Discussion Paper No. 97-07 E

Foul or Fair?
The Heckman Correction for Sample Selection and Its Critique
A Short Survey

Patrick A. Puhani

Foul or Fair?
The Heckman Correction for Sample Selection and Its Critique
A Short Survey

by

Patrick A. Puhani

Zentrum für Europäische Wirtschaftsforschung (ZEW), Mannheim
Centre for European Economic Research
SELAPO, University of Munich

April 1997

Abstract: This paper gives a short overview of Monte Carlo studies on the usefulness of Heckman's (1976, 1979) two–step estimator for estimating a selection model. It shows that exploratory work to check for collinearity problems is strongly recommended before deciding on which estimator to apply. In the absence of collinearity problems, the full–information maximum likelihood estimator is preferable to the limited–information two–step method of Heckman, although the latter also gives reasonable results. If, however, collinearity problems prevail, subsample OLS (or the Two–Part Model) is the most robust amongst the simple–to–calculate estimators.

Patrick A. Puhani
ZEW
P.O. Box 10 34 43
D–68034 Mannheim, Germany
e–mail: puhani@zew.de

Acknowledgement
I thank Klaus F. Zimmermann, SELAPO, University of Munich, Viktor Steiner, ZEW, Mannheim and Michael Lechner, University of Mannheim, for helpful comments. All remaining errors are my own.

1 Introduction

Selection problems occur in a wide range of applications in econometrics. The basic problem is that sample selection usually leads to a sample being unrepresentative of the population we are interested in. As a consequence, standard ordinary least squares (OLS) estimation will give biased estimates. Heckman (1976, 1979) has proposed a simple practical solution, which treats the selection problem as an omitted variable problem. This easy–to–implement method, which is known as the two–step or the limited information maximum likelihood (LIML) method, has been criticised recently, however. The debate around the Heckman procedure is the topic of this short survey. The survey makes no claim on completeness, and the author apologises for the possible omission of some contributions.

The paper is structured as follows. Section 2 outlines Heckman's LIML as well as the FIML estimator. Section 3 summarises the main points of criticism, whereas Section 4 reviews Monte Carlo studies. Section 5 concludes.

2 Heckman's Proposal

Suppose we want to estimate the empirical wage equation [1a] of the following model:

$$y_{2i}^{*} = \mathbf{x}_{2i}^{'}\boldsymbol{\beta}_2 + u_{2i}\,. \qquad [1a]$$

$$y_{1i}^{*} = \mathbf{x}_{1i}^{'}\boldsymbol{\beta}_1 + u_{1i} \qquad [1b]$$

$$y_{2i} = y_{2i}^{*} \qquad if \quad y_{1i}^{*} > 0$$

$$y_{2i} = 0 \qquad if \quad y_{1i}^{*} \le 0. \qquad [1c]$$

One of the $\mathbf{x}_2$–variables may be years of education. As economists we will be interested in the wage difference an extra year of education pays in the labour market. Yet we will not observe a wage for people who do not work. This is expressed in [1c] and [1b], where [1b] describes the propensity to work.

Economic theory suggests that exactly those people who are only able to achieve a comparatively low wage given their level of education will decide not to work, as for them, the probability that their offered wage is below their reservation wage is highest. In other words, u_1 and u_2 can be expected to be positively correlated. It is commonly assumed that u_1 and u_2 have a bivariate normal distribution:

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \sim BN\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}\right] \qquad [2]$$

Given this assumption, the likelihood function of model [1] can be written (Amemiya, 1985, p.386): [3]

$$L = \prod_{y_2=0} 1 - \Phi\left(\frac{\mathbf{x}_1'\beta_1}{\sigma_1}\right) \prod_{y_2>0} \Phi\left(\frac{\beta_1}{\sigma_1} + \frac{\sigma_{12}}{\sigma_2^2}(y_2 - \mathbf{x}_2'\beta_2)\right) \Big/ \sqrt{\sigma_1^2 - \frac{\sigma_{12}^2}{\sigma_2^2}} \times \frac{1}{\sigma_2}\phi\left(\frac{y_2 - \mathbf{x}_2'\beta_2}{\sigma_2}\right)$$

As the maximisation of this likelihood (full–information maximum likelihood, FIML) took a lot of computing time until very recently, Heckman (1979) proposed to estimate likelihood [3] by way of a two–step method (limited–information maximum likelihood, LIML).

It is obvious that for the subsample with a positive y_2^* the conditional expectation of y_2^* is given by:

$$E(y_{2i}^* | \mathbf{x}_{2i}, y_{1i}^* > 0) = \mathbf{x}_{2i}'\beta_2 + E(u_{2i} | u_{1i} > -\mathbf{x}_{1i}'\beta_1). \qquad [4]$$

It can be shown that, given assumption [2], the conditional expectation of the error term is:

$$\mathrm{E}\left(\mathrm{u}_{2i} | \mathrm{u}_{1i} > -\mathbf{x}_{1i}'\beta_1\right) = \frac{\sigma_{21}}{\sigma_1} \frac{\phi\left(-\left(\mathbf{x}_{1i}'\beta_1/\sigma_1\right)\right)}{1 - \Phi\left(-\left(\mathbf{x}_{1i}'\beta_1/\sigma_1\right)\right)}, \qquad [5]$$

where $\phi(.)$ and $\Phi(.)$ denote the probability and cumulative density functions of the standard normal distribution, respectively. Hence we can rewrite the conditional expectation of y_2^* as

$$E(y_{2i}^* | \mathbf{x}_{2i}, y_{1i}^* > 0) = \mathbf{x}_{2i}'\beta_2 + \frac{\sigma_{21}}{\sigma_1} \frac{\phi(\mathbf{x}_{1i}'\beta_1/\sigma_1)}{1 - \Phi(\mathbf{x}_{1i}'\beta_1/\sigma_1)}. \qquad [6]$$

