0.00127	0.00034	− 0.0030	0.00733
		(1.383)	(322.01)	(0.0019)	(0.00004)	(0.0086)	(0.001)
	N2P	11.9245	1631.90	− 0.00074	0.00018	− 0.0057	0.0060
		(1.214)	(177.66)	(0.0014)	(0.00003)	(0.0077)	(0.0009)
	DA2P	11.2889	2033.24	0.00645	0.00260	0.0008	0.0073
		(1.380)	(296.35)	(0.0051)	(0.0007)	(0.0086)	(0.001)
0.5	LIML	2.3069	41.490	− 0.00338	0.00138	0.08899	0.3343
		(0.139)	(4.665)	(0.00372)	(0.0002)	(0.057)	(0.051)
	FIML	2.2981	41.073	− 0.00385	0.00140	0.1118	0.3624
		(0.138)	(4.533)	(0.00374)	(0.0002)	(0.06)	(0.055)
	N2P	1.5183	22.2057	− 0.002359	0.00064	0.0596	0.1498
		(0.122)	(2.877)	(0.00254)	(0.0001)	(0.038)	(0.025)
	DA2P	1.5243	24.4951	− 0.004188	0.02128	0.0570	0.2177
		(0.125)	(3.181)	(0.01465)	(0.0053)	(0.047)	(0.039)
0.25	LIML	0.3176	0.8760	− 0.0115	0.0390	0.2621	5.2400
		(0.013)	(0.094)	(0.02)	(0.006)	(0.229)	(0.872)
	FIML	0.2825	0.7055	− 0.0576	0.0398	0.3340	4.6069
		(0.016)	(0.065)	(0.019)	(0.005)	(0.213)	(0.733)
	N2P	0.2201	0.3678	− 0.0103	0.0027	0.2397	4.1699
		(0.010)	(0.03)	(0.005)	(0.0004)	(0.204)	(0.821)
	DA2P	0.2211	0.4287	− 0.0178	0.4624	0.2253	5.8789
		(0.011)	(0.037)	(0.068)	(0.107)	(0.243)	(1.082)

Table 11
LIML estimates of σ (standard errors in parentheses)

Design	P_+	Parameter bias	Parameter squared error
[1]	0.75	0.0285 (0.00928)	0.00934 (0.00299)
	0.5	0.08056 (0.01736)	0.03633 (0.00997)
	0.25	0.48408 (0.06691)	0.67753 (0.1827)
[2]	0.75	0.00438 (0.00291)	0.00085 (0.00009)
	0.5	0.00541 (0.00327)	0.00109 (0.00014)
	0.25	0.02059 (0.00881)	0.0081 (0.00186)

extreme and negative remarks against the sample selection model made by Duan et al. (1983, 1984, 1985) are unwarranted and misleading.[11]

The sample selection model is susceptible to collinearity problems. In ordinary linear regression models, collinearity does not bias the parameter estimates and the predictions. Table 2 verifies that the parameter estimates of β are not affected by collinearity because the parameter biases (in absolute values) of LIML are roughly the same for all three values of P_+ and are smaller than those of N2P even when there are serious collinearity problems in the sample selection model. However, the mean prediction biases of LIML in Table 2 are larger than those of N2P (especially when $P_+ = 0.25$), indicating that collinearity biases the predictions. The reason is that MPB is a function of $E(y)$ which in turn depends on ρ and σ in the sample selection model [see (5)]. As shown in Table 11, the parameter bias of the LIML estimate of σ increases with collinearity. The exponentiation of σ^2 in the formula of $E(y)$ in (5) further exacerbates the bias in the prediction of LIML. Thus collinearity biases the prediction in this type of nonlinear model. As the expression for the elasticity contains σ [see (7)], this also explains why the elasticity bias of the LIML increases with collinearity and is higher than that of the N2P in Table 2.

