Algorithmic Errors in the Estimation of Tobit II Models and the Corresponding Failure To Recognize Selection Bias

Thomas W. Zuehlke [a] and Anthony Kassekert [b]

[a] *Department of Economics, Florida State University, Tel (850) 644-5001*

[b] *Askew School of Public Administration and Policy, Florida State University, Tel (850) 644-3525*

Abstract

Tobit II models are a standard statistical tool for detecting and correcting selection bias. ML estimation is complicated by the possibility of multiple roots to the score equations. Most software packages ignore this problem and may fail to converge to the global MLE even when consistent starting values are used. Convergence to the global MLE can be insured by use of a two-step algorithm which conducts a grid search over the bounded space of the error correlation, and then uses the conditional ML estimates as starting values for simultaneous estimation. The nature of the problem is illustrated using Monte Carlo simulation. Major software packages are then compared and found to suffer from the same algorithmic errors. Finally, replication of estimates for a sample of published data sets finds that roughly half of the studies report inaccurate estimates.

Key words:

Tobit II, Heckman Selection Model, Simultaneous Estimation

Email addresses: tzuehlke@fsu.edu (Thomas W. Zuehlke), ajk05f@fsu.edu (Anthony Kassekert).

Preprint submitted to Elsevier *15 August 2008*

1 The Type II Tobit Model

The models considered in this paper is classified as "Type 2" Tobit models by Amemiya (1984). This bivariate model has the following structure:

$$\begin{aligned} Y_{1i} &= X_{1i}\beta_1 + \sigma\varepsilon_{1i} \\ Y_{2i} &= X_{2i}\beta_2 + \varepsilon_{2i} \end{aligned} \tag{1}$$

where (ϵ_{1i},ϵ_{2i}) is bivariate standard normal with correlation ρ. The regressors in each equation, X_{1i} and X_{2i}, are observed. Observation of the dependent variables is incomplete, however. Only qualitative information is available for the dependent variable in the selection equation, Y_{2i}. This is recorded as a binary variable, J_i, that takes the value one when Y_{2i} is positive. In addition, the dependent variable in the regression equation, Y_{1i}, is observed only when Y_{2i} is positive. [1]

The log-likelihood function for this model is

$$\ln L(\alpha,\beta,\sigma,\rho) = \sum_{i=1}^{n} \{J_i[-\ln(\sigma) + \ln\phi(Z_i) + \ln\Phi(W_i)] + (1-J_i)\ln[1-\Phi(X_{2i}\alpha)]\} \tag{2}$$

where Z_i=$(Y_{1i}-X_{1i}\beta)/\sigma$, W_i=$(X_{2i}\alpha+\rho Z_i)/\sqrt{1-\rho^2}$, and where ρ is restricted to the open interval (-1,1). This likelihood function is highly non-linear, and a solution to its score equations must be obtained with numerical methods. Unfortunately, the likelihood function is not globally concave. Gradient methods may converge to a local MLE. One can only be assured of obtaining a global MLE, if the estimation processes is started in the neighborhood of the global

[1] In many economic applications, the regression equation is a pricing or expenditure function, and the selection equation is a decision function that governs the occurrence of the transaction.

maximum.

Starting values for numerical solution of the score equations are typically provided by the two-stage method of Heckman (1976) and Lee (1976). [2] The small sample performance of this estimator is erratic, often providing estimates of ρ that exceed one in absolute value. Zuehlke (1991) show that the mean square error of the sub-sample OLS estimator is often superior to that of the HL estimator. [3] There is little reason to believe that the HL estimator will provide starting values that are in the neighborhood of the global maximum.

There is a solution to this problem, however. Olsen (1982) shows that the likelihood function of the Type II Tobit model is globally concave conditional on ρ. He suggests that a grid search over the bounded parameter ρ, in conjunction with the conditional MLEs, can be used to trace the profile of the maximized value of $\ln L(\alpha, \beta, \sigma, \rho)$ over the space of ρ. The location of any local or global maxima are easily determined, and a fully simultaneous maximization can be started in the neighborhood of the global MLE. Unfortunately, this algorithm has not yet been incorporated into "canned" software.

The possibility of distinct global and local maxima is not the only problem. As noted in (Olsen 1982), a global maximum need not exist. Since the parameter space of ρ is the open interval (-1,1), a maximizing value of will not exist if the conditional likelihood function is increasing (or decreasing) in ρ right up to the boundary of the parameter space. The possible outcomes of the maximization

[2] Many applications simply report the estimates obtained using the two-stage method. The performance of the two stage estimator is comparable to maximum likelihood only for small values of ρ; conditions under which selection bias is minimal. Moreover, Nawata (1994) shows that as the degree of correlation increases, the maximum likelihood estimator is "much more efficient than Heckman's estimator."
[3] Zuehlke (1996) shows that this result is not restricted to models that are "identified by non-linearity."

process include: no root to the score equations; a unique root that is not a global MLE; multiple roots to the score equations; a unique root that is a global MLE. These cases are illustrated in Figures 1 through 4, respectively. Each of these figures was generated by re-sampling the error terms while using exactly the same set of parameter values and regressors.

