On Testing Sample Selection Bias under the Multicollinearity Problem

Takashi Yamagata

May 2005

CWPE 0522

Not to be quoted without permission

On testing sample selection bias under the multicollinearity problem

Takashi Yamagata*
Faculty of Economics, University of Cambridge, Sidgwick Avenue,
Cambridge CB3 9DE, UK

15 December 2004

Abstract

This paper examines and compares the finite sample performance of the existing tests for sample selection bias, especially under the multicollinearity problem pointed out by Nawata (1993). The results show that under such multicollinearity problem, (i) the t-test for sample selection bias based on the Heckman and Greene variance estimator can be unreliable; (ii) the standard t-test (Heckman 1979) and the asymptotically efficient Lagrange multiplier test (Melino 1982) have correct size but very little power; (iii) however, the likelihood ratio test following the maximum likelihood estimation remains powerful.

Key Words: Sample selection bias; t-test; Wald test, likelihood ratio test, Lagrange multiplier test

JEL Classification: C12, C24

1 Introduction

Sample selection models are widely used in economics (e.g. labour economics). For these models, testing for sample selection bias ($H_0 : \rho = 0$) is always important, since the model can be estimated easily without taking the selection bias into account.

Recently Nawata & McAleer (2001) investigated the tests for sample selection bias in Maximum Likelihood (ML) context. They compared the empirical size of the Lagrange multiplier (LM) test, the Likelihood Ratio (LR) test, and the Wald test. They found that the Wald test tends to be oversized severely, and the LM test often yields negative value when the multicollinearity problem pointed out by Nawata (1993) is severe. On the other hand, the empirical size of the LR test is much better than those, even though the LR test tends to be slightly oversized for small sample. However, their investigation for testing sample selection bias is limited and incomplete.

Firstly, the LM test statistics that often took negative value in Nawata & McAleer (2001) were computed based on the "information matrix" being estimated by the Hessian matrix. However, the past research has shown that

*Corresponding author: Tel: +44-1223-335-273; fax: +44-1223-335-299; E-mail address: ty228@econ.cam.ac.uk

the LM test statistic based on the asymptotically efficient estimator for the information matrix has better finite sample performance (Orme(1990), Chesher & Spady (1991)). Melino (1982) proposed the asymptotically efficient LM test, which is always numerically positive, and conjectured that it has an optimal property in testing sample selection bias. Also, as the standard t-test for sample selection bias proposed by Heckman (1979) is very similar to this asymptotically efficient LM test, Melino recommended to use this standard t-test.

Secondly, As Olsen (1980) shows that unlike maximum likelihood estimator, Heckman two-step estimation is consistent when errors in structural equation are non-normal, as long as its conditional expectation upon errors in selection equation is linear. If the asymptotically efficient LM test performs as good as the LR test, the latter is less attractive since the full ML estimation required in the LR test is much more computationally expensive and more restrictive in terms of distributional assumptions.

Thirdly, Nawata & McAleer (2001) only investigated the size of the tests for sample selection bias, however, power properties are an equally (or more) important issue, under the multicollinearity problem. There is some evidence that the standard t-test for selectivity bias has very little power under such multicollinearity problem; see Leung & Yu (1996). If the asymptotically efficient LM test has correct size, the only advantage of the LR test following the full ML estimation would be that the LR test should have more power than the asymptotically efficient LM test, particularly under the multicollinearity problem.

The plan of this paper is as follows. The model and estimation methods are described in Section 2. Various tests for sample selection bias are described in Section 3. The design and the results of Monte Carlo simulation are discussed in Section 4. An empirical example is given in Section 5. Finally, Section 6 contains some concluding remarks.

2 The model and estimation

Consider the model

$$\begin{aligned} y_{1i}^* &= \mathbf{x}_{1i}'\boldsymbol{\beta}_1 + u_{1i} && (1)\\ y_{2i} &= \mathbf{x}_{2i}'\boldsymbol{\beta}_2 + u_{2i} && (2)\\ y_{1i} &= I(y_{1i}^* > 0),\ i = 1, ..., n \end{aligned}$$

where $(y_{1i}^*, \mathbf{x}_{1i}', y_{2i}, \mathbf{x}_{2i}') \in \mathbb{R}^1 \times \mathbb{R}^{k_1} \times \mathbb{R}^1 \times \mathbb{R}^{k_2}$ are independently and identically distributed (iid), $\mathbf{x}_{1i}'$ and $\mathbf{x}_{2i}'$ are strictly exogenous, $I(A)$ is an indicator function, where as $I(A)$ is one if A is true, and zero otherwise. y_{1i}^* is not observable, only its sign. $(y_{2i}, \mathbf{x}_{2i}')$ is observed only when $y_{1i} = 1$. We define n_1 being the number of positive observations of y_{1i}^*. Initially we assume that (u_1, u_2) have a bivariate normal distribution

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} 1 & \rho\sigma_2 \\ \rho\sigma_2 & \sigma_2^2 \end{bmatrix}\right),$$

where ρ is correlation coefficient σ_{12}/σ_2 with σ_{12} being covariance of u_1 and u_2, without loss of generality.[1]

[1] When we assume that the variance of u_{1i} is σ_1^2; $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^*/\sigma_1$ and $\boldsymbol{\beta}_1^*$ is not identifiable.

