BOOTSTRAP INFERENCES IN HETEROSCEDASTIC SAMPLE SELECTION MODELS: A MONTE CARLO INVESTIGATION

LEE C. ADKINS AND R. CARTER HILL

Abstract. Several methods of estimating heteroscedastic sample selection models are presented and the sampling properties of bootstrap estimators of test statistics are studied using Monte Carlo simulations.

1. Introduction

Many researchers use Heckman's (1979) 2-step estimator to estimate the parameters of a linear regression model whose dependent variable is only observed when the latent variable in a 'selection' model is positive. While the 2-step estimator is easy to implement, computation of the proper standard errors and the resulting t-statistics and confidence intervals is not as simple. This problem was investigated by Hill, Adkins & Bender (2003).

If the errors of the selection equation, the regression equation, or both are heteroscedastic, it is well-known that the usual two-step and maximum likelihood estimators are inconsistent. Donald (1995) has studied this problem and suggested a semiparametric estimator that is consistent in heteroscedastic selectivity models. Chen & Khan (2003) has also proposed a semiparametric estimator of this model. More recently, Lewbel (2003) has proposed an alternative that is both easy to implement and robust to heteroscedastic misspecification of unknown form.

In this paper we propose a simple estimator that is easily computed using standard regression software and study its performance in a small set of Monte Carlo simulations. To obtain test statistics we use the bootstrap to estimate standard errors and use these to form test statistics that are close to being pivotal when heteroscedasticity is accounted

Date: November 15, 2004.

Key words and phrases. Heckit, heteroscedasticity, bootstrap, sample selection bias.

Presented at the 74th Meeting of the Southern Economic Association Meetings, New Orleans, LA, 22 November 2004.

for in the regression equation. Unlike Fernandez-Sainz, Rodriguez-Poo & Martin (2002), who consider heteroscedasticity in the selection equation, our focus is on heteroscedasticity in the regression equation.

2. The Selectivity Model

Following Greene (1997) consider a model consisting of two equations. The first is the "selection equation," defined

$$z_i^* = w_i'\gamma + u_i, \quad i = 1, \ldots, N \tag{2.1}$$

where z_i^* is a latent variable, γ is a $Kx1$ vector of parameters, w_i' is a $1xK$ row vector of observations on K exogenous variables, and u_i is a random disturbance. The latent variable is unobservable, but we do observe the dichotomous variable

$$z_i = \begin{cases} 1 & z_i^* > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

The second equation is the linear model of interest. It is

$$y_i = x_i'\beta + e_i, \quad i = 1, \ldots, n, \quad N > n \tag{2.3}$$

where y_i is an observable random variable, β is an $Mx1$ vector of parameters, x_i' is a $1xM$ vector of exogenous variables, and e_i is a random disturbance. It is assumed that the random disturbances of the two equations are distributed as

$$\begin{bmatrix} u_i \\ e_i \end{bmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho\sigma_e(q_i) \\ \rho\sigma_e(q_i) & \sigma_e^2(q_i) \end{pmatrix} \right] \tag{2.4}$$

where q_i is a $Px1$ vector of independent variables that 'cause' the regression part of the model to be heteroscedastic. A selectivity problem arises when y_i is observed only when $z_i = 1$ and $\rho \neq 0$. In this case the ordinary least squares estimator of β in (2.3) is biased and inconsistent. In the homoscedastic case, $\sigma_e^2(q_i) = \sigma_e^2, i = 1, \ldots, n$, a consistent estimator has been suggested by Heckman (1979) and is commonly referred to as Heckman's two-step estimator, or more simply, *Heckit*. The presence of heteroscedasticity in the errors of the regression complicates things, however.

The conditional mean of y_i given that it is observed is

$$E[y_i | z_i > 0] = x_i'\beta + \lambda_i\beta_{\lambda i} \tag{2.5}$$

where $\lambda_i = \phi(w_i'\gamma)/\Phi(w_i'\gamma)$ is the *inverse Mill's ratio*, $\phi(\cdot)$ is the standard normal probability density function evaluated at the argument, $\Phi(\cdot)$ is the cumulative density function of the standard normal random

variable evaluated at the argument and $\beta_{\lambda i} = \rho\sigma_e(q_i) \equiv \rho\sigma_{ei}$. Adding a random disturbance yields:

$$y_i = x_i'\beta + \lambda_i\beta_{\lambda i} + \nu_i \tag{2.6}$$

The random disturbance ν_i had conditional mean and variance given by

$$\begin{aligned} E[\nu_i|z_i > 0] &= 0 \\ var(\nu_i|z_i > 0) &= \sigma_{ei}^2(1 - \rho^2\delta_i) \end{aligned} \tag{2.7}$$

where $\delta_i = \lambda_i(\lambda_i + w_i'\gamma)$. The implication is that (2.6) is heteroscedastic.

Even if λ_i were known and nonstochastic, then selectivity corrected model (2.6) could not be estimated by generalized least squares. This follows from the fact that heteroscedasticity in the regression equation introduces n terms to estimate into the second step of Heckman's estimator, i.e., $\beta_{\lambda i}$. Solving this problem holds the key to the estimators used below.

In this particular parameterization the unknown λ_i can be estimated using the sample in the usual way. As in the usual selectivity model, the stochastic nature of λ_i in (2.6) makes the automatic use of White's heteroscedasticity consistent covariance estimator in the second step uncertain in this context.

3. Estimators

There have been a number of interesting estimators proposed that can tackle this problem. Most of them rely on semiparametric of $\beta_{\lambda i}$. Ahn & Powell (1993) propose a two-step estimator that uses nonparametric regression to estimate the selection equation followed by a weighted instrumental variables estimator for the regression equation. Although the estimator is robust to some distributional misspecification, it is not robust to the presence of heteroscedasticity. Chen & Khan (2003) extend the Ahn and Powell estimator for use in heteroscedastic models by introducing another step that is carried out between the two steps of the Ahn and Powell estimator. Thus, the first step consists of nonparametric estimation of the so-called propensity score (i.e., the probability of being selected given the values of w_i in the selection equation). The second step requires the estimation of two quantile functions for the regression equation using local polynomial estimators and taking their difference to obtain the so-called 'interquartile range.' The third stage is the same as Ahn and Powell's second step. There are similarities between the Chen and Khan estimator and the one proposed by Donald (1995), who uses just two steps, both nonparametric regressions.

