

A Multilevel Sample Selection Probit Model with an Application to Contraceptive Use

*Un Modello Probit Multilivello con Selezione Campionaria
con un'Applicazione all'Uso di Metodi Contraccettivi*

Riccardo Borgoni, Francesco C. Billari

Max Planck Institute for Demographic Research

Doberaner Str. 114

D-18057 Rostock, Germany

E-mail: borgoni@demogr.mpg.de, billari@demogr.mpg.de

Riassunto: Nel presente lavoro discutiamo l'estensione multilivello di un modello di selezione campionaria di tipo probit. Stimiamo il modello secondo il metodo della massima verosimiglianza e presentiamo un'applicazione all'uso di metodi contraccettivi durante il primo rapporto sessuale.

Keywords: Multilevel statistical models, probit selection model, contraceptive use.

1. The multilevel sample selection probit model

In some situations, a binary outcome is observed only for a specific part of a sample. The idea that factors affecting selection into the sample may simultaneously affect the binary outcome of interest has been the motivation for the introduction of the probit sample selection model (van De Ven and van Praag, 1981). This is a specification for a binary outcome of the well-known Heckman sample selection model (1979). The probit sample selection model can be expressed as follows. Let Y_1 and Y_2 be two binary variables such that Y_1 is observable only if $Y_2 = 1$. Binary outcomes can be seen in terms of propensity or latent utility. If we assume Y_2^* to be an unobservable outcome, say the propensity or the expected utility that individuals attach to a binary choice, we hypothesize that $Y_2 = 1$ (the choice is observed) only if $Y_2^* \geq 0$, with $Y_2 = 0$ if $Y_2^* < 0$. If $Y_2 = 1$, individuals are faced with a second binary choice, Y_1 . Also to the second binary choice (usually the outcome of main interest), a latent propensity random variable Y_1^* is attached, so that $Y_1 = 1$ if $Y_1^* \geq 0$, with $Y_1 = 0$ if $Y_1^* < 0$. Introducing two sets of predictors, X_1 and X_2 , to explain latent propensities, we define a two-equation system. The first equation describes the probability of experiencing the selecting event: $\text{Probit}(Y_2 = 1 | X_2) = X_2 h$. The second equation is defined only if $Y_2 = 1$, and it describes the outcome of interest: $\text{Probit}(Y_1 = 1 | X_1) = X_1 b$. In the same way, the system can be stated linearly in terms of the unobservable propensities. The first equation describes the propensity to be selected: $Y_2^* = X_2 h + e_2$. The second equation, defined only if $Y_2^* \geq 0$, describes the propensity toward the outcome of interest: $Y_1^* = X_1 b + e_1$. b and h are suitable vectors of unknown regression parameters. (e_1, e_2) is a zero-mean unit-variance bivariate normal random variable with $\text{corr}(e_1, e_2) = r$. Estimating the first equation, without taking into account the selection equation, induces biases in the inference results on the parameters (van De Ven and van Praag, 1981; Vella, 1998).

It this paper, we aim at extending the probit sample selection model to a multilevel setting. Specifically, we assume that the data are embedded in a hierarchical structure so that the behavior of individuals sharing some common characteristics (in the case study presented in this paper the municipality of residence during youth) is correlated. A similar problem arises for the estimation of sample selection models based on panel data with non-random drop-out (Vella, 1998; Hausman and Wise, 1979). The multilevel approach (Goldstein, 1995) provides a convenient framework for studying such forms of hierarchically structured data. In this approach, one or more random effects are introduced to control for the correlation in the data, decomposing the total variability according to the different levels. A two-level (individuals and areas) model for individual i in area g is specified as:

$$\text{Probit}(Y_{1gi}=1 | X_{1g}, U_{1g}, V_{1gi}) = X_{1gi} \mathbf{b} + U_{1g} + V_{1gi} \quad Y_{1gi} \text{ observed only if } Y_{2gi}=1 \quad (1)$$

$$\text{Probit}(Y_{2gi}=1 | X_{2g}, U_{2g}, V_{2gi}) = X_{2gi} \mathbf{h} + U_{2g} + V_{2gi} \quad (2)$$

