Dipartimento di Politiche Pubbliche e Scelte Collettive – POLIS
Department of Public Policy and Public Choice – POLIS

Working paper n. 156

November 2009

Sample selection correction in panel data models when selectivity is due to two sources

Cinzia Di Novi

UNIVERSITA' DEL PIEMONTE ORIENTALE "*Amedeo Avogadro*" ALESSANDRIA

Periodico mensile on-line "POLIS Working Papers" - Iscrizione n.591 del 12/05/2006 - Tribunale di Alessandria

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

Cinzia Di Novi†

Abstract

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 practise. 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.

1 Introduction

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. Since Heckman (1979)'s seminal paper, the problem of sample selection bias has been extensively studied in economics literature with empirical applications. Sample selection has been commonly treated in cross-sectional studies but it has not been often considered a concern in panel data. In fact when the selection process is time constant, panel data estimator may eliminate most forms of unobserved heterogeneity (Vella, 1998; Dustmann and Rochina-Barrachina, 2000; 2007). However, the selection process in

*The author wish to thank M. Piacenza and the participants of the Health Economics Association (AIES), 14th Annual Conference, Bergamo (2009) for helpful comments.

†Polis- Dipartimento di Politiche Pubbliche e Scelte Collettive, Università del Piemonte Orientale, Alessandria, Italy. e-mail: cinzia.dinovi@sp.unipmn.it.

many economic applications is not time constant. Wooldridge has proposed a panel estimator for sample selection models which also accounts for heterogeneity across individuals. In this note we present a new characterization of the Wooldridge's two-steps estimation method: we apply the model to the case in which selectivity is due to two sources rather than one. Then, we apply the proposed model to a test for adverse selection in the private health insurance markets. The data for the empirical application is drawn from the 2003/2004 Agency for Healthcare Research Quality's Medical Expenditure Panel– Household Component (MEPS-HC)[1] in conjunctions with the 2002 National Health Interview Survey (NHIS)[2]. We use a subsample of 496 individuals followed for two years resulting in 992 observations; the subsample includes single individuals of working age (from 18 to 65 years old), they get health insurance through individual markets or through their employers or organizations (such as unions, professional associations, or other groups). For the employers-sponsored private coverage we include in the sample individuals who have the possibility to choose between several plans[3]. 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.

The rest of the paper is organized as follows. Section 2 extends Wooldridge(1995)'s model to the case in which selectivity is due to two sources. Subsections 2.1 and 2.2 present the empirical illustration of the model in detail. Section 3 concludes the paper with a discussion. The definition of the variables, descriptive statistics and tables with estimation coefficients are in Appendix .

[1]MEPS is an annual survey whose main purpose is to examine insurance trends and healthcare utilization among the non-institutionalized population in the United States.

[2]National Centers for Health Statistics (NCHS), (Center for Disease Control and Prevention)- NHIS provides rather detailed information about health status, diseases, life-style, education and other individual characteristics.

[3]In U.S. it is quite common that employers provide health insurance as part of the benefits package for employees. In many employer-sponsored private coverage, employers allow employees to choose between several plans, including both indemnity insurance and managed care. Other employers offer only one plan. Only if employers allow insurers flexibility in designing health insurance plans adverse selection may occur.

2 Wooldridge Estimation With Two Selection Criteria

We start by sketching Wooldridge's (1995) sample selection model with one selection criterion, then we present a specification of the model in which the selection process is based on two selection criteria rather than one. According to Rochina-Barrachina (1999) we consider the following problem:

$$\begin{array}{c} d_{it}^{*} = z_{it}\gamma + \mu_i + u_{it} \\ d_{it} = 0 \quad if \quad d_{it}^{*} \leq 0 \\ d_{it} = 1 \quad if \quad d_{it}^{*} > 0 \end{array} \tag{1}$$

$$\begin{array}{c} y_{it}^{*} = x_{it}\beta + \alpha_i + \varepsilon_{it} \\ y_{it} = y_{it}^{*} \quad if \quad d_{it} = 1 \\ y_{it} \ not\ observed\ otherwise \end{array} \tag{2}$$

where equation (1) defines the selection rule while equation (2) is the primary equation. i $(i = 1, ...n)$ denotes the individuals while t $(t = 1, ..., t)$ denotes the panel. x_{it} and z_{it} are vector of exogenous variables with possibly common elements and definitely with an exclusion restriction. γ and β are unknown parameter vectors to be estimated. Terms μ_i and α_i are unobservable time invariant fixed effects[4] which are possibly correlated with each other. u_{it} and ε_{it} are unobserved disturbances, possibly correlated with each other. The dependent variable in the primary equation(1), y_{it}, is observed only for the observations satisfying the selection rule i.e. only if the indicator variable $d_{it} = 1$.

