Indirect estimation of a simultaneous limited dependent variable model for patient costs and outcome

Per Hjertstrand
Research Institute of Industrial Economics (IFN), Stockholm, Sweden – Per.Hjertstrand@ifn.se

Gunnar Rosenqvist*
Hanken School of Economics, Helsinki, Finland – gunnar.rosenqvist@hanken.fi

Abstract

This paper proposes applying indirect inference to estimate simultaneous equation limited dependent variable models. The model under study is motivated by the important problem of studying cost and quality, their determinants and their interrelationship in health care with patient level data. Models of this type are commonly estimated by two-step estimators suggested in the literature. In our simulations the indirect inference estimator outperforms its two-stage competitors.

Keywords: indirect inference; health economics; simultaneous equation system.

1. Introduction

This paper proposes applying indirect inference (II) (Gourieroux *et al.*, 1993) to estimate simultaneous equation limited dependent variable models. II estimators have been shown to be highly useful in many different situations when the likelihood is difficult to deal with, but the model is simple to simulate, (e.g. Westerlund & Hjertstrand, 2014). Its primary advantage is its generality: unlike other methods that may require optimizing of a complicated criterion function, as is the case with maximum likelihood estimation in many econometric models, the indirect inference technique applies a simplified approximate model, and basically updates its estimates via simulations from the more complicated underlying structural model to obtain consistent estimates. Consequently, II is applicable in a broad range of model specifications including nonlinear models, making it very suitable for the kind of models considered in this paper.

1.1. Motivation

This paper originates from the important area within health economics of studying costs and quality in health care. An increasing literature in health economics is aimed at measuring patient costs and quality of care and determining their determinants and their interrelationship. Achieving high quality of care and containing costs are both important goals for the health care. Hence performance evaluation has to consider both costs and quality of care. However, these may be conflicting goals, and there may be a trade-off between them. If there is no relation between costs and quality of care, costs may be contained without impairing quality or alternatively quality may be improved without increasing costs. The relationship and possible trade-off between outcome and use of resources in health care has recently been studied by among others Hussey *et al.* (2013), Häkkinen *et al.* (2014), Stargard *et al.* (2014). The particular type of simultaneous equation limited dependent variable model we consider in the following is one possible approach to model the relation between costs and quality in health care with detailed data on patient level.

1.2. Summary of contributions

A number of difficulties arise when studying cost data in health care. First, cost data typically exhibits a distribution characterized by non-negativity, right-skewness and a heavy right tail (e.g. Mihaylova *et al.*, 2011, Iversen *et al.* 2015). This is apparent for the cost data collected within the EUROHOPE project (See Häkkinen *et al.*, 2013 and the EuroHOPE homepage) as in several other projects on cost data. A number of approaches to modelling costs have therefore been suggested, see e.g. Mihaylova *et al.*, (2011), Iversen *et al.* (2015), Jones (2010) and Jones *et al.* (2014). Second, further challenges arise when quality of health care is considered together with costs. Quality is often measured as the time-to-death or as a survival intensity rate. Since these variables are by construction censored, quality will enter non-linearly in the model. This calls for rather flexible multivariate models which are able to handle nonlinearities.

In this paper, we formulate a bivariate simultaneous equation limited dependent variable model, where one (nonlinear) equation expresses the time-to-death dependent on demographic and other covariates while the other equation expresses costs as a function of the time-to-death and covariates. Since the likelihood function is difficult to handle in practice, other estimation procedures may be more appealing. Because the model is simple to simulate, estimators based on indirect inference seem well suited for the purpose of estimating parameters in the model. Simulation results show that the II estimator performs much better than well-known two-step estimation techniques suggested in the literature.

2. Model

As discussed above, cost data typically exhibits a distribution characterized by non-negativity, right-skewness and heavy right tail. Hence, logarithmizing costs is one way of proceeding, and the same transformation is also applied to survival times, as e.g. in Lee *et al.* (2007). We consider the simultaneous equation system

$$\ln y_i = \begin{cases} \ln y_i^* \text{ if } \ln y_i^* < 365 \\ \ln 365 \text{ if } \ln y_i^* \geq 365 \end{cases} \tag{1}$$