In the cases illustrated by Figures 1 and 2, there is no global MLE. A "near MLE" can be defined, as in (Rao 1973), but the statistical properties of such an estimator are unknown. In the cases illustrated by Figures 2 and 3, there is a local maximum that is not a global maximum. The relative performance the local MLE and the corresponding "near MLE" or global MLE are also unknown. Useful asymptotic theory is available only in the case illustrated by Figure 4, where the local and global MLE correspond.

When encountered in application, the case illustrated in Figure 1 is often interpreted as a sign of model mis-specification. As this simulation illustrates, however, this conclusion may be erroneous. Likewise, "canned" software that does not scan the space of ρ may report a local maximum when the global maximum is either different or does not exist. The focus of this paper is on the relative frequency of the problems illustrated in Figures 1 through 3.

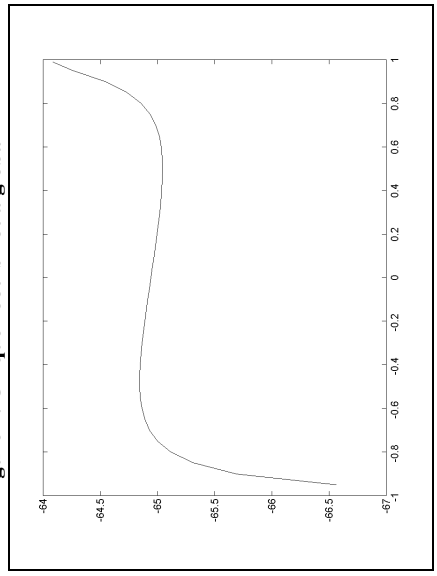


Figure 2: Unique root is not a global MLE

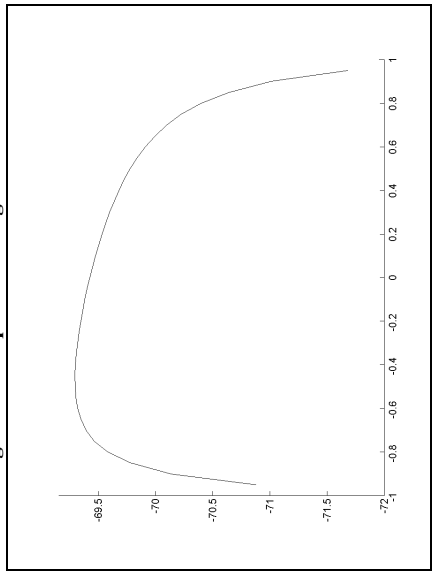


Figure 4: Unique root is a global MLE

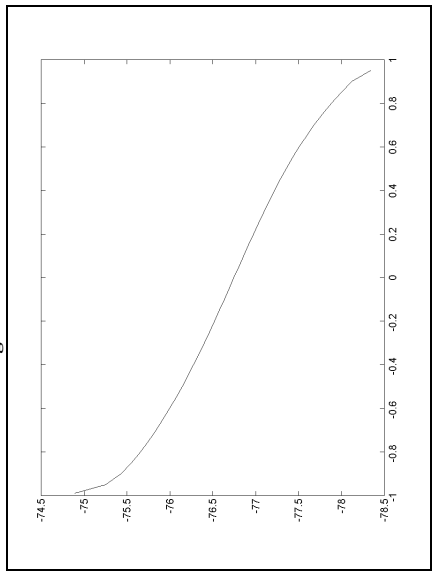


Figure 1: No root

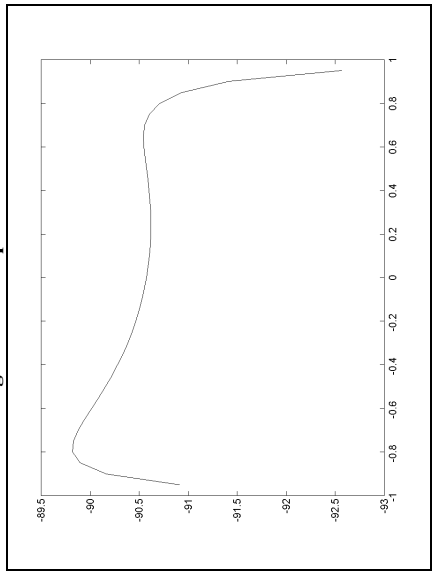


Figure 3: Multiple roots

2 Data Generation

The purpose of the Monte Carlo portion of this study is to examine the relative frequency of multiple roots to the score equations. In order to focus on this topic, the structure of the model is kept as simple as possible.[4] Both the regression equation and the selection equation contain an intercept and a single regressor. The regressors, X_1 and X_2, are random draws of a bivariate standard normal with correlation ρ_X. The assumption of a zero mean and unit variance for the regressors involves no loss in generality. In practice, standardizing a regressor will simply scale the coefficient estimate without affecting the precision of the estimate or the fit of the model. The degree of independent variation in the regressors of the selection and regression equations is controlled by the parameter ρ_X. The regressors will be fixed in repeated sampling, and are statistically independent of the disturbances. The disturbances, ϵ_1 and ϵ_2, are a statistically independent sequence of bivariate standard normals with correlation ρ_ϵ. The degree of selection bias is increasing in absolute value with increases in the absolute value of ρ_ϵ.

