Sample Selection Correction in Panel Data Models When Selectivity Is Due to Two Sources

Cinzia Di Novi

POLIS - Department of Public Policy and Choice

University of Eastern Piedmont, Alessandria.

Objective

This paper proposes a specification of Wooldridge's (1995) two step estimation method in which selectivity bias is due to two sources rather than one.

The main objective of the paper is to show how the method can be applied in practice.

The application concerns an important problem in health economics: the presence of adverse selection in the private health insurance markets on which there exists a large literature.

The data for the empirical application is drawn from the 2003/2004 Medical Expenditure Panel Survey in conjunction with the 2002 National Health Interview Survey.

Rational

In many applied economic problems, it is possible to observe data only for a subset of individuals from the overall population.

When observations are selected in a process that is not independent of the outcome of interest a problem of sample selection may arise.

Statistical Problem

The key idea of the application is to test whether the individuals who are more exposed to health risks also buy insurance contracts with more coverage or higher expected payments.

The critical statistical problem is that the extension of insurance is only measured for those who are insured and face positive health care expenditure. So there is a possible sample selection bias effect.

Methodology

In order to test for differences in insurance purchases by high and low risk profile individuals we use as a measure of completeness of coverage the natural logarithm of health insurance reimbursement (i.e. of health care expenditure paid by private insurance) as a share of total health expenditures (Keeler et al., 1977, Browne and Doerpinghaus, 1993).

However, it is only defined for those who participate in insurance and face positive health care expenditure, thus, it is only defined for a subset of individuals from the overall population.

Methodology

Thus, the model may suffer from sample selection bias and straightforward regression analysis may lead to inconsistent parameters estimate.

Another problem that arises from the estimation is the presence of unobserved heterogeneity in the equations of interest.

If the selection process is time constant, panel estimators solve both problems. But this is not the case.

Methodology

Wooldridge (1995) has proposed an estimator which deals with both sources of estimation bias.

We extend this estimation method to the case in which selectivity is due to two sources rather than one (participation in insurance and participation in health care expenditure).

Wooldridge's(1995) two-step estimation

We consider the following characterization of the Wooldridge's sample selection model where the selectivity bias is a function of two indices

$$d^*_{it_1} = z_{it_1}\gamma_1 + \mu_{i_1} + u_{it_1}$$

$$d_{it_1} = 0 \quad if \quad d^*_{it_1} \leq 0$$

$$d_{it_1} = 1 \quad if \quad d^*_{it_1} > 0$$

Selection equation$_{(1)}$. Let d_{it_1} be an unobserved variable denoting the insurance participation decision.

Wooldridge's(1995) two-step estimation

$$d^{*}_{it_2} = z_{it_2}\gamma + \mu_{i_2} + u_{it_2}$$

$$d_{it_2} = 0 \quad if \quad d^{*}_{it_2} \leq 0$$

$$d_{it_2} = 1 \quad if \quad d^{*}_{it_2} > 0$$

Selection equation$_{(2)}$. Let d_{it_2} be an unobserved variable denoting the health care expenditure participation. decision.

$$y^{*}_{it} = x_{it}\beta + \alpha_i + \varepsilon_{it}$$

$$y_{it} = y^{*}_{it} \quad if \quad d_{it} = 1$$

$$y_{it} \quad not \quad observed \quad otherwise$$

Outcome/Primary equation. Let y_{it} be an unobserved variable the natural logarithm of health insurance reimbursement (i.e. of health care expenditure paid by private insurance) as a share of total health expenditures

Wooldridge's(1995) two-step estimation

By substituting Chamberlain characterization into the selection equations yields:

$$d^*_{it_1} = z_{it_1}\gamma + z_{i1_1}\delta_{1_1} + \ldots + z_{it_1}\delta_{t_1} + v_{it_1}$$

$$d^*_{it_2} = z_{it_2}\gamma + z_{i1_2}\delta_{1_2} + \ldots + z_{it_2}\delta_{t_2} + v_{it_2}$$

Wooldridge's(1995) two-step estimation

The sample selection is now based on two criteria.

The method of estimation relies crucially on the relationship between v_{it_1} and v_{it_2} .

In particular, the estimation depends on whether the two error terms are independent or correlated, that is whether or not $Cov(v_{it_1}, v_{it_2}) = 0$.

The simplest case is when the disturbances are uncorrelated (Maddala,1983, Vella, 1998).

In that cases we can easily extend the Wooldridge's two-step estimation method to this model.

Wooldridge's(1995) two-step estimation

The correction term to include as regressor in the primary equation is:

$$E\left[\varepsilon_{it}\left|z_{it}, d_{it_1}=1, d_{it_2}=1\right.\right]$$
$$=\rho_{t_t}\lambda_1\left(z_{i1_1}\gamma_{1_1}+\ldots+z_{it_1}\gamma_{t_1}\right)+\rho_{t_2}\lambda_2\left(z_{i1_2}\gamma_{1_2}+\ldots+z_{it_2}\gamma_{t_2}\right)$$

We estimate the following model

$$y_{it}=x_{i1}\psi_1+\ldots+x_{it}\psi_t+x_{it}\beta+$$
$$+(\phi_{t_1}+\rho_{t_1})\hat{\lambda}_1(\bullet)+\left(\phi_{t_2}+\rho_{t_2}\right)\hat{\lambda}_2(\bullet)+e_{it}$$

Wooldridge's(1995) two-step estimation

The procedure consists in first estimating, for each period, by two single cross-sectional probit models, the selection equation one and the selection equation two.

