Università degli Studi di Napoli Federico II
Scuola delle Scienze Umane e Sociali
Quaderni
11

ASMOD 2018
Proceedings of the International Conference on Advances in Statistical Modelling of Ordinal Data

Naples, 24-26 October 2018

Editors
Stefania Capecchi, Francesca Di Iorio, Rosaria Simone

Federico II University Press
fedOA

Università degli Studi di Napoli Federico II
Scuola delle Scienze Umane e Sociali
Quaderni
11

ASMOD 2018

Proceedings of the International Conference on Advances in Statistical Modelling of Ordinal Data

Naples, 24-26 October 2018

Editors

Stefania Capecchi, Francesca Di Iorio, Rosaria Simone

Federico II University Press

fedOA Press

ASMOD 2018 : Proceedings of the Advanced Statistical Modelling for Ordinal Data Conference : Naples, 24-26 October 2018 / editors Stefania Capecchi, Francesca Di Iorio, Rosaria Simone. – Napoli : FedOAPress, 2018. – (Scuola di Scienze Umane e Sociali. Quaderni ; 11).

Accesso alla versione elettronica:
http://www.fedoabooks.unina.it

ISBN: 978-88-6887-042-3
DOI: 10.6093/978-88-6887-042-3
ISSN Collana: 2499-4774

Università degli Studi di Napoli Federico II
Centro di Ateneo per le Biblioteche “Roberto Pettorino”
Piazza Bellini 59-60
80138 Napoli, Italy
http://www.fedoapress.unina.it/

Published in Italy

A latent variable model for a derived ordinal response accounting for sampling weights, missing values and covariates

Fulvia Pennoni*, Miki Nakai**

Abstract: We consider a latent class model especially tailored for an ordinal response derived by comparing two continuous variables. We propose a general method to estimate the model parameters with survey data when there are missing responses and survey weights. First, we estimate the model with the missing responses without covariates with a weighted likelihood function maximised through the Expectation-Maximization algorithm. In order to determine the suitable number of latent classes we rely on the Akaike Information Criterion. Second, by fixing the parameters of the measurement model we estimate the remaining parameters by adding the full set of covariates. We make predictions on the basis of the maximum a posteriori probability. In the application, we consider Japanese survey data collected at four waves covering 40 years with the aim to study changes on couples' breadwinning patterns.

Keywords: Akaike information criterion, Expectation-Maximization algorithm, Gender inequality, Household income composition.

1. Introduction

The latent class model (Lazarsfeld and Henry, 1968) has been considered for the analysis of data arising in different contexts by many authors since it is a flexible model to account for the heterogeneity among responses provided by different individuals which cannot be explained by means of the observable covariates. This model is especially tailored for an ordinal response variable when it has been derived for example by comparing values of two or more continuous variables. It is a model-based approach that properly accounts for the underlying latent continuous responses and allow us to investigate the associations with the covariates as well as to dispose of data driven typologies of individuals (see, among others, Pennoni, 2014). Another advantage is that

*Department of Statistics and Quantitative Methods, University of Milano-Bicocca, fulvia.pennoni@unimib.it

**College of Social Sciences, Ritsumeikan University, mnakai@ss.ritsumei.ac.jp

it is possible to elaborate the model in many ways and to assess the tenability of the underlying hypothesis.

Maximum likelihood estimation of the model parameters is well established and it is carried out through the Expectation-Maximization algorithm (see, among others Bartolucci *et al.*, 2013). However, the use of weighted methods for the estimation of the parameters with missing responses and covariates still deserves research. In the current proposal, instead of performing listwise deletion we rely on the missing at random assumption and we retain the missing responses for the outcome, while the values of the missing covariates are imputed through multivariate imputation by chained equations. In this way, we allow the allocations on the latent variables at individual level also for individuals not providing a response.

In Section 2 we introduce the model and the steps of the maximum likelihood estimation. In Section 3 we describe the data collected within the Japanese Stratification and Social Psychology Survey and in Section 4 we show the main results.

2. *The proposed model*

In the following, we deal with a derived response variable and we introduce the latent class model to account for the missing responses assuming that they are conditionally independent given the latent variable and the observed covariates as well as for survey weights for the representativeness of each unit in the population.