In deriving a test of the normality assumption commonly used in selectivity models der Klaauw and Koning (2003) use a semiparametric method inspired by Gallant and Nychka (1987). They approximate the the unknown density of the model's errors using a flexible parametric form based on Hermite series. Based on the results from simulations, this flexible parametric form is found to be useful in reducing bias attributable to heteroscedastic errors.

Lewbel (2004) suggests using GMM estimation to efficiently estimate the parameters of the standard selectivity model and extends that idea to models that contain an endogenous regressor in the regression equation. A simpler estimator is also proposed that requires what Lewbel refers to as a 'very exogenous regressor' that provides the basis for linear two stage least squares estimation of a sample selection model with endogenous regressors. An attractive feature of Lewbel's estimator is that the selection equation does not have to be estimated. Lewbel states that these simple estimators are in fact special cases of nonparametric estimators in Lewbel (2003).

Fernandez-Sainz et al. (2002) study the finite sample behavior of usual parametric Heckit and Ahn and Powell's (1993) semiparametric two step estimators of homoscedastic sample selection models when the errors of the selection equation are actually heteroscedastic. Note, in their study both estimators are being used for misspecified models. They find that even though the semiparametric estimator is robust to heteroscedasticity in the selection equation, the usual Heckit performs better than the semiparametric estimator in finite samples in most cases, especially for highly truncated samples and those where the level of correlation between errors is high.

In a recent study Jolliffe (2002) uses a technique suggested by Honoré, Kyriazidou & Udry (1997) for the estimation of heteroscedastic Type III Tobit (T3T) models. In Type III Tobit models, the dependent variable in the selection equation is censored rather than binary. Many applications of the Heckit procedure could actually be estimated as T3T, but the censored variable is transformed into a dummy variable (referred to as a Type II Tobit). In the Honoré et al. approach trimmed least squares can be used in both steps to ensure ensure robustness against heteroscedasticity in either the selection or regression equations. The Jolliffe paper is interesting because he in fact obtains standard errors and percentiles of the bias corrected empirical density. This is evidence that the Type III Tobit-Trimmed Least Squares estimator he uses can be automated.

Quite frankly, the semiparametric and nonparametric estimators are difficult to use in practice. They rely on kernel estimation in all cases,

quartile regression in some, and as (Fernandez-Sainz et al. 2002) have shown, may not outperform the usual Heckit estimator in a misspecified model. Their absence in canned software programs means that they rely on user written software. The difficulties involved with nonparametric and semiparametric approaches are no doubt responsible for their limited use. A search of the Social Sciences Citation Index reveals no references to applications of Donald (1995). The specific number of applications of Ahn & Powell (1993) is more difficult to determine. There are 53 total citations, most of these are either theoretical in nature (e.g., Chen & Khan (2003)) or only peripherally related (e.g., Gordon (2003)). One exception is a recent paper by Blundell, Reed & Stoker (2003) which appears to use the Ahn and Powell estimator.

4. Proposed Estimators

The estimators proposed in this paper are simple ones and we make no theoretical claims about their consistency. On the other hand, nonparametric estimators that are known to have good asymptotic properties are devilishly difficult to use by common practitioners and hence, are not widely used. So, we employ an intuitively appealing estimators that are easy to compute and study their small sample properties in a small Monte Carlo exercise. In some respects this is similar in spirit to Fernandez-Sainz et al. (2002) who use estimators that are likewise misapplied.

We use the fact that the selectivity parameter in the regression model, $\beta_{\lambda i}$ is a function of the heteroscedasticity. We then use all available sample information, x_i and w_i to model changing variances. For simplicity, we use a second order polynomial to approximate this function, that is, we take all unique variables in x_i and w_i along with their squares and cross products to construct regressors q_i. Then,

$$\beta_{\lambda i} = q_i'\gamma \tag{4.1}$$

This is substituted into (2.6) where it interacts with each λ_i. The model becomes

$$y_i = x_i'\beta + \hat{\lambda}_i q_i'\gamma + \nu_i \tag{4.2}$$

where $\hat{\lambda}_i$ is estimated using probit estimation in the first stage, multiplied times each element of q_i, and OLS is used on (**??**) in the second stage.

The advantage of this estimator is that it is linear and is easy to compute and to bootstrap. The disadvantage is that it offers no elegant way to test the selectivity hypothesis. As an alternative, we also

consider modeling

$$\beta_{\lambda i} = \rho\{\sigma^2 \exp(q_i'\alpha)\}^{\frac{1}{2}} = \exp(q_i'\gamma)\beta_\lambda \tag{4.3}$$

which yields the model

$$y_i = x_i'\beta + \hat{\lambda}_i \exp(q_i'\gamma)\beta_\lambda + \nu_i \tag{4.4}$$

which is estimated using nonlinear least squares. This form is particularly useful for testing various hypotheses about the model. If $\gamma = 0$, then the model reduces to the usual homoscedastic selectivity model. If $\beta_\lambda = \rho\sigma = 0$ then there is no sample selection problem and OLS can be consistently used for the parameters of the regression function. The biggest problem with this model is that it requires nonlinear estimation. In particular, the estimator is prone to nonconvergence and can be sensitive to starting values. In any given application, this is not so much of a problem. It does make a simulation study of the estimator difficult, however. This problem is compounded if a bootstrapping layer is added to the Monte Carlo.