$$\ln c_i = \theta \ln y_i + z_i' \beta_2 + \varepsilon_i \tag{2}$$

where

$$\ln y_i^* = x_i' \beta_1 + \eta_i$$

with the error terms η_i and ε_i assumed to follow a bivariate normal distribution with variances σ_η^2 and $\sigma_\varepsilon^2$and correlation ρ, i=1,…,n. In the model y_i can be seen as denoting survival after the start of for example acute myocardial infarction (AMI), with censoring at 365 days, while c_i stands for costs of patient i. The endogenous variable y is observed via a nonlinear transformation of a latent structural continuous variable y^*. This transformation consists of a censoring mechanism in which we observe the structural continuous variable if it is below some known threshold (in this case 365 days follow-up), and observe the value of the threshold if above. The degree of censoring may be high (Lee *et al.*, 2007). The vectors of explanatory variables x_i and z_i contain patient characteristics like age, gender, comorbidities and perhaps also hospital characteristics. Regression coefficients θ and the parameter vectors β_1 and β_2 are to be estimated. An obvious problem of this model is the endogeneity as a consequence of the simultaneity of c and y.

3. Two-step estimation approaches from the literature

Because of the joint problem of censoring and simultaneity, the likelihood is difficult to deal with. Consequently, alternatives to full information maximum likelihood have been derived and proposed in the literature. For simultaneous equation models with limited dependent variables two-step estimators have been suggested (Nelson & Olson, 1978; Heckman, 1978; Amemiya, 1979; Blundell and Smith, 1989). For the kind of models considered in this paper, with a censored endogenous regressor, Vella (1993) present a simple two-step estimator, in which a generalized residuals approach is employed to adjust for the inconsistency caused by the endogeneity of the censored regressor.

Nelson & Olson's (1978) estimator. Equation (1) taken alone constitutes a Tobit model. With this approach it is estimated by maximum likelihood (ML). In the second stage $\ln y_i$ on the right hand side of equation (2) is replaced by its predicted value from equation (1), and equation (2) is then estimated by least squares. This procedure is analogous to traditional two-stage least squares for systems of linear equations.

Vella's (1993) estimator. Following Vella (1993), the expected value of (2) conditional upon y_i is

$$E(\ln(c_i)|y_i) = \theta \ln y_i + z_i'\beta_2 + E(\epsilon_i|y_i) \qquad (3)$$

where $E(\epsilon_i|y_i) = \lambda\tilde{\eta}_i$,

$$\tilde{\eta}_i = I_i\hat{\eta}_i + (1 - I_i)\hat{\sigma}_\eta \frac{\hat{\phi}_i}{1-\hat{\Phi}_i}, \qquad (4)$$

$\hat{\phi}_i$ and $\hat{\Phi}_i$ are the probability distribution function (pdf) and the cumulative distribution function (cdf) of the standardized normal $\hat{\phi}_i = \phi((\ln(365) - x_i'\hat{\beta}_1)/\hat{\sigma}_\eta)$ and $\hat{\Phi}_i = \Phi((\ln(365) - x_i'\hat{\beta}_1)/\hat{\sigma}_\eta)$,

$$\lambda = Cov(\eta_i, \varepsilon_i)/\sigma_\eta^2,\ \hat{\eta}_i = \ln y_i - x_i'\hat{\beta}$$

and I_i is an indicator for non-censoring (= 1 for uncensored observations, 0 otherwise). Again, equation (1) is first estimated by Tobit ML. In the second stage equation (2) is estimated with least squares including the generalized residual, as motivated by (3) and (4).

Terza *et al.* (2008) found that residual inclusion (as suggested by Vella, 1993) performs better than predictor substitution (as suggested by Nelson & Olson, 1978).

4. Indirect inference

Indirect inference (Gourieroux *et al.*, 1993) consists of an auxiliary model which is used for estimation, and a simulation routine to simulate data under the structural model. The idea is that the auxiliary model should be easier to estimate than the structural model. In our case the structural model is given by (1) and (2).

As auxiliary model we use equations (1) and (3) which we estimate with a simplified version of Vella's approach: we estimate each of equations (1) and (3) with least squares, treating the sample as uncensored, and with ordinary residuals in the second stage. This does not give consistent estimates of the parameters in the structural model, but it is computationally simple and fast.