Heckman's (1979) two–step proposal is to estimate the so–called inverse Mills ratio $\lambda(\mathbf{x}_{1i}'\beta_1/\sigma_1) = \frac{\phi(\mathbf{x}_{1i}'\beta_1/\sigma_1)}{1 - \Phi(\mathbf{x}_{1i}'\beta_1/\sigma_1)}$ by way of a Probit model and then estimate equation [7]:

$$y_{2i} = \mathbf{x}_{2i}'\boldsymbol{\beta}_2 + \frac{\sigma_{21}}{\sigma_1}\lambda(\mathbf{x}_{1i}'\boldsymbol{\beta}_1 / \sigma_1) + \varepsilon_2 \qquad [7]$$

in the second step. Hence, Heckman (1979) characterised the sample selection problem as a special case of the omitted variable problem with λ being the omitted variable if OLS were used on the subsample for which $y_2^* > 0$. As long as u_1 has a normal distribution and ε_2 is independent of λ, Heckman's two step estimator is consistent.[1] However, it is not efficient as ε_2 is heteroscedastic. To see this, note that the variance of ε_2 is given by

$$V(\varepsilon_{2i}) = \sigma_2^2 - \frac{\sigma_{12}^2}{\sigma_1^2}\mathbf{x}_{1i}'\boldsymbol{\beta}\left(\frac{\mathbf{x}_{1i}'\boldsymbol{\beta}}{\sigma_1} + \lambda\left(\frac{\mathbf{x}_{1i}'\boldsymbol{\beta}}{\sigma_1}\right)\right) \qquad [8]$$

Clearly, $V(\varepsilon_{2i})$ is not constant, but varies over i, as it varies with $\mathbf{x}_{1i}$. In order to obtain a simple and consistent estimator of the asymptotic variance–covariance matrix, Lee (1982, p.364f.) suggests to use White's (1980) method. Under the null hypothesis of no selectivity bias, Heckman (1979, p.158) proposes to test for selectivity bias by way of a t–test on the coefficient on λ. Melino (1982) shows that the t–statistic is the Lagrange multiplier statistic and therefore has the corresponding optimality properties.[2]

Although there are other LIML estimators of model [1] than the one proposed by Heckman (*e.g.* Olsen, 1980; Lee, 1983), this paper will use LIML to denote Heckman's estimator unless stated otherwise. Heckman (1979) considered his estimator to be useful for 'provid(ing) good starting values for maximum likelihood estimation'. Further, he stated that '(g)iven its simplicity and flexibility, the procedure outlined in this paper is recommended for exploratory empirical work.' (Heckman, 1979, p.160). However, Heckman's estimator has become a standard way to obtain final estimation results for models of type [1]. As we will see in the following section, this habit has been strongly criticised for various reasons.

---

[1] However, as Rendtel (1992) points out, the orthogonality of $\mathbf{x}_2$ to u_2 does not imply that $\mathbf{x}_2$ be orthogonal to ε_2.

[2] Olsen (1982) proposes a residuals-based test to test for selectivity.

3 The Critique of Heckman's Estimator

Although LIML has the desirable large–sample property of consistency, various papers have investigated and criticised its small–sample properties. The most important points of criticism can be summarised as follows:

1) It has been claimed that the predictive power of subsample OLS or the Two–Part Model (TPM) is at least as good as the one of the LIML or FIML estimators. The debate of sample–selection versus two–part (or multi–part) models was sparked off by Duan *et al.* (1983, 1984, 1985). The Two–Part Model (see also Goldberger, 1964, pp.251ff.; and Cragg, 1971, p.832) is given by

$$y_{2i}^{*} \left| y_{1i}^{*} > 0 = \mathbf{x}_{2i}^{'} \boldsymbol{\beta}_2 + u_{2i} \right. \qquad [9a]$$

$$y_{1i}^{*} = \mathbf{x}_{1i}^{'} \boldsymbol{\beta}_1 + u_{1i}, \text{ and} \qquad [9b]$$

$$y_{2i} = y_{2i}^{*} \qquad \textit{if} \quad y_{1i}^{*} > 0, \text{ and}$$

$$y_{2i} = 0 \qquad \textit{if} \quad y_{1i}^{*} \leq 0. \qquad [9c]$$

The point is that [9a] models y_2^* *conditional* on y_1^* being positive. The expected value of y_2^* is then

$$E\left(y_{2i}^{*}\right) = \Phi\left(\mathbf{x}_{1i}\boldsymbol{\beta}_1/\sigma_1\right)\left(\mathbf{x}_{2i}\boldsymbol{\beta}_2\right) \qquad [10]$$