Similar to Chamberlain (1980), Wooldridge (1995) assumes the fixed effects in the equation (2) have the following relationship:

$$\mu_i = z_{i1}\delta_1 + ... + z_{it}\delta_t + c_i \tag{3}$$

where c_i is a random component. By substituting Chamberlain characterization into the selection equation yields:

$$d_{it}^{*} = z_{it}\gamma + z_{i1}\delta_1 + ... + z_{it}\delta_t + v_{it} \tag{4}$$

where $v_{it} = c_i + u_{it}$. v_{it} is distributed independently of z_{it} and it is normally distributed with zero mean and σ^2 variance. The regression function of α_i on z_{it} and v_{it} is linear, accordingly:

$$E\left[\alpha_i \left| z_{it}, v_{it}\right.\right] = x_{i1}\psi_1 + ... + x_{it}\psi_t + \phi_t v_{it} \tag{5}$$

[4]The individual effects are assumed to be the fixed effets rather than the random effects.

We do not observe v_{it}, but only the binary indicator d_{it}. Then, we replace $E[\alpha_i | z_{it}, v_{it}]$ with:

$$E[\alpha_i | z_{it}, d_{it} = 1] = x_{i1}\psi_1 + ... + x_{it}\psi_t + \phi_t E[v_{it} | z_{it}, d_{it} = 1] \quad (6)$$

Wooldridge assumes that ε_{it} is mean independent of z_{it} conditional on v_{it} and its conditional mean is linear on v_{it}:

$$E[\varepsilon_{it} | z_{it}, v_{it}] = E[\varepsilon_{it} | v_{it}] = \rho_t v_{it} \quad (7)$$

By the Law of Iterated Expectation:

$$E[\varepsilon_{it} | z_{it}, d_{it} = 1] = \rho_t E[v_{it} | z_{it}, d_{it} = 1] \quad (8)$$

From the above assumption, Wooldridge derives an explicit expression for

$$\begin{aligned} E[\alpha_i + \varepsilon_{it} | z_{it}, d_{it} = 1] &= E[\alpha_i | z_{it}, d_{it} = 1] + E[\varepsilon_{it} | z_{it}, d_{it} = 1] = \\ &= x_{i1}\psi_1 + ... + x_{it}\psi_t + (\phi_t + \rho_t) E[v_{it} | z_{it}, d_{it} = 1] \end{aligned} \quad (9)$$

where

$$E[v_{it} | z_{it}, d_{it} = 1] = \lambda (z_{i1}\gamma_1 + ... + z_{it}\gamma_t) \quad (10)$$

So, for each period, Wooldridge suggests to estimate a cross-sectional probit model for participation and compute the Inverse Mills Ratio (IMR), then, estimate the structural equation:

$$y_{it} = x_{i1}\psi_1 + ... + x_{it}\psi_t + x_{it}\beta + (\phi_t + \rho_t) \lambda (z_{i1}\gamma_1 + ... + z_{it}\gamma_t) \quad (11)$$

by using fixed effect OLS or pooled OLS for the sample for which $d_{it} = 1$ (Vella, 1998).

In the following we will propose a new specification of Wooldridge two step estimation method extended to the case in which selectivity is based on two indices. We apply the method to a test for adverse selection. The test is based on the hypothesis that there exists a positive correlation between the high risk profile individuals and the extension of health insurance plan. In order to test for differences in insurance purchases by high and low risk profile individuals we use as indicator of completeness of coverage the natural logarithm of health insurance reimbursement (i.e. of healthcare expenditure paid by private insurance) as a share of total health expenditures (Keeler et al., 1977, Browne and Doerpinghaus, 1993). Health insurance reimbursement is only defined for those who participate in insurance and face positive health care expenditure. So, we consider the following characterization of Wooldridge's sample selection model where selectivity bias is a function of two indices:

$$\begin{array}{c} d_{it_1}^* = z_{it_1}\gamma_1 + \mu_{i_1} + u_{it_1} \\ d_{it_1} = 0 \quad if \quad d_{it_1}^* \leq 0 \\ d_{it_1} = 1 \quad if \quad d_{it_1}^* > 0 \end{array} \tag{12}$$

$$\begin{array}{c} d_{it_2}^* = z_{it_2}\gamma_2 + \mu_{i_2} + u_{it_2} \\ d_{it_2} = 0 \quad if \quad d_{it_2}^* \leq 0 \\ d_{it_2} = 1 \quad if \quad d_{it_2}^* > 0 \end{array} \tag{13}$$

$$\begin{array}{c} y_{it}^* = x_{it}\beta + \alpha_i + \varepsilon_{it} \\ y_{it} = y_{it}^* \quad if \quad d_{it} = 1 \\ y_{it} \ not\ observed\ otherwise \end{array} \tag{14}$$

where d_{it_1} is an unobserved variable denoting insurance participation decision and d_{it_2} an unobserved variable denoting health care expenditure participation decision. z_{it_1} , z_{it_2} and x_{it} are vector of exogenous variables with possibly common elements and definitely with an exclusion restriction. y_{it} denotes the natural logarithm of health insurance reimbursement as share of total healthcare expenditure. y_{it} is observed only for the sample for which $d_{it_1} = 1$ and $d_{it_2} = 1$. Terms μ_{i_1}, μ_{i_2} and α_i are fixed effects. u_{it_1}, u_{it_2} and ε_{it} are unobserved disturbances, possibly correlated with each others.