2.1 Heckman two-step estimation

The conditional expectation of y_{2i} conditional upon $\mathbf{x}_{1i}, \mathbf{x}_{2i}$, and $y^*_{1i} > 0$ is

$$E\left(y_{2i}|\mathbf{x}_{1i}, \mathbf{x}_{2i}, y_{1i} = 1\right) = \mathbf{x}'_{2i}\boldsymbol{\beta}_2 + E\left(u_{2i}|\mathbf{x}_{1i}, \mathbf{x}_{2i}, y_{1i} = 1\right)$$

so that

$$y_{2i} = \mathbf{x}'_{2i}\boldsymbol{\beta}_2 + E\left(u_{2i}|\mathbf{x}_{1i}, \mathbf{x}_{2i}, y_{1i} = 1\right) + \varepsilon_{2i}$$

where $\varepsilon_{2i} = y_{2i}-\ E\left(y_{2i}|\mathbf{x}_{1i}, \mathbf{x}_{2i}, y_{1i} = 1\right)$. By the properties of the bivariate normal distribution,

$$E\left(u_{2i}|\mathbf{x}_{1i}, \mathbf{x}_{2i}, y_{1i} = 1\right) = \rho\sigma_2 E\left(u_{1i}|\mathbf{x}_{1i}, \mathbf{x}_{2i}, y_{1i} = 1\right).$$

Noting that $y^*_{1i} > 0$ implies $u_{1i} > -\mathbf{x}'_{1i}\boldsymbol{\beta}_1$,

$$E\left(u_{1i}|\mathbf{x}_{1i}, \mathbf{x}_{2i}, u_{1i} > -\mathbf{x}'_{1i}\boldsymbol{\beta}_1\right) = \lambda_i \text{ with } \lambda_i = \frac{\phi(-\mathbf{x}'_{1i}\boldsymbol{\beta}_1)}{1 - \Phi(-\mathbf{x}'_{1i}\boldsymbol{\beta}_1)} = \frac{\phi(\mathbf{x}'_{1i}\boldsymbol{\beta}_1)}{\Phi(\mathbf{x}'_{1i}\boldsymbol{\beta}_1)},$$

where $\phi(c) = (2\pi)^{-1/2}\, e^{-c^2/2}$ is the standard normal density and $\Phi(c)$ is its cumulative distribution function. Therefore,

$$y_{2i} = \mathbf{x}'_{2i}\boldsymbol{\beta}_2 + \rho\sigma_2\lambda_i + \varepsilon_{2i}, \tag{3}$$

where $\lambda_i = \phi(\mathbf{x}'_{1i}\boldsymbol{\beta}_1)/\Phi(\mathbf{x}'_{1i}\boldsymbol{\beta}_1)$. This result leads to Heckman's (1976) two-step estimation method. Firstly, $\boldsymbol{\beta}_1$ is estimated by the probit ML method to obtain $\check{\boldsymbol{\beta}}_1$, then the λ_i of (3) is replaced by $\check{\lambda}_i = \phi(\mathbf{x}'_{1i}\check{\boldsymbol{\beta}}_1)/\Phi(\mathbf{x}'_1\check{\boldsymbol{\beta}}_1)$, and obtain

$$y_{2i} = \check{\mathbf{z}}'_{2i}\gamma_2 + v_{2i} \tag{4}$$

where $\check{\mathbf{z}}_{2i} = (\mathbf{x}'_{2i}, \check{\lambda}_i)'$ and $\boldsymbol{\gamma}_2 = (\boldsymbol{\beta}'_2, \rho\sigma_2)'$. Secondly, (4) is estimated by the ordinary least square (OLS) method to obtain $\hat{\boldsymbol{\gamma}}_2 = (\hat{\boldsymbol{\beta}}'_2, \widehat{\rho\sigma_2})'$.

For later usage, stacking (4) for all i, we have

$$\mathbf{y}_2 = \check{\mathbf{Z}}_2\gamma_2 + \mathbf{v}_2.$$

The Heckman-Greene consistent variance estimator (Heckman (1979) and Greene (1981)), which take the estimation effects and heteroskedasticity of v_{2i} into account, is

$$\check{\mathbf{V}}_{HG} = \hat{\sigma}^2_2 \left(\check{\mathbf{Z}}'_2\check{\mathbf{Z}}_2\right)^{-1} \left[\check{\mathbf{Z}}'_2 \left(\mathbf{I}_{n_1} - \hat{\rho}^2\hat{\boldsymbol{\Delta}}\right)\check{\mathbf{Z}}_2 + \hat{\rho}^2\check{\mathbf{Z}}'_2\hat{\boldsymbol{\Delta}}\underline{\mathbf{X}}_1 \left(n\check{\mathbf{V}}_p\right)^{-1}\underline{\mathbf{X}}'_1\hat{\boldsymbol{\Delta}}\check{\mathbf{Z}}_2\right] \left(\check{\mathbf{Z}}'_2\check{\mathbf{Z}}_2\right)^{-1} \tag{5}$$