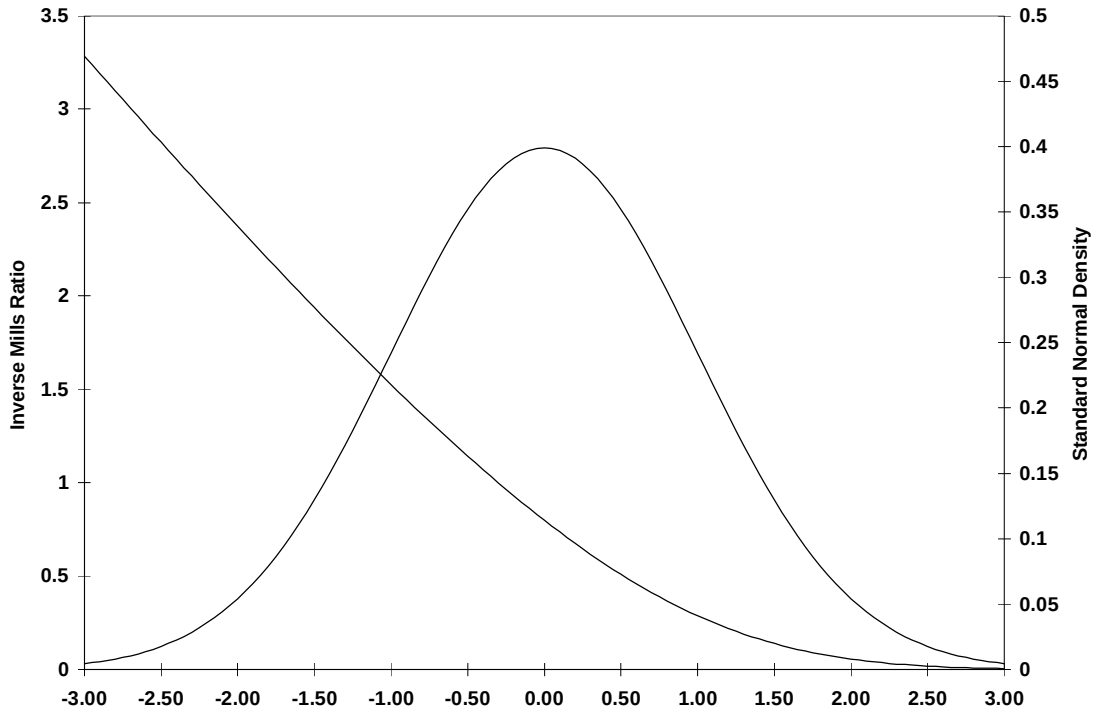
Marginal effects on the expected value of y_2^* of a change in an $\mathbf{x}_2$–variable would thus have to be calculated by differentiating [10] with respect to the variable of interest.

We argue that there are three main ways to interpret the TPM. The first is to claim that it is not the unconditional, but rather the conditional expectation of y_2^* that is of interest to us. This approach is taken by Duan *et al.* (1983, 1984, 1985). The other approach is to stress the behavioural structure of the model (Maddala, 1985a, 1985b), to which the selection process is central. In this case, LIML and TPM estimate the same behavioural relation. The TPM, however, then makes an implicit distributional assumption for the unconditional distribution, which will be a mixing distribution also depending on the distribution driving the selection mechanism. This approach, however, seems unsatisfactory from a theoretical point of view (see Hay and Olsen, 1984; and Maddala, 1985a, p.14). A third and very crude interpretation, which is not explicitly stated in the literature on the TPM, is to interpret the coefficients of [9a]

also as the ones of the unconditional equation [1a]. This is tantamount to estimating [1a] by subsample OLS. As we will see below, the justification for the latter two interpretations will be given on statistical rather than theoretical grounds.

2) In practical problems, $\mathbf{x}_1$ and $\mathbf{x}_2$ often have a large set of variables in common. In some cases, they are even identical. One says that there are no exclusion restrictions if no variables that are in $\mathbf{x}_1$ are excluded from $\mathbf{x}_2$. In these cases, equation [7] is only identified through the nonlinearity of the inverse Mills ratio λ. However, collinearity problems are likely to prevail as $\lambda(\cdot)$ is an approximately linear function over a wide range of its argument. This is illustrated in Figure 1.

Figure 1: The Quasi–Linearity of the Inverse Mills Ratio



Note that the probability to work for a person with characteristics $\mathbf{x}_1$ is given by $\Phi(\mathbf{x}_1'\beta_1/\sigma_1)$. Only if this probability is higher than about 97.5 percent will $\mathbf{x}_1'\beta_1/\sigma_1$ be higher than 2. If most cases in a particular sample are not such extreme examples, most observations will lie within the quasi–linear range of the inverse Mills ratio, as demonstrated in Figure 1. It follows that regression [7] is likely to yield rather unrobust results due to collinearity problems. Therefore, Little and Rubin (1987, p.230) state that 'for the (Heckman) method to work in practice, variables are needed in ($\mathbf{x}_1$) that are good predictors of (y_1^*) and do not appear in ($\mathbf{x}_2$), that is, are not associated with (y_2) when other covariates are controlled.' Unfortunately, it is often very difficult to find such variables in practice. In our wage example, theory would

suggest that household variables like children and the income of the spouse are likely to influence the reservation wage, but unlikely to influence the gross offered wage and hence should only be included in $\mathbf{x}_1$. However, these household data are not always available, and even if they are, it is not guaranteed that these variables are good predictors of the propensity to work y_1^*. But even if they are, the household variables may well be also associated with the offered wage y_2^*, especially if the after–tax wage is being observed, as children and the income of other family members have an impact on the tax rate in many tax regimes.

3) Another line of criticism stresses the sensitivity of the estimated coefficients with respect to the distributional assumptions placed on the error terms in [1a] and especially in [1b] (Little and Rubin, 1987, pp.225ff.). Instead of making strong distributional assumptions, some authors suggest semi–parametric or non–parametric procedures (see, for example, Chamberlain, 1986; Duncan, 1986; Powell, 1986; Robinson, 1988; Newey, Powell, and Walker, 1990; Cosslett, 1991; Ichimura and Lee, 1991; Ahn and Powell, 1993; Lee, 1996; Stern, 1996). These studies will not be surveyed here.