5. Experimental Design

The Monte Carlo we employ is very similar to that of Hill et al. (2003) which is based on Zuehlke & Zeman (1991), and modified by Nawata & Nagase (1996). Heteroscedasticity in the regressions errors are then added. The performance of two-step estimator of a regression slope parameter is examined under various circumstances likely to affect performance. The sample size, severity of censoring, degree of selection bias, correlation between independent variables in the selection and regression equations are all varied within the Monte Carlo.

The vector of explanatory variables, x_i contains a constant and one continuous explanatory variable. Likewise, w_i, the vector or explanatory variables in the first stage selection equation also contains a constant and 1 explanatory variable. Thus, the selection equation and regression equation are

$$z_i^* = \gamma_1 + \gamma_2 w_i + u_i \tag{5.1}$$

$$y_i = \beta_1 + \beta_2 x_i + e_i \tag{5.2}$$

Another variable, v_i is introduced that is correlated with the regressors, x_i of the second equation. These are used to generate the desired heteroscedasticity. Our goal in creating a new variable that is only partially correlated with regressors is to purposely omit an offending variable in our estimator of heteroscedasticity. Indeed, these variables are seldom known to the researcher and it is unlikely that all variables

that cause heteroscedasticity will be available to the user. As the correlation between the x_i and v_i diminish, we would expect a deterioration in the performance of the proposed estimators.

The correlation between x_i and w_i. denoted ρ_{xw}, is set at 0, .5, or .95. For identification, it is seldom a good idea for this correlation to be perfect (as in $w_i = x_i$). The Monte Carlo evidence in Hill et al. (2003) speaks to this. In addition, many of the theoretical papers in this literature use restrictions to identify parameters in the regression (e.g., (Lewbel 2003), Ahn & Powell (1993), and Fernandez-Sainz et al. (2002)). The severity of censoring is controlled by varying the constant, γ_1, in the first stage probit model. Following Zuehlke & Zeman (1991), γ_1=[-.96 0 .96] which corresponds to expected subsamples of 25%, 50%, and 75% given u_i iid $N(0,1)$. The degree of selection bias is controlled by deviating ρ, which takes on values of 0, .5, and .99. The other parameter of the selection equation, $\gamma_2 = 1$. The specification is completed by the following parameter choices: $\beta' = [100\ 1]$ and $\sigma_e = 1$.

Heteroscedasticity enters the regression equation as

$$\sigma_i^2 = \exp\left(\alpha_1 + \alpha_2 x_i + \alpha_3 w_i + \alpha_4 v_i\right) \tag{5.3}$$

$\alpha = \{0\ 0.3\ 0.3\ 0.2\}$ and v_i is chosen to be correlated with x_i but otherwise omitted from either x_i and w_i. This value of α was chosen to keep the overall mean variance in the equation close to 1. The correlation between v_i and x_i is set to -0.5.

Our goal in this paper is to develop reliable means for testing hypotheses about the slope parameters in a heteroscedastic regression equation that is incidentally truncated. Specifically, we we test the null hypothesis Ho: $\beta_2 = 0$ against the alternative $\beta_2 \neq 0$ using the t-statistic $t = (\hat{\beta}_2 - \beta_2)/\hat{\sigma}_{\hat{\beta}_2}$ where $\hat{\beta}_2$ is one of our proposed estimators of β_2 and $\hat{\sigma}_{\hat{\beta}_2}$ is an estimated standard error for $\hat{\beta}_2$. The statistic is computable in the Monte Carlo since the true value of β_2 is completely known in the experiments. If the distribution of t is symmetric, the null hypothesis is rejected if $|t| \geq t_c$ where t_c is the $\alpha/2$ critical value from the distribution of t. When t is not symmetric, then the hypothesis is rejected if $t \leq t_{lc}$ or $t \geq t_{uc}$ where t_{lc} and t_{uc} are the lower and upper $\alpha/2$ critical values, respectively, from the distribution of t. The problem here is that the exact or approximate distribution of t is not known in this model. Even if a consistent estimator of β_2 is found, a consistent estimator of $\hat{\sigma}_{\hat{\beta}_2}$ is unavailable due the heteroscedasticity in the errors of the regression.

Specifically, in each Monte Carlo sample m we resample, with replacement, from the rows of the data matrix y, x, z, w to obtain a bootstrap sample $\hat{\beta}_b$ of size N. A pivotal statistic is obtained for each of

400 bootstrap samples by computing the t-statistic value for the (true in the sample) hypothesis Ho: $\beta_2 = 0$ against the alternative $\beta_2 \neq 0$. That is, we compute for each bootstrap sample $t_b = (\hat{\beta}_b - \hat{\beta}_m)/\hat{\sigma}_{\hat{\beta}_b}$ where $\hat{\beta}_b$ is estimate of β_2 from the i^{th} bootstrap sample and $\hat{\beta}_m$ is the estimate from the underlying m^{th} iteration of the monte carlo; $\hat{\sigma}_{\hat{\beta}_b}$ is the sample standard error from the bootstrap iterations. The values of these t-statistics are sorted by magnitude and the lower and upper t-critical values t_{lc} and t_{uc} are chosen to be the lower and upper 5% values. We will report the size of the rejection region from these tests in the tables that follow.

A total of 400 samples are drawn using each of the combinations of the parameter above for sample size of 400. Where employed, the number of bootstrap samples is 400.

6. Results

In the first set of experiments, $\alpha = 0$ which makes the regression equation homoscedastic. Two different estimators are examined. In column 4 the usual Heckit results appear. In column 5, the results for the linear heteroscedastic estimator that uses x, w and their cross product appear. In column 6 results appear for the linear heteroscedastic Heckit model that includes x, w, and their cross products and squares. In the last column, the nonlinear Heckit model results are tabled where w, x, and their cross product appears in the heteroscedasticity function. The first 3 columns describe the basic experimental design. The degree of truncation varies between 75%, 50%, and 25%. Collinearity describes the degree of linear association between the regressors of the selection equation and of the regression. Correlation refers to ρ in (2.4) which is the degree of correlation between the errors of the two equations.