Given this structure, the unconditional mean of Y_2, and consequently, the degree of censoring, is controlled by the intercept of the selection equation, α_1. The explanatory power of the selection equation is determined by the slope coefficient of the selection equation, α_2. This parameter is chosen to give an expected R^2 of 50 percent in the uncensored selection equation. The explanatory power of the selection equation increases as α_2 increases in absolute value.

Nelson (1984) shows that the variance of the coefficient estimates of a Type II Tobit model are affected proportionally by a change in σ. When comparing the

[4] Henceforth, the observational subscript is omitted in order to simplify notation.

relative performance of the estimators, the choice of σ is arbitrary. Changes in σ affect the absolute, but not the relative scale of the variances. If the normalization $\sigma = \sqrt{1 - \beta_2^2}$ is adopted, then the fit of the regression equation is controlled by the single parameter, β_2. For this choice of σ, the slope coefficient, β_2, corresponds to the correlation coefficient between Y_1 and X_1.

A brief summary of the Monte Carlo process is as follows:

(1) The regressors, X_1 and X_2, are drawn. They are fixed across repetitions.
(2) The disturbances, ϵ_1 and ϵ_2, are drawn. Given values for the regressors and disturbances, and the parameters α, β, σ, and ρ_ϵ, the values of J and Y_1 are computed.
(3) The parameters are estimated, and any local or global maximum are determined.
(4) Steps 2 and 3 are repeated on successive repetitions, and sample moments are compiled across repetitions.

The data generation process described above was carefully structured in order to limit intra-experiment random variation. [5] Each estimator is applied to the same sequence of data sets for any given parameter combination, $(\alpha, \beta, \sigma, \rho_\epsilon)$. This will limit random variation in comparisons across estimators. In addition, the same sequence of independent standard normal pairs, (ϵ_1,ϵ_2), will be used to construct the data sequence (J,Y_1) required for each distinct parameter combination. [6] This will limit random variation in comparisons across parameter values.

The plots in Figure 6 summarize the results from the MC simulation. In most

[5] See, Hendry (1984).
[6] The independent standard normals were obtained with the algorithm of Forsythe (1977).