In the following, we give summaries of important Monte Carlo studies on the performance of the LIML estimator in their historical order.

4 Monte Carlo Studies

4.1 Nelson (1984)

Nelson investigates the bias and efficiency of the LIML, FIML, and subsample OLS estimators dependent on both the coefficient of determination (R^2) of the regression of the inverse Mills ratio λ on $\mathbf{x}_2$ (consisting of two variables) and the correlation ρ between u_1 and u_2. With R^2 taking on the values 0, 0.35, 0.641, 0.95, and 0.999 and the correlation between u_1 and u_2 varying between –0.5, 0, 0.25, 0.5, 0.75, and 0.95, Nelson analyses 30 different specifications. The sample size chosen is 2,000 observations in the selected sample. The estimators of β_{21} and ρ are compared in terms of their variances when applying different estimation methods. Nelson finds that the relative efficiency of the FIML over the LIML estimators of both β_{21} and ρ rises both with a higher correlation between $\mathbf{x}_2$ and λ and with a higher correlation between u_1 and u_2. Hence, the author concludes that '(t)he conditions under which the OLS bias is largest are precisely the conditions under which the dominance of the (FIML) over the (LIML) estimator is greatest.' (p.195) Nelson's suggestions for applied research are to calculate the correlation of $\mathbf{x}_2$ with the estimated inverse Mills ratio. When this correlation is very low, he suggests using subsample OLS

because of its small bias. If the correlation is very high, Nelson recommends the FIML estimator. LIML may be used in intermediate cases, but Nelson's results show that even there FIML is no worse than LIML.

4.2 Paarsch (1984)

Paarsch investigates model [1] with no exclusion restrictions and with identical errors in [1a] and [1b] which amounts to the type 1 Tobit model in Amemiya's (1985) classification. He estimates LIML, FIML(Tobit), subsample OLS, and Powell's (1992) Least Absolute Deviation estimator (LAD). The true models are distinguished by the error distributions (normal, Laplace, and Cauchy) as well as the degree of censoring (25 and 50 percent). Further, the sample size is varied between 50, 100, and 200. Thus, 18 different models are looked at with 100 repetitions each. The parameter estimates are judged on the basis of their means, standard deviations, medians, as well as upper and lower quartiles. In all experiments, subsample OLS turns out to be the worst estimator. Also, the Tobit estimator performs poorly when the errors have a Cauchy distribution. In that case, the LAD estimator is robust. When the assumption of normal errors is fulfilled, the LIML estimator is much less efficient than the FIML(Tobit) estimator. The Tobit estimator also performs well when the true error distribution is Laplace.

4.3 Hay, Leu, and Rohrer (1987)

Hay, Leu, and Rohrer compare the LIML estimator with another LIML estimator which has a Logit model in the selection equation [1b], and the Two–Part Model. There are no exclusion restrictions. The authors use data from the Swiss Socio–Medical Indicator System for the Population of Switzerland (SOMIPOPS) to perform a Monte Carlo simulation experiment. The true model is obtained from initial estimates, whereby high and low parameter value groups are determined for the simulation. Then Monte Carlo data sets are obtained by generating three different error structures: bivariate normal, bivariate logistic, and bivariate Cauchy. The correlations between u_1 and u_2 are chosen to be 0, 0.33, 0.66, 0.90, and 1.00. The sample size is varied between 300, 1,500, and 3,000, of which around 20 percent are censored. The estimators are evaluated on the basis of predictive performance as well as the mean squared error of parameter estimates. Hay, Leu, and Rohrer's findings are that for the case of no exclusion restrictions analysed here, the TPM is the most robust of the three estimators investigated. This is especially true when the error distributions in the selection equation are normal or logistic. In the Cauchy case, none of the models can establish a superiority over the others. The authors therefore conclude that for specifications similar to theirs, the small–sample inefficiency of the LIML estimator effects the obtained estimates more gravely than the theoretical

deficiencies and the inconsistency of the TPM. The TPM is therefore seen as a robust practical solution to the problem of estimating a selection model like [1].

4.4 Manning, Duan, and Rogers (1987)

Manning, Duan, and Rogers compare the LIML, FIML, TPM and Data–Analytic TPM[3] in a Monte Carlo study with 1,000 observations, where the true models are selection models with a correlation between u_1 and u_2 of 0.5 and 0.9, and 25, 50, and 75 percent censoring. $\mathbf{x}_1$ and $\mathbf{x}_2$ each contain one variable, but two cases are distinguished. In the first case, $\mathbf{x}_1$ and $\mathbf{x}_2$ are identical (no exclusion restrictions), in the second, they are orthogonal to each other. There are 100 repetitions for each model specification. The estimators are judged solely on their predictive power, not on their merits concerning bias and efficiency. Manning, Duan, and Rogers show that in the case of no exclusion restrictions, LIML estimation yields the worst predictions of all four models investigated. The Data–Analytic TPM and FIML estimation turn out to be the best in this case. The authors find that both LIML and FIML perform especially badly when the degree of censoring is very high. This is daunting, of course, as in this case, correction for selection bias would be most needed. When there are effective exclusion restrictions, however (*i.e.* in the design where $\mathbf{x}_1$ and $\mathbf{x}_2$ are orthogonal), the LIML estimator performs best of all (FIML is not investigated). Yet the difference in predictive power between LIML and the Data–Analytic TPM turns out to be very small. Manning, Duan, and Rogers therefore conclude that in general, the Data–Analytic TPM is a robust estimator, 'as long as analysts are concerned about the response surface, rather than particular coefficients' (p.82). Of course, in many economic applications, the consistent and robust estimation of the coefficients is of more interest than the predictive power of a model. As the authors mention, the good predictive performance of the Data–Analytic TPM stems from the fact that a large share of the variation of the inverse Mills ratio λ can be explained by higher–order terms of $\mathbf{x}_2$.