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

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. Under 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 7. 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. Figs. 4 and 5 plot the t-ratio against the condition number for all three values of P_+. In Fig. 4, 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 Fig. 5, 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 $\lambda(\underset{\sim}{x}_1\hat{\alpha})$ 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 $\lambda(\underset{\sim}{x}_1\hat{\alpha})$ will be highly collinear. A high degree of censoring can also generate near collinearity because more censoring reduces the range of the argument of $\hat{\lambda}$. This has been shown in Table 5, 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

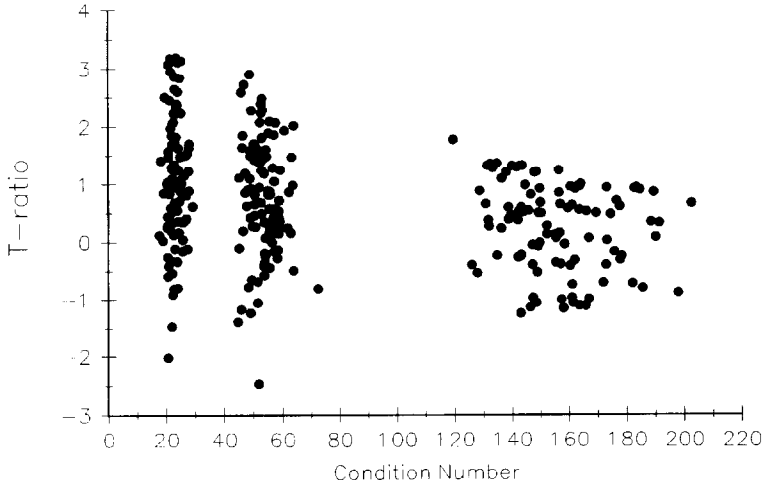

Fig. 4. T-ratio versus condition number: Design [1].

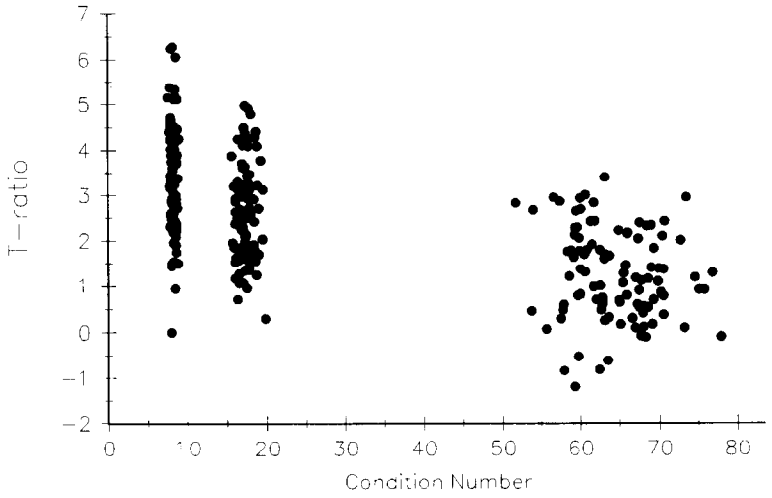

Fig. 5. T-ratio versus condition number: Design [2].

because the range of the argument of the inverse Mills' ratio decreases with the standard error of u_1.[12] 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.

In two recent studies, Nawata (1993, 1994) also points out the collinearity problem of Heckman's two-step estimator. Nevertheless, he only examines the effect of varying the correlation coefficient of x_1 and x_2 on Heckman's estimator. More importantly, our results reveal that Nawata's central claim is misleading. He asserts that Heckman's two-step estimator 'is expected to perform poorly when there is a high degree of collinearity between $w_i'\hat{\alpha}[\underset{\sim}{x}_1\hat{\alpha}]$ and $x_i[x_2]$' (Nawata, 1993, p. 24; 1994, p. 40). This claim is faulty because, even when x_2 and $\underset{\sim}{x}_1\hat{\alpha}$ are perfectly correlated, Heckman's estimator will still perform well if the range of $\underset{\sim}{x}_1$ is sufficiently large or if the proportion of censored observations is small. For example, x_2 and $\underset{\sim}{x}_1\hat{\alpha}$ are perfectly correlated in our designs [1] and [2] (because $x_1 = x_2$), yet Table 3 shows that Heckman's estimator behaves well because there is enough variability in x_1. Table 2 also shows that Heckman's estimator performs relatively well when censoring is not serious ($P_+ = 0.75$). Hence, a high correlation between x_2 and $\underset{\sim}{x}_1\hat{\alpha}$ does not necessarily lead to