where $\hat{\sigma}^2_2 = \hat{\sigma}^2_{v_2} + \hat{\bar{\delta}}\,(\widehat{\rho\sigma_2})^2$ with $\hat{\sigma}^2_{v_2} = n_1^{-1}\sum_{i=1}^{n_1}\hat{v}^2_{2i}$, $\hat{\bar{\delta}} = n_1^{-1}\sum_{i=1}^{n_1}\hat{\delta}_i$, $\hat{\delta}_i = \check{\lambda}_i\left(\check{\lambda}_i + \mathbf{x}'_{1i}\check{\boldsymbol{\beta}}_1\right)$, $\check{\mathbf{Z}}_2 = \left(\mathbf{X}_2, \check{\boldsymbol{\lambda}}\right)$ with $\mathbf{X}_2 = (\mathbf{x}_{21}, ..., \mathbf{x}_{2n_1})'$ and $\check{\boldsymbol{\lambda}} = (\check{\lambda}_1, ..., \check{\lambda}_{n_1})'$, $\hat{\rho}^2 = (\widehat{\rho\sigma_2})^2/\hat{\sigma}^2_2$, $\hat{\boldsymbol{\Delta}} = diag(\hat{\delta}_i)$, $\underline{\mathbf{X}}_1 = (\mathbf{x}_{11}, ..., \mathbf{x}_{1n_1})'$, and $\check{\mathbf{V}}_p$ is any consistent estimator for the asymptotic variance of the score of probit ML.

Olsen (1980) shows that the bivariate normality assumption can be relaxed to the assumption of normality of u_1 and the linearity of the conditional expectation of u_2 upon u_1. Thus, unlike maximum likelihood estimator, the Heckman two-step estimation is consistent when u_2 has non-normal distribution, as long as the conditional expectation of u_2 given u_1 is linear.

Nawata (1993, 1994) shows that under certain conditions[2] the Heckman (1979) two-step estimator can suffer from a multicollinearity between the inverse Mill's ratio $\check{\lambda}_i$ and $\mathbf{x}_{2i}$ in the augmented structural equation, due to its construction. On the other hand, Nawata (1994), Nawata and Nagase (1996) show that the full ML estimator is robust to such multicollinearity problem, and can produce more reliable estimator in the same circumstances. Hereafter, we call this "the multicollinearity problem".

2.2 The scanning maximum likelihood estimation

The log-likelihood of the model (1) and (2) is

$$\begin{aligned} l_n(\boldsymbol{\theta}) &= \sum_{i=1}^{n} l_i(\boldsymbol{\theta}), \\ l_i(\boldsymbol{\theta}) &= (1-y_{1i}) \ln [1-\Phi_{1i}] \\ &\quad + y_{i1} \{\ln \Phi(g_i) - \ln \sigma_2 + \ln \phi(h_i)\} \end{aligned}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \sigma_2, \rho)'$, $\Phi_{1i} = \Phi(\mathbf{x}_{1i}'\boldsymbol{\beta}_1)$, $g_i = \left(\mathbf{x}_{1i}'\boldsymbol{\beta}_1 + \frac{\rho}{\sigma_2} u_{2i}\right) / \left(1-\rho^2\right)^{1/2}$, and $h_i = u_{2i}/\sigma_2$.

Olsen (1982) shows that the maximum likelihood function is not globally concave in $\boldsymbol{\theta}$, however, given a value of ρ it is globally concave in $(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \sigma_2)'$, because after deleting all columns and rows involving partials with respect to ρ, the Hessian matrix is negative semi-definite (Olsen 1982 p.238).

Nawata (1994, 1995) points out; 1) when $|\rho|$ is close to unity the full maximum likelihood estimation does not converge; 2) because of the potential existence of local maxima, the estimation result may not be correct even if the procedure converges. Then, Nawata (1994) proposes a scanning Maximum likelihood method; see Nawata (1994) for details. Basically one obtains ML estimator for $\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \sigma_2$ using each value of ρ varying from 0 to 0.99 (resp. -0.99), increasing (resp. decreasing) by 0.01, and find the value of ρ which maximises the log-likelihood with associated ML estimators. By this method, as the log-likelihood function is continuous in ρ, the neighbourhood of the global maximum is always found when the model is correctly specified. Nawata and Nagase (1996) shows that in terms of mean square errors, the Nawata's MLE outperforms Heckman two-step estimator, especially when the multicollinearity problem exists.

Given the superiority of Nawata's scanning ML estimation method to conventional ML estimation method, we adopt it and "ML" signifies scanning ML, hereafter.

3 Tests for sample selection bias, $H_0 : \rho = 0$

The test for $\rho = 0$ is important, since the model can be estimated easily without taking the selection bias into account. We review the tests particularly referring to the multicollinearity problem.

[2] Leung & Yu (2000) point out that there are two conditions to satisfy in order for the Heckman two-step estimation method to suffer from the multicollinearity problem pointed out by Nawata (1993), and Leung & Yu (2000) criticise Nawata (1993) for his Monte Carlo design fixing one of these conditions and exagerate the possibility of the multicollinearity problem.

First of all, we can use the t-test for $\rho = 0$ following Heckman two-step estimation, based on the Heckman-Greene heteroskedastic variance estimator defined by (5),

$$t_{HG} = \frac{\widehat{\rho\sigma_2}}{\sqrt{\left(\check{V}_{HG}\right)_{\rho}}} \tag{6}$$

where $\left(\check{V}_{HG}\right)_{\rho}$ is the bottom-right diagonal element of $\check{\mathbf{V}}_{HG}$.