In Table 1, bias results are reported for each of the 27 designs. In Table 2 are the t-ratios computed based on the Monte Carlo standard errors. These results can be used to indicate whether the biases reported in Table 1 are statistically significant or not.

It is clear from these two tables that bias is small and in nearly every instance, not significantly different from zero at the 5% level.

The results from the next set of experiments appear in Tables 3 and 4. Heteroscedasticity is introduced in these designs by setting $\alpha = \{0\ 0.3\ 0.3\ 0.2\}$ and v_i was generated to be correlated with x_i (but not included in w_i or x_i. The same set of estimators employed above were used. That is, in column 4 the usual Heckit results appear. In column 5 (`Het x,w`), the results for the linear heteroscedastic estimator that uses x, w and their cross product appear. In column 6 (`Het2 x,w`)

TABLE 1. Estimated Bias of Estimators: Homoscedastic Errors

Design			Bias			
			(4)	(5)	(6)	(7)
Trunc	Collin	Corr	Heckit	Het x,w	Het2 x,w	Nlin w,w
0.75	0.00	0.00	0.0016	0.0038	0.0081	0.0097
0.75	0.50	0.00	0.0031	0.0162	0.0174	0.0058
0.75	0.95	0.00	0.0009	0.0047	0.0696	0.0015
0.75	0.00	0.50	-0.0012	-0.0147	-0.0136	-0.0087
0.75	0.50	0.50	0.0015	-0.0048	-0.0060	-0.0056
0.75	0.95	0.50	0.0103	-0.0045	0.0337	0.0236
0.75	0.00	0.99	0.0010	0.0047	0.0043	-0.0046
0.75	0.50	0.99	-0.0049	-0.0107	0.0043	-0.0230
0.75	0.95	0.99	0.0094	0.0103	0.0477	0.0196
0.50	0.00	0.00	-0.0018	0.0097	0.0088	0.0017
0.50	0.50	0.00	-0.0041	-0.0049	-0.0047	-0.0080
0.50	0.95	0.00	0.0129	0.0301	0.0620	0.0270
0.50	0.00	0.50	0.0030	0.0158	0.0143	0.0057
0.50	0.50	0.50	-0.0013	-0.0057	-0.0028	-0.0066
0.50	0.95	0.50	0.0057	0.0102	0.0242	0.0312
0.50	0.00	0.99	0.0012	-0.0094	-0.0076	-0.0050
0.50	0.50	0.99	0.0019	0.0054	0.0028	-0.0066
0.50	0.95	0.99	-0.0012	-0.0035	-0.0087	0.0035
0.25	0.00	0.00	-0.0017	0.0007	0.0008	-0.0019
0.25	0.50	0.00	-0.0051	-0.0006	-0.0048	-0.0056
0.25	0.95	0.00	0.0035	-0.0048	0.0192	-0.0086
0.25	0.00	0.50	-0.0011	0.0054	0.0050	0.0004
0.25	0.50	0.50	-0.0013	0.0036	0.0024	0.0021
0.25	0.95	0.50	-0.0014	-0.0067	-0.0170	0.0171
0.25	0.00	0.99	0.0023	-0.0032	-0.0029	-0.0028
0.25	0.50	0.99	-0.0048	-0.0154	-0.0125	-0.0206
0.25	0.95	0.99	0.0077	0.0032	0.0012	0.0251

results appear for the linear heteroscedastic Heckit model that includes x, w, and their cross products and squares. In the last column, the nonlinear Heckit model results are tabled where w, x, and their cross product appears in the heteroscedasticity function.

The biases in Table 3 are all relatively small, but significantly different from zero at the 5% level in 15 of the 27 designs according to the t-ratios in Table 4. For the linear heteroscedastic estimators things

TABLE 2. T-ratios for Estimated Bias: Homoscedastic Errors

Design			T-ratio			
			(4)	(5)	(6)	(7)
Trunc	Collin	Corr	Heckit	Het x,w	Het2 x,w	Nlin w,w
0.75	0.00	0.00	0.3457	0.1941	0.4057	1.0243
0.75	0.50	0.00	0.5594	0.7830	0.7107	0.6177
0.75	0.95	0.00	0.0611	0.1618	1.3924	0.0542
0.75	0.00	0.50	-0.2938	-0.7799	-0.7049	-0.8855
0.75	0.50	0.50	0.2949	-0.2383	-0.2496	-0.5106
0.75	0.95	0.50	0.8036	-0.1719	0.6719	0.9415
0.75	0.00	0.99	0.3160	0.3187	0.2819	-0.4422
0.75	0.50	0.99	-1.4496	-0.6250	0.2072	-2.0885
0.75	0.95	0.99	0.9994	0.4601	1.0992	0.8439
0.50	0.00	0.00	-0.5386	1.0698	0.9459	0.2875
0.50	0.50	0.00	-1.1441	-0.4775	-0.3961	-1.3184
0.50	0.95	0.00	1.4332	1.9878	2.4113	1.7897
0.50	0.00	0.50	0.9264	1.8536	1.6643	0.8661
0.50	0.50	0.50	-0.3606	-0.5545	-0.2304	-0.9908
0.50	0.95	0.50	0.7155	0.7006	1.0513	2.0876
0.50	0.00	0.99	0.4816	-1.2845	-1.0200	-0.7560
0.50	0.50	0.99	0.7310	0.6504	0.2713	-0.8987
0.50	0.95	0.99	-0.1787	-0.2650	-0.4004	0.2516
0.25	0.00	0.00	-0.6126	0.1304	0.1419	-0.4603
0.25	0.50	0.00	-1.7504	-0.1085	-0.6864	-1.1918
0.25	0.95	0.00	0.5478	-0.4937	1.3701	-0.8155
0.25	0.00	0.50	-0.4272	1.0442	0.9742	0.0843
0.25	0.50	0.50	-0.4134	0.5934	0.3536	0.4093
0.25	0.95	0.50	-0.2445	-0.7248	-1.3487	1.8497
0.25	0.00	0.99	0.9975	-0.6470	-0.5970	-0.5987
0.25	0.50	0.99	-1.8122	-2.8356	-1.9797	-4.1124
0.25	0.95	0.99	1.4082	0.3776	0.0945	2.6999

improve. For `Het x,w` and `Het2 x,w` only 2 of 27 are significantly biased. For the nonlinear heteroscedastic model only w, x and the cross product is included in the heteroscedasticity function; the result is that 5 of 27 are biased.[1]