[12] When σ_1 (standard error of u_1) is not necessarily unity, the inverse Mills' ratio is given by $\lambda(\underset{\sim}{x}_1\alpha/\sigma_1)$. Other things being equal, the range of $\underset{\sim}{x}_1\alpha/\sigma_1$ decreases with σ_1.

collinearity problems. Similar to Manning, Duan, and Rogers, Nawata's faulty assertion is derived from a deficient design in his experiments. In Nawata's design, $\alpha_1 = -1$, $\alpha_2 = 0.1$, the range of x_1 is (0, 20], and the variance of u_1 is 1. These specifications imply that the range of $\alpha_1 + \alpha_2 x_1$ is only $(-1, 1]$. From Nawata's (1993) Table 1, the correlation coefficient of $\lambda(z)$ and z when z is distributed uniformly over $(-1, 1]$ is -0.9959. With such a high correlation, it is not surprising to find that Heckman's estimator performs poorly when there are no exclusion restrictions on x_1 and x_2. Nawata is apparently unaware of this design problem because he attributes the poor performance of Heckman's estimator solely to the high correlation between x_2 and $x_1\hat{\alpha}$. The actual causes are the high correlation *and* the small range of $\alpha_1 + \alpha_2 x_1$. As shown in Tables 3, 8, and 9, when x_1 follows $U(0, 10)$, Heckman's estimator performs well regardless of the correlation between x_1 and x_2 because the range of $\alpha_1 + \alpha_2 x_1$ is larger than Nawata's $(-1, 1]$.

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 5: (i) $P_+ = 0.9$ in design [1] and (ii) $P_+ = 0.3$ in design [2]. 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 squares 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 any collinearity problems whereas the condition number in case (ii) indicates serious collinearity problems. In contrast, given the same sample correlation coefficient (each about -0.91) in cases (i) and (ii), one cannot tell whether there are collinearity problems. This example illustrates that the condition number is superior to the sample correlation coefficient in detecting collinearity.

The condition numbers reported in Table 4 are the averages of 100 replications. The distribution of the condition numbers (not shown here) is fairly symmetric and concentrated around the mean; see Leung and Yu (1992) for details. The standard error is small relative to the mean, so the condition number is a reliable indicator of collinearity. In view of this and the previous results, and the fact that condition numbers (or eigenvalues) can readily be obtained from many computer software packages (such as GAUSS, LIMDEP, SAS, SPSS), we recommend that they be used to detect collinearity.

Collinearity problems provide an additional (or alternative) explanation for the anomalous results that many have found in their applications of Heckman's two-step procedure. For example, a large number of empirical studies find (somewhat surprisingly) that the coefficient estimates of 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 (Duncan, 1983). However, 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. To substantiate this conjecture, we employ a popular data set compiled by Mroz (1987) to illustrate the possibility of collinearity problems in the two-step procedure.

Mroz's (1987) sensitivity analysis of empirical models of female labor supply has been widely cited and his data set has been used for pedagogy and for further research by many others.[13] As there are many sample selection models in Mroz's study, we only choose the two main ones which are also summarized and reviewed in Berndt (1991). In Mroz's Tables IX and X (pp. 782–783), he tests for self-selection bias in the labor supply equation by comparing the estimates between two models: a two-stage least squares model (specification 6) and a multi-stage sample selection model (specification 3). Without controlling for the possibility of selection bias, specification 6 estimates the hours of work equation by standard two-stage least squares using a set of variables to instrument the wife's wage rate. Specification 3 controls for possible selection bias by first applying Heckman's two step procedure to the wage equation and then including the inverse Mills' ratio (along with the fitted log wage rate) into the set of regressors in the third and final stage of the multi-stage model. Without using the wife's working experience as an instrument (Table IX), Mroz's test fails to reject the null hypothesis that there is no self-selection bias. However, when the wife's working experience and its square are added to the set of instruments (Table X), the test rejects the null of no self-selection bias. Mroz offers some explanations for the puzzling test results. We believe that collinearity problems provide an alternative account for his findings.

