A note on Heckman-type corrections in models for zero expenditures

Frederic Vermeulen*
Center for Economic Studies
Katholieke Universiteit Leuven

July, 2001

Abstract

In Heien and Wessells (1990), a two-step estimation procedure, that makes use of Heckman-type corrections, is proposed to estimate consumption on household budget surveys. It is shown that this approach, which draws from switching regressions models, leads to inconsistent estimates.

1. Introduction

In Heien and Wessells (1990), a two-step estimation procedure, that makes use of Heckman-type corrections, is proposed to estimate demand systems on household budget surveys. The latter are characterised by the presence of zero expenditures.

[*] Center for Economic Studies, Katholieke Universiteit Leuven, Naamsestraat 69, B-3000 Leuven, Belgium, Tel.: +32-(0)16326806, Fax: +32-(0)16326796, frederic.vermeulen@econ.kuleuven.ac.be. Financial support of this research by the Fund for Scientific Research - Flanders (contract FWO G.0327.97) is gratefully acknowledged. I would like to thank Marno Verbeek and Jeffrey Wooldridge for helpful comments. Any errors remaining are my own.

The proposed two-step estimation procedure however, draws from a wrong interpretation of Lee's (1981) generalised framework, which nests, among others, the switching regressions model. Moreover, it can easily be shown that the approach results in the same inconsistent estimates as if ordinary least squares (in the case of a single equation) or, for example, seemingly unrelated regressions (in the case of a demand system) would have been used on all the observations, including the zeroes.

2. The switching regressions model

In Lee (1981) a unified simultaneous equations model is presented that consists of observable continuous endogenous variables, limited dependent variables, unobservable latent endogenous variables with binary indicators and censored dependent variables. A special case of this model is the switching regressions model, that was put into practice by, for example, Lee (1978).

Suppose that an agent can be subject to two mutually exclusive and exhausted regimes. More specifically, his behaviour can be described by two different regression equations. Each observation in the sample either belongs to the first regime or to the second, but never to both. A selection mechanism, which determines whether an observation will be found in the first regime or in the second, is captured by a standard probit model :

$$z_i^* = W_i'\gamma + u_i, i = 1, 2, ..., n \tag{1}$$

$$I_i = 1 \text{ if } z_i^* > 0 \tag{2}$$

$$I_i = 0 \text{ if } z_i^* \leq 0 \tag{3}$$

where z_i^* is an unobservable latent variable, W_i is a vector of exogenous variables, I_i is a binary indicator and n is the number of observations in the sample. The normalisation restriction $\sigma_u^2 = 1$ is assumed. Each regime is characterised by an own equation with (possibly the same) vectors of explanatory variables X_{1i} and X_{2i} :

$$y_{1i} = X_{1i}'\beta_1 + \varepsilon_{1i} \tag{4}$$

for observations belonging to the first regime ($I_i = 1$), and

$$y_{2i} = X_{2i}'\beta_2 + \varepsilon_{2i} \tag{5}$$

for i belonging to the second regime ($I_i = 0$). It is assumed that u_i, ε_{1i} and ε_{2i} follow a trivariate normal distribution, independent of all regressors. The problem now is, that $E\left[\varepsilon_{1i} \mid I_i = 1\right]$ and $E\left[\varepsilon_{2i} \mid I_i = 0\right]$ do not equal zero, if u_i is correlated with ε_{1i} or ε_{2i}, respectively. Consequently, in that case, ordinary least squares (OLS) estimation of equations (4) and (5) would yield inconsistent estimates of respectively β_1 and β_2. In particular, it follows that

$$E\left[\varepsilon_{1i} \mid I_i = 1\right] = E\left[\varepsilon_{1i} \mid u_i > -W_i'\gamma\right] = \sigma_{\varepsilon_1 u} E\left[u_i \mid u_i > -W_i'\gamma\right] \qquad (6)$$

and

$$E\left[\varepsilon_{2i} \mid I_i = 0\right] = E\left[\varepsilon_{2i} \mid u_i \leq -W_i'\gamma\right] = \sigma_{\varepsilon_2 u} E\left[u_i \mid u_i \leq -W_i'\gamma\right] \qquad (7)$$

where $\sigma_{\varepsilon_i u}$ denotes the covariance between the error term of regime i and that of the selection mechanism. It is possible to use the above results to adjust the regime regressions, such that they have a mean zero error term uncorrelated with the explanatory variables. Writing

$$y_{1i} = X_{1i}'\beta_1 + \sigma_{\varepsilon_1 u}\frac{\phi\left(W_i'\gamma\right)}{\Phi\left(W_i'\gamma\right)} + \eta_{1i} \qquad (8)$$