We hypothesize that in (1) and (2), $U_g=(U_{1g}, U_{2g})$ is a 0-mean bivariate normal random variable with variances (t_1^2, t_2^2) and a correlation coefficient $\mathbf{q}$, $g=1, \dots, G$, where G is the number of areas. $\mathbf{q}$ is thus the area-level correlation. Furthermore, $V_{gi}=(V_{1gi}, V_{2gi})$ is also a 0-mean bivariate normal random variable with variances (s_1^2, s_2^2) and a correlation coefficient $\mathbf{r}$. In this model, $\mathbf{r}$ is the individual level correlation. In the case of one observation per individual, the variances (s_1^2, s_2^2) cannot be identified, and as usual with probit functions, they can be fixed at 1. Formula (1), and (2) although they could be potentially simplified in the special case of the probit link function, allow to have different functions that link predictors to outcome variables (i.e. logit, or complementary log-log). We estimate the model using full-information maximum likelihood. Residuals at both the individual and area level are integrated out numerically (details are given in Lillard and Panis, 2000).

2. An application to contraceptive use in Italy

In reproductive health literature, authors often place a emphasis on the behavior of adolescents, as they are more exposed to risk of contracting sexually transmitted diseases, including HIV/AIDS. Adolescents are also at risk of having unplanned pregnancies. Particular attention has been placed on contraceptive use during the first sexual intercourse (Hogan et al., 2000). We apply the model sketched in Section 1 to contraceptive use at first sexual intercourse. The analyzed data comes from the Fertility and Family Survey (FFS) (De Sandre et al., 1997). We selected a subset of 1,011 women aged between 20 and 25 years at the time of the interview, that is 1,011 . Municipality of main residence during the first 15 years is used as a territorial unit, as we hypothesize that context matters in contraceptive choice (see also Borgoni and Billari, 2001).

Contraceptive use during the first sexual intercourse can be observed only for those who have already experienced intercourse. Therefore, the model described in the previous section fits this application naturally. For the individual i in the municipality g , having had the first sexual intercourse is the response variable for the selection equation ($Y_{2gi}=1$) and having used some contraceptive methods at first intercourse ($Y_{1gi}=1$) is the

response variable for the behavioralequation. In the sample we used, 62.8% had experienced the selecting event and, among them, 73.5% had used some contraceptive methods. The predictors covered by the selection equation are: educational level (binary, low level, i.e. age at leaving school is less than 16-years-old, as baseline), the dimension of the municipality of residence at age 15 (binary with larger municipalities, i.e. more then 50,000 inhabitants, as the baseline), geographical location (binary, with southern Italy as baseline), cohort (binary with older cohort, i.e. aged more than 23-years-old, as baseline) and the presence of siblings (no siblings as baseline). In the behavioral equation we considered geographical location, the age of the person at their first sexual intercourse (4 classes with age 17-18 as baseline) and the presence of the person's until he or she was 15-years-old. (both parents as baseline).