Let the parameters in the structural model be $\Phi = (\beta_1', \beta_2', \theta, \sigma_\eta^2, \sigma_\varepsilon^2, \rho)$ where β_1, β_2 and θ are regression coefficients, σ_η^2 and σ_ε^2 are variances of the error terms in (1) and (2) and ρ is their correlation coefficient. Let $\Psi = (\alpha_1', \alpha_2', \gamma, \sigma_w^2, \sigma_v^2, \rho_{wv})$ be the corresponding parameters in the auxiliary model, and $\hat{\Psi}$ its estimate based on the actual data.

Now we simulate S independent data sets from the 'true' underlying model (1) and (2), say $\{y_1^s, y_2^s, \dots, y_n^s\}$ and $\{c_1^s, c_2^s, \dots, c_n^s\}$, s = 1, 2, ..., S. For each of these simulated data sets we compute $\tilde{\Psi}_s(\Phi)$ by solving the bivariate sequence of least squares criterion functions that constitutes the auxiliary model.

Thus, for each generated data set we get an estimate $\tilde{\Psi}_s(\Phi)$ of the parameters in the auxiliary model, the whole series of such estimates being $\{\tilde{\Psi}_1(\Phi), \dots, \tilde{\Psi}_S(\Phi)\}$. Finally the indirect inference estimator of Φ is obtained as

$$\hat{\Phi} = \arg\min_{\Phi}(\hat{\Psi} - \tfrac{1}{S}\textstyle\sum_{s=1}^{S} \tilde{\Psi}_s(\Phi))'(\hat{\Psi} - \tfrac{1}{S}\textstyle\sum_{s=1}^{S} \tilde{\Psi}_s(\Phi)).$$

5. Monte Carlo simulation

5.1 Simulation set up

Data are generated from the model

$$\mathrm{y_i} = \begin{cases} x_i'\beta_1 + \eta_i & \text{if } x_i'\beta_1 > -\eta_i \\ 0 \text{ otherwise} \end{cases} \tag{5}$$

$$c_i = \theta y_i + z_i'\beta_2 + \varepsilon_i \tag{6}$$

with η_i and ε_i bivariate normal $(0,0,\sigma_\eta^2,\sigma_\varepsilon^2,\rho)$. The coefficients are given fixed values, $\beta_1 = \beta_2 = (1,1,1,1,1)'$ and θ = $\sigma_\eta^2 = \sigma_\varepsilon^2$ = 1. For the correlation coefficient we use four different values, $\rho \in$ {0; 0.25; 0.5; 0.75}. Four degrees of censoring are applied (by adjusting the constant in the first equation): 20%, 40%, 60% and 80%. Four sample sizes are applied, n ϵ {250; 500; 1500; 5000}. This gives altogether 64 simulation settings, with 1000 replications of each. For each replication the number of simulated II-data sets from the assumed 'true' structural model (5) and (6) is S = 250.

We also need values for the explanatory variables. In each of the vectors $\mathrm{x_i}$ and $\mathrm{z_i}$ we choose to have two continuous and two discrete variables in addition to a constant. The discrete variables are drawn from independent Bernoulli distributions with success probability 0.5. The continuous variables are drawn from independent chi-squared distributions with one degree of freedom. One discrete and one continuous variable common for $\mathrm{x_i}$ and $\mathrm{z_i}$.

5.2 Results

The Nelson & Olson estimator is outperformed by the Vella estimator and the indirect inference estimator. We therefore report results only for the latter two estimators. To save space we focus in the following on root mean square error (RMSE). As an example Table 1 gives RMSE for estimated parameters for sample size n = 500. Similar results are obtained for the other sample sizes. Rows in the table correspond to different values of the correlation ρ between the error terms of the model, while pairs of columns correspond to different censoring rates (20%, 40%, 60%, 80%).

Estimator	II	Vella	II	Vella	II	Vella	II	Vella
Corr.\Cens.rate	20%		40%		60%		80%	
0	0.0679	0.0818	0.0751	0.0925	0,0875	0.1123	0.1138	0.1635
0.25	0.0689	0.082	0.0736	0.0933	0.0864	0.1136	0.1145	0.1665
0.5	0.0684	0.0832	0.0733	0.0939	0.0866	0.1134	0.1142	0.1696
0.75	0.0658	0.0819	0.071	0.0935	0.0851	0.114	0.1142	0.1708

Table 1. Total root mean square error of the estimated regression parameters for sample size n = 500.

