19th International Workshop on Statistical Modelling
Florence, Italy, 4 - 8 July 2004

A Polytomous Response Multilevel Model with a Non Ignorable Selection Mechanism

Leonardo Grilli – Carla Rampichini
Dipartimento di Statistica "G. Parenti", Firenze
grilli@ds.unifi.it, carla@ds.unifi.it

Main goals

- Investigation of the factors which influence the way of acquisition of the professional skills;
- Comparative evaluation of the degree programmes on the basis of the adequacy of the skills they give to their graduates.

Methodological issues

- The data have a hierarchical structure, so that the responses are correlated;
- The question of interest has a polytomous response;
- The response might be affected by selection bias due to the design of the questionnaire.

Outline of the presentation

- Data description
- The model
- Main results
- Work in progress
- Conclusions

The data

- Univ. of Florence, graduates of year 2000
- 3148 CATI interviews about two years after the degree
- For **employed** graduates: questions on the skills required by the present job

Do you have a job?	Frequency	Percent
1 Yes, now	**2543**	**80.86**
2 Yes, but not now	219	6.96
3 No	383	12.18

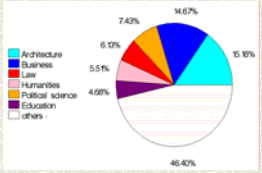
Frequency Missing = 3

Graduates by job condition at the interview

Hierarchical structure of the data

Employed graduates: distribution among the 56 degree programmes (quite unbalanced)

Min	Max	Med
4	386	21



Hierarchical nature of the data (graduates within course programmes)

MULTILEVEL ANALYSIS

Outcome of interest

Where have you acquired the professional and technical skills?

(if used in the current job)

% Used	Way of acquisition %		
	UNIV	WORK	OTHER
91.0	43.4	36.2	11.4

The selection issue

due to the design of the questionnaire:

the outcome is missing if the skills are not used

- potential bias if the selection mechanism depends on unobserved variables correlated with the principal model's error terms
- To avoid selection bias: insert an equation that explicitly models the selection mechanism (Heckman, 1979) .
- Applications rare in the multilevel framework (Borgoni e Billari, 2002).

Model with selection

(1) Selection equation Y_{ij}^{S}:

do you USE the skills?

(2) Acquisition equations $Y_{ij}^{P(m)}$:

WHERE have you acquired the skills?

$Y_{ij}^{P(m)}$ not observed if $Y_{ij}^{S}=0$ m=1,…,M alternatives

Not ignorable selection mechanism **IF** the two set of equations have correlated residuals

Selection equation: *do you USE the skills?*

Dichotomous multilevel logistic model

$$p(Y_{ij}^{S}=1\,|\,\mathbf{x}_{ij}^{S},\xi_{j}^{S},\delta_{ij}^{S})=\frac{\exp(\alpha^{S}+\boldsymbol{\beta}^{S\prime}\mathbf{x}_{ij}^{S}+\xi_{j}^{S}+\delta_{ij}^{S})}{1+\exp(\alpha^{S}+\boldsymbol{\beta}^{S\prime}\mathbf{x}_{ij}^{S}+\xi_{j}^{S}+\delta_{ij}^{S})}$$

i =1,…,n_j graduates; j =1,…,56 degree programmes

subjects clusters

x graduate or degree programme characteristics
ξ random errors at degree programme level
δ random errors at graduate level

Acquisition equations: *WHERE did you acquired the skills?*

Multinomial multilevel logistic model

observable only if Y^{S}=1

$$p(Y_{ij}^{P}=m\,|\,\mathbf{x}_{ij}^{P},\boldsymbol{\xi}_{j}^{P},\boldsymbol{\delta}_{ij}^{P})=\frac{\exp(\eta_{ij}^{(m)})}{1+\sum_{l=2}^{M}\exp(\eta_{ij}^{(l)})}$$

$$\eta_{ij}^{(m)}=\alpha^{P(m)}+\boldsymbol{\beta}^{P(m)}\mathbf{x}_{ij}^{P}+\xi_{j}^{P(m)}+\delta_{ij}^{P(m)}$$

i =1,…,n_j graduates; j =1,…,56 degree programmes
m = 1,2,3 alternatives (university, at work, other)

Nb: all parameters equal to 0 for m=1 (university)

x graduate or degree programme characteristics
ξ random errors at degree programme level
δ random errors at graduate level

Assumptions on the error terms

- Errors at different levels are independent
- At each level:
- $\left[\xi_{j}^{S},\xi_{j}^{P(2)},\ldots,\xi_{j}^{P(M)}\right]' \sim \mathrm{MN}(0,\boldsymbol{\Sigma}_{\xi})$ degree programme
- $\left[\delta_{ij}^{S},\delta_{ij}^{P(2)},\ldots,\delta_{ij}^{P(M)}\right]' \sim \mathrm{MN}(0,\boldsymbol{\Sigma}_{\delta})$ graduate