4.5 Stolzenberg and Relles (1990)

Stolzenberg and Relles focus their Monte Carlo study on the impacts of the correlation between $\mathbf{x}_1$ and $\mathbf{x}_2$ (both consist of only one variable) as well as the correlation between u_1 and u_2 on estimating model [1] by LIML and subsample OLS. In addition, the variance of both u_1 and u_2 is varied. They censor 90 percent of the

[3] A TPM is called *Data-Analytic* if, for example, transformations of the regressors (*e.g.* higher-order terms) are also included in order to improve the fit of the regression (Manning, Duan, and Rogers, 1987, p.64f.).

observations noting that severe censoring is common in sociological situations. Furthermore, they want to create a case which is sufficiently distinct from a non–censored regression in order to facilitate comparability of the subsample–OLS and the LIML estimator. The data set size chosen is 500. The authors obtain 144 different true models by varying the squared correlations between $\mathbf{x}_1$ and $\mathbf{x}_2$ between 0, 0.25, 0.5, and 0.75, varying the squared correlations between u_1 and u_2 between 0, 0.25, 0.5, and 0.75, varying the variance of u_1 between 1/9, 1, and 9, and also varying the variance of u_2 between 0.25, 1, and 4. Each of the 144 models is replicated 100 times. The estimators are judged on the basis of the mean absolute error of the estimated coefficient on $\mathbf{x}_2$. In sum, LIML is superior to subsample OLS in one half of the estimates and the average absolute error of both estimators is roughly the same. The simulations show no clear relationship between the variances of u_1 and u_2 and the performance of the two estimators in question. Yet it is found that a high correlation between u_1 and u_2 and simultaneously a high correlation between $\mathbf{x}_1$ and $\mathbf{x}_2$ render LIML to be superior to subsample OLS in terms of parameter *bias*. However, even in those cases, the absolute *error* of the estimates is larger than in subsample OLS in over a third of the simulations, which confirms the view that the LIML estimator is not robust. Stolzenberg and Relles conclude that the LIML estimator is generally not recommendable.

4.6 Zuehlke and Zeman (1991)

Zuehlke and Zeman investigate the sensitivity of subsample OLS, LIML, and Lee's (1982, p.359f.) robust estimator with respect to the joint distribution of the error terms u_1 and u_2. The distributions implemented in their Monte Carlo study are bivariate normal, bivariate t_5, and bivariate χ^2_5. There are no exclusion restrictions, and there is only one regressor. The degree of censoring is varied between 25, 50, and 75 percent. Further, the correlation between u_1 and u_2 takes on the values 0, 0.5, and 1. The full sample size is chosen to be 100, and there are 1,000 repetitions on each model. The models are compared on their merits concerning the mean bias and the mean squared error. Zuehlke and Zeman find that whereas the LIML estimator reduces the bias compared to subsample OLS, its parameter estimates have very large standard errors due to the high degree of collinearity between $\mathbf{x}_1$ and the inverse Mills ratio λ. The authors find that the collinearity problem is exacerbated with a high degree of censoring. The Lee (1982) estimator does not perform very well. These results are robust with respect to variations in the joint distribution of u_1 and u_2. Thus Zuehlke and Zeman suggest employing the more robust subsample OLS estimator if the subsample is very small.

4.7 Rendtel (1992)

Rendtel compares the LIML estimator with the total–sample OLS and the FIML estimator, respectively. He mainly focuses on the issue of exclusion restrictions by observing the sensitivity of the estimates of the equation of interest [1a] with respect to the adding and dropping of variables in the selection equation [1b]. Rendtel's initial Monte Carlo design has three variables in both $\mathbf{x}_1$ and $\mathbf{x}_2$ with no exclusion restrictions, one of them being a dummy variable. Rendtel points out that '(d)ummy variables rather frequently occur in empirical work but rather seldom in simulation studies.' (p.9). He mentions that the coefficient of the dummy variable in [1a] may be very susceptible to selectivity bias if this variable is important in the selection process [1b], as there may be severe collinearity problems with the constant term. In Rendtel's setup, the total sample size is 400 with a third of the y_2–values being censored. Each of the specifications is replicated 100 times. In the first case, the variables in $\mathbf{x}_1$ and $\mathbf{x}_2$ are identical. Secondly, an additional variable is placed into the $\mathbf{x}_2$–vector of the selection equation [1b]. Thirdly, a variable is omitted from [1b]. Finally, one variable is omitted from [1a] and one is added to [1b]. The models are evaluated on the bias and the variance of the estimated coefficients. In the first case with no exclusion restrictions, it shows that although LIML and FIML yield less biased estimates than total sample OLS, the mean squared error of the LIML (FIML) estimates is about 4 (1.3) times the OLS mean squared error. So LIML is shown to be very unrobust. This is especially true for the coefficient on the dummy variable. In the second design, Rendtel investigates the common procedure of including an additional variable into the selection equation to obtain exclusion restrictions as suggested, for example, by Little and Rubin (1987, p.230). The author distinguishes between four statistical cases which may occur employing this method. The added variable may be uncorrelated with both y_1^* and y_2^*, or correlated with either or both of y_1^* and y_2^*. Rendtel uses this setup because '(i)n empirical work one is rather seldom in a position to know a priori the relationship of the added variable to (y_2^*) and (y_1^*).' (p.20). As expected, when the added variable is uncorrelated with both y_1^* and y_2^*, the relative efficiency of the OLS, FIML, and LIML estimators is similar to the case with no exclusion restrictions. If, on the other hand, the added variable is only correlated with y_1^*, the exclusion restriction makes both the LIML and the FIML estimators very effective by reducing the FIML and LIML mean squared error remarkably to about one half of the OLS mean squared error. If the variable added to the selection equation is correlated with both y_1^* and y_2^*, the FIML, but especially the LIML parameter estimates in equation [1a] are strongly downward biased. The worst case, however, is the one where the added variable is solely correlated with y_2^*. Rendtel shows that the distribution of the FIML dummy parameter estimate in [1a] becomes bimodal and very dissimilar to a normal distribution. The mean squared error of the FIML estimate becomes 5 times, and the mean squared error of the LIML estimate