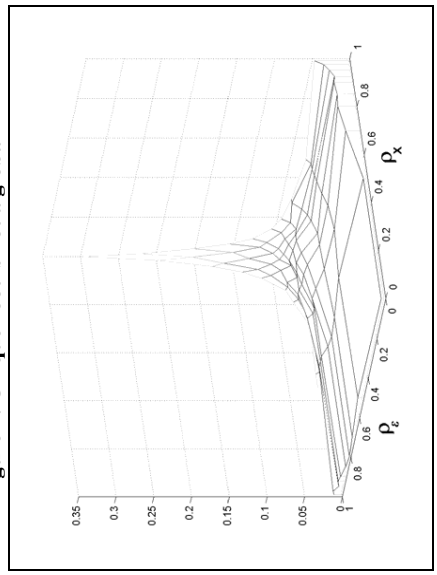


Figure 2: Unique root is not a global MLE

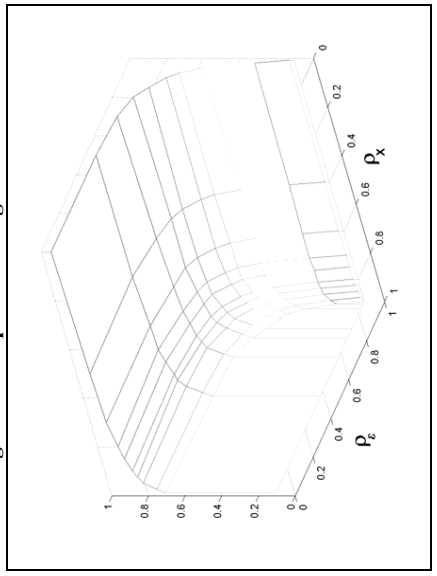


Figure 4: Unique root is a global MLE

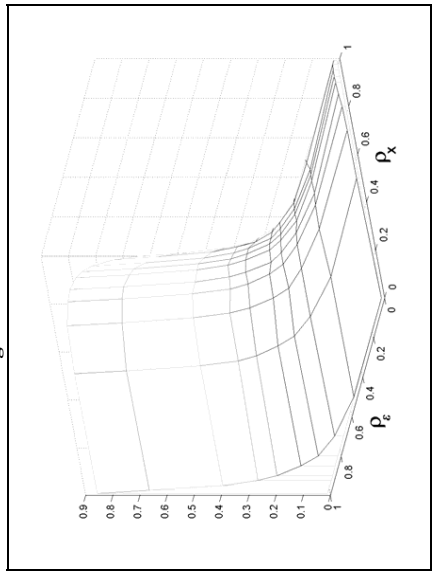


Figure 1: No roots

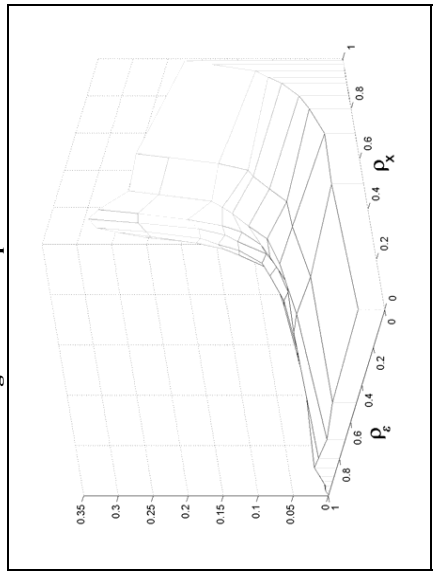


Figure 3: Multiple roots

instances, a unique root is discovered. The absence of any root is associated with extremely high values of ρ_e. Presences of a single root that was not a MLE is more likely in situations where both ρ_x and ρ_e is large. Finally, multiple roots is most influenced by high values of ρ_x. These MC simulation results suggest that the possibility for Tobit II estimation problems is limited to data with high values of ρ_e and ρ_x. The next section addresses how frequently these situations arise in published data.

3 Application

The Monte Carlo results of the previous section suggest that, in most applications, a local MLE found with gradient methods is likely to be a global MLE. While this is of some consolation, it does nothing to diminish the magnitude of the problem that may occur on a case by case basis. To illustrate this point, we provide examples from three published data sets using selection models. The data for this project come from published works in peer review journals (Mroz 1987; Kenkel and Terza 2001; Martins 2001)[1]. For each data set, the results from estimation using a grid search are compared to the estimates from the commonly used statistical packages SAS, STATA, and R.

3.1 Methods

In order to produce results for the global MLE, the authors wrote R code provided in Appendix A. The code uses subsample OLS for starting values with ρ equal to zero. The initial values were then used to estimate the entire grid by increments of .05. In other words, the maximized values at rho equal to zero were used as initial guesses for ρ equal to .05, then the maximized values at .05 were used as beginning values for ρ equal to .10 and so on. Once the maximum was found on the .05 increment scale, a second grid search was conducted with .01 increments to get two digits of accuracy for the correlation parameter.

The statistical programs reviewed are SAS, STATA and R. Other commonly used packages, such as Minitab and SPSS do not offer Tobit II models directly, although macros are sometimes available online. Each set of data was

run through the corresponding selection model estimation for each statistical package. For the first run, no initial values were specified and the package default determined estimates[2]. If the initial run failed to return the global MLE, a second attempt was made with starting values from a model using a correlation of .05 less than the true value. For example, a true value of ρ=.90 would be given starting values based on .85. This is done in order to check if specifying accurate initial values improved software performance. The code for each package is located in Appendix B.

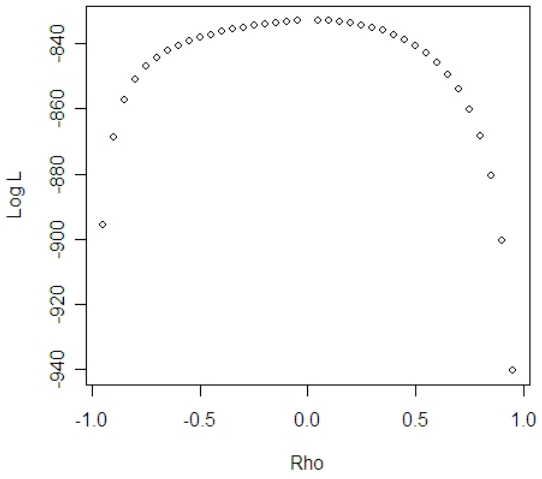
The choice of data sources is not random and was done based on availability. Principally, the data come from journals with online archives. The three cases presented here were chosen to represent a variety of likelihood behaviors for ρ. The Mroz (1987) data is contained in most software as an example, partly because the likelihood function for rho is a well behaved parabolic curve and simultaneous estimation without a grids search provides accurate results (Figure 1a). The Martins (2001) data has both a local and a global maximum, making it more difficult to identify (Figure 1b). The Kenkel and Terza data set is large with numerous variables and no global MLE (Figure 1c).

3.2 Results

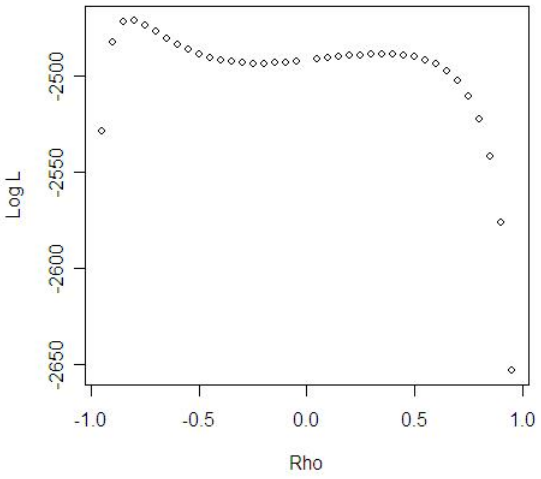
The condensed results of each analysis are presented below in Table 3.2 along with a plot of the likelihood for each data set in Figure 1. The global MLE values for ρ and the log likelihood are given for each model in the column titled "Author". Second, the maximum difference for a non-intercept t-statistic is given. The t-statistic difference is used because it incorporates both the parameter estimate and its standard error into a single measurement, which summarizes the data more efficiently[3]. Finally, the number of significant vari-

ables at the 95 percent confidence level is counted. The statistical package results reported below are all based off of the default settings.