The method of estimation relies crucially on the relationship between v_{it_1} and v_{it_2}[5], in particular, the estimation depends on whether the two error terms are independent or correlated, that is whether or not $Cov\left(v_{it_1}, v_{it_2}\right) = 0$. The simplest case is when the disturbances are uncorrelated (Maddala, 1983; Vella, 1998). If $Cov\left(v_{it_1}, v_{it_2}\right) = 0$ we can easily extend Wooldridge's two-step estimation method to our model. The correction term to include as regressor in the primary equation is:

$$\begin{array}{c} E\left[\varepsilon_{it}\left|z_{it}, d_{it_1} = 1, d_{it_2} = 1\right.\right] = \rho_{t_1}\lambda_1\left(z_{i1_1}\gamma_{1_1} + ... + z_{it_1}\gamma_{t_1}\right) + \\ + \rho_{t_2}\lambda_2\left(z_{i1_2}\gamma_{1_2} + ... + z_{it_2}\gamma_{t_2}\right) \end{array} \tag{15}$$

Then, we estimate the following model:

$$\begin{array}{c} y_{it} = x_{i1}\psi_1 + ... + x_{it}\psi_t + x_{it}\beta + \left(\phi_{t_1} + \rho_{t_1}\right)\lambda_1\left(z_{i1_1}\gamma_{1_1} + ... + z_{it_1}\gamma_{t_1}\right) + \\ + \left(\phi_{t_2} + \rho_{t_2}\right)\lambda_2\left(z_{i1_2}\gamma_{1_2} + ... + z_{it_2}\gamma_{t_2}\right) \end{array} \tag{16}$$

The procedure consists in first estimating, for each period, by two single a cross-sectional probit model, the selection equation one and the selection equation two. Than, the two corresponding Inverse Mills

[5]From Chamberlain trasformation of the individual effects: $v_{it_1} = c_{i_1} + u_{it_1}$ and $v_{it_2} = c_{i_2} + u_{it_2}$

Ratio can be imputed and included as correction terms in the primary equation. Thus, by fixed effect or pooled OLS[6], estimate of the resulting primary equation corrected for selection bias can be done for the sample for which $d_{it_1} = 1$ and $d_{it_2} = 1$.

In the case v_{it_1} and v_{it_2} are correlated, so that $Cov\left(v_{it_1}, v_{it_2}\right) = \sigma^2$ we have to use for each period cross-sectional bivariate probit methods to estimate γ_{it_1} and γ_{it_2}. Further,

$$E\left[\varepsilon_{it}\left|z_{it_1}, z_{it_2} d_{it_1} = 1, d_{it_2} = 1\right.\right] = \rho_{t_1} M_{12} + \rho_{t_2} M_{21} \tag{17}$$

where $M_{ij} = \left(1 - \sigma_{12}\right)^{-1}\left(P_i - \sigma_{12} P_j\right)$ and $P_j = \frac{\int_{-\infty}^{z_{it_1}\gamma_{t_1}} \int_{-\infty}^{z_{it_2}\gamma_{t_2}} v_{it_1} v_{it_2}\ f\left(v_{it_1}, v_{it_2}\right) dv_{it_1} dv_{it_2}}{F\left(z_{it_1}\gamma_{t_1}, z_{it_2}\gamma_{t_2}\right)}$.

2.1 Bivariate Probit Model for Care Expenditure and Insurance

In order to test whether v_{it_1} and v_{it_2} are correlated we run for each year a "preliminary" bivariate probit between insurance and health care expenditure participation. In our model the dependent variable employed to predict the probability of facing positive health care expenditure is a binary variable that takes value one if individuals incur in positive health care expenditure during the year of interview, and zero otherwise. The independent variables employed can been categorized into three dimensions: need for care (need to see a specialist, need to have treatments or tests), predisposition to use health services (age, sex, race) and enabling factors (education, insurance, income, employment status, region and residential location). Among enabling factor, we consider insurance participation. An insured individual, in fact, may consume more medical services and have a greater expenditure compared to an uninsured one (Arrow, 1963; Pauly, 1974). In this application, the situation is further complicated by the fact that insurance participation itself may be affected by the likelihood of having positive health expenditure. The choice of insurance coverage may be affected by planned medical expenditure and expectations about medical care utilization. Thus, in order to test the potential endogeneity of health insurance and at the same time whether the covariance between health insurance choice and health expenditure participation is significantly different of zero, we run for