even 10 times the OLS mean squared error. In the third design, a variable is omitted from the selection equation, a case which might arise if this variable is unobserved for the censored cases or if the researcher thinks to be able to stabilise his or her estimates in [1a] this way (p.21). However, it turns out that if a relevant variable is missing in the selection equation, both the LIML and the FIML estimators lose their power of correcting for selectivity bias when estimating the coefficient of this variable in the equation of interest [1a]. Hence, one should be very careful in these situations. Rendtel concludes that in general subsample OLS yields more robust results than both FIML or LIML. Only if effective exclusion restrictions can be found and implemented, FIML or LIML are clearly superior to subsample OLS. For this reason, researchers are advised to look at the correlation of the variables excluded in [1a] with the dependent variables in [1a] and [1b]. If these variables are mainly correlated with y_1^*, FIML estimation is proposed as it is generally more stable and efficient than LIML estimation.

4.8 Nawata (1993) and Nawata (1994)

Nawata compares the LIML estimator with the subsample OLS (1993) and the FIML estimator (1994), respectively. The FIML estimator uses a modified maximisation procedure (1994, p.35) to facilitate convergence to the global maximum. In both papers, the issues under investigation are the sensitivities of the estimates with respect to the correlations between u_1 and u_2 and between $\mathbf{x}_1$ and $\mathbf{x}_2$ (both $\mathbf{x}_1$ and $\mathbf{x}_2$ consist of only one variable). For u_1 and u_2, the correlation coefficients considered are 0, 0.2, 0.4, 0.6, 0.8, and 1 (1993) and 0, 0.4, and 0.8 (1994). For $\mathbf{x}_1$ and $\mathbf{x}_2$, the correlations generated are 0, 0.4, 0.8, 0.9, 0.95, and 1 in both papers. 50 percent of the observations are censored. The sample size is 200 and each model is estimated 500 times in (1993) and 200 times in (1994), respectively. The results are compared on the basis of the mean, standard deviation, median, as well as the upper and lower quartiles of the parameter estimates. Comparing subsample OLS and FIML with LIML, both the 1993 and 1994 papers show that LIML is less efficient the higher the degree of correlation between $\mathbf{x}_1$ and $\mathbf{x}_2$. This result is consistent with the *a priori* reasoning on collinearity outlined in Section 3 above. Similarly, a correlation between u_1 and u_2 above 0.9, which is the case where correction for sample selection is most needed, renders the LIML estimator very unstable. In this case, Nawata suggests applying the FIML estimator. But subsample OLS would also give more robust results than LIML.

4.9 Leung and Yu (1996)

Leung and Yu evaluate the predictive power, the parameter bias and error, and the elasticity bias and error of the LIML, FIML, TPM, and the Data–Analytic TPM

estimators in specification [1]. In their Monte Carlo designs, $\mathbf{x}_1$ and $\mathbf{x}_2$ consist of one variable each. 25, 50, and 75 percent of the observations are censored from a sample size of 1,000 observations with 100 repetitions on each specification. There are five different designs of the true model. One is a TPM, the other four are sample selection models. The sample selection model without exclusion restrictions is specified both with a small [0,3] and a large [0,10] range of the exogenous variable. In the case with exclusion restrictions Leung and Yu consider both a correlation of 0.5 and a correlation of zero between $\mathbf{x}_1$ and $\mathbf{x}_2$. The authors demonstrate that the crucial determinant of the performance of the LIML (and FIML) estimator is the presence or absence of collinearity problems. Three main causes of collinearity problems are identified, *viz.* the lack of exclusion restrictions, a too small range of the argument of the inverse Mills ratio, and a too high degree of censoring. The first cause follows from the quasi–linearity of the inverse Mills ratio as depicted in Figure 1 above. Similarly does the second cause. The third cause is less obvious but can easily be made plausible. An increase in the share of censored observations is achieved by increasing the share of observations with $\mathbf{x}_1'\boldsymbol{\beta}_1 < 0$. So if the distribution of $\mathbf{x}_1'\boldsymbol{\beta}_1$ is shifted to the left and everybody with $\mathbf{x}_1'\boldsymbol{\beta}_1 < 0$ is censored, the remaining observations will exhibit a smaller range of $\mathbf{x}_1'\boldsymbol{\beta}_1$ thus causing collinearity problems (*cf.* Figure 1 above). The authors show that collinearity not only effects the accuracy of the estimates of $\boldsymbol{\beta}_2$ and the predictive power, but also the power of the t–test on the coefficient of the inverse Mills ratio for the presence of selectivity bias. However, in the absence of collinearity problems, the t–test turns out to be a good discriminator between the TPM and the selection model. In particular, if the variation of the argument of the inverse Mills ratio is small and there are no exclusion restrictions, LIML estimation yields less robust results than the TPM. However, when the range of the argument of the inverse Mills ratio is increased, Leung and Yu find that even with no exclusion restrictions the LIML estimator is preferable to the TPM estimator. When $\mathbf{x}_1$ and $\mathbf{x}_2$ are correlated, but not highly correlated, the findings are that the LIML and FIML estimators are superior to both the TPM and the Data–Analytic TPM. These results hold even more strongly if $\mathbf{x}_1$ and $\mathbf{x}_2$ are uncorrelated. In this latter case, Leung and Yu found that FIML slightly outperforms LIML. However, the superiority of FIML over LIML could not be established as a general result. Roughly speaking, although FIML turns out to be more efficient, LIML is often less biased. Leung and Yu also consider the case where the TPM is the true model. Although LIML and FIML perform worse than the TPM, so does the Data–Analytic TPM, the reason being collinearity problems due to the inclusion of irrelevant variables. To sum up, the results indicate that the crucial criterion for model selection in applied research should be the issue of collinearity. Leung and Yu suggest to test for collinearity by calculating the condition number for the regressors in [7] (a *LIMDEP*