$$y_{2i} = X_{2i}'\beta_2 - \sigma_{\varepsilon_2 u}\frac{\phi\left(W_i'\gamma\right)}{1-\Phi\left(W_i'\gamma\right)} + \eta_{2i} \qquad (9)$$

yields zero conditional mean error terms η_{1i} and η_{2i}. Both equations are thus adjusted with a Heckman-type correction term, where $\phi\left(.\right)$ and $\Phi\left(.\right)$ are the normal probability density function and the normal cumulative distribution function, respectively. The two-stage estimation procedure suggested by Lee (1978, 1981) is as follows. In a first step the selection mechanism is estimated by making use of probit maximum likelihood in order to obtain estimates of γ. These $\hat{\gamma}$ coefficients are used to obtain consistent estimates of the correction terms $\phi\left(W_i^{\prime}\gamma\right)/\Phi\left(W_i^{\prime}\gamma\right)$ and $-\phi\left(W_i^{\prime}\gamma\right)/\left[1-\Phi\left(W_i^{\prime}\gamma\right)\right]$. In a second step the regressions (8) and (9) are estimated consistently by OLS.

Heien and Wessells (1990) draw from the above switching regressions framework to estimate demand systems on household budget surveys. Implicitly (and incorrectly), they suppose that positive expenditures belong to the first regime of the model, while zero expenditures are captured by the second regime. In first instance, a selection mechanism is estimated by probit maximum likelihood to derive the Heckman-type correction terms for both regimes. In the second stage of their approach, a demand system is estimated by seemingly unrelated regressions (SUR) by merging both regimes and estimating the system on all observations. More specifically, positive expenditures are given the correction term of equation (8) while zero expenditures are corrected by the extra regressor of equation (9). The other regressors are the same for both positive and zero expenditures. This approach however, results in inconsistent estimates of the coefficients of the demand system. A characterisation of this inconsistency is given in the next section.

3. The nature of the bias

Without loss of generality, we will concentrate on a single equation which explains the demand for a certain commodity. The observed demand is drawn from a censored distribution with censoring at zero. Such a censored regression can be adequately estimated by a tobit model or by Heckman's two-stage estimation procedure (see Heckman, 1979)[1]. Let us now formalise the approach followed by Heien and Wessells (1990), which resembles the Heckman two-stage procedure.

We have again a selection mechanism which determines whether the expenditures will be positive or equal to zero. Depending on the binary indicator I_i, associated with the latent selection variable z_i^*, a limit or a continuous observation y_i will be observed :

$$z_i^* = W_i'\gamma + u_i, i = 1, 2, ..., n \tag{10}$$

$$y_i^* = X_i'\beta + \varepsilon_i, i = 1, ..., n \tag{11}$$

$$y_i = y_i^*,\ I_i = 1 \text{ if } z_i^* > 0 \tag{12}$$

$$y_i = 0,\ I_i = 0 \text{ if } z_i^* \leq 0 \tag{13}$$

under the assumption of a bivariate normal distribution of the error terms u_i

[1] Note that in the tobit model, zeroes are assumed to be the result of a corner solution. In Heckman's model it is assumed that consumers of a certain commodity in a budget survey are a nonrandom sample of the population, so that there is a sample selection problem.

and ε_i and the normalisation $\sigma_u^2 = 1$. Note that in terms of Lee's (1978) model, both equations of the switching regressions model collapsed. For the positive observations we have the following conditional expectation (remark that the latter is also conditional on the explanatory variables, but to keep notation simple this is not formalised here) :

$$E\left[y_i \mid I_i = 1\right] = X_i'\beta + \sigma_{\varepsilon u} E\left[u_i \mid u_i > -W_i'\gamma\right] = X_i'\beta + \sigma_{\varepsilon u} \frac{\phi\left(W_i'\gamma\right)}{\Phi\left(W_i'\gamma\right)}. \tag{14}$$

In the trivial case of zero expenditures, $E\left[y_i \mid I_i = 0\right] = y_i = 0$. So, in general, the following expectation of y_i can be derived :

$$E\left[y_i\right] = E\left[y_i \mid I_i = 1\right] \cdot P\left(I_i = 1\right) + E\left[y_i \mid I_i = 0\right] \cdot P\left(I_i = 0\right) \tag{15}$$

$$E\left[y_i\right] = \left[X_i'\beta + \sigma_{\varepsilon u} \frac{\phi\left(W_i'\gamma\right)}{\Phi\left(W_i'\gamma\right)}\right] \Phi\left(W_i'\gamma\right) + 0 \cdot \left[1 - \Phi\left(W_i'\gamma\right)\right] \tag{16}$$

$$E\left[y_i\right] = \Phi\left(W_i'\gamma\right) \cdot X_i'\beta + \sigma_{\varepsilon u} \phi\left(W_i'\gamma\right). \tag{17}$$