With reference to a random unit drawn from the population of interest let Y_{ij} be the observed derived variable with $j, j = 1, \ldots, r$ ordered categories for individual $i, i = 1, \ldots, n$. This response is obtained by comparing two or more continuous variables. We assume that the observed response depends on the underlying unobserved latent variable denoted as U_i which has a distribution with k support points assuming finite discrete values. The observed responses are independent one another conditionally to this latent variable.

The first set of parameters in the model is related to the probability to belong to each latent class. These probabilities may be influenced by time-specific individual covariates arranged in the vector $\boldsymbol{X}$ where $\boldsymbol{x}$ is a corre-

sponding realization. We use a baseline category logit model for the following parameters

$$\log \frac{p(U=u|\boldsymbol{X}=\boldsymbol{x})}{p(U=1|\boldsymbol{X}=\boldsymbol{x})} \quad = \quad \log \frac{\pi_{u|\boldsymbol{x}}}{\pi_{1|\boldsymbol{x}}} = \beta_{0u} + \mathbf{x}'\beta_{\mathbf{1u}}, \quad u = 2, \ldots, k, \ (1)$$

where β_0 is an intercept specific of each latent class and $\boldsymbol{\beta}_{1u}$ is the vector of parameters that define the influence of the covariates on the distribution of the latent variable.

Another set of parameters is referred to the manifest part of the model and is given by the conditional probability of each response category given the latent variable denoted as

$$\phi_{j|u} = p(Y_j = y|U = u), \quad u = 1, \ldots, k, \ j = 1, \ldots, r.$$

To account for individual sampling weights denoted as w_i, $i = 1, \ldots, n$ such as that provided with survey data we propose to estimate the model through a weighted log-likelihood. The latter is determined given a sample of n independent individuals for which we observe the responses $y_1, \ldots, y_n$ as

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} w_i \ell_i(\boldsymbol{\theta}), \quad \ell_i(\boldsymbol{\theta}) = \log p(y_i, \ldots, y_n),$$

where $\boldsymbol{\theta}$ denotes the overall vector of free parameters arranged in a suitable way. The above quantity is maximized through the Expectation-Maximization (EM) algorithm (Dempster *et al.*, 1977). It is based on the *complete data log-likelihood* and it represents the main tool to estimate the models with latent variables. For more details see Bartolucci, Farcomeni and Pennoni (2013).

To avoid that parameters referred to the covariates are biased we perform a two step estimation procedure. First, we estimate the model with the missing responses and sampling weights excluding the covariates and we execute a model selection strategy to choose the proper number of latent classes. We perform the model estimation several times to check for local maxima and we rely on the AIC criterion (Akaike, 1973). The latter is a measure of the relative goodness of it of a model, accounting simultaneously for the accuracy

and complexity of the model since it is defined on the basis of the following index

$$\mathrm{AIC} = -2\,\hat{\ell}(\boldsymbol{\theta}) + 2\#\mathrm{par},$$

where $\hat{\ell}$ denotes the maximum of the log-likelihood and $\#\mathrm{par}$ denotes the number of free parameters of the model. Then, by fixing the parameters of the measurement model we estimate the remaining parameters by considering the full set of covariates. Standard errors for the parameters estimates are obtained according to the observed information matrix computed through numerical methods.

Once all the parameters have been estimated, the estimated *a-posteriori* probability to be assigned to a latent class is determined as

$$\hat{q}_u = \frac{\prod_{j=1}^{r} \hat{\phi}_{j|u} \hat{\pi}_{u|\boldsymbol{x}}}{\sum_{u=1}^{k} \prod_{j=1}^{r} \hat{\phi}_{j|u} \hat{\pi}_{u|\boldsymbol{x}}}, \quad u = 1, \ldots, k, \quad j = 1, \ldots, r. \tag{2}$$

In this way, we dispose of a suitable allocation rule for each individual to be assigned to the latent class having the maximum *a-posteriori* probability.

3. Data

The proposed model is applied to explore the coherent breadwinning arrangement classes and to estimate the effects of the covariates on the underlying latent variable. The data are related to spouses within the households and were obtained from the past three decades (1985, 1995, 2005) of Japanese cross-sectional data of the Social Stratification and Social Mobility (SSM) surveys, and the last decade (2015) of the Japanese Stratification and Social Psychology (SSP) survey.