As Heckman (1979) discusses, under the null hypothesis the variance of $\hat{\boldsymbol{\gamma}}_1$ can be consistently estimated by $\check{\mathbf{V}}_{HG0} = \breve{\sigma}^2_{v_2}(\check{\mathbf{Z}}'\check{\mathbf{Z}})^{-1}$. Then, we can use the conventional t-test statistic for $\rho = 0$ in the regression of y_{2i} on $\check{\mathbf{Z}}$ (Heckman 1979, p.158-9). We define this t-test as

$$t_1 = \frac{\widehat{\rho\sigma_2}}{\sqrt{\left(\check{V}_{HG0}\right)_{\rho}}} \tag{7}$$

where $\left(\check{V}_{HG0}\right)_{\rho}$ is the bottom-right element of $\check{\mathbf{V}}_{HG0}$. For the Monte Carlo simulation below, the squared t-test statistics, t^2_{HG} and t^2_1 are used, so that the results are directly comparable to other tests.

Nawata & McAleer (2001) compare the finite sample behaviour of the Wald test, the LR test, and the LM test for $\rho = 0$ following the maximum likelihood estimation. They use the negative Hessian matrix as the variance estimator of the restricted and unrestricted score. Then, they find that; 1) under no multicollinearity problem, the LR test and the LM test perform adequately; 2) under the multicollinearity problem, the LR test performs adequately but the LM test perform badly (almost half of the LM test statistics in their experiments were negative[3]); 3) the Wald test tends to reject the null too often, and such tendency gets worse under the multicollinearity problem.

Before examining their results, we define the test statistics for $\rho = 0$. Firstly the Wald test statistic is

$$Wald = \frac{\tilde{\rho}^2}{n\tilde{I}_{\rho}} \tag{8}$$

under the null hypothesis, where $\tilde{I}_{\rho}$ is the bottom-right element of the inverse of the average information matrix estimator, $\mathcal{I}_n(\tilde{\boldsymbol{\theta}})^{-1}$, evaluated at the unrestricted ML estimator, $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\beta}}'_1, \tilde{\boldsymbol{\beta}}'_2, \tilde{\sigma}_2, \tilde{\rho})$.[4]

The likelihood ratio test statistic is defined as

$$LR = 2\sum_{i=1}^{n}\left(l_i(\tilde{\boldsymbol{\theta}}) - l_i(\breve{\boldsymbol{\theta}})\right) \tag{9}$$

where $\breve{\boldsymbol{\theta}} = (\breve{\boldsymbol{\beta}}'_1, \breve{\boldsymbol{\beta}}'_2, \breve{\sigma}_2, 0)'$ is the restricted maximum likelihood estimator, where $\breve{\boldsymbol{\beta}}'_1$ is the probit ML estimator of (1), $\breve{\boldsymbol{\beta}}'_2$ and $\breve{\sigma}_2$ are OLS estimator from regressing $\mathbf{y}_2$ on $\mathbf{X}_2$.

[3] In their experiments, the LM test accepts the null too often, because they accept the null hypothesis whenever the LM test statistic is negative. Their decision rule is to "reject the null hypothesis if the test statistic is larger than the critical value (Nawata and McAleer (2001, p.110)".

[4] In our simulation, he negative average Hessian matrix estimator for $\mathcal{I}_n(\boldsymbol{\theta})$ is used.

Now let us consider the Lagrange multiplier test. Under the null hypothesis, it is easily seen that the score indicator for $\rho = 0$ is

$$\left.\frac{\partial l_i(\boldsymbol{\theta})}{\partial \rho}\right|_{\rho=0} = y_{1i}\sigma_2^{-1}\lambda_i u_{2i}.$$

This basically tests the 'omitted variable' λ_i in the structural equation, and it is the same test indicator as that of the t_1 for $\rho = 0$ in the Heckman two-step model. The LM test statistic for $\rho = 0$ is

$$LM = \frac{1}{n}\frac{\partial \sum_{i=1}^{n} l_i(\breve{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}'}\left(\mathcal{I}_n(\breve{\boldsymbol{\theta}})\right)^{-1}\frac{\partial \sum_{i=1}^{n} l_i(\breve{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}}$$

where $\mathcal{I}_n(\breve{\boldsymbol{\theta}})$ is any asymptotically valid average information matrix estimator evaluated at $\breve{\boldsymbol{\theta}}$.

There are various consistent estimators for $\mathcal{I}_n(\breve{\boldsymbol{\theta}})$. Firstly, we can use an average negative Hessian matrix, evaluated at the null hypothesis $\rho = 0$ as Nawata & McAleer (2001) do. In their experiments this estimator often produces negative variance estimates, causing negative value of LM. Next, an average of outer product of gradients (OPG) estimator is another option, however, it is well-known that it is not a good estimator (Orme (1990), Chesher & Spady (1991)). Thirdly, the asymptotically efficient estimator is also a choice, and which is known to perform much better than OPG. From Melino (1982), the information matrix evaluated at $\rho = 0$ is always positive definite matrix. Also, it can be shown easily that the LM using this information matrix is $\mathbf{u}_2'\boldsymbol{\lambda}\left(\boldsymbol{\lambda}'\mathbf{M}_{X_2}\boldsymbol{\lambda}\right)^{-1}\boldsymbol{\lambda}'\mathbf{u}_2/\sigma_2^2$ and the asymptotically efficient LM test statistic can be defined as

$$LM_{AE} = \frac{\breve{\mathbf{u}}_2'\breve{\boldsymbol{\lambda}}\left(\breve{\boldsymbol{\lambda}}'\mathbf{M}_{X_2}\breve{\boldsymbol{\lambda}}\right)_2^{-1}\breve{\boldsymbol{\lambda}}\breve{\mathbf{u}}_2}{\breve{\sigma}_2^2} \qquad (10)$$

where $\breve{\mathbf{u}}_2$ is the OLS residual vector obtained by regressing $\mathbf{y}_2$ on just $\mathbf{X}_2$, and $\hat{\sigma}_2^2 = \breve{\mathbf{u}}_2'\breve{\mathbf{u}}_2/n_1$. As can be seen, if we replace the denominator of LM_{AE}, $\breve{\sigma}_2$, with $\hat{\sigma}_{v_2}^2$, which is defined just below the (5), we have t_1^2. As LM_{AE} is expected to have optimal properties (Melino 1982), it is recommended to use t_1 to test for $\rho = 0$ (Heckman 1979, Melino 1982). In this paper t_1 and LM_{AE} are separately treated, in order to emphasise that the LM_{AE} is the better choice than the LM test statistic used in Nawata and McAleer (2001).