This equation can be estimated in two steps using all observations, including the zeroes (see, for example, Maddala, 1991). On the other hand, in the Heien and Wessells approach, it is implicitly assumed that :

$$E\left[y_i \mid I_i = 0\right] = X_i'\beta + \sigma_{\varepsilon u} E\left[u_i \mid u_i \leq -W_i'\gamma\right] = X_i'\beta - \sigma_{\varepsilon u} \frac{\phi\left(W_i'\gamma\right)}{1 - \Phi\left(W_i'\gamma\right)}. \tag{18}$$

Consequently, if we substitute the conditional expectations (14) and (18) into

(15), we obtain the following :

$$E\left[y_i\right] = \left[X_i'\beta + \sigma_{\varepsilon u}\frac{\phi\left(W_i'\gamma\right)}{\Phi\left(W_i'\gamma\right)}\right]\Phi\left(W_i'\gamma\right) + \left[X_i'\beta - \sigma_{\varepsilon u}\frac{\phi\left(W_i'\gamma\right)}{1-\Phi\left(W_i'\gamma\right)}\right]\cdot\left[1-\Phi\left(W_i'\gamma\right)\right] \tag{19}$$

$$E\left[y_i\right] = X_i'\beta. \tag{20}$$

This result corresponds to OLS estimation on all the observations (including the zeroes), so that this approach yields inconsistent estimates of the regression coefficients. The estimation procedure was also applied by Heien and Durham (1991), Warnaar and Van Praag (1997) and Saha, Capps and Byrne (1997), in the context of systems of regression equations. Byrne, Capps and Saha (1996) used the approach for the estimation of a single equation.

4. A simple illustration

To illustrate the above result, an Engel curve for expenditures on public transport is estimated on the Belgian Household Budget Survey of 1995-1996. Public transport was chosen because it may be the case that users of it form a nonrandom sample of the population (remark that 62% of all households do not report any expenditures on this commodity). Consequently, a sample selection model may be adequate to use. In what follows, the Engel curve is estimated by OLS on all observations, by the Heien and Wessells approach and by Heckman's two-stage estimation procedure. A Working-Leser Engel curve was chosen as functional form (Leser, 1963) :

$$w_{ih} = \alpha_{ih} + \beta_{ih} \log x_h \tag{21}$$

where w_{ih} is the budget share of commodity i, x_h are total expenditures and α_{ih} and β_{ih} are coefficients which may depend on other household characteristics. In this illustration, we assume that household characteristics enter only in the constant.

Table 1 shows the estimated parameters of the selection mechanism with their standard errors and t-ratios. It is clear from the table, that the probability of observing positive expenditures on public transport decreases significantly with the logarithm of total expenditures. Due to the positive coefficient associated with the squared of the latter, this probability decreases less than proportional with the logarithm of total expenditures. Further, the probability of having non-limit observations increases significantly with the number of household members and with the age of the head of the family. Also living in the Brussels and Walloon region increases this probability[2]. As to social status, households with a white-collar head of the family, or a head of the family which is unemployed, are more likely to have positive expenditures on public transport.

[2] As to the dummy variables, Flemish region and self-employed head of the family are chosen as the base.

Table 1 : Parameter estimates of the selection mechanism

Variable	**coeff.**	**s.e.**	***t*-ratio**
Constant	42.9868	13.6284	3.154
$\log x$	-6.6391	2.0009	-3.318
$(\log x)^2$	0.2497	0.0734	3.401
Number of earners	-0.0404	0.0566	-0.714
Number of household members	0.0954	0.0248	3.847
Age head of the family	0.0772	0.0302	2.557
Brussels region	0.7784	0.0830	9.383
Walloon region	0.1328	0.0553	2.399
White-collar head of the family	0.4299	0.0922	4.665
Blue-collar head of the family	0.1618	0.1005	1.610
Retired head of the family	0.0294	0.1365	0.215
Other non-employed head of the family	0.4089	0.1405	2.910

The above results are now used to obtain the correction terms for the second stage of both the Heien and Wessells approach and Heckman's two-stage estimation procedure. In table 2, the results of both procedures are shown, together with the OLS results on all observations. It is clearly seen from the table, that the Heien and Wessells approach and the OLS estimator obtain the same inconsistent parameter estimates, except for the correction term of course.