[1] The experiments were repeated for nonlinear heteroscedastic model with squared and cross products of w and x added to the model. This slowed the Monte Carlo simulations to a crawl and the estimation was terminated after a few days which yielded results for only the first 5 designs. These results, though limited, were no better, if slightly worse than those for the nonlinear model above. At this

TABLE 3. Estimated Bias of Estimators: Heteroscedastic Errors

Design			Bias			
			(4)	(5)	(6)	(7)
Trunc	Collin	Corr	Heckit	Het x,w	Het2 x,w	Nlin w,w
0.75	0.00	0.00	0.0022	-0.0049	0.0007	0.0058
0.75	0.50	0.00	0.0044	0.0251	0.0131	0.0134
0.75	0.95	0.00	0.0142	0.0555	0.0891	0.0424
0.75	0.00	0.50	0.0451	-0.0065	0.0034	-0.0224
0.75	0.50	0.50	0.0509	0.0081	0.0143	0.0039
0.75	0.95	0.50	0.0009	-0.0073	-0.0333	-0.0072
0.75	0.00	0.99	0.0862	0.0199	0.0166	0.0592
0.75	0.50	0.99	0.0922	-0.0299	0.0065	0.0113
0.75	0.95	0.99	-0.0007	0.0461	0.0558	0.0973
0.50	0.00	0.00	-0.0029	0.0063	0.0022	0.0031
0.50	0.50	0.00	-0.0052	-0.0052	-0.0005	-0.0072
0.50	0.95	0.00	0.0264	0.0619	0.0702	0.0744
0.50	0.00	0.50	0.0322	0.0235	0.0193	0.0140
0.50	0.50	0.50	0.0217	-0.0059	0.0000	-0.0075
0.50	0.95	0.50	-0.0109	0.0237	0.0403	0.0287
0.50	0.00	0.99	0.0545	-0.0148	-0.0019	-0.0001
0.50	0.50	0.99	0.0527	0.0048	0.0118	0.0002
0.50	0.95	0.99	-0.0523	-0.0155	-0.0041	-0.0078
0.25	0.00	0.00	0.0000	0.0024	0.0039	-0.0006
0.25	0.50	0.00	-0.0048	-0.0004	-0.0040	-0.0052
0.25	0.95	0.00	0.0016	-0.0007	0.0122	0.0063
0.25	0.00	0.50	0.0130	0.0070	0.0073	-0.0062
0.25	0.50	0.50	0.0077	0.0025	0.0013	-0.0041
0.25	0.95	0.50	-0.0112	-0.0015	-0.0028	0.0120
0.25	0.00	0.99	0.0266	-0.0014	0.0002	-0.0026
0.25	0.50	0.99	0.0178	-0.0136	-0.0063	-0.0152
0.25	0.95	0.99	-0.0252	0.0095	0.0105	0.0437

Sizes of 10% nominal tests based on the usual Heckit estimator and the linear heteroscedastic estimators are computed in the Monte Carlo. The results are based on 400 Monte Carlo samples and 400 bootstrap

point we cannot recommend its use. If one is determined to test a hypothesis about the occurrence of selectivity, then this estimator may still be the best way to go. The key to its use is to keep the nonlinear heteroscedasticity function as simple as possible.

TABLE 4. T-ratios for Estimated Bias: Heteroscedastic Errors

Design			T-ratio			
			(4)	(5)	(6)	(7)
Trunc	Collin	Corr	Heckit	Het x,w	Het2 x,w	Nlin w,w
0.75	0.00	0.00	0.3860	-0.1915	0.0433	0.4285
0.75	0.50	0.00	0.6275	0.7719	0.5191	0.8672
0.75	0.95	0.00	0.6773	1.1529	1.4149	0.8026
0.75	0.00	0.50	8.8447	-0.2677	0.2302	-1.7919
0.75	0.50	0.50	8.0293	0.2668	0.6113	0.2505
0.75	0.95	0.50	0.0473	-0.1626	-0.5551	-0.1404
0.75	0.00	0.99	22.2382	1.0403	1.3980	4.1777
0.75	0.50	0.99	19.3049	-0.9980	0.2842	0.6634
0.75	0.95	0.99	-0.0451	1.0697	0.9249	2.0524
0.50	0.00	0.00	-0.7422	0.5358	0.2514	0.3931
0.50	0.50	0.00	-1.1850	-0.3817	-0.0426	-0.8172
0.50	0.95	0.00	2.0679	2.5958	2.2879	2.4170
0.50	0.00	0.50	8.7451	2.1599	2.3517	1.7501
0.50	0.50	0.50	4.9041	-0.4024	-0.0032	-0.8157
0.50	0.95	0.50	-0.9120	0.9999	1.2892	0.9649
0.50	0.00	0.99	18.0333	-1.4621	-0.2565	-0.0124
0.50	0.50	0.99	15.3009	0.3546	1.0631	0.0205
0.50	0.95	0.99	-4.5644	-0.6646	-0.1346	-0.2704
0.25	0.00	0.00	0.0120	0.3487	0.6733	-0.1129
0.25	0.50	0.00	-1.4467	-0.0511	-0.5939	-0.8480
0.25	0.95	0.00	0.1854	-0.0492	0.7456	0.3658
0.25	0.00	0.50	4.4851	1.0983	1.3650	-1.1821
0.25	0.50	0.50	2.1725	0.3153	0.2005	-0.6060
0.25	0.95	0.50	-1.3771	-0.1154	-0.1832	0.6989
0.25	0.00	0.99	10.1242	-0.2242	0.0369	-0.4541
0.25	0.50	0.99	5.6186	-1.7848	-0.9582	-2.2347
0.25	0.95	0.99	-3.1202	0.7441	0.6857	2.6133