7.0 programme is given in the appendix to this paper). If the condition number exceeds 20, the TPM is more robust, otherwise, FIML (or LIML) is recommended.[4]

5 Conclusions

The general conclusions which may be drawn from the surveyed Monte Carlo studies as well as the theoretical considerations cast doubt on the omnipotence implicitly ascribed by many applied researchers to Heckman's (1976, 1979) two–step estimator. Indeed, Heckman himself is confirmed when he writes that the purpose of his estimator is only to 'provide() good starting values for maximum likelihood estimation' and 'exploratory empirical work.' (Heckman, 1979, p.160).

The cases where the need to correct for selectivity bias are largest are those with a high correlation between the error terms of the selection and the main equation, and those with a high degree of censoring. Unfortunately, though, as the Monte Carlo analyses show, in exactly those cases Heckman's estimator is particularly inefficient and subsample OLS may therefore be more robust. In addition, empirical researchers are often confronted with a high correlation between the exogenous variables in the selection and the main equation. Because the inverse Mills ratio is approximately linear over wide ranges of its argument, such high correlation is also likely to make Heckman's LIML, but also the FIML estimator very unrobust due to the collinearity between the inverse Mills ratio and the other regressors.

The practical advice one may draw from these results is that the estimation method should be decided upon case by case. A first step should be to investigate whether there are collinearity problems in the data. This can be done by calculating R^2 of the regression of the inverse Mills ratio on the regressors of the main equation or by calculating the corresponding condition number (a short *LIMDEP 7.0* programme is given in the appendix). If collinearity problems are present, subsample OLS (or the Two–Part Model) may be the most robust and simple–to–calculate estimator. If there are no collinearity problems, Heckman's LIML estimator may be employed, but given the constant progress in computing power, the FIML estimator is recommended, as it is usually more efficient than the LIML estimator.

[4] This differs from Belsley, Kuh, and Welsch's (1980, p.105) suggestion of taking 30 as the critical value. Leung and Yu (1996, p.224) believe that chosing 20 as the critical value gives fairly accurate results, as the standard error of the condition number is quite small relative to its mean.

References

Ahn, H, and Powell, J. L. (1993): Semiparametric Estimation of Censored Selection Models with a Non–Parametric Selection Mechanism, *Journal of Econometrics*, Vol.58, pp.3–29.

Amemiya, T. (1985): Advanced Econometrics, Basil Blackwell, Oxford.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980): Regression Diagnostics, Identifying Influential Data and Sources of Collinearity, John Wiley & Sons, New York.

Chamberlain, G. (1986): Asymptotic Efficiency in Semi–Parametric Models with Censoring, *Journal of Econometrics*, Vol.32, pp.189–218.

Cosslett, S. (1991): Semiparametric Estimation of a Regression Model with Sample Selectivity, in: Barnett, W. A., Powell, J., and Tauchen, G. (eds.): Nonparametric and Semiparametric Methods in Econometrics and Statistics, Cambridge University Press.

Cragg, J. G. (1971): Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods, *Econometrica*, Vol.39, No.5, pp.829–844.

Duan, N., Manning, W. G., Morris, C. N., and Newhouse, J. P. (1983): A Comparison of Alternative Models for the Demand for Medical Care, *Journal of Business & Economic Statistics*, Vol.1, No. 2, pp.115–126.

Duan, N., Manning, W. G., Morris, C. N., and Newhouse, J. P. (1984): Choosing Between the Sample–Selection Model and the Multi–Part Model, *Journal of Business & Economic Statistics*, Vol.2, No. 3, pp.283–289.

Duan, N., Manning, W. G., Morris, C. N., and Newhouse, J. P. (1984): Comments on Selectivity Bias, *Advances in Health Economics and Health Services Research*, Vol.6, pp.19–24.

Duncan, G. M. (1986): A Semi–Parametric Censored Regression Estimator, *Journal of Econometrics*, Vol.32, pp.5–34.

Goldberger, A. S. (1964): Econometric Theory, John Wiley & Sons, New York.

Hay, J. W., Leu, R., and Rohrer, P. (1987): Ordinary Least Squares and Sample–Selection Models of Health–Care Demand: Monte Carlo Comparison, *Journal of Business & Economic Statistics*, Vol.5, pp.499–506.

Hay, J. W., Olsen, R. J. (1984): Let Them Eat Cake: A Note on Comparing Alternative Models of the Demand for Medical Care, *Journal of Business & Economic Statistics*, Vol.2, No.3, pp.279–282.