[6] In this analysis fixed effect however presents a significant limitation with the respect to pooled OLS : we can not assess the effect of variables that do not vary very much within group: i.e. degree of education, race, region, etc. that can impact significantly the health insurance reimbursement. Also, explanatory variables whose change across time is constant – e.g. age – can not be included.

each year a cross sectional recursive bivariate probit models (Maddala, 1999). For each period, the recursive structure builds on a first reduced form equation for the potentially endogenous dummy measuring insurance participation and a second structural form equation determining the expenditure participation:

$$d^*_{it_1} = z_{i1_1}\gamma_{1_1} + ... + z_{it_1}\gamma_{t_1} + v_{it_1} \tag{18}$$

$$\begin{aligned} d^*_{it_2} &= z_{i1_2}\gamma_{1_2} + ... + z_{it_2}\gamma_{t_2} + v_{it_2} = \\ &= z_{i1_2}\gamma_{1_2} + ... + d_{it_1}\zeta + w_{it}\xi + v_{it_2} \end{aligned} \tag{19}$$

where $d^*_{it_1}$ and $d^*_{it_2}$ are latent variables, and d_{it_1} and d_{it_2} are dichotomous variables observed according to the rule:

$$\begin{cases} d_{it_j} = 0 & if \quad d^*_{it_j} \leq 0 \\ d_{it_j} = 1 & if \quad d^*_{it_j} > 0 \end{cases} ; \; j = 1, 2 \tag{20}$$

z_{it_j} and w_{it} are vectors of exogenous variable with possibly common elements, γ and ξ are parameter vectors, ζ is a scalar parameter. The dependent variable d_{it_1} used to predict the probability of being insured is again a dummy variable that takes value one if respondents are insured and zero otherwise. The vector of explanatory variables z_{it_1} used to predict the probability of being insured includes both exogenous variables that are determinants of health expenditure and personal attributes that are only determinative of health insurance choice[7] (i.e. employment status, union status, insurance attitude[8]).

We assume that, for each period, the error terms v_{it_1} and v_{it_2} are distributed as bivariate normal, with zero mean and variance covariance

[7] Estimation of a recursive bivariate probit model requires some considerations for the identification of the model parameters: at least one of the insurance equation exogenous variables has not to be included in the expenditure equation as explanatory variable (Maddala, 1983). Following Maddala's approach we include among explanatory variables in the insurance equation a measure of attitude toward health insurance participation and the indicator of employment status and union status.

[8] MEPS contains a self-administered questionnaire (SAQ) with questions that ascertain health-related attitudes; respondents were asked if they agree strongly, or disagree with the following statements "Health insurance is not worth the money it costs"; " I am more likely to take risks than the average person". The first statement is directly related to an individual's preferences for health insurance: respondent is asked to directly assess the value of health insurance relative to his perception of its cost. In contrast, the second statement provides indirect measures that are likely to be associated with attitudes toward health insurance. While individual's preferences for health insurance may affect the extent of insurance purchase, attitude toward health insurance might influence decisions to purchase health insurance. Hence we include the first indicator in the structural equation for insurance reimbursement, and the second one in the insurance participation equation.

matrix Σ. Σ has values of 1 on the leading diagonal and correlations $\rho_{12} = \rho_{21}$ as off-diagonal elements:

$$\begin{pmatrix} v_{it_1} \\ v_{it_2} \end{pmatrix} \sim IIDN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{12} \\ \rho_{21} & 1 \end{bmatrix}\right) \tag{21}$$

In the above setting, the exogeneity condition is stated in terms of the correlation coefficient, which can be interpreted as the correlation between the unobservable explanatory variables of the two different equations. The two selection equations can be estimated separately as single probit models only in the case of independent error terms v_{it_1} and v_{it_2} i.e. the coefficient ρ_{jk} is not significantly different of zero ($k = 1, 2$). If the error terms v_{it_1} and v_{it_2} are independent we can deal with the above model as independent equations (Maddala, 1983) and apply the model in the equation (16).

Table 3 shows the correlation coefficients and the p-value for each year sample: the null hypothesis of $Cov\left(v_{it_1}, v_{it_2}\right) = 0$ is not rejected; hence, we can deal with the model in the equation (16) and compute Inverse Mills Ratio by using the two selection equations as single probit models. Tables 4 and 5 show coefficients for insurance choice and expenditure participation equation estimated using bivariate probit specification.