samples. To compute valid bootstrap test statistics we require consistent estimators of the parameters, namely β_2 and its estimated standard error. Based on the initial experiments the linear heteroscedastic estimator may satisfy this. If heteroscedasticity is not too severe or not completely unrelated to x then it is unbiased and it appears from initial experiments that its variance shrinks as sample size increases. Consistent estimation of the standard error is more problematic. Since there is no known heteroscedastic consistent covariance for this estimator the bootstrap standard error of the estimates themselves is used.

That is,

$$\hat{\sigma}^2_{b_s} = \frac{1}{N}\sum(\hat{\beta}_{bi} - \overline{\hat{\beta}_{bi}})^2 \tag{6.1}$$

Bootstrapping the nonlinear model is simply not feasible in the context of the monte carlo; with a 3-GHz Pentium 4 with 2 gigabytes of ram we estimated over 200 days to complete all 27 designs! Hence, we focus on test statistics for the linear models, which as it turns out, behave reasonably well in our simulations.

In Table 5 the sizes of nominal 10% t-tests of the hypothesis that the regression slope is equal to zero are reported. The data are homoscedastic and the usual Heckit estimator is asymptotically unbiased and the usual t-test should be pivotal and be approximately normally distributed.

As can be seen, the usual Heckit estimator in columns (4) and (5) results in tests that are close to the nominal 10% size. In some cases bootstrapping the standard error of the homoscedastic Heckit helps (column 5) and sometimes not (column 4). In the last row, the average distance between the actual test size and the nominal size (0.1) is given. According to the summary measure, the results in column 5 that are based on bootstrap standard errors are closer to nominal 10% level than the usual asymptotic test. The test sizes for the heteroscedastic estimators are not quite as close to the nominal 10% level as the others, but their performance–even when the model is homoscedastic–is reasonably good.

In Table 6, the experiments were repeated with heteroscedastic errors. In this case the heteroscedasticity is generated according to equation (5.3) with $\alpha = \{0\ 0.3\ 0.3\ 0.2\}$.

In this case, the test sizes for the proposed estimators are very close to the nominal test size. Clearly the heteroscedastic estimator and the bootstrap test outperforms the usual tests, especially when correlation between the 2 model's errors is high. The data are fairly heteroscedastic in this design. In a random sample, the variances of the regression errors ranged from .25 to 6.5.

Another round of experiments were conducted that increased the degree of heteroscedasticity. These results appear in Table 7. In this case, α in (5.3) was doubled to $\alpha = \{0\ 0.6\ 0.6\ 0.4\}$. The purpose is to show the large size distortion associated with doing nothing about heteroscedasticity vs. accounting for it in our admittedly *ad hoc* way. The results for the various t-tests appear in the table above. In this case, and v_i was generated to be correlated with x_i (but not included in w_i or x_i. With $\alpha = \{0\ 0.6\ 0.6\ 0.4\}$, the data are highly heteroscedastic. For instance, variances in a typical sample ranged in value from

TABLE 5. Size of 10% t-tests: Homoscedastic Errors

Design			asy t	Bootstrap		
			(4)	(5)	(6)	(7)
Trunc	Collin	Corr	Heckit	Heckit	Het x,w	Het2 x,w
0.75	0.00	0.00	0.1025	0.0875	0.0400	0.0425
0.75	0.50	0.00	0.1175	0.1100	0.0700	0.0600
0.75	0.95	0.00	0.1075	0.1200	0.0700	0.0475
0.75	0.00	0.50	0.0950	0.1075	0.0600	0.0550
0.75	0.50	0.50	0.1025	0.0925	0.0700	0.0650
0.75	0.95	0.50	0.0775	0.0850	0.0850	0.0600
0.75	0.00	0.99	0.1025	0.1100	0.0725	0.0750
0.75	0.50	0.99	0.1075	0.1050	0.1000	0.1000
0.75	0.95	0.99	0.0725	0.1175	0.1300	0.0900
0.50	0.00	0.00	0.1150	0.1150	0.0800	0.0725
0.50	0.50	0.00	0.1075	0.1050	0.0850	0.0800
0.50	0.95	0.00	0.0675	0.0825	0.0875	0.1075
0.50	0.00	0.50	0.0975	0.1050	0.0900	0.1000
0.50	0.50	0.50	0.1175	0.1175	0.0975	0.0875
0.50	0.95	0.50	0.0875	0.1025	0.1125	0.1000
0.50	0.00	0.99	0.0875	0.0800	0.0675	0.0600
0.50	0.50	0.99	0.0725	0.0725	0.0675	0.0650
0.50	0.95	0.99	0.0825	0.1200	0.1250	0.0950
0.25	0.00	0.00	0.1100	0.1250	0.0825	0.0775
0.25	0.50	0.00	0.1050	0.1050	0.0850	0.0925
0.25	0.95	0.00	0.0700	0.1150	0.1175	0.0850
0.25	0.00	0.50	0.1225	0.1175	0.0925	0.0850
0.25	0.50	0.50	0.0750	0.0800	0.0725	0.0875
0.25	0.95	0.50	0.0800	0.1100	0.0850	0.0675
0.25	0.00	0.99	0.0850	0.0875	0.0875	0.0825
0.25	0.50	0.99	0.1075	0.1150	0.0775	0.0750
0.25	0.95	0.99	0.0400	0.0775	0.0625	0.0525
Avg distance from .1			0.0203	0.0154	0.0254	0.0291