As to the more appropriate Heckman's two-stage estimation procedure, there is a significant (with corrected standard errors) negative relationship between the share of public transport and the logarithm of total expenditures. Further, the share declines significantly with the age of the head of the family. A test of the null hypothesis of no sample selection bias ($\sigma_{\varepsilon u} = 0$ in equation (14)) has been performed, using the standard t-test on the uncorrected standard errors (see, Heckman, 1979). The alternative hypothesis of sample selection bias could indeed not be rejected, so that the take up of the correction term was appropriate. Finally, in order to illustrate the importance of using the correct parameter estimates, income elasticities for both Heckman's two-stage estimation procedure and OLS were calculated. Evaluated at the average budget share of those households having positive expenditures on public transport, the income elasticities equal 0.035 and 0.739, respectively. As to the character of the commodity, public transport appears to be a necessity under both estimation methods. The difference in magnitude between both income elasticities however, is rather large. According to Heckman's two-stage estimation procedure, public transport even tends to an inferior commodity.

Table 2 : Engel curve estimates

	Heckman		**H & W**		**OLS**	
Variable	**coeff.**	**s.e.**	**coeff.**	**s.e.**	**coeff.**	**s.e.**
Constant	0.2369	0.0359	0.0548	0.0069	0.0548	0.0083
$\log x$	-0.0139	0.0019	-0.0037	0.0005	-0.0037	0.0006
Number of earners	0.0002	0.0013	-0.0002	0.0004	-0.0002	0.0005
Number of household members	-0.0011	0.0008	0.0005	0.0002	0.0005	0.0002
Age head of the family	-0.0026	0.0009	-0.0004	0.0002	-0.0004	0.0003
Brussels region	-0.0087	0.0050	0.0053	0.0006	0.0053	0.0008
Walloon region	-0.0023	0.0015	0.0005	0.0004	0.0005	0.0005
White-collar head of the family	-0.0046	0.0034	0.0026	0.0007	0.0026	0.0008
Blue-collar head of the family	-0.0029	0.0025	0.0002	0.0008	0.0002	0.0009
Retired head of the family	-0.0023	0.0031	-0.0000	0.0010	-0.0000	0.0013
Other non-employed head of the fam.	-0.0038	0.0043	0.0045	0.0011	0.0045	0.0013
Correction term	-0.0185	0.0096	0.0089	0.0003		

5. Conclusion

In this paper, it is shown that the two-step estimation procedure proposed by Heien and Wessells (1990) to estimate demand on budget surveys, obtains inconsistent parameter estimates. This procedure, which makes use of Heckman-type correction terms for both positive and zero expenditures, draws from a wrong interpretation of the switching regressions model that is nested in Lee's (1981) generalised framework. It can be shown that this approach gives the same inconsistent estimates as if ordinary least squares would be applied on all observations, including the zeroes.

References

[1] Byrne, P., Capps O., Saha A. (1996) Analysis of food-away-from-home expenditure patterns for U.S. households, 1982-89, *American Journal of Agricultural Economics,* **78**, 614-627.

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

[3] Heien, D., Durham, C. (1991) A test of the habit formation hypothesis using household data, *The Review of Economics and Statistics*, **73**, 189-199.

[4] Heien, D., Wessells, C. (1990) Demand systems estimation with microdata : a censored regression approach, *Journal of Business & Economic Statistics,* **8**, 365-371.

[5] Lee, L.-F. (1978) Unionism and wage rates : a simultaneous equations model with qualitative and limited dependent variables, *International Economic Review,* **19**, 415-433.

[6] Lee, L.-F. (1981) Simultaneous equations models with discrete and censored dependent variables, in : Manski, C., McFadden, D. (Eds.), *Structural analysis of discrete data with econometric applications*, MIT Press, Cambridge Mass., pp. 346-364.

[7] Leser, C. (1963) Forms of Engel functions, *Econometrica*, **31**, 694-703.

[8] Maddala, G. (1991) *Limited-dependent and qualitative variables in econometrics*, Cambridge University Press, Cambridge.

[9] Saha, A., Capps, O., Byrne, P. (1997) Calculating marginal effects in models for zero expenditures in household budgets using a Heckman-type correction, *Applied Economics*, **29**, 1311-1316.

[10] Warnaar, M., Van Praag, B. (1997) How Dutch teenagers spend their money, *De Economist*, **145**, 367-397.