Let $Z_{IX,3}$ and $Z_{IX,6}$ denote the data matrices in the last stage of the regression in specifications 3 and 6 in Table IX, respectively. The differences between the two matrices are that $Z_{IX,3}$ contains one more regressor (the inverse Mills' ratio) than $Z_{IX,6}$ and that the fitted log wage rate in $Z_{IX,3}$ includes the inverse Mills' ratio as one of the regressors (see Berndt, 1991, pp. 641–642, for a nice summary). We find that the condition numbers of $Z_{IX,6}$ and $Z_{IX,3}$ are 31.66 and 57.95, respectively. In other words, adding the inverse Mills' ratio nearly doubles the condition number of the data matrix. With such a high condition number, it is not surprising to find that the coefficient estimate of the inverse Mills' ratio in specification 3 is insignificant [t-value = 0.63; see Eq. (11.53) in Berndt, 1991] because collinearity problems prevail. Hence, Mroz's test fails to reject specification 6, which means that there is no evidence of self-selection bias. For the corresponding data matrices in specifications 3 and 6 in Table X, we find that

[13] For instance, Newey, Powell, and Walker (1990) utilize the data set to evaluate some new semiparametric estimation methods and Berndt (1991) uses it extensively to introduce the sample selection literature. Our version of Mroz's data was obtained from the diskette accompanying Berndt's textbook.

the condition numbers are 28.50 and 30.44, respectively. When the wife's working experience and its square are included in the set of instruments, adding the inverse Mills' ratio does not generate collinearity problems because it only slightly increases the condition number of the data matrix. The coefficient estimate of the inverse Mills' ratio in specification 3 is strongly significant [t-value = 4.43; see Eq. (11.52) in Berndt, 1991], which explains why Mroz's test rejects specification 6 and hence there is evidence of self-selection bias. We also find that the ranges of the argument of the inverse Mills' ratio are $[-1.178, 1.635]$ and $[-1.565, 2.107]$ for specification 3 in Tables IX and X, respectively. Hence, the larger range of values shows that adding the wife's working experience increases the variation of the inverse Mills' ratio, thereby reducing the problems of collinearity. In contrast to Mroz (1987, p. 790) who thinks that the wife's working experience 'appears to be an instrument which does little to increase the accuracy of the estimates while complicating the required model', we believe that the opposite is true because the experience variables increase the variation of the data and improve the precision of the estimate of the inverse Mills' ratio. His conclusion that controls for self-selection are unimportant when experience is treated as endogenous (excluded from the set of instruments) seems premature because his tests may have been impaired by collinearity problems due to insufficient variations in the data or inadequate exclusion restrictions in the models.

6. Conclusion

Since the sample selection model and the two-part model perform well under different simulation setups, we believe that a balanced view of the merits of the two models is more appropriate. The two-part model does provide a useful alternative to the sample selection model that has dominated the literature. The sample selection model was primarily designed to identify the unconditional parameters and the unconditional (potential) outcome, whereas the two-part model was designed solely to predict the conditional (actual) outcome. They are competing models because the sample selection model can also be used to predict the conditional (actual) outcome. The choice between the two models depends partly on what parameters and outcomes one wants to identify.