2.2 Empirical Illustration of Structural Equation

In order to perform the correlation test, first we classify individuals as being high and low risk profile individuals. Individuals are classified as being high-risk if their health status is not good. As a measure of health status we use SAH (self-assessed health)[9], which is a five category variable rating from poor to excellent. We construct a binary variable (*high_risk*) with the value one if individuals report that their health status is fair or poor and zero otherwise (excellent, very good, good). Then, we classify as high-risk individuals those whose self-reported health is fair or poor. In addition to the health indicator, the independent variables, used to control for differences in policy, can be grouped in the following categories: demographic variables (age, sex), socioeconomic variables (education, income[10]) individual's prefer-

[9]SAH is supported by a large literature that shows the strong predictive relationship between people's self rating of their health and mortality or morbidity (Idler and Beyamini, 1997; Kennedy et al. 1998). Moreover, self assessed health correlates strongly with more complex health indices such as functional ability or indicators derived from health service use (Unden and Elofosson, 2006).

[10]We do not include in the structural equation employment and union status among socioeconomic variables to avoid multicollinearity problems since they are strictly

ences for health insurance, health insurance plan characteristics (out-of-pocket annual premium, co-payment, whether insurance plan covers prescription drug costs and dental bills, whether respondents get their insurance through their employers or other organizations), observable risk (whether individuals suffer from any form of disabilities that limit their activities[11]).Moreover, we control for the healthcare expenditure paid by other sources different of insurance company.

Table 6 shows the coefficients for the structural insurance reimbursement equation estimated using pooled OLS specification. We find evidence for adverse selection: table 6 shows that the coefficient estimate for the variable "high_risk" is positively and significantly correlated with the health insurance reimbursement. Other than regular variables, two independent variables here are the IMR (Inverse Mills Ratio) which have been estimated from the first and second probit selection equations. When added to the outcome equation as additional regressors, they measure the sample selection effect due to lack of observations on the non-health insurance purchasers and non-health expenditure participants. These variables should be statistically significant to justify the use of Wooldridge two-step estimation. Since in our models they are statistically significant there may be sample selection problem in the data and we need to use the extension of Wooldridge method.

3 Summary and Conclusions

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 $Cov\left(v_{it_1}, v_{it_2}\right) = 0$. 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 included as a regressor in the equation of

correlated with the variable that measure whether respondents have an employer or union-sponsored private coverage.

[11]The variable that we use as indicator of limited activity controls for the portion of risk observable to the insurer. The activity limitations indicator is expected to be positively related to the generosity of the health insurance plan, because being limited increases the likelihood of need for medical care

interest (estimated using pooled ordinary least squares regression) are simple extensions of those proposed by Wooldridge (1995).

Since not many studies exist that use this method in practise, we have applied the proposed model. The application concerns an important problem in health economics: the presence of adverse selection in the private health insurance markets. We have tested whethere there exists a positive 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.

References

[1] G.Akerlof (1970), “The Market for 'Lemons': Quality Uncertainty and the Market Mechanism”, *Quarterly Journal of Economics*, **84**: 488-500.

[2] K.J.Arrow (1985), ”The Economics of Agency,” in John W. Pratt and Richard J.Zeckhauser (Eds.), Principals and Agents: The Structure of Business, Boston, MA: Harvard Business School Press: 37-51.

[3] P. Bajari, H.Hong, Khwaja (2006), ”Moral Hazard, Adverse Selection and Health Expenditure: A Semiparametric Approach Analysis” National Bureau of Economics Research, Working Paper N.12445.

[4] M.J. Browne, H.I.Doerpinghaus (1993), ”Information Asymmetries and Adverse Selection in the Market for Individual Medical Expense Insurance”, *The Journal of Risk and Insurance*, 60: 300-312.

[5] J.H. Cardon; I.Hendel (2001), ”Asymmetric Information in Health Insurance: Evidence from the National Medical Expenditure Survey”, *The RAND Journal of Economics*, **32**: 408-42.

[6] G.Chamberlain, (1980), ”Analysis with Qualitative Data”, *Review of Economic Studies*, **47**: 225-238.

[7] P.A Chiappori, B. Salanié (2000), ”Testing Contract Theory: A Survey of Some Recent Work”, invited lecture World Congress of the Econometric Society Seattle, August 2000.

[8] P.A Chiappori, B. Salanié (2000), ”Testing for Asymmetric Information in Insurance Markets”, *The Journal of Political Economy*, **108**: 56-78.

[9] P.Contoyannis, A. M. Jones, (2004), ”Socio-Economic Status, Health and Lifestyle”, *Journal of Health Economics*, **23**: 965-995.

[10] D.M.Cutler, R.J.Zeckhouser (1997), ”Adverse Selection in Health Insurance”, National Bureau of Economics Research, Working Paper N.6107.