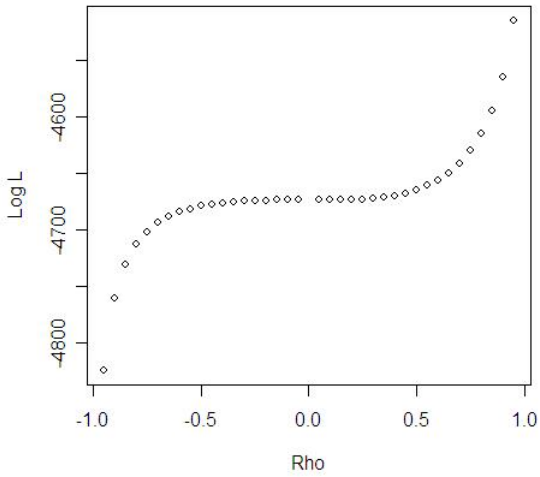
Fig. 1. Likelihood Plots Over ρ



(a) Mroz



(b) Martins



(c) Kenkel and Terza

Overall, there seems to be very little difference in the estimates produced by the major statistical softwares. Even in error, the programs generate consistent results. The degree of agreement is even more surprising considering that nonlinear regression routines have been shown to spawn divergent answers

		Author	SAS	STATA	R
Mroz (1987)	Rho	0.03	0.03	0.03	0.03
	LogL	−832.89	−832.89	−832.89	−832.88
	Max t-Statistic Difference		0.36	0.36	0.36
	Number of Significant Variables (.05)	10.00	10.00	10.00	10.00
Martins (2001)	Rho	−0.81	0.35	0.35	0.35
	LogL	−2470.74	−2488.00	−2488.36	−2488.48
	Max t-Statistic Difference		− ∞	8.32	8.32
	Number of Significant Variables (.05)	10.00	6.00	6.00	6.00
Kenkel and Terza (2001)	Rho	∞	−0.04	−0.04	−0.04
	LogL	∞	−4674.00	−4673.70	−4673.70
	Max t-Statistic Difference		∞	∞	∞
	Number of Significant Variables (.05)	0.00	8.00	8.00	8.00

Table 1. Comparison of Tobit II Model Results

when comparing results among programs for less complex mathematical systems (McCullough 1999). The well behaved Mroz data set provided similar results for each program. The differences that do occur are probably linked to the correlation only being calculated out to two significant digits for the answers generated by the author. Additionally, the different methods of estimation could have played a small part, but this is unlikely since other methods were tried and all provided the same response for this particular problem.

For the Martins data set, SAS estimates the standard error for the interaction term between potential experience and age squared to be 0. The variable appears to have too much collinearity for SAS, which forces the variance to zero and the t statistic to negative infinity. The second largest difference in SAS is 8.32, which is in line with the other programs. None of the other methods provide an answer for the interaction term either. (There is likely a singularity tolerance option in the nonlinear optimization commands that I have been unable to find and adjust to solve this problem.)

Beyond SAS's individual issues, all of the major packages misestimate the correlation and consequently the parameter values for the Martins' data. The commonplace statistical softwares all find the same local maximum. The same result is found regardless of method of estimation in SAS and STATA, which indicates that the solution is not as simple as changing the default options. All of the software packages do achieve the correct answer when given close initial values from the author's grid search. This example simply confirms the fact that Tobit II models can have multiple roots and that a grid search over the range of the correlation is the only way to verify a global MLE exists.

The Kenkel and Terza data set has no global root, yet all of the packages produce a similar answer. Again, this does not appear to be a trouble with

the default options because the answer does not change significantly based on any adjustments. Just as in the last problem, if starting values are given farther away from zero with a correlation approaching one, the computer algorithms are able to correctly identify the model.

4 Conclusions

Tobit II models are frequently used in the social sciences as a check against selection bias. The problematic nature of ML estimation for Heckman selection models is displayed using a Monte Carlo simulation. The simulations and the reviewed data sets clearly show the possibility of multiple roots and a singular root which is not a MLE. Simultaneous estimation of ρ and the regression parameters is only accurate when the starting values are in close proximity to the global MLE. Since the global MLE is rarely known, a two-stage estimator is proposed here which incorporates a grid search over rho in order to find accurate starting values.

The Monte Carlo simulations demonstrate that the correct global root will often be identified under the seemingly reasonable conditions of non-extreme values of ρ_e and ρ_x. The problematic cases of multiple roots or singular local roots are more frequent when their is high censoring, high correlation between error terms, and little unique information between regression and selection equations. These situations frequently occur in the literature and are the main reasons for using a Tobit II model in the first place, which is why we also examine actual published data sets.