Our findings on the collinearity problem of Heckman's two-step procedure also shed some light on the distinctive identification problem in semiparametric estimation of censored regression models (Chamberlain, 1986). Without specifying the probability distributions of u_1 and u_2, Eq. (2) can be written as $m = \underset{\sim}{x}_2\beta + \psi(\underset{\sim}{x}_1\alpha) + v$ for $m > 0$, where $\psi(\underset{\sim}{x}_1\alpha) = E(u_2 | \underset{\sim}{x}_2, u_1 > -\underset{\sim}{x}_1\alpha)$ and $E(v | \underset{\sim}{x}_2, u_1 > -\underset{\sim}{x}_1\alpha) = 0$. Treating ψ as an unknown function, several methods have been proposed to estimate β (e.g., Robinson, 1988; Cosslett, 1991; Ichimura and Lee, 1991; Ahn and Powell, 1993). Despite differences in the construction of

the estimators, all these semiparametric estimation methods have to impose some kind of exclusion restrictions on $\underset{\sim}{x}_2$ in order to identify β. For instance, Cosslett (1991) requires that at least one of the regressors in $\underset{\sim}{x}_1$ cannot be included in $\underset{\sim}{x}_2$, whereas Robinson (1988) and Ichimura and Lee (1991) require that $\underset{\sim}{x}_1$ and $\underset{\sim}{x}_2$ cannot have any regressors in common. In contrast, parametric methods do not require these kinds of identifying exclusion restrictions because ψ is known and is usually a nonlinear function. Although *in principle* the nonlinearity of ψ or exclusion restrictions on $\underset{\sim}{x}_2$ are sufficient to identify β, our results indicate that the performance of the estimators depends crucially on the degree of collinearity between $\psi(\underset{\sim}{x}_1\alpha)$ and $\underset{\sim}{x}_2$. If there are few exclusion restrictions and little variability in $\underset{\sim}{x}_1$ and $\underset{\sim}{x}_2$, then collinearity problems may be serious and the prospect of semiparametric estimation will likely to be poor.

Finally, we remark that the design problem in Manning, Duan, and Rogers is actually pervasive in the sample selection literature. In a separate paper (Leung and Yu, 1993), we demonstrate that the data generating processes in a large number of Monte Carlo studies also produce serious 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 $U(-1.7, 1.7)$, which is basically 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 variance, the effective range of the regressor becomes $[0, 2]$. Consequently, Heckman's two-step estimator does not behave well in Paarsch's simulations. Hence, there are many notable misleading results in the sample selection literature because of the inadvertent bias against Heckman's two-step procedure.

References

Ahn, H. and J. Powell, 1993, Semiparametric estimation of censored selection models with a nonparametric selection mechanism, Journal of Econometrics 58, 3–29.

Amemiya, T., 1985, Advanced econometrics (Harvard University Press, Cambridge, MA).

Belsley, D., E. Kuh, and R. Welsch, 1980, Regression diagnostics: Identifying influential data and sources of collinearity (Wiley, New York, NY).

Berndt, E., 1991, The practice of econometrics: Classic and contemporary (Addison–Wesley, New York, NY).

Chamberlain, G., 1986, Asymptotic efficiency in semi-parametric models with censoring, Journal of Econometrics 32, 189–218.

Cosslett, S., 1991, Semiparametric estimation of a regression model with sample selectivity, in: W. Barnett, J. Powell, and G. Tauchen, eds., Nonparametric and semiparametric methods in econometrics and statistics (Cambridge University Press, New York, NY).

Cragg, J., 1971, Some statistical models for limited dependent variables with application to the demand for durable goods, Econometrica 39, 829–844.

Duan, N., 1983, Smearing estimate: A nonparametric retransformation method, Journal of the American Statistical Association 78, 605–610.

Duan, N., W. Manning, C. Morris, and J. Newhouse, 1983, A comparison of alternative models for the demand for medical care, Journal of Business and Economic Statistics 1, 115–126.

Duan, N., W. Manning, C. Morris, and J. Newhouse, 1984, Choosing between the sample-selection model and the multi-part model, Journal of Business and Economic Statistics 2, 283–289.

Duan, N., W. Manning, C. Morris, and J. Newhouse, 1985, Comments on selectivity bias, Advances in Health Economics and Health Services Research 6, 19–24.

Dudley, L. and C. Montmarquette, 1976, A model of the supply of bilateral foreign aid, American Economic Review 66, 132–142.

Duncan, G., 1983, Sample selectivity as a proxy variable problem: on the use and misuse of Gaussian selectivity corrections, Research in Labor Economics 2, 333–345.