Then, the two corresponding Inverse Mills Ratio can be imputed and included as correction terms in the primary equation.

Thus, by fixed effect or pooled OLS, estimate of the resulting primary equation corrected for selection bias can be done for the sample for which $d_{itj} = 1$.

Wooldridge's(1995) two-step estimation

In the case v_{it_1} and v_{it_2} are correlated, so that , $Cov\left(u_{it_1}, u_{it_2}\right) = \sigma_{12}$ "… the expression get very messy…" (Maddala, 1983) and we have to use for each period cross-sectional bivariate probit methods to estimate γ_{it_1} and γ_{it_2}. Further,

$$E\left[v_{it} \middle| z_{it}, d_{it_1} = 1, d_{it_2} = 1\right] = \rho_{t_1} M_{12} + \rho_{t_2} M_{21}$$

where:

$$M_{ij} = \left(1 - \sigma_{12}\right)^{-1}\left(P_i - \sigma_{12} P_j\right)$$

$$P_j = \frac{\int_{-\infty}^{z_{it1}\gamma_{t1}} \int_{-\infty}^{z_{it2}\gamma_{t2}} u_{it_j} f\left(u_{it_1}, u_{it_2}\right) du_{it_1} du_{it_2}}{F\left(z_{it_1}\gamma_{t_1}, z_{it_2}\gamma_{t_2}\right)}$$

Before starting the estimation we run a preliminary cross-sectional bivariate probit

Bivariate Probit Models for Health Expenditure and Insurance Participation

- The null hypothesis of $Cov(v_{it_1}, v_{it_2}) = 0$ is not rejected. The table below shows the correlation coefficients and the p-value for each year:

Dependent Variables	pho	p-value
Positive Expenditure/ Be Insured 2003	-0.1340	0.893
Positive Expenditure/ Be Insured 2004	-0.3727	0.446

- Since the error terms are independent we can deal with the above model as independent equations (Maddala, 1983).

Main Results

Source	SS	df	MS		
Model	6.33909527	21	.30186168	Number of obs	895
Residual	18.9630248	873	.021721678	F(21, 873)	13.90
Total	25.3021201	894	.028302148	Prob > F	0.0000
				R-squared	0.2505
				Adj R-squared	0.2325
				Root MSE	.14738

reimb_prop	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
high_risk	.0777558	.0198539	3.92	0.000	.0387888	.1167228
ac_limit	.0405989	.0121352	3.35	0.001	.0167812	.0644165
age	.0007162	.0005176	1.38	0.167	-.0002998	.0017322
male	-.0028701	.0134017	-0.21	0.830	-.0291735	.0234333
msa	-.0244871	.0146245	-1.67	0.094	-.0531904	.0042161
northeast	.0043527	.0156153	0.28	0.781	-.0262952	.0350006
midwest	.0264188	.0129779	2.04	0.042	.0009473	.0518902
west	-.0101534	.0146484	-0.69	0.488	-.0389036	.0185968
black	-.0015825	.02237	-0.07	0.944	-.0454878	.0423229
other_race	-.0284706	.0246451	-1.16	0.248	-.0768412	.0199
income	-4.64e-07	1.73e-07	-2.68	0.008	-8.03e-07	-1.24e-07
high_educ_~l	-.0265496	.0242308	-1.10	0.274	-.0741071	.0210078
group_insu~e	.0781624	.0205714	3.80	0.000	.0377873	.1185376
lntot_expe~p	-.0160395	.0054431	-2.95	0.003	-.0267226	-.0053563
lncopayment	-.0383916	.0039923	-9.62	0.000	-.0462273	-.0305559
annual_pre~m	-8.42e-07	2.10e-06	-0.40	0.689	-4.97e-06	3.28e-06
dental_ins	.0438746	.0114497	3.83	0.000	.0214025	.0663467
drugs_ins	.0916885	.022535	4.07	0.000	.0474595	.1359175
insurance_~e	.0462061	.0120624	3.83	0.000	.0225314	.0698807
mills1	-.1566046	.0735553	-2.13	0.034	-.3009704	-.0122388
mills2	-.0899416	.0510747	-1.76	0.079	-.1901851	.0103019
_cons	.5308791	.0531492	9.99	0.000	.4265641	.6351942

Summary

In this paper we discuss Wooldridge's (1995) two step estimator that address the problem of sample selection and correlated individual heterogeneity in selection and outcome equation simultaneously.

We show how it can be extended to the case in which selectivity bias is due to two sources rather than one.

The appropriate selection correction depends on whether the error terms for the two selection equations are independent. Thus we have run, for each year, a "preliminary" cross-sectional bivariate probit to test if $Cov\left(v_{it_1}, v_{it_2}\right) = 0$.

Summary

The bivariate probit indicated that the hypothesis $Cov\left(v_{it_1}, v_{it_2}\right) = 0$ could not be rejected.

Thus, we have estimated the selection equations and constructed the estimate of the selection correction terms using two separated standard probit model estimates for each year in order to calculate the correction terms (IMRs).

The selectivity terms that we have included as a regressor in the equation of interest which is estimated using pooled ordinary least squares (OLS) regression, are simple extensions of those proposed by Wooldridge (1995).

Summary

We have conducted a positive correlation test which estimates the correlation between the amount of insurance an individual buys and his ex-post risk experience.

As indicator of generosity and completeness of health plan, we have employed the natural logarithm of health insurance reimbursement (i.e. of health care expenditure paid by private insurance) as a share of total health expenditures.

Our findings support the hypothesis of a systematic relation between illness of individuals and insurance choice.