[11] D.M.Cutler, R.J.Zeckhouser (2000),” The Anatomy of Health Insurance ”, Handbook of Health Economics. A.J. Culyer and J. P. Newhouse. North Holland, Elsevier Science B.V. 1A: 563-643.

[12] D.M.Cutler, S.J.Reber (1998), ”Paying for Health Insurance: the Trade-Off between Competition and Adverse Selection”, *The Quartely Journal of Economics,* **113**: 433-466.

[13] Dustmann C., and M.E. Rochina-Barrachina, (2000), Selection Correction in Panel Data Models: an Application to Labour Supply and Wages, IZA Discussion Paper 162 (IZA, Bonn).

[14] A.M.Gill, R.J. Michels (1992), ”Does Drug Use Lower Wage?”, *Industrial and Labor Relation Review,* **45**: 419-434.

[15] J. Heckman, (1979) "Sample Selection Bias as a Specification Error." *Econometrica*, **47**: 153-61.
[16] E. L.Idler, Y.Benyamini, (1997) "Self- Rated Health and Mortality: A Review of Twenty-Seven Community Studies." *Journal of Health and Social Behavior*, **38**:21-37.
[17] Keeler, E. B., J. P. Newhouse, and C. E. Phelps (1977), "Deductibles and the Demand for Medical Care Services: The Theory of a Consumer Facing a Variable Price Schedule Under Uncertainty", *Econometrica*, **4**: 641-656.
[18] B.P.Kennedy, et al. (1998), "Income Distribution, Socio-Economic Status, and Self Rated Health in the United States: Multilevel Analysis, *British Medical Journa*l 317: 917–921.
[19] G.S.Maddala (1983), "Limited Dependent and Qualitative Variables in Econometrics", Cambridge University Press.
[20] S.S.Makki.;A. Somwaru (2001), "Evidence of Adverse Selection in Crop Insurance Markets", *The Journal of Risk and Insurance*, **68**: 685-708.
[21] M.S.Marquis (1992), "Adverse Selection with Multiple Choice among Health Insurance Plans: a Simulation Analysis" *Journal of Health Economics*, **11**:129-151.
[22] P. Olivella, M.V.Hernández, "Testing for Adverse Selection into Private Medical Insurance" , The Institute for Fiscal Studies, Working Paper N. 06/02.
[23] M. V.Pauly (1974). "Overinsurance and Public Provision of Insurance: The Roles of Moral Hazard and Adverse Selection." *Quarterly Journal of Economics* **88**: 44-62..
[24] M.Rochina-Barrachina (1999) "A New Estimator for Panel Data Sample Selection Models," Annales. d'Economie et de Statistique, 55/56.
[25] M.Rothschild, J.Stiglitz (1976), "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information", *The Quarterly Journal of Economics*, **90**: 629-649.
[26] A.L. Undon, S.Elofsson (2006), "Do Different Factors Explain Self-Rated Health in Men and Women?" *Gender Medicine*, **3**, No 4.
[27] F.Vella (1988), "Generating Conditional Expectations from Models with Selectivity Bias." *Economics Letters* , **28**(1):97-103.
[28] F.Vella (1992),. "Simple Tests for Sample Selection Bias in Censored and Discrete Choice Models." *Journal of Applied Econometrics* **7**:413-21.
[29] F.Vella (1993)"A Simple Estimator for Models with Censored Endogenous Regressors." *International Economic Review* **34**:441-57.
[30] F.Vella (1998), "Estimating Models with Sample Selection Bias: A

Survey", *The Journal of Human Resources*, **33**: 127-169.
[31] J.Wooldridge (1995), "Selection Correction for Panel Data Models under Conditional Mean Independence Assumptions", *Journal of Econometrics*, **68**: 115-132.

4 Appendix

Table 1: Variables Name and Definition

Variables Name	*Variables Definition*
age	age in years
male	1 if male, 0 otherwise
white	1 if white, 0 otherwise
black	1 if black, 0 otherwise
other_race	1 if other race, 0 otherwise
northeast	1 if lives in Northeast region, 0 otherwise
midweast	1 if lives in Midweast region, 0 otherwise
west	1 if lives in West region, 0 otherwise
south	1 if lives in South region, 0 otherwise
msa	1 if lives in Metropolitan Statistical Area, 0 otherwise
income	total annual income
union	1 if union status, 0 otherwise
employed	1 if employed, 0 otherwise
education	1 if had high_school, master or PhD degree , 0 otherwise
expenditure	total annual health care expenditure
lnreimbursement	natural logarithm of reimbursement paid by insurance
share_reimbursement	natural logarithm of reimbursement paid by insurance as share of total annual health care expenditure
lnexp_paid_other_sources	natural logarithm of expenditure paid by other sources
family size	family size
high_risk	1 if current health is poor or fair, 0 otherwise
activity limitations	1 if has limited in any activities because health problems, 0 otherwise
need care	1 if needs for care during the year of interview, 0 otherwise
need specialist	1 if needs for specialist during the year of interview, 0 otherwis
insured	1 if insured, 0 otherwise
insurance_preference	1 if agree with "Health insurance is not worth the money it costs", 0 otherwise
insurance_attitude	1 if is likely to take risk, 0 otherwise
dental_bills	1 if plan covers dental bills, 0 otherwise
drug_costs	1 if plan covers drug costs, 0 otherwise
group_insurance	1 if gets insurance through their employers or organizations,
lncopayment	natural logarithm of copayment
mills1	mills ratio insurance partecipation
mills2	mills ratiohealth care expenditure partecipation