The respondents are interviewed and asked a wide range of questions such as respondents' socioeconomic background. The derived response variable of interest is couple's income provision-role type consisting of five ordinal categories obtained by comparing the declared incomes (earned and investment incomes) and it has been constructed on whether a dominant provider exists and who s/he may be. This response is of primary importance since marriage between man and women in Japan has been considered the only way to form a family until recently, and a necessary way for women's financial sur-

Table 1. Observed and missing frequencies in 2015 for the response variable weighted with the survey weights: (1) "husband sole provider", (2) "husband provides majority", (3) "equal providers", (4) "wife provides majority", (5) "wife sole provider".

Response categories (%)	1	2	3	4	5	Missing
Income provision-role type	22.9	42.2	11.8	5.3	0.6	17.2

vival, social interaction and personal well-being. Moreover, the trend towards dual-earner families can be detected in recent years but gender division of labor has been accepted as "normal" and still strong in Japan. Many studies, see Sorensen and McLanahan (1987), argued that women's economic dependency on men is an important attribute of stratification systems and essential force in the maintenance of gender inequality.

In Table 1 we report the observed frequencies for the last wave concerning 2,497 couples.[5] We notice that despite the continuing rise in Japanese women's participation in the economy as well as in many Western societies, husbands until recently have been the sole or the primary breadwinner in 65% of the couples and equal-provider couples have been only 11.8%.

The available covariates are chosen according to subject matter knowledge for example couples' relative education-level between spouse has been considered to measures whether wife has equal, higher or lower education level than husband.[6]

4. Results

We report the results of the model estimated on the data collected in 2015 due to space limitations. First, we performed a multivariate imputation for the missing values reported for age and husband income by a using weighted

[5] The missing values for the response are due to *missing household* and/or wife's income information.

[6] List of covariates and corresponding categories: *wife's age*: $\leq$ 32, (32,37], (37,40]; (40,44]; (44,47]; (47,51]; (51,55]; (55,58]; (58,61]; $>$ 61; *husband's age:* $\leq$ 34; (34,39]; (39,43]; (43,46]; (46,50]; (50,54]; (54,58]; (58,61]; (61,64]; $>$ 64; *husband's income in ten thousands yen*: $\leq$ 175, (175,275], (275,325]; (325,375]; (375,425]; (425,500]; (500,600]; (600,700]; (700,900]; >900; *size of the place of living*: major cities; $\geq$ 200,000; [100,000,200,000]; $<$ 100,000; small towns and villages; number of children: 0,1,2,$>$ 3; *preschool children*: no, yes; *wife's educational level*: less than high school, high school, college degree, higher; *wife's relative education*: equal, lower, higher than the husband level.

Table 2. Estimated conditional probabilities ($\hat{\phi}_{j|u}$) under the selected model of the responses given the latent classes.

Conditional Probabilities ($\hat{\phi}_{j|u}$)	1	2	3	4	5
latent class 1 (U_T)	1.000	0.000	0.000	0.000	0.000
latent class 2 (U_N)	0.179	0.578	0.161	0.074	0.008

mean matching method according with the sampling weights and with the other covariates as predictive variables. Then, we estimated the latent class model without covariates with a number of latent classes ranging from 1 to 4 by accounting for different initializations of the EM algorithm[7]. The model with two latent classes has the highest maximum log-likelihood at convergence equal to $\hat{\ell} = -2,456.7$, and a lowest AIC value equal to 4,935.4 with 11 free parameters. The two latent subpopulations are disentangled on the basis of the estimated probabilities of the manifest model that are reported in Table 2. According to the results we define the first latent class as that of Traditional couple (U_T) and the second latent class as that of New couple (U_N). The first one is characterized by a high degree of gender role specialization, strong gender based division of work where the husband specialize in market-work and wife in domestic work and caregiver. The second one is comprised mainly of couples where the husband is not the unique provider. A small proportion of the class is husband sole provider 17% and 58% are the couples where the husband provide majority. It is important to note that only 16% is the probability of equal providers.