If at least one of the correlations among the couples

$(\xi_{j}^{S},\xi_{j}^{P(m)})$ $(\delta_{ij}^{S},\delta_{ij}^{P(m)})$

is not null ➔ not ignorable selection mechanism

IIA property

The odds for two alternatives *m* and *l* for subject *i* of cluster *j* are:

$$\frac{P(Y_{ij}=m \mid \mathbf{x}_{ij},\boldsymbol{\xi}_j^P,\boldsymbol{\delta}_{ij}^P)}{P(Y_{ij}=l \mid \mathbf{x}_{ij},\boldsymbol{\xi}_j^P,\boldsymbol{\delta}_{ij}^P)}=\frac{\exp[\eta_{ij}^{(m)}]}{\exp[\eta_{ij}^{(l)}]}$$

The odds depends only on the linear predictors η of the two involved alternatives and does not depend on the other alternatives.

The IIA property holds conditionally on all the covariates and **error terms** at all levels.

Interpretation of the acquisition equations

- An alternative specification of the multinomial logit model is based on the random utility model (McFadden, 1973):

$$U_{ij}^{(m)}=\eta_{ij}^{(m)}+\varepsilon_{ij}^{(m)}$$

$\varepsilon^{(m)}_{ij}$ are iid errors following the Gumbel distribution.

- This leads to a definition of the Intraclass Correlation Coefficient of the *m*-th equation analogous to the usual definition for dichotomous logit models:

$$\mathrm{ICC}^{(m)}=\frac{\mathrm{Var}(\xi_j^{(m)})}{\mathrm{Var}(\xi_j^{(m)})+\mathrm{Var}(\delta_{ij}^{(m)})+\pi^2/3}$$

Model identification

- The parameters of the cluster level covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}$ are all identified;
- For the subject level covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\delta}}$:
  - the variance of δ_{ij}^{S} in the selection equation is obviously not identifiable
  - the other parameters are in principle identified, but prone to empirical underidentification, unless some alternative specific covariate is included in the model (Skrondal and Rabe-Hesketh, 2003).

Model likelihood

$$L(\boldsymbol{\theta} \mid \mathbf{Y}_{ij})=\prod_{j=1}^{J}\int\prod_{i=1}^{n_j}\left\{\int P(\mathbf{Y}_{ij} \mid \mathbf{x}_{ij},\boldsymbol{\xi}_j,\boldsymbol{\delta}_{ij})f(\boldsymbol{\delta}_{ij})d\boldsymbol{\delta}_{ij}f(\boldsymbol{\xi}_j)d\boldsymbol{\xi}_j\right\}$$

$$\boldsymbol{\theta}'=\left(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\Sigma}_{\boldsymbol{\xi}},\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\right)$$

An approximated method is needed to solve the integrals

Maximization achieved by means of the **GLLAMM** command of STATA (adaptive numerical quadrature)

The subject level covariance parameters are empirically not identified (high condition number and convergence difficulties)
➔ the random errors δ_{ij} are omitted from the models.

Covariates selection strategy

The criterion for choosing the relevant covariates is the likelihood ratio test, with a *p*-value threshold of 5%.

1. Selection of the covariates separately for the Selection model and for the Acquisition model;
2. Refinement using the joint model, trying to reinsert in the Acquisition equations the variables which were previously discarded from the Acquisition model, but retained in the Selection model.

Models comparison

	Model 1 *joint* S&P	*Model* 2 *unrelated* S. P
log*L*	-2840.28	-2843.04
n. of parameters	26	24
Random parameters		
$\mathrm{Var}(\xi^S)$	0.1510	0.1448
$\mathrm{Var}(\xi^{P(2)})$	0.1637	0.1531
$\mathrm{Var}(\xi^{P(3)})$	0.4274	0.4129
$\mathrm{Corr}(\xi_j^S,\xi_j^{P(2)})$	-0.2401	.
$\mathrm{Corr}(\xi_j^S,\xi_j^{P(3)})$	-0.6772	.
$\mathrm{Corr}(\xi_j^{P(2)},\xi_j^{P(3)})$	0.8768	0.8479