However, it is easily seen that under the multicollinearity problem, t_1 and LM_{AE} will lack power, because the variance estimator of $\widehat{\sigma_2\rho}$ inflates due to $\left(\boldsymbol{\lambda}'\mathbf{M}_{X_2}\boldsymbol{\lambda}\right)^{-1}$ being nearly singular/zero. Indeed, there is evidence that this efficient t_1 lacks the power under the multicollinearity problem stated above. Leung & Yu (1996) show this lack of power in their limited simulation result (Leung & Yu (1996), Table 7, p.215), and by using the results of Mroz (1987).

Now, the following question arises: how to test $\rho = 0$ under the multicollinearity problem? Nawata & McAleer (2001) are only interested in the empirical size of the Wald, the LR and the LM test, and do not investigate the empirical power of these tests. However, as discussed above, the power properties of these tests are of great interest. In addition, the LR and the Wald test utilise the information under the alternative, where the ML estimation appears less affected by the multicollinearity problem. Therefore, the LR and the Wald test may be powerful under the multicollinearity problem.

The finite sample performance of these tests are considered next.

4 Monte Carlo design and results

The finite sample performance of the tests t_{HG}^2, t_1^2, LM_{AE}, LR, and $Wald$ defined by (6), (7), (10), (9), and (8) is examined. All tests are assumed to tend to χ_1^2 distribution under the null hypothesis, and hypothesis testing is conducted accordingly. As the finite sample behaviour of the tests especially under the multicollinearity problem is of interest, the experimental design from Nawata (1993) is adopted.

4.1 Design

Consider a DGP corresponding to (1) and (2)

$$\begin{aligned} y_{1i}^* &= \alpha_1 + x_{1i}\beta_1 + u_{1i} \\ y_{2i} &= \alpha_2 + x_{2i}\beta_2 + u_{2i} \\ y_{1i} &= I(y_{1i}^* > 0),\ i, ..., n \end{aligned}$$

where x_{1i} and x_{2i} are scalars. To control the multicollinearity, we draw x_{1i} and x_{2i} such that

$$\begin{aligned} x_{1i} &\sim iidU(0, 20) \\ z_{2i} &\sim iidU(0, 20) \\ x_{2i} &= \frac{\pi_0 x_{1i} + (1-\pi_0) z_{2i}}{\sqrt{\pi_0^2 + (1-\pi_0)^2}}, \end{aligned}$$

where $iidU(0, 20)$ denotes iid uniform random variables over 0 to 20. The correlation coefficient between x_{1i} and x_{2i} is $\pi = \frac{\pi_0}{\sqrt{\pi_0^2+(1-\pi_0)^2}}$ and $\pi_0 = 0.0, 0.5, 0.8,$ and 1.0 (or $\pi = 0.00,\ 0.71, 0.97, 1.00$) are considered. To control ρ, u_{1i} and u_{2i} are drawn such that

$$\begin{aligned} u_{1i} &\sim iidN(0, 1) \\ u_{2i} &= \frac{\sigma_2 \left[\rho_0 u_{1i} + (1-\rho_0) \upsilon_{2i}\right]}{\sqrt{\rho_0^2 + (1-\rho_0)^2}}, \end{aligned} \tag{11}$$

where υ_2 is drawn from i) $iidN\,(0, 1)$; ii) $\left(iid\chi^2(2) - 2\right)/2$, to see how the ML method is robust to non-normality, even they are not justified. The correlation coefficient between u_{1i} and u_{2i} is $\rho = \frac{\rho_0}{\sqrt{\rho_0^2+(1-\rho_0)^2}}$, and $\rho_0 = 0.0, 0.2, 0.4, 0.6, 0.8$ (or $\rho = 0.00, 0.24, 0.55, 0.83, 0.97$) are considered. Also σ_2 is set to 10. As π_0 increases, the multicollinearity problem becomes severer.

We set $\alpha_1 = -1$ and $\beta_1 = 0.1$, $\alpha_2 = -10$ and $\beta_2 = 1$. Also under this design, the degree of censoring is maintained around 50%.

The number of replications is 5000 for each experiment.[5] The sample size is also set to $N = 200$ and 400.

Note that t_{HG}^2 and $Wald$ can be negative. We reject the null hypothesis in such case, and report the proportion to the number of replications.

[5] All computations were performed using Gauss 6.0 for Windows (Aptech Systems Inc., 2004).