Table 2: Summary Statistics

	All	Insured	Uninsured
Age	44	44.04	43.61
Male	0.3306	0.3333	0.2973
Income	42519.25	44452.26	18539.39
Total health care expenditure	35000.09	3592.092	2357.689
Copayment		879.3203	
Group Insurance		0.9223	
Annual premium		1821.522	
Northeast	0.1532	0.1634	0.0270
South	0.3679	0.2897	0.5676
West	0.1966	0.3518	0.2162
Midwest	0.2823	0.1949	0.1892
White	0.8568	0.8758	0.6216
Black	0.0968	0.0806	0.2973
Other Race	0.0464	0.0436	0.0810
Metropolitan statistical area	0.8145	0.83	0.6216
High Risk Individuals	0.0776	0.0708	0.1622
Activity limitations	0.2520	0.2462	0.3243
Low Insurance Attitude	0.2218	0.2233	0.2027
Low Insurance Preferences	0.2429	0.2321	0.3783
Number of observations	992	918	74

Table 3: Preliminary BivariateProbit Correlation Coefficients
(p-value in parentheses)

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

Note: sample size 496.

Table 4: Cross-Sectional Bivariate Probit Estimation Coefficients
(p-value in parentheses)

	Expenditure 2003	Be Insured 2003
intercept	0.5013 (0.659)	-1.6287 (0.032)
age	0.0264 (0.075)	0.0076(0.378)
male	-1.1982(0.000)	0.0939(0.699)
black	-0.3491(0.449)	-0.9542(0.000)
other_race	-0.2243(0.754)	-0.5702(0.204)
family size	-0.1871(0.109)	0.2500(0.012)
msa	-0.0803(0.849)	0.6041(0.007)
northeast	0.0537(0.893)	0.7778(0.113)
midwest	0.5476(0.224)	0.0891(0.741)
west	1.1711(0.082)	-0.0963(0.721)
insured	1.2838(0.485)	
income	4.0600(0.453)	0.0001(0.008)
union		0.3602(0.486)
employed		0.4195(0.149)
education	0.0765(0.908)	0.7719(0.009)
need care	-0.2017(0.560)	
need specialist	0.8533(0.160)	
insurance attitude		-0.4376(0.068)

Note: sample size 496.

Table 5: Cross-Sectional Bivariate Probit Estimation Coefficients (p-value in parentheses)

	Expenditure 2004	Be Insured 2004
intercept	1.6613(0.145)	-2.823(0.008)
age	0.0112(0.441)	0.0133(0.137)
male	-1.4139(0.000)	-0.0372(0.880)
black	-0.3407(0.472)	-0.9401(0.001)
other_race	0.5758(0.448)	-0.6887(0.129)
family size	-0.2696(0.012)	0.2954(0.002)
msa	-0.0089(0.981)	0.6012(0.014)
northeast	-0.4157(0.406)	0.9329(0.061)
midwest	-0.3945(0.367)	0.1165(0.664)
west	-0.5889(0.177)	-0.0947(0.733)
insured	1.0708(0.256)	0.0002(0.000)
income	4.9400(0.306)	
union		0.3671(0.449)
employed		0.3262(0.270)
education	0.1199(0.827)	0.6830(0.030)
need care	0.8899(0.010)	
need specialist	-1.1089(0.061)	
insurance attitude		-0.2287(0.410)

Note: sample size 496

Table 6: Pooled OLS Regression Results.
Risk Variable: Self-Assessed Health.