The other step of the analysis is performed by adding the covariates. The estimates of the logit regression parameters as in Equation (1) affecting the transition from the traditional couples (U_T) to the new couples (U_N) are reported in Table 3 only for the coefficients which resulted to be significant due to space limitations (see footnote 2 for the categories of each covariate). The estimated intercept is positive indicating a general tendency towards the new type of family U_N. We observe that having preschool children shows the highest estimated coefficient whose negative sign indicates that wife hav-

[7] The model is estimated by adapting the functions of the `R` package `LMest` (Bartolucci, Pandofi and Pennoni, 2017)

Table 3. Estimates of significant logit regression parameters. (Income in ten thousands yen; significance at 10%(†), 5%*,1%**).

Estimates	U_N
$\hat{\beta}_0$	3.777^{**}
wife age ≤ 32	-1.039^{*}
husband's income (600,700]	-1.479^{**}
husband's income (700,900]	-1.221^{**}
husband's income >900	-1.265^{**}
wife'education less than high school	$-1.085^{\dagger}$
wife'education higher than husband'edu.	$0.887^{\dagger}$
preschool children	-2.485^{**}
one child	$-0.479^{\dagger}$

ing preschool children tend to belong to the cluster of traditional couples (the estimated odd ratio for them is equal to 0.08). Interestingly, husband's top incomes determine a lower probability towards the U_N.

The spouses's allocation to each latent class is performed through the estimated maximum *a-posteriori* probability, determined as in Equation (2). The percentage of couples predicted in the traditional family structure U_T is 11.21%. For this subpopulation in Table 4 we show the covariates configuration that can be compare with that obtained for the couples assigned to latent class U_N. We notice that none has husband's income less then 1,750,000 yen a year and that 61% of them has preschool children. We expected that the younger couples support gender egalitarian values more and this would be reflected in gender equality in couples earnings structures. However, we found negative association between age and the probability of being in U_N. It is still not normative for young married women to share equal financial responsibilities within household. This is partly because of the chronic shortages of regular childcare arrangements.

Comparing these results with those obtained for the data collected at previous years we found an increase in the proportion of couples in non-traditional families. One of the reasons why the new families has been more represented in the last past decade is that being a conventional single-income household has becoming more difficult due to the recent financial crisis.

Table 4. Weighted frequencies with survey weights of the relevant covariates for couples allocated in latent class U_T (h.s. high school, h.e. husband'education, see footnote 2 for categories).

Covariates (%)	1	8	9	10
wife'age	29.4			
husband's income	0.0	12.5	11.9	15.2
wife'edu. $<$ h.s.	8.4			
wife'edu. $>$ h.e.	6.4			
preschool	61.1			
one child	40.9			

Acknowledgements: Fulvia Pennoni acknowledges the financial support of the grant "Finite mixture and latent variable models for causal inference and analysis of socio-economic data" (FIRB – Futuro in ricerca) funded by the Italian Government (RBFR12SHVV_004). Miki Nakai acknowledges the financial support of the JSPS Grant-in-Aid for Scientific Research (No. 26380658, No. 17K04103 and No. 16H02045 as part of the SSP project). She thanks both the SSM and the SSP committee for the permission to use their data.

References

Akaike H. (1973) Information Theory as an Extension of the Maximum Likelihood Principle. In BN Petrov, C F (Eds.), *Second International Symposium on Information Theory*, 267-281. Budapest: Akademiai Kiado.

Bartolucci F., Farcomeni A., Pennoni F. (2013) *Latent Markov Models for Longitudinal Data*, Boca Raton: Chapman and Hall/CRC press.

Bartolucci F., Pandolfi S., Pennoni F. (2017) `LMest`: An `R` Package for Latent Markov Models for Longitudinal Categorical Data, *Journal of Statistical Software*, 81, 1-38.

Dempster A.P., Laird N.M., Rubin D.B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion), *Journal of the Royal Statistical Society B*, 39, 1-38.

Lazarsfeld P.F., Henry N.W. (1968) *Latent Structure Analysis, Houghton Mifflin, Boston.*

Nakai M. (2017) Changes in couples' breadwinning patterns and wife's economic role in Japan. In: Greselin, F. *et al.* (Eds.), *Proceedings of the conference of the CLAssification and Data Analysis Group*, Universitas Studiorum, Mantova, 1-6.

Pennoni F. (2014) *Issues on the Estimation of Latent Variable and Latent Class Models*, Scholars'Press, Saarbücken.

Sorensen A., McLanahan S. (1987) Married women economic dependency, 1940-1980, *American Journal of Sociology*, 93, 659-687.