4.2 Results

Table 1a shows the size ($\rho_0 = 0.0$) and the power ($\rho_0 = 0.2, 0.4, 0.6, 0.8$) of the tests varying π_0 to control the degree of multicollinearity problem, under bivariate normal errors. First of all the rejection frequencies of t_1^2 and LM_{AE} are very similar across the experiments, but the former is always slightly larger than the latter. Size of t_1^2, and LM_{AE} is correct across the experiments. When $N = 200$, LR tends to overreject the null slightly, giving the rejection frequencies between 6.36% and 7.38%, but LR has correct size when $N = 400$. $Wald$ rejects the null far too often even when there is no multicollinearity problem at all, $\pi_0 = 0$, which is consistent to Table 4 of Nawata and McAleer (2001). The size of t_{HG}^2 is correct for $\pi_0 = 0.0, 0.5, 0.8$, but when $\pi_0 = 1.0$, t_{HG}^2 becomes unreliable, rejecting the null too often. From Table 1b, 4.48% of t_{HG}^2 in the replications are rejected due to negative t_{HG}^2 when $N = 200$. Even when subtracting 4.48 from 13.62, 9.14% of t_{HG}^2 is rejected purely because the test statistics exceeded the critical values. When $N = 400$, t_{HG}^2 tends to overreject the null, but with less negative statistics.

As conjectured, the power of t_1^2 and LM_{AE} reduces substantially as the multicollinearity problem becomes severer. When $N = 200$ and $\pi_0 = 1.0$, t_1^2 and LM_{AE} exhibit almost no power even when $\rho_0 = 0.8$. On the other hand, LR maintains the power, even when $\pi_0 = 1.0$. For example, when $\rho_0 = 0.6$ and $N = 400$, the rejection frequency of LR is 61.80%, while that of LM_{AE} is 7.14%. $Wald$ seems to have power even when $\pi_0 = 1.0$, however, it can produce negative test statistic, especially when N is small (here 200), and the value of ρ is high. In addition, as the size of $Wald$ is heavily distorted, $Wald$ is not recommended to be used.

[**Table 1a about here**]

[**Table 1b about here**]

Of course, the LR test is justified only when the ML estimation is valid. Table 2a shows the size of the tests when we relax the bivariate normal assumption, by drawing v_{2i} in (11) from the standardised χ_2^2 distribution. Although t_{HG}^2, t_1^2, and LM_{AE} show similar rejection frequencies to that of Table 1a, the size of LR and $Wald$ are more than 90%.

Given these finite sample evidence, an empirical example is illustrated next.

[**Table 2a about here**] [**Table 2b about here**]

5 An empirical example

In this section, we provide an empirical example to illustrate the importance of our previous discussion. The example is based on the famous Mroz's (1987) data set, which contains earnings statistics of 753 white married women extracted from the 1976 Panel Study of Income Dynamics (PSID). In the selected data set, 428 women are working. We focus on the determinants of the wage, following Mroz (1987), or a part of the exercise 6(b), chapter 11 of Berndt (1991). Data set are available from the accompanied diskette of Berndt (1991). We use 17 regressors: constant, KL6, KL618, WA, WE, WA2, WE2, WAWE, WA3, WE3, WA2WE, WAWE2, WFED, WMED, UN, CIT, and PRIN; see the appendix for descriptions of these variables. In the context of model (1) and (2), $y_{1i} = 1$ if a woman works and 0 otherwise. y_{2i} is natural logarithm of the woman's wage rate, called LWW, if she works. The same set of regressors are used for the selection equation and the structural equation.

[Table 3 about here][Table 4 about here]

Table 3 shows the estimation results of structural equation only. The adjusted R^2 of the regression of $\check{\lambda}_i$ on all regressors is 0.979, which shows the multicollinearity problem is very severe, and the Heckman two-step estimation result may be unreliable. The Nawata's scanning maximum likelihood estimation yields $\tilde{\rho} = -0.8$ with the value of the log-likelihood being -872.3384.[6]

Table 4 shows the tests for sample selection bias. As our analysis predicts, the standard t-test and the asymptotically efficient LM test fail to reject the null of no sample selection bias. The t-test based on the Heckman-Greene variance estimator also fail to reject the null hypothesis. Reflecting the tendency of the Wald test to reject the null too much, the value of the Wald test statistic is extremely high. Finally the LR test rejects the null hypothesis at 1% significance level. Therefore, under this model specification, there is strong evidence of existence of selection bias.[7]

6 Concluding remarks

Sample selection models are widely used in economics (e.g. labour economics). For these models, testing for sample selection bias is always important, since the model can be estimated easily without taking the selection bias into account.

This paper examined and compared the finite sample performance of the existing tests for sample selection bias, especially under the multicollinearity problem pointed out by Nawata (1993). The results show that under such multicollinearity problem; (i) the t-test for sample selection bias based on the Heckman-Greene variance estimator can be unreliable; (ii) the standard t-test for selectivity bias and the asymptotically efficient Lagrange multiplier test (Heckman 1979, Melino 1982) has correct size but very little power, which is consistent to the results of Leung & Yu (1996), however; (iii) the likelihood ratio test following the full maximum likelihood estimation remains powerful, even when the standard t-test and the asymptotically efficient Lagrange multiplier test exhibit no power; (iv) the Wald test is very unreliable for all circumstances, and should not be used, which is consistent to the results of Nawata and McAleer (2001).

The empirical example, which is shown in section 5, illustrated the importance of using the likelihood ratio test for sample selection bias under the multicollinearity problem. The standard t-test, the t-test based on the Heckman-Greene variance estimator and the asymptotically efficient Lagrange multiplier test all fail to reject the null of no sample selection bias, on the other hand, the likelihood ratio test rejects the null soundly. The Wald test also rejects the null, but its test result is unreliable.