The use of published data sets in this paper is meant to emphasize two points. First, that all of the major statistical packages suffer from the same algorithmic shortcomings. SAS, STATA, and R all rely upon simultaneous estimation of all parameters, which can lead to incorrect results unless the likelihood is globally concave over ρ. Second, the three articles demonstrate that improper Tobit II model estimation is not a statistical artifact that happens rarely in practice. Inaccurate results from these models are frequently reported in journals, or

possibly worse not reported when researchers find ρ to be insignificant and revert to a model that ignores selection bias altogether. Fixing the major statistical packages is serious issue that demands rectification.

The Martins (1987) and Kenkel and Terza (2001) results have an interesting aspect in that not conditioning on ρ provides unpredictable bias. The Martins publication would have been helped (based on number of significant variables) by implementing a more accurate routine. Contrarily, the Kenkel and Terza data would probably not have been published (at least in its current form) if the model was estimated correctly. These examples are indicative of the fact that failure to condition of ρ frequently leads to erroneous estimates with unpredictable biases.

The next step in this research is to gather more published data sets using Heckman selection models and test them using a two-stage estimation technique. Of the data sets that have been examined thus far, about half of the results from published articles are inaccurate. In addition to gauging how widespread the problem is, the authors are also working on R code that will be made freely available for public use. Hopefully, the implementation of the more accurate two-stage estimator in R will encourage other statistical platforms to update their algorithms.

In conclusion, the existence of this error in statistical packages is complectly unnecessary in this case because there is a relatively simple fix to eliminating this problem. Almost all programs provide a warning that accuracy is dependent on starting values, but researchers rarely have informative prior information regarding the size of coefficients and almost never have a estimate of ρ. Selection bias is a prominent threat to the validity to inquiry in survey research and properly estimated Tobit II models will never offer a complete

solution to these issues, but proper modeling is a start.

Notes

[1]We are in the process of collecting more published data sets. So far, seven additional data sets have been analyzed but only two authors have granted permission to publish the results. Therefore, we limited the results reported here to freely available online data.

[2]Adjusting the default options, i.e. changing from Newton-Raphson to BHHH optimization, had little impact on the results