.07 to 43. In the column 4 are results for the usual t test based on the asymptotic covariance of the usual Heckit estimator. As the correlation between the equations' errors increases, the test size becomes highly distorted. This is to be expected given the large degree of bias in the usual estimator under heteroscedasticity. The bootstrap procedures improve things remarkably. Even when the bootstap is used to obtain the standard error of the usual Heckit estimator, things improve

TABLE 6. Size of 10% t-tests: Heteroscedastic Errors

Design			asy t		Bootstrap	
			(4)	(5)	(6)	(7)
Trunc	Collin	Corr	Heckit	Heckit	Het x,w	Het2 x,w
0.75	0.00	0.00	0.1100	0.1100	0.0525	0.0475
0.75	0.50	0.00	0.1375	0.1225	0.0800	0.0800
0.75	0.95	0.00	0.1750	0.1550	0.1425	0.1175
0.75	0.00	0.50	0.1300	0.1125	0.0550	0.0525
0.75	0.50	0.50	0.1175	0.1225	0.0875	0.0850
0.75	0.95	0.50	0.1225	0.0900	0.1175	0.1050
0.75	0.00	0.99	0.3500	0.3450	0.0825	0.0600
0.75	0.50	0.99	0.2500	0.2325	0.1250	0.1200
0.75	0.95	0.99	0.1850	0.1300	0.1825	0.1775
0.50	0.00	0.00	0.1125	0.1025	0.0675	0.0725
0.50	0.50	0.00	0.1275	0.1150	0.0850	0.0875
0.50	0.95	0.00	0.1425	0.1150	0.1325	0.1450
0.50	0.00	0.50	0.1525	0.1550	0.0825	0.0925
0.50	0.50	0.50	0.1525	0.1425	0.1000	0.0900
0.50	0.95	0.50	0.1825	0.1375	0.1375	0.1525
0.50	0.00	0.99	0.2100	0.1925	0.0800	0.0875
0.50	0.50	0.99	0.2100	0.1875	0.0850	0.0825
0.50	0.95	0.99	0.2275	0.1500	0.1625	0.1475
0.25	0.00	0.00	0.1300	0.1250	0.0925	0.0875
0.25	0.50	0.00	0.1125	0.1000	0.0875	0.0750
0.25	0.95	0.00	0.1750	0.1300	0.1350	0.1100
0.25	0.00	0.50	0.1375	0.1200	0.0825	0.0850
0.25	0.50	0.50	0.0875	0.0775	0.0650	0.0700
0.25	0.95	0.50	0.1725	0.1275	0.1025	0.0850
0.25	0.00	0.99	0.1450	0.1200	0.0800	0.0875
0.25	0.50	0.99	0.1450	0.1375	0.0950	0.0800
0.25	0.95	0.99	0.1800	0.1100	0.1000	0.0775
Avg distance			0.0817	0.0648	0.0314	0.0309

(column 5). There is still large size distortion when truncation is high and correlation between errors is high, but in many cases the bootstrap is able to improve performance of the t-test. The test sizes in column 6 are based on the estimator that uses x, w and their cross product in

TABLE 7. Size of 10% t-tests: Highly Heteroscedastic Errors

Design			asy t		Bootstrap	
			(4)	(5)	(6)	(7)
Trunc	Collin	Corr	Heckit	Heckit	Het x,w	Het2 x,w
0.75	0.00	0.00	0.1250	0.1050	0.0575	0.0575
0.75	0.50	0.00	0.1575	0.1400	0.0975	0.1150
0.75	0.95	0.00	0.2625	0.1825	0.2475	0.2000
0.75	0.00	0.50	0.2300	0.1850	0.0600	0.0700
0.75	0.50	0.50	0.1800	0.1700	0.0850	0.0775
0.75	0.95	0.50	0.1625	0.1150	0.1425	0.1225
0.75	0.00	0.99	0.6475	0.6050	0.0725	0.0675
0.75	0.50	0.99	0.4975	0.4600	0.1475	0.1450
0.75	0.95	0.99	0.3125	0.1550	0.2850	0.2700
0.50	0.00	0.00	0.1225	0.1000	0.0750	0.0700
0.50	0.50	0.00	0.1350	0.1150	0.0800	0.0900
0.50	0.95	0.00	0.2725	0.1500	0.1650	0.2050
0.50	0.00	0.50	0.2200	0.2150	0.0800	0.0800
0.50	0.50	0.50	0.1900	0.1850	0.0975	0.0900
0.50	0.95	0.50	0.3100	0.1400	0.1700	0.2050
0.50	0.00	0.99	0.4100	0.3325	0.0950	0.0825
0.50	0.50	0.99	0.3400	0.2925	0.1025	0.0800
0.50	0.95	0.99	0.3375	0.1650	0.2000	0.2025
0.25	0.00	0.00	0.1550	0.1200	0.0875	0.0850
0.25	0.50	0.00	0.1375	0.1050	0.0875	0.0725
0.25	0.95	0.00	0.2650	0.1400	0.1525	0.1300
0.25	0.00	0.50	0.1650	0.1275	0.0950	0.0975
0.25	0.50	0.50	0.1250	0.0800	0.0600	0.0575
0.25	0.95	0.50	0.3225	0.1300	0.1300	0.1325
0.25	0.00	0.99	0.2300	0.2025	0.0900	0.0900
0.25	0.50	0.99	0.2350	0.1600	0.1075	0.0850
0.25	0.95	0.99	0.3250	0.1175	0.1025	0.1050
Avg distance			0.1963	0.1423	0.0579	0.0563

the linear heteroscedastic Heckit model. In column 7 are the sizes for the the linear heteroscedastic Heckit model that includes x, w, cross products and squares. In each case, the largest size for the nominal 10% test is .20. While high, this is a vast improvement over the usual t-tests that ignore heteroscedasticity altogether.