[6] The maximum likelihood estimation procedure of Stata 8.0 stopped at $\hat{\rho} = 0.075$ with the value of the log-likelihood being -876.7991, which is a local maximum.

[7] It may be worth noting that if $\mathbf{x}_{1i}$ includes some variables in $\mathbf{x}_{2i}$ or variables highly correlated with those in $\mathbf{x}_{2i}$, given that $\lambda\left(-\mathbf{x}_{1i}'\boldsymbol{\beta}_1\right)$ is a non-linear function of $\mathbf{x}_{1i}$, $\lambda\left(-\mathbf{x}_{1i}'\boldsymbol{\beta}_1\right)$ may pick up any non-linear terms omitted in (2), such as a non-linear function of women's working experience here, and $\lambda\left(-\mathbf{x}_{1i}'\boldsymbol{\beta}_1\right)$ could be significant even though there may be no selection bias. See Maddala (1983), p.269-270.

The cost of using the likelihood ratio test are the imposition of bivariate normal assumption on the model, and the expensive computation. The standard t-test and asymptotically efficient Lagrange multiplier test are valid when errors in the structural equation are non-normal, as long as its conditional expectation upon errors in selection equation is linear (Olsen (1980)). Therefore, in general, the standard t-test procedure proposed by Heckman (1979) and Melino (1982) is recommended. When the maximum likelihood estimation is justified, the likelihood ratio test should be the choice, particularly under the multicollinearity problem.

Acknowledgement

The author received financial support from the Graduate School of Social Sciences and the School of Economic Studies, University of Manchester. I am deeply indebted to my PhD supervisor Chris D. Orme. Also I thank to Denise Osborn and Joao Santos-Silva for their helpful comments on the earlier version of this paper. Usual disclaimer holds.

References

[1] Berndt, E. R. (1991). *The practice of econometrics: Classic and contemporary.* Addison-Wesley publishing company.

[2] Chesher. A. and Spady, R. (1991). Asymptotic expansions of the information matrix test statistic. *Econometrica.* 59: 787-815.

[3] Greene, W. H. (1981). Sample selection bias as a specification error: Comment. *Econometrica.* 49: 795-798.

[4] 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 and Social Measurement.* 5: 475-492.

[5] Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica.* 47: 153-161.

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

[7] Leung, S. F., Yu, S. (2000). Collinearity and two-step estimation of sample selection models: Problems, origins, and remedies. *Computational Economics.* 15: 173-199.

[8] Maddala, G. S. (1983). *Limited-dependent and qualitative variables in econometrics.* Cambridge University Press.

[9] Melino, A. (1982). Testing for sample selection bias. *Review of Economic Studies.* 49: 151-153.

[10] 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.

[11] Nawata, K. (1993). A note on the estimation of models with sample-selection biases. *Economics Letters.* 42: 15-24.

[12] 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.

[13] Nawata, K. (1995). Estimation of sample-selection models by the maximum likelihood method. *Mathematics and Computers in Simulation.* 39: 299-303.

[14] Nawata, K., McAleer, M. (2001). Size characteristics of tests for sample selection bias: A Monte Carlo comparison and empirical example. *Econometric Reviews.* 20: 105-112.

[15] Nawata, K., Nagase, N. (1996). Estimation of sample selection bias models. *Econometric Reviews.* 15: 387-400.

[16] Olsen, R. J. (1980). A least squares correction for selectivity bias. *Econometrica.* 48: 1815-1820.

[17] Olsen, R. J. (1982). Distributional tests for selectivity bias and a more robust likelihood estimator. *International Economic Review.* 23: 223-240.

[18] Orme, C. D. (1990). The small sample performance of the information matrix test. *Journal of Econometrics.* 46: 309-331.

Appendix

The variables used in the empirical example section

LWW	woman's logarithm of estimated wage
KL6	number of kids less than 6 years old
KL618	number of kids between 6-18 years old
WA	woman's age in years
WE	woman's years of schooling
WA2	WA squared
WE2	WE squared
WAWE	WA×WE
WA3	WA cubed
WE3	WE cubed
WA2WE	WA2×WE
WAWE2	WA×WE2
WFED	woman's farther's years of schooling
WMED	woman's mother's years of schooling
UN	unemployment rate in county of residence
CIT	1 if she lives in SMSA
PRIN	(family income - wage×hours)/1000