Table 1: *Estimated parameters of multilevel sample selection probit model.*

	Model 1	Model 2	Model 3	Model 4
<i>First Sexual Intercourse</i>				
<i>Intercept</i>	-0.391 *** (0.089)	-0.415 *** (0.097)	-0.540 *** (0.125)	-0.542 *** (0.131)
<i>Education: left school after age 16</i>	0.176 * (0.087)	0.177 * (0.088)	0.229 (0.122)	0.178 (0.120)
<i>Municipality size less than 50000 inhabitants</i>	0.297 *** (0.085)	0.245 * (0.108)	0.392 ** (0.122)	0.251 (0.144)
<i>Northern/Central Italy</i>	0.638 *** (0.090)	0.704 *** (0.105)	0.905 *** (0.127)	0.987 *** (0.144)
<i>Aged more than 23 at the time of the interview</i>	0.531 *** (0.092)	0.544 *** (0.097)	0.763 *** (0.130)	0.789 *** (0.137)
<i>Presence of siblings</i>	-0.026 (0.161)	-0.053 (0.170)	-0.070 (0.230)	-0.169 (0.237)
<i>Contraceptive Use</i>				
<i>Intercept</i>	0.525 *** (0.101)	0.472 *** (0.140)	0.561 (0.373)	-0.017 (0.407)
<i>Northern/Central Italy</i>	0.442 *** (0.107)	0.570 *** (0.148)	0.700 *** (0.205)	1.095 *** (0.233)
<i>Age at intercourse <17</i>	-0.283 (0.148)	-0.342 * (0.171)	-0.396 (0.208)	-0.439 (0.229)
<i>Age at interc. 19-20</i>	-0.117 (0.154)	-0.145 (0.167)	-0.157 (0.218)	-0.151 (0.223)
<i>Age at interc. >=20</i>	-0.108 (0.155)	-0.146 (0.171)	-0.134 (0.221)	-0.103 (0.237)
<i>Co-residence with less than 2 parents</i>	-0.565 * (0.259)	-0.579 * (0.279)	-0.784 * (0.362)	-0.797 * (0.379)
<i>Individual-level Residuals</i>				
<i>s₁</i>		1	1	1
<i>s₂</i>		1	1	1
<i>r</i>		0	0.321 (0.620)	0.927 (0.575)
<i>Municipality-level residuals</i>				
<i>t₁</i>		0.256 * (0.111)		0.332 * (0.160)
<i>t₂</i>		0.490 ** (0.180)		0.714 ** (0.232)
<i>q</i>		0		0.875 (0.517)
<i>ln-L</i>	-966.35	-959.64	-966.31	-958.05
<i>LRT vs. Model 1 (p)</i>		0.00	0.78	0.00

Asymptotic standard errors in parentheses

P-value: *=5%, **=1%; ***=0.1%.

The results of the analysis are shown in table 1. Model 1 is without correlation at any level and without consideration of the multilevel structure. Model 2 introduces the variance components at the territorial level. Model 3 is similar to a standard probit sample selection model, without accounting for the hierarchical structure in the data. Model 4 is the complete multilevel sample selection probit model. According to the likelihood ratio test, model 2 and model 4 are significantly better than model 1. The same test applied to model 4 versus model 2 gives a p-value of .20, which indicates that

accounting selection on unobservables, given the high variability of the correlation coefficients, does not significantly improve the fit of the model. Variance components at the area level also tell us that unobserved area-level variables significantly affect both behaviors. The positive correlation coefficients may be interpreted from a behavioral point of view: an unobserved individual characteristic that favors having first sexual intercourse early (in a relative sense) also favors the use of contraception. Nevertheless, the high value of correlation coefficients is not statistically significant when using asymptotic tests. If we look at the results on the use of contraception at first intercourse, we see that geographical differences are even more emphasized when controlling for sample selection. Age does not have a significant effect. The effect of co-residing with both parents on the propensity to use contraception is also significant.

References

- Borgoni R., Billari F.C. (2001) Bayesian spatial analysis of demographic survey data with an application to contraceptive use at first sexual intercourse, *mimeo*, Max Planck Institute for Demographic Research, Rostock.
- De Sandre, P., Ongaro, F., Rettaroli, R., Salvini, S. (1997) *Matrimonio e figli: tra rinvio e rinuncia*, il Mulino, Bologna.
- Goldstein, H. (1995) *Multilevel Statistical Models*. 2nd edition, Edward Arnold, London.
- Hausman J.A., Wise D.A. (1979) Attrition bias in experimental and panel data: The Gary income maintenance experiment, *Econometrica*, 47, 455-473.
- Heckman, J.J. (1979) Sample selection bias as a specification error, *Econometrica*, 47, 153-161.
- Hogan, D.P., Sun, R., Cornwell, G.T. (2000) Sexual and Fertility Behaviors of American Females Aged 15-19 Years: 1985, 1990 and 1995, *American Journal of Public Health*, 90, 1421-1425.
- Lillard, L., Panis, C.W.A. (2000) *aML Multilevel Multiprocess Statistical Software*. EconWare, Los Angeles, California.
- van De Ven, W.P.M.M., van Praag, B.M.S. (1981) The demand of deductibles in private health insurance: A probit model with sample selection, *Journal of Econometrics*, 17, 229-252.
- Vella F. (1998). Estimating Models with Sample Selection Bias: A Survey, *Journal of Human Resources*, 33 (1), 127-169.