Preidictor Variables	*Coefficients*	*p-values*
intercept	0.5309	0.000
age	0.0007	0.167
male	-0.0029	0.830
msa	-0.0245	0.094
northeast	0.0044	0.781
midwest	0.0264	0.042
west	-0.0102	0.488
black	-0.0016	0.944
other race	-0.0285	0.248
education	-0.0265	0.274
income	-4.64e-07	0.008
group_insurance	0.07812	0.000
lnpremium	-8.42e-07	0.689
lncopayment	-0.0384	0.000
lnexp_paid_other_sources	-0.0160	0.003
dental_bills	0.0439	0.000
drug_ costs	0.0917	0.000
high_risk	0.0776	0.000
activity limitations	0.0406	0.001
insurance preferences	-0 .0462	0.000
mills1	-0.1566	0.034
mills2	-0.0899	0.079

Note: sample size 895; R^2 = 0.2505; Adjusted R^2 =0.2325

Recent working papers

The complete list of working papers is can be found at http://polis.unipmn.it/pubbl

***Economics Series**	****Political Theory Series**	**ε Al.Ex Series**
T Territories Series	**t Transitions Series**	

2009 n.156* Cinzia Di Novi: *Sample selection correction in panel data models when selectivity is due to two sources*

2009 n.155* Michela Martinoia: *European integration, labour market dynamics and migration flows*

2009 n.154* Massimo Pasquariello and GianMarco Chiesi: *Valore aggiunto e tipologia di spesa in Piemonte. Un confronto tra Alessandria e gli altri capoluoghi di provincia*

2009 n.153* Massimo Pasquariello: *Produttività, sistemi locali del lavoro, specializzazione produttiva e scenari futuri. Studio preliminare in preparazione del Piano strategico per il comune di Alessandria*

2009 n.152* Massimo Pasquariello and GianMarco Chiesi: *L'analisi dei determinanti di un'area territoriale. Studio preliminare in preparazione del Piano strategico per il comune di Alessandria*

2009 n.151* Cristina Elisa Orso: *Formal and informal sectors: Interactions between moneylenders and traditional banks in the rural Indian credit market*

2009 n.150* Michele Giuranno: *The logic of party coalitions with political activism and public financing*

2009 n.149* Matteo Migheli: *Sharing the pie: the Lutheran is neither opportunistic nor generous*

2009 n.148* Amedeo Fossati and Marcello Montefiori: *Migrants and mafia as global public goods*

2009 n.147* Alberto Cassone: *L'impatto economico dell'Universita' del Piemonte Orientale "Amedeo Avogadro": un aggiornamento al 2008*

2009 n.146* Daron Acemoglu, Davide Ticchi and Andrea Vindigni: *Persistence of civil wars*

2009 n.145* Daniele Bondonio: *Impact identification strategies for evaluating business incentive programs*

2009 n.144* Barry Eichengreen, Rachita Gullapalli and Ugo Panizza: *Capital account liberalization, financial development and industry growth: a synthetic view*

2009 n.143* Emma Galli and Roberto Ricciuti: _Sulla_ political economy _del deficit pubblico nell'Italia liberale_

2009 n.142* Matteo Migheli: _Religiosity and happiness: an ever-winning couple? An answer from India_

2009 n.141** Stefano Parodi: _I media dell'Alessandrino e l'Unione Europea_

2009 n.140* Matteo Migheli: _The two sides of a ghost: Twenty years without the wall_

2009 n.139[ε] Matteo Migheli and Francesco Scacciati: _How does labor supply react to different tax rates? A field enquiry_

2009 n.138[ε] Matteo Migheli and Guido Ortona: _Majority, proportionality, governability and factions_

2009 n.137** Noemi Podestà: _Strumenti di mediazione per la risoluzione di conflitti. L'esperienza dell'Osservatorio per il collegamento ferroviario Torino-Lione_

2009 n.136** Noemi Podestà and Alberto Chiari: _Esperimenti di democrazia deliberativa. Informazioni, preferenze e stili di conduzione in tre giurie di cittadini._

2009 n.135** Andrea Lanza: _1848 comme reconfiguration des discours politiques._

2009 n.134* Rongili Biswas, Nicolas Gravel and Rémy Oddou: _The segregative properties of endogenous jurisdictions formation with a welfarist central government_

2009 n.133[ε] Matteo Migheli: _Assessing trust through social capital? A possible experimental answer_

2009 n.132* Piero Cavaleri, Michael Keren, Giovanni B. Ramello and Vittorio Valli: _Publishing an E-journal on a shoe string: is it a sustainable project?_

2009 n.131* Alberto Cassone: _L'impatto economico e sociale dell'Universita' del Piemonte Orientale Amedeo Avogadro_

2009 n.130* Alberto Cassone and Pietro Zaccarella: _Il bilancio sociale delle universita'. Inventario dei problemi e analisi comparata delle esperienze italiane_

2009 n.129[ε] Matteo Migheli, Guido Ortona and Ferruccio Ponzano: _A preliminary simulative assessment of disproportionality indices_

2008 n.128* Fabio Privileggi: _On the transition dynamics in endogenous recombinant growth models_

2008 n.127* Roberto Zanola: _Who likes circus animals?_

2008 n.126* Michele Giuranno: _Regional income disparity and the size of the Public Sector_