Table 1a
Size and power of the tests for $\rho = 0$ at 5% level

	$N = 200$				$N = 400$			
$\rho_0 \backslash \pi_0$	0.0	0.5	0.8	1.0	0.0	0.5	0.8	1.0
t^2_{HG}								
0.0	5.22	4.74	5.36	13.62	5.02	4.40	5.20	9.60
0.2	11.02	8.14	5.62	13.42	19.88	13.72	6.48	9.78
0.4	45.54	27.96	8.22	13.48	75.96	51.14	12.92	10.68
0.6	90.50	68.46	15.42	14.06	99.74	93.82	27.14	12.04
0.8	99.30	91.02	23.50	15.08	100.00	99.64	43.50	14.42
t^2_1								
0.0	6.00	5.30	5.14	5.42	5.28	4.58	5.04	5.46
0.2	12.42	9.16	5.46	5.22	20.56	14.32	6.44	5.40
0.4	48.28	29.48	8.16	5.00	76.58	51.94	12.66	6.14
0.6	91.44	70.38	15.20	5.64	99.78	93.96	26.96	7.32
0.8	99.34	91.66	22.94	6.52	100.00	99.68	43.40	8.78
LM_{AE}								
0.0	5.82	5.18	5.08	5.34	5.22	4.52	4.98	5.40
0.2	12.16	9.02	5.38	5.16	20.44	14.26	6.36	5.28
0.4	47.90	29.22	8.04	4.92	76.50	51.72	12.54	6.14
0.6	91.26	70.12	15.00	5.52	99.78	93.92	26.82	7.14
0.8	99.34	91.52	22.68	6.48	100.00	99.68	43.18	8.74
LR								
0.0	6.36	6.62	7.38	7.08	5.24	5.18	5.98	5.46
0.2	13.12	10.30	7.88	7.54	21.04	14.42	7.16	6.42
0.4	51.42	33.36	12.24	10.74	78.92	56.32	17.62	12.36
0.6	95.92	85.10	46.42	37.30	99.92	98.00	73.12	61.80
0.8	100.00	99.86	97.22	95.82	100.00	100.00	99.92	99.88
$Wald$								
0.0	12.68	20.02	42.56	48.52	7.88	11.64	36.10	43.44
0.2	21.54	25.86	43.36	49.56	26.32	24.62	38.98	45.86
0.4	63.84	55.10	54.26	58.26	83.84	69.24	57.70	59.42
0.6	97.90	94.38	85.98	85.64	99.98	99.28	95.08	93.72
0.8	100.00	100.00	99.82	99.64	100.00	100.00	100.00	100.00

Table 1b

Rejection rate due to negative statistics

	$N = 200$				$N = 400$			
$\rho_0 \backslash \pi_0$	0.0	0.5	0.8	1.0	0.0	0.5	0.8	1.0
t^2_{HG}								
0.0	0.00	0.00	0.00	4.48	0.00	0.00	0.00	0.50
0.2	0.00	0.00	0.00	4.20	0.00	0.00	0.00	0.62
0.4	0.00	0.00	0.00	4.14	0.00	0.00	0.00	0.70
0.6	0.00	0.00	0.00	4.34	0.00	0.00	0.00	0.96
0.8	0.00	0.00	0.00	4.88	0.00	0.00	0.00	1.36
$Wald$								
0.0	0.00	0.04	0.08	0.08	0.00	0.00	0.00	0.00
0.2	0.00	0.00	0.06	0.14	0.00	0.00	0.00	0.00
0.4	0.00	0.00	0.02	0.10	0.00	0.00	0.00	0.00
0.6	0.26	0.22	0.36	0.42	0.00	0.00	0.02	0.02
0.8	9.90	9.76	8.20	7.82	1.00	0.98	0.72	0.74

Notes: Figures are computed as $\left(\frac{\text{The number of negative statistics in the replications}}{\text{The number of replications}}\right) \times$ 100.

Table 2a

Size of the tests under non-normal errors of the structural equation: $iid\left(\chi^2_2 - 2\right)/2$

	$N = 200, \rho_0 = 0.0$			
π_0	0.0	0.5	0.8	1.0
t^2_{HG}	4.38	4.62	6.08	13.92
t^2_1	5.28	5.36	5.90	5.14
LM	5.12	5.18	5.86	5.04
LR	91.56	96.58	99.96	100.00
$Wald$	95.30	98.52	99.96	100.00

Table 2b

Rejection rate due to negative statistics

	$N = 200, \rho_0 = 0.0$			
π_0	0.0	0.5	0.8	1.0
t^2_{HG}	0.00	0.00	0.00	4.74
$Wald$	87.66	79.92	85.40	99.22

Notes: see notes to Table 1b.

Table 3
Estimation results by Heckman two-step method and Nawata's scanning maximum likelihood method

	Heckman two-step		Nawata's ML	
	Estimates	Standard errors	Estimates	Standard errors
constant	-3.24850	(13.77500)	8.19215	(11.63580)
KL6	-0.43228	(0.39578)	0.21624	(0.10569)
KL618	-0.06862	(0.04292)	-0.01939	(0.03278)
WA	0.40627	(0.66249)	-0.14703	(0.55151)
WE	-0.47387	(1.53412)	-1.26841	(1.45407)
WA2	-0.01180	(0.01293)	-0.00146	(0.01074)
WE2	-0.00555	(0.08393)	0.02994	(0.08065)
WAWE	0.02302	(0.03334)	0.03939	(0.03207)
WA3	0.00007	(0.00009)	0.00001	(0.00008)
WE3	0.00206	(0.00183)	0.00139	(0.00179)
WA2WE	0.00010	(0.00027)	0.00001	(0.00027)
WAWE2	-0.00131	(0.00072)	-0.00169	(0.00066)
WFED	-0.02108	(0.01410)	-0.01270	(0.01300)
WMED	-0.00740	(0.01381)	-0.00979	(0.01362)
UN	-0.00534	(0.01300)	0.00347	(0.01207)
CIT	0.09248	(0.08165)	0.07473	(0.07971)
PRIN	-0.00583	(0.01033)	0.00981	(0.00376)
$\rho\sigma_2$	0.62905	(0.78086)	—	—
ρ	0.78746	—	-0.80000	(0.05406)
σ_2	0.79883	—	0.80258	(0.04342)

Table 4
Test results for sample selection bias

	Statistic	p-value
t_{HG}^2	0.649	[0.4205]
t_1^2	0.652	[0.4195]
LM	0.679	[0.4099]
LR	8.981	[0.0027]
$Wald$	219.029	[0.0000]