It is seen in Table 1 that RMSE for the indirect inference estimator in all cases falls below that of the Vella estimator. There is (considering also the results for the other sample sizes, although not reported here) a tendency of improvement of RMSE of the II estimator relative to that of the Vella estimator as correlation increases. Over sample sizes (200, 500, 1500, 5000) and over correlation coefficients the RMSE for Vella's estimator exceeds that of the indirect inference estimator as described by Table 2.

Share of censoring	RMSE of Vella above that of II
20%	19-25%
40%	20-32%
60%	21-38%
80%	27-60%

Table 2. Degree of excess RMSE of Vellas estimator in comparison with the indirect inference estimator over sample sizes and correlation coefficients.

In Table 2 we see that the difference in performance between the two estimators increases with the truncation. For absolute bias the difference between the results for the II and Vella estimators are much in line with those for RMSE, and for bias as such the results are even more in favour of the II estimator. Concerning bias it turns out that the indirect estimator has much lower bias than Vella’s estimator, and that this result is stable over varying degrees of correlation between the error terms.

6. Conclusions

In empirical economics two-stage estimators as suggested e.g. by Vella (1993) are commonly applied. In our simulation results the indirect inference estimator outperforms its two-stage competitors in terms of bias and root mean square error. The II procedure is of interest whenever the model of primary interest is easy to simulate from but perhaps difficult to estimate directly.

References

Amemiya, T. (1979). The estimation of a simultaneous equation Tobit model. *International Economic Review*, **20**, 169-181.

Blundell, R. & Smith, R, (1989). Estimation in a class of simultaneous equation limited dependent variable models. *Review of Economic Studies*, **56**, 37-58.

EuroHOPE homepage http://www.eurohope.info/

Gourieroux, C., Monfort, A. & Renault, E. (1993). Indirect inference. *Journal of Applied Econometrics*, **8**, S85-S118.

Häkkinen, U., Iversen, T., Peltola, M., Seppälä, M., Malmivaara, A., Belicza, E., Heijink, R., Fattore, G., Numerato, D., Medin, E. & Rehnberg, C. (2013). Health care performance comparison using a disease-based approach: the EuroHOPE project. *Health Policy*, **112,** 100-109.

Häkkinen, U., Rosenqvist, G., Peltola, M., Kapiainen, S., Rättö, H., Cots, F.,Geissler, A., Or, Z., Serden, L. & Sund R. (2014). Quality, cost, and their trade-off in treating AMI and stroke patients in Eurpean hospitals. *Health Policy*, **117**, 15-27.

Heckman, J. (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrica*, **46**, 931-959.

Hussey, P., Werthmeier, S. & Mehrotra, A. (2013). The association between health care quality and costs. *Annals of Internal Medicines*, **158**, 27-34.

Iversen, T., Aas, E., Rosenqvist, G. & Häkkinen, U. (2015). Cost analysis and treatment costs in EuroHOPE. Working paper. Earlier version available as EuroHOPE Discussion Papers No 2 at http://www.eurohope.info/doc/EHDP2_Cost.pdf

Jones A., (2010). *Models for health care*. HEDG Working Paper 10/01. The University of York.

Jones A., Lomas J. & Rice N. (2014). Applying beta-type size distributions to healthcare cost regressions. *Journal of Applied Econometrics*, **29**, 649–670.

Lee, M., Häkkinen, U. & Rosenqvist, G. (2007). Finding the best treatment under heavy censoring and hidden bias. *Royal Statistical Society. Journal. Series A: Statistics in Society*, **170**, 133-147.

Mihaylova, M., Briggs, A., O'Hagan, A. & Thompson, S.G. (2011). Review of statistical methods for analysing healthcare resources and costs. *Health Economics*, **20**, 897-916.

Nelson, F. & L. Olson (1978). Specification and estimation of a simultaneous equation model with limited dependent variables. *International Economic Review*, **19**, 659-705.

Stargardt, T., Schreyögg, J. & Kondofersky, I. (2014). Measuring the relationship between costs and outcomes: the example of acute myocardial infarction in German hospitals. *Health Economics*, ***23***, 653-669.

Terza, J. V., Basu, A., Rathouz, .P. J. (2008). Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modelling. *Journal of Health Economics* **27**, 531-543.

Vella, F. (1993). A simple estimator for simultaneous models with censored endogenous regressors. *International Economic Review*, **34**, 441-457.

Westerlund, J. & Hjertstrand, P. (2014). Indirect estimation of semiparametric binary choice models. *Oxford Bulletin of Economics and Statistics*, **76**, 298-314.