[3]The full set of results is available from the authors

5 Appendix

A R Code for Tobit II Estimation

```
### This code still has bugs and will be posted online when a correct version is a

tobit2<- function(y1,c1,x1,x2,theta,rho) {
y<<-cbind(y1)   # This makes sure the data is in matrix form
x1<<-cbind(x1)
x2<<-cbind(x2)
c1<<-cbind(c1)  #need to change c to c1

nobs<-as.numeric(dim(y)[1])
km1<<-as.numeric(dim(x1)[2])+1
km2<<-as.numeric(dim(x2)[2])+1
df<-nobs-km1-km2

ones<-matrix(1,nobs,1)
colnames(ones)<-'Intercept'
x1<<-cbind(ones,x1)
x2<<-cbind(ones,x2)

logl0<-tobit2_ll(theta,rho)          # LogL at starting values.
tol<-0.001                          # Convergence tolerance.
maxit<-500                          # Iteration Limit.
md<-1
i<-1

while (md >= tol & i<=maxit) {      # Start iteration loop.
#####################################################################
#tobit2_gr(theta,rho)                # Compute BHHH update.

beta1<- theta[1:km1,]
beta2<- theta[(km1+1):(km1+km2),]
ls<- theta[(km1+km2+1),]
sig<- as.numeric(exp(ls))

a1<- 1/(cos(asin(rho)))
a2<- tan(asin(rho))
xb1<- (x1%*%beta1)
xb2<- (x2%*%beta2)
z <- (y-xb2)/sig
w<- (a1*xb1) + (a2*z)
```

```
fl0<- dnorm(xb1)  # dnorm calls normal PDF (density)
flw<- dnorm(w)    # dnorm calls normal PDF (density)
fbw<- pnorm(w)  #  pnorm calls  normal CDF. (distribution)
fbc<- pnorm(xb1, lower.tail = F)  #  pnorm calls  normal CDF. (distribution)
r2<- fl0/fbc
rw<- flw/fbw

q1<- ((c1*rw*a1-(1-c1)*r2) %*% matrix(1,1,km1))*x1
q2<- ((c1*(z-c1*rw*a2)/sig) %*% matrix(1,1,km2))*x2
q3<- c1*z*z-c1*(rw*a2*z)-c1
q<- cbind(q1,q2,q3)


sc<-(apply(q,2,sum))                              # Computes gradient (score).
m<-t(q)%*%q
vc<-solve(m)                                  # Estimated covariance matrix.
d<-vc%*%sc                                    # BHHH directional update.
md<-max(abs(d)) # Convergence criterion.
#########################################################################

j<-0
deltf<- -1

          while (deltf<0 & j<=4) {          # Start step reduction loop.
           step<- 1 *(0.5)^j                      # Reduce step by half.
           gamma<- theta+(step*d)                # Try new parameter vector.
           logl<- tobit2_ll(gamma,rho)          # Evaluate Probit logL.
           deltf<- logl-logl0                    # Determine function change.
           j<-j+1                                # Update step counter.
          }                                     # End step reduction loop.
      logl0<-logl                                # Update value of logL.
      theta<-gamma    # Update beta0.


if (i==1) int<-cbind(i,md, logl, step)
if (i>1) int<-rbind(int,cbind(i,md, logl, step))
colnames(int)<-list('Iteration','Grad:','LogL:','Size:')

     i<-i+1;                                      # Update iteration counter.
}


stderr <- sqrt(diag(vc))                        # Compute standard errors.
t <- theta/stderr                               # Compute t-statistics.
```

```
pvt<-(1-pf(t^2,1,df))                          # Compute p-values.

if (i>maxit) print('Iteration Limit Exceeded.')

ind<-round(cbind(theta,stderr,t,pvt),5)
colnames(ind)<-c('Coefficient','Std. Error','t-Stat','Prob>|t|')
ind1<-ind[1:km1,]
ind2<-ind[(km1+1):(km1+km2),]

cat('Tobit II Estimates for dependent variable:', colnames(y),'\n')
cat('\n')
v<-dim(int)[1]
if (v<=10) print(round(int,4))
if (v>10)  print(round(int[(v-10):v,],4))
cat('\n\n')
cat('Selection Equation Estimates: \n\n')
print(ind1)
cat('Regression Equation Estimates: \n\n')
print(ind2)
ls<- theta[km1+km2+1,]
sig<- exp(ls)
cat('Sigma      ',sig,'\n')
cat('Rho          ',rho,'\n')
}



GridSearch <- function(theta,rho){
for (k in 1:19){
rho<-(0 - k/20)
# print (rho) }

logl0<-tobit2_ll(theta,rho)          # LogL at starting values.
tol<-0.001                          # Convergence tolerance.
maxit<-500                          # Iteration Limit.
md<-1
i<-1

while (md >= tol & i<=maxit) {      # Start iteration loop.
#####################################################################
#tobit2_gr(theta,rho)                # Compute BHHH update.

beta1<- theta[1:km1,]
beta2<- theta[(km1+1):(km1+km2),]
ls<- theta[(km1+km2+1),]
```

```
sig<- as.numeric(exp(ls))

a1<- 1/(cos(asin(rho)))
a2<- tan(asin(rho))
xb1<- (x1%*%beta1)
xb2<- (x2%*%beta2)
z <- (y-xb2)/sig
w<- (a1*xb1) + (a2*z)

fl0<- dnorm(xb1)  # dnorm calls normal PDF (density)
flw<- dnorm(w)    # dnorm calls normal PDF (density)
fbw<- pnorm(w)  #  pnorm calls  normal CDF. (distribution)
fbc<- pnorm(xb1, lower.tail = F)  #  pnorm calls  normal CDF. (distribution)
r2<- fl0/fbc
rw<- flw/fbw

q1<- ((c1*rw*a1-(1-c1)*r2) %*% matrix(1,1,km1))*x1
q2<- ((c1*(z-c1*rw*a2)/sig) %*% matrix(1,1,km2))*x2
q3<- c1*z*z-c1*(rw*a2*z)-c1
q<- cbind(q1,q2,q3)


sc<-(apply(q,2,sum))                              # Computes gradient (score).
m<-t(q)%*%q
vc<-solve(m)                                  # Estimated covariance matrix.
d<-vc%*%sc                                    # BHHH directional update.
md<-max(abs(d)) # Convergence criterion.
#########################################################################

j<-0
deltf<- -1

          while (deltf<0 & j<=4) {          # Start step reduction loop.
           step<- 1 *(0.5)^j                      # Reduce step by half.
           gamma<- theta+(step*d)                # Try new parameter vector.
           logl<- tobit2_ll(gamma,rho)         # Evaluate Probit logL.
           deltf<- logl-logl0                   # Determine function change.
           j<-j+1                               # Update step counter.
          }                                     # End step reduction loop.
      logl0<-logl                                  # Update value of logL.
      theta<-gamma    # Update beta0.


if (i==1) int<-cbind(i,md, logl, step)
if (i>1) int<-rbind(int,cbind(i,md, logl, step))
```

```
colnames(int)<-list('Iteration','Grad:','LogL:','Size:')

     i<-i+1;                                    # Update iteration counter.
}

if (i>maxit) print('Iteration Limit Exceeded.')
maxlog<-cbind(rho, logl)
if (k==1) test_neg<-maxlog
if (k>1) test_neg<-rbind(test_neg,maxlog)
#plot(test[,1],test[,2])
colnames(theta)<-round(rho,5)
if (k==1) theta_neg<-cbind(theta)
if (k>1)  theta_neg<-cbind(theta_neg,theta)
}

##############################################################################
##############################################################################
##############################################################################
theta<-t0
for (k in 1:19){
rho<-(0+k/20)
# print (rho) }

logl0<-tobit2_ll(theta,rho)          # LogL at starting values.
tol<-0.001                          # Convergence tolerance.
maxit<-500                          # Iteration Limit.
md<-1
i<-1

while (md >= tol & i<=maxit) {      # Start iteration loop.
#####################################################################
#tobit2_gr(theta,rho)                # Compute BHHH update.

beta1<- theta[1:km1,]
beta2<- theta[(km1+1):(km1+km2),]
ls<- theta[(km1+km2+1),]
sig<- as.numeric(exp(ls))

a1<- 1/(cos(asin(rho)))
a2<- tan(asin(rho))
xb1<- (x1%*%beta1)
xb2<- (x2%*%beta2)
z <- (y-xb2)/sig
w<- (a1*xb1) + (a2*z)
```

```
fl0<- dnorm(xb1)  # dnorm calls normal PDF (density)
flw<- dnorm(w)    # dnorm calls normal PDF (density)
fbw<- pnorm(w)  #  pnorm calls  normal CDF. (distribution)
fbc<- pnorm(xb1, lower.tail = F)  #  pnorm calls  normal CDF. (distribution)
r2<- fl0/fbc
rw<- flw/fbw

q1<- ((c1*rw*a1-(1-c1)*r2) %*% matrix(1,1,km1))*x1
q2<- ((c1*(z-c1*rw*a2)/sig) %*% matrix(1,1,km2))*x2
q3<- c1*z*z-c1*(rw*a2*z)-c1
q<- cbind(q1,q2,q3)


sc<-(apply(q,2,sum))                              # Computes gradient (score).
m<-t(q)%*%q
vc<-solve(m)                                  # Estimated covariance matrix.
d<-vc%*%sc                                    # BHHH directional update.
md<-max(abs(d)) # Convergence criterion.
#####################################################################

j<-0
deltf<- -1

          while (deltf<0 & j<=4) {          # Start step reduction loop.
           step<- 1 *(0.5)^j                      # Reduce step by half.
           gamma<- theta+(step*d)                # Try new parameter vector.
           logl<- tobit2_ll(gamma,rho)          # Evaluate Probit logL.
           deltf<- logl-logl0                    # Determine function change.
           j<-j+1                               # Update step counter.
          }                                     # End step reduction loop.
      logl0<-logl                               # Update value of logL.
      theta<-gamma    # Update beta0.


if (i==1) int<-cbind(i,md, logl, step)
if (i>1) int<-rbind(int,cbind(i,md, logl, step))
colnames(int)<-list('Iteration','Grad:','LogL:','Size:')

     i<-i+1;                                    # Update iteration counter.
}

if (i>maxit) print('Iteration Limit Exceeded.')

maxlog<-cbind(rho, logl)
if (k==1) test_pos<-maxlog
```

```
if (k>1) test_pos<-rbind(test_pos,maxlog)
#plot(test[,1],test[,2])
colnames(theta)<-round(rho,5)
if (k==1) theta_pos<-cbind(theta)
if (k>1)  theta_pos<-cbind(theta_pos,theta)
}
test_all<-rbind(test_neg,test_pos)
theta_all<-cbind(theta_neg,theta_pos)
#plot(test_all[,1],test_all[,2])
}
```