It is particularly striking how bad the usual tests perform under high levels of truncation and correlation between errors. Surely tests here have high power, but a size near 50% makes them unusable. The bootstrap heteroscedastic tests at their worst have a size near 28%. This occurs when the x and w are highly collinear. With the addition of extra regressors to the model that are also highly collinear, it is not all that surprising that the proposed estimator performs badly here. One can speculate that proper specification of the heteroscedasticity function in the estimator should improve performance.[2]

7. Conclusion

It is pretty clear from the results that ignoring heteroscedasticity in a selectivity model can be dangerous. The usual 2-step Heckit estimator is seriously biased and subsequent t-tests of regression coefficients can suffer from large "size distortion." Although there are several nonparametric and semiparametric estimators for this model, they are not nearly as easy to estimate as the ones proposed here. The estimators here can be computed using standard regression software equipped with what has come to be known as X,Y bootstrap algorithms and includes Limdep and STATA. The performance of the estimators in our limited Monte Carlo exercise is promising. Size distortion is greatly reduced and nearly disappears when the data are not too severely heteroscedastic. When the data are not heteroscedastic, the test still works in the sense that it has the desired size. On the other hand, for point estimation, these estimators leave much to be desired. The linear model adds terms to the usual Heckit that are highly collinear with the original regressors and amongst themselves. This causes precision to suffer, and in most cases, to suffer badly. However, tests based on the more precise, yet badly biased, homoscedastic Heckit are not acceptable.

An obvious weakness of the simple estimators proposed here is that the model to which they are applied is probably not very realistic. To us, it is hard to believe that the regression equation is heteroscedastic, but the underlying selection equation is not. It stands to reason that either both are homoscedastic or neither is. The underlying decisions are being made by the same individual, why would the error variances be different for one kind of decision and not the other? The solution

[2]Actually, we tried this. Recall that the DGP function includes a variable, v that is omitted from estimation. To test this thesis we run another set of experiments where v is omitted from the model (5.3). While results improved some, the improvement was not dramatic enough to convince us that better specification of q would help. On the other hand, heteroscedasticity is very severe at this point and probably exceeds that which is found in most data.

would be to relax homoscedasticity of the selection equation and that would require either a semiparametric approach like the ones we have tried to avoid using or the use of a traditional MLE estimator of the heteroscedastic probit model that requires knowledge of the selection equation's deterministic scedasticity function. The good news about this prospect is that such models are easily estimated in STATA and other econometric software programs and our simple estimator could still be applied as a second step.

References

Ahn, H. & J. L. Powell (1993), 'Semiparametric estimation of censored selection models with a nonparametric selection mechanism', *Journal of Econometrics* **58**, 3–29.

Blundell, R., H. Reed & T. M. Stoker (2003), 'Interpreting aggregate wage growth: The role of labor market participation', *American Economic Review* **93**(4), 1114–1131.

Chen, Songnian N. & Shakeeb Khan (2003), 'Semiparametric estimation of a heteroskedastic sample selection model', *Econometric Theory* **19**(6), 1040–1064.

Donald (1995), 'Two step estimation of heteroskedastic sample selection models', *Journal of Econometrics* **65**, 347–380.

Fernandez-Sainz, Ana, Jaun M. Rodriguez-Poo & Inmaculada Villanua Martin (2002), 'Finite sample behavior of two step estimators in selection models', *Computational Statistics* **17**, 1–16.

Gallant, A. R. & D. W. Nychka (1987), 'Semi-parametric maximum likelihood estimation', *Econometrica* **55**, 363–390.

Gordon, T. M. (2003), 'Crowd out or crowd in?: The effects of common interest developments on political participation in california', *Annals of Regional Science* **37**(2), 203–233.

Greene, William H. (1997), *Econometric Analysis*, 3rd edn, Prentice-Hall, Upper Saddle River, NJ.

Heckman, James J. (1979), 'Sample selection bias as a specification error', *Econometrica* **47**(1), 153–161.

Hill, R. Carter, Lee C. Adkins & Keith A. Bender (2003), *Test Statistics and Critical Values in Selectivity Models*, Vol. 17 of *Advances in Econometrics*, Elsevier Science.

Honoré, Bo E., Ekaterini Kyriazidou & Christopher Udry (1997), 'Estimation of type 3 tobit models using symmetric trimming and pairwise comparisons', *Journal of Econometrics* **76**(1-2), 107–128.

Jolliffe, Dean (2002), 'The gender wage gap in bulgaria: A semiparametric estimation of discrimination', *Journal of Comparative Economics* **30**(2), 276–295.

Lewbel, Arthur (2003), Endogenous selection or treatment model estimation. Department of Economics, Boston College, 140 Commonwealth Ave., Chestnut Hill, MA, 02467, USA., lewbel@bc.edu.

Lewbel, Arthur (2004), 'Simple estimators for hard problems: Endogeneity and dependence in binary choice related models', Department of Economics, Boston College, 140 Commonwealth Ave., Chestnut Hill, MA, 02467, USA., lewbel@bc.edu.

Nawata, Kazumitsu & Nobuko Nagase (1996), 'Estimation of sample selection bias models', *Econometric Reviews* **15**(4), 387–400.

van der Klaauw, B. & R. H. Koning (2003), 'Testing the normality assumption in the sample selection model with an application to travel demand', *Journal of Business & Economic Statistics* **21**(1), 31–42.

Zuehlke, Thomas W. & Allen R. Zeman (1991), 'A comparison of two-stage estimators of censored regression models', *Review of Economics and Statistics* **73**(1), 185–188.

Lee Adkins, Professor of Economics, College of Business Administration, Oklahoma State University, Stillwater OK 74078
E-mail address: ladkins@okstate.edu

Carter Hill, Thomas J. Singletary Professor of Economics, College of Business Administration, Louisiana State University Baton Rouge, LA 70803
E-mail address: eohill@lsu.edu