LR Test
χ^2=5.52, df=2
p-value=0.0633

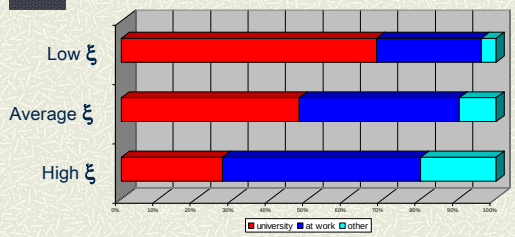
⬇

without selection model

Estimation with GLLAMM

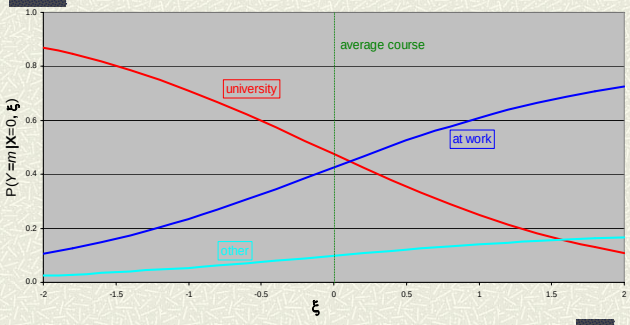
Fixed parameters quite similar in Models 1 and 2!

Probabilities of acquisition by degree programme (baseline graduate)

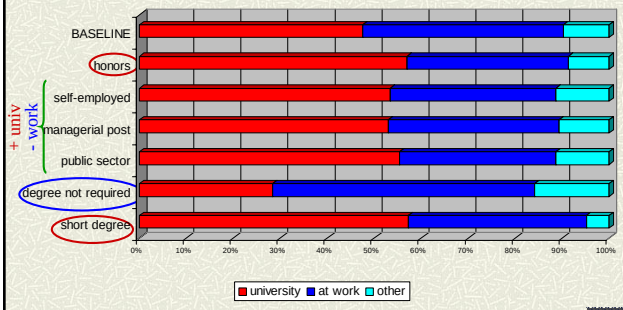


Given graduate and degree programme observed characteristics, the probability of each alternative vary a lot among degree programmes!

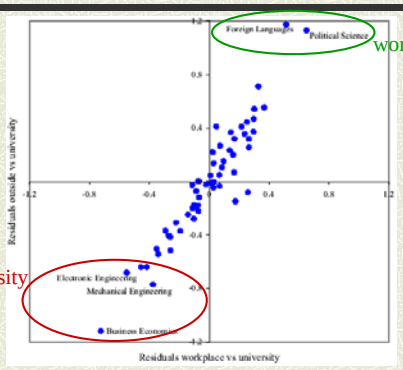
Acquisition probabilities and ξ values ($\xi^{P(2)}=\xi^{P(3)}$, baseline graduate)



Acquisition probabilities by graduate characteristics (average degree prog.)



Degree programme level EB residuals from the acquisition equations



workplace and other

university

Where did the graduates acquired the skills?

Work in progress…

- Further work is needed to fully understand the implications of selection mechanisms that act in a hierarchical framework;
- Some sensitivity analysis is required to assess how the results are affected by the distributional assumptions.

Work in progress (cont'd)

Simulation study:

- How do the selection mechanism interact with the hierarchical structure of the data?
- What are the consequences of ignoring the hierarchy?
- How change the power of the LR Test for presence of selection?
- What is the influence of the distributional assumptions on error terms?

Preliminary results …

We start considering a two-level bivariate probit model with selection mechanism and covariates at both levels (some are common).

Given a certain total correlation between the two latent variables, things are better if the correlation is mainly due to the clusters.

Estimation

- The ML estimation algorithm based on adaptive numerical quadrature, used in the application, is accurate and flexible, but it requires long computational times, which rapidly increase with the model complexity.
- Many alternative estimation methods are possible, e.g. Bayesian MCMC and Maximum Simulated Likelihood (Train, 2003).

Concluding remarks

- We have shown how to build a complex polytomous response model for the analysis of graduates' skills, taking into account:
  - the hierarchical structure of the phenomenon
  - the adjustment for a possible selection bias.

  In the application the hierarchical structure has a crucial role, while selection bias results negligible.
- The model allows to:
  - characterize ways of acquisition of the skills
  - find out extreme degree programmes to be further investigated.

Thanks for your attention…

References

Heckman J.J. (1979) Sample selection bias as a specification error, *Econometrica*, 47, 153–161.

McFadden D. (1973) Conditional logit analysis of qualitative choice behaviour, in: *Frontiers in Econometrics*, Academic Press, New York.

Rabe-Hesketh S., Pickles A. and Skrondal A. (2001) GLLAMM manual, Technical Report 2001/01, Department of Biostatistics and Computing, Institute of Psychiatry, King's College, London.

Skrondal A. and Rabe-Hesketh S. (2003) Multilevel logistic regression for polytomous data and rankings, *Psychometrika*, 68, 267–287.

Train K. (2003) *Discrete Choice Methods with Simulation*, Cambridge University Press, New York.