B Sample Code for Commercial Statistical Software

B.1 SAS

proc qlim data = martins ; model sel = child ychild hw edu age age2 /discrete;
model wage = edu pexp pexp2 pexpchd pexpchd2 /select(sel=1); run;

B.2 STATA

heckman wage edu pexp pexp2 pexpchd pexpchd2, sel(sel = child ychild hw
edu age age2)

B.3 R

```
library(sampleSelection) # Adds Heckman model to base package

mod <- selection (sel ~ child+ychild+hw+edu+age+age2 , wage~edu+pexp+pexp2+pexpchd
summary(mod)
```

6 Bibliography

Amemiya, T. (1984). Tobit Models: A Survey. Journal of Econometrics, 24, 3-61.

Forsythe, G. E., Malcolm, M. A., & Moler, C. B. (1977). Computer Methods for Mathematical Computations. Englewood Cliffs, New Jersey: Prentice-Hall.

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.

Hendry, D. F. (1984). Monte Carlo Experimentation in Econometrics. Handbook of Econometrics, pp. 937-976.

Kenkel, D. S., & Terza, J. V. (2001). The Effect of Physician Advice on Alcohol Consumption: Count Regression with an Endogenous Treatment Effect. Journal of Applied Econometrics, 16(2), 165-184.

Lee, L.-F. (1976). Estimation of Limited Dependent Variable Models by Two-Stage Methods. University of Rochester: Doctoral Dissertation.

Martins, M. F. O. (2001). Parametric and Semiparametric Estimation of Sample Selection Models: An Empirical Application to the Female Labour Force in Portugal. Journal of Applied Econometrics, 16, 23-39.

McCullough, B. D. (1999). Assessing the Reliability of Statistical Software: Part II. The American Statistician, 53(2), 149-159.

Mroz, T. A. (1987). The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions. Econometrica, 55(4), 765-799.

Nawata, K. (1994). Estimation of Sample Selection Bias Models by the Maximum Likelihood Estimator and Heckman's Two-Stage Estimator. Economics Letters, 45, 33-40.

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

Olsen, R. J. (1982). Distributional Tests for Selectivity Bias and a More Robust Likelihood Estimator. International Economic Review, 23, 223-240.

Rao, C. R. (1973). Linear Statistical Inference and Its Application. New York: John Wiley $&$ Sons.

Zuehlke, T. W. (1996). Identification by Non-Linearity in Censored Regression Models. Journal of Statistical Computation and Simulation, 54, 289-304.

Zuehlke, T. W., & Zeman, A. R. (1991). A Comparison of Two-Stage Estimators of Censored Regression Models. Review of Economics and Statistics, 73, 185-188.