Goldberger, A., 1964, Econometric theory (Wiley, New York, NY).

Grossman, M. and T. Joyce, 1990, Unobservables, pregnancy resolutions, and birth weight production functions in New York City, Journal of Political Economy 98, 983–1007.

Hay, J. and R. Olsen, 1984, Let them eat cake: A note on comparing alternative models of the demand for medical care, Journal of Business and Economic Statistic 2, 279–282.

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

Heckman, J., 1976, 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, 475–592.

Heckman, J., 1979, Sample selection bias as a specification error, Econometrica 47, 153–161.

Heckman, J., 1990, Varieties of selection bias, American Economic Review Papers and Proceedings 80, 313–318.

Ichimura, H. and L.F. Lee, 1991, Semiparametric estimation of multiple index models: Single equation estimation, in: W. Barnett, J. Powell, and G. Tauchen, eds., Nonparametric and semi-parametric methods in econometrics and statistics (Cambridge University Press, New York, NY).

Leung, S.F. and S. Yu, 1992, On the choice between sample selection and two-part models, Rochester Center for Economic Research working paper no. 337 (University of Rochester, Rochester, NY).

Leung, S.F. and S. Yu, 1993, Collinearity and Heckman's two-step estimator, Working paper (Department of Economics, University of Rochester, Rochester, NY).

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, 3–18.

Maddala, G.S., 1985b, Further comments on selectivity bias, Advances in Health Economics and Health Services Research 6, 25–26.

Mallows, C., 1973, Some comments on C_p, Technometrics 15, 661–675.

Manning, W., L. Blumberg, and L. Moulton, 1995, The demand for alcohol: The differential response to price, Journal of Health Economics 14, 123–148.

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

Manning, W., C. Morris, and J. Newhouse, 1981, 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 (North-Holland, Amsterdam).

Manning, W., H. Bailit, B. Benjamin, and J. Newhouse, 1985, The demand for dental care: Evidence from a randomized trial in health insurance, Journal of the American Dental Association 110, 895–902.

Manning, W., A. Leibowitz, G. Goldberg, W. Rogers, and J. Newhouse, 1984, A controlled trial of the effect of a prepaid group practice on the use of services, New England Journal of Medicine 310, 1505–1510.

Melino, A., 1982, Testing for sample selection bias, Review of Economic Studies 49, 151–153.

McLaughlin, K., 1991, A theory of quits and layoffs with efficient turnover, Journal of Political Economy 99, 1–29.

Mroz, T., 1987, The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions, Econometrica 55, 765–799.

Nawata, K., 1993, A note on the estimation of models with sample-selection biases, Economics Letters 42, 15–24.

Nawata, K., 1994, Estimation of sample selection bias models by the maximum likelihood estimator and Heckman's two-step estimator, Economics Letters 45, 33–40.

Nelson, F., 1984, Efficiency of the two-step estimator for models with endogenous sample selection, Journal of Econometrics 24, 181–196.

Newey, W., J. Powell, and J. Walker, 1990, Semiparametric estimation of selection models: some empirical results, American Economic Review Papers and Proceedings 80, 324–328.

Newhouse, J., W. Manning, and C. Morris et al., 1981, Some interim results from a controlled trial of cost sharing in health insurance, New England Journal of Medicine 305, 1501–1507.

Paarsch, H., 1984, A Monte Carlo comparison of estimators for censored regression models, Journal of Econometrics 24, 197–213.

Peters, S. and R. Smith, 1991, Distributional specification tests against semiparametric alternatives, Journal of Econometrics 47, 175–194.

Powell, J., 1986, Symmetrically trimmed least squares estimation for Tobit models, Econometrica 54, 1435–1460.

Robinson, P., 1988, Root-N-consistent semiparametric regression, Econometrica 56, 931–954.

van de Ven, W. and B. van Praag, 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 (North-Holland, Amsterdam).

van de Ven, W. and B. van Praag, 1981b, The demand for deductibles in private health insurance: A probit model with sample selection, Journal of Econometrics 17, 229–252.