Heckman, J. 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 Social Measurement*, Vol.5, No.4, pp.475–492.

Heckman, J. J. (1979): Sample Selection Bias as a Specification Error, *Econometrica*, Vol.47, No.1, pp.153–161.

Ichimura, H., and Lee, L.–F. (1991): Semiparametric Least Squares Estimation of Multiple Index Models: Single Equation Estimation, in: Barnett, W. A., Powell, J., and Tauchen, G. (eds.): Nonparametric and Semiparametric Methods in Econometrics and Statistics, Cambridge University Press.

Lee, L.–F. (1982): Some Approaches to the Correction of Selectivity Bias, *Review of Economic Studies*, Vol.49, pp.355–372.

Lee, L.–F. (1983): Generalized Econometric Models with Selectivity Bias, *Econometrica*, Vol.51, No.2, pp.507–512.

Lee, M.–J. (1996): Nonparametric Two–Stage Estimation of Simultaneous Equations with Limited Endogenous Regressors, *Econometric Theory*, Vol. 12, pp.305–330.

Leung, S., F., Yu, S. (1996): On the Choice Between Sample Selection and Two–Part Models, *Journal of Econometrics*, Vol.72, pp.197–229.

Little, R. J. A. (1985): A Note About Models for Selectivity Bias, *Econometrica*, Vol.53, No.6, pp.1469–1474.

Little, R. J. A., and Rubin, D. B. (1987): Statistical Analysis with Missing Data, John Wiley & Sons, New York.

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*, Vol.6, pp.3–18.

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

Manning, W. G., Duan, N., Rogers, W. H. (1987): Monte Carlo Evidence on the Choice Between Sample Selection and Two–Part Models, *Journal of Econometrics*, Vol.35, pp.59–82.

Melino, A. (1982): Testing for Sample Selection Bias, *Review of Economic Studies*, Vol.49, pp.151–153.

Nawata, K. (1993): A Note on the Estimation of Models with Sample–Selection Biases, *Economics Letters*, Vol.42, pp.15–24.

Nawata, K. (1994): Estimation of Sample Selection Bias Models by the Maximum Likelihood Estimator and Heckman's Two–Step Estimator, *Economics Letters*, Vol.45, pp.33–40.

Nelson, F. D. (1984): Efficiency of the Two–Step Estimator for Models with Endogenous Sample Selection, *Journal of Econometrics*, Vol.24, pp..181–196.

Newey, W. K., Powell, J. L., and Walker, J. R. (1990): Semiparametric Estimation of Selection Models: Some Empirical Results, *American Economic Review Papers and Proceedings*, Vol.80, pp.324–328.

Olsen, R. J. (1980): A Least Squares Correction for Selectivity Bias, *Econometrica*, Vol.48, No.7, pp.1815–1820.

Olsen, R. J. (1982): Distributional Tests for Selectivity Bias and a More Robust Likelihood Estimator, *International Economic Review*, Vol.23, No.1, pp.223–240.

Paarsch, H. J. (1984): A Monte Carlo Comparison of Estimators for Censored Regression Models, *Journal of Econometrics*, Vol.24, pp.197–213.

Powell, J. L. (1986): Symmetrically Trimmed Least Squares Estimation for Tobit Models, *Econometrica*, Vol.54, No.6, pp.1435–1460.

Powell, J. L. (1992): Least Absolute Deviations Estimation for censored and truncated Regression Models, unpublished Ph.D. Dissertation, Stanford University, CA, U.S.A.

Rendtel, U. (1992): On the Choice of a Selection–Model when Estimating Regression Models with Selectivity, DIW–Discussion Paper, No.53, Berlin.

Robinson, P. M. (1988): Root–N–Consistent Semiparametric Regression, *Econometrica*, Vol.56, pp.931–954.

Stern, S. (1996): Semiparametric Estimates of the Supply and Demand Effects of Disability on Labor Force Participation, *Journal of Econometrics*, Vol.71, pp.49–70.

Stolzenberg, R. M., and Relles, D. A. (1990): Theory Testing in a World of Constrained Research Design, The Significance of Heckmans' Censored Sampling Bias Correction for Nonexperimental Research, *Sociological Methods and Research*, Vol.18, No.4, pp.395–415.

White, H. (1980): A Heteroskedasticity–Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity, *Econometrica*, Vol.48, No.4, pp.817–828.

Zuehlke, T. W., and Zeman, A. R., (1990): A Comparison of Two–Stage Estimators of Censored Regression Models, *The Rewiev of Economics and Statistics*, Vol.72, pp.185–188.

Appendix

A Limdep Programme Which Calculates a Condition Number

```
LOAD         ; FILE = c:\data\example.sav $
OPEN         ; OUTPUT = c:\out\cond#.out  $

? List of variables for which the condition number
? is to be calculated
NAMELIST     ; X = var1, var2, var3 $

? Compute the normalised moment matrix
MATRIX       ; XX = X'X
             ; D = DIAG(XX); D = ISQR(D)
             ; XX = D * XX * D $

? Find the highest and lowest eigenvalues
MATRIX       ; E = ROOT(XX) $
CALCULATE    ; r = ROW(E) $
MATRIX       ; EH = PART(E,1,1) $
MATRIX       ; EL = PART(E,r,r) $

? Calculate and display the condition number
CALCULATE    ; Cond = EH/EL $
MATRIX       ; LIST ; Cond $
```

Note: I thank B. Greene for help with the normalisation (conversation by electronic mail).