ESTUDIOS DE ECONOMÍA APLICADA VOL. 20 - II, 2002. PÁGS. 499-508

Correcting Classification Error in Income Mobility

JESÚS PÉREZ MAYO y MIGUEL A. FAJARDO CALDERA.
Departamento de Economía Aplicada y Organización de Empresas. Facultad de CC. Económicas y Empresariales. Universidad de Extremadura.
Avda. de Elvas, s/n-06071 Badajoz. Fax: 924 24 25 09. E-mail: jperez@unex.es.

SUMMARY

The mobility of a categorical variable can be a mix of two different parts: true movement and measurement or classification error. For instance, observed transitions can be hiding a real immobility and, therefore, these changes are caused by measurement error.

The Latent Mixed Markov Model is proposed to solve this problem in this paper.

Income mobility is a well-known example of categorical variables mobility in Economics. So, the authors think that the Latent Mixed Markov Model is a good option for measuring the actual income mobility.

In this paper, manifest or observed mobility is analysed firstly and, afterwards, Latent Mixed Markov Model is applied to improve the reliability of the study.

Finally, the latent transition matrix shows that mobility is a "virtual" phenomenon, but not a real one due to the high latent probabilities in the main diagonal cells of the matrix.

Income data are from Spanish Household Panel Survey (ECPF) for 1995 and the categories where the households are located every quarter are compared.

Keywords: mobility, latent variables, Markov models, panel data.

RESUMEN

A la hora de estudiar la movilidad de una variable categórica nos hallamos ante dos fenómenos conjugados: el cambio real y el error de clasificación o de medida. Así, los movimientos observados pueden ocultar una situación real de inmovilidad, por lo que los cambios se deberían al error antes citado.

Se propone el modelo latente de Markov para corregir este problema. Entre los distintos ejemplos de variables categóricas en la economía, uno de los más conocidos y estudiados es el análisis de la movilidad de la renta. Por esta razón, los autores piensan que el modelo presentado es una buena alternativa para medir correctamente tal movilidad.

En primer lugar, se muestra en este artículo la movilidad observada o manifiesta y, más tarde, se aplica el modelo latente de Markov para mejorar la fiabilidad del estudio.

Finalmente, la matriz de transición latente muestra que la movilidad es más un fenómeno "virtual" que real, puesto que son muy elevadas las probabilidades correspondientes a la diagonal principal.

AMS classification: 60J20, 91C20, 62P20, 62H17.

Artículo recibido el 14 de febrero de 2001. Aceptado el 1 de marzo de 2002.

Los datos analizados provienen de la Encuesta Continua de Presupuestos Familiares de España para 1995 y se comparan las categorías que los hogares ocupan cada trimestre de este año.

Palabras clave: movilidad, variables latentes, modelos de Markov, datos de panel.

1. INTRODUCTION

Although the majority of models to analyse income mobility assume that income data is reliable, these data presents a low reliability level due to the existence of measurement error. Its cause is the lack of sincerity for declaring the real amount of income. Besides, it is well known that the measurement error causes an over-estimation of mobility, showing transitions that do not take place between the true states.

Therefore, a model to correct this error is needed.

2. METHODOLOGY

2.1 Latent class model

The variable of interest is usually assumed to be measured without error. However, since in most situations such an assumption is unrealistic, it is important to be able to consider measurement error when statistical models are specified. This problem of measurement error has caused that a family of models called latent structure models, based on the assumption of local independence, has been proposed. This means that the observed variables, which are used to measure the unobserved variable of interest, are assumed mutually independent for a particular value of the latent variable.

Latent structure models can be classified according to the measurement level of the latent variable(s) and the measurement level of the manifest variables (Bartholomew, 1987; Heinen, 1993). Continuous manifest variables are used as indicators for one or more continuous latent variables in factor analysis. In latent trait models, normally one continuous latent variable is assumed to underlie a set of categorical indicators. Finally, when both the manifest and the latent variables are categorical, we have a latent class model (Lazarsfeld and Henry, 1968; Goodman, 1974; Haberman, 1979).

A latent class model is assumed with one latent variable W with index w and three indicators A, B, and C with indices a, b, and c. Moreover, let W* denote the number of latent classes, and A*, B*, and C* the number of categories of A, B, and C, respectively. The basic equations of the latent class model are

$$p_{abc} = \sum_{w=1}^{W^*} p_{wabc} \qquad \textbf{(1)}$$

where

$$p_{wabc} = p_w p_{a|w} p_{b|w} p_{c|w} \quad \textbf{(2)}$$

Here, π_{wabc} denotes a probability of belonging to cell *(w, a, b,* c) in the joint distribution including the latent dimension W. π_w is the proportion of the population belonging to latent class *w.* The other p's-parameters are conditional response probabilities. For instance, $\pi_{a|w}$ is the probability of being in class a on A given that one belongs to latent class *w.*

Thus, the population is divided into W* exhaustive and mutually exclusive classes. Therefore, the joint probability of the observed variables can be obtained by summation over the latent dimension. The classical parameterisation of the latent class model, as proposed by Lazarsfeld and Henry (1968) and as used by Goodman (1974), is given in Equation 2. It can be seen that the observed variables A, *B,* and C are assumed to be mutually independent given a particular class on the latent variable W.

2.2 Latent Markov model

If you combine a usual Markov model and the latent class model given in Equation 1, you obtain a model which can be used for analysing change, but in which the states occupied at different points in time may be measured with error. This model, originally proposed by Wiggins (1973), is called a latent Markov model. Poulsen (1982), Van de Pol and De Leeuw (1986), and Van de Pol and Langeheine (1990) contributed to its practical applicability.

It is well known that measurement error attenuates the relationships between variables. This means that the relationship between two observed variables that are subject to measurement error will generally be weaker than their true relationship. When mobility is analysed, this fact implies that the strength of the relationships among the true states occupied at two subsequent points in time will be underestimated, or in other words, the amount of mobility will be overestimated when the observed states are subject to measurement error. When the data are subject to measurement error, the observed transitions are, in fact, a mixture of true mobility and spurious change resulting from measurement error (Van de Pol and De Leeuw, 1986; Hagenaars, 1992). The latent Markov model makes it possible to separate both.

To be able to formulate the latent Markov model, the notation has to be extended. Let W_t be the latent or true state at T = t having two indicators denoted by A_t and B_t. Assume again that, one has observations for two occasions, that is, *T** = 2. So, the data are collected in a four-way frequency table with cell counts $n_{a_1 a_2 b_1 b_2}$ and the probability of belonging to a particular cell in the joint distribution of the two latent variables and the four indicators is denoted by $p_{w_1 a_1 b_1 w_2 a_2 b_2}$. The latent Markov model for two points in time and two indicators per occasion can be defined as

$$p_{w_1 a_1 b_1 w_2 a_2 b_2} = p_{w_1} p_{a_1|w_1} p_{b_1|w_1} p_{w_2|w_1} p_{a_2|w_2} p_{b_2|w_2} \quad \textbf{(3)}$$

It is possible to identify the latent Markov model with only one indicator (manifest variable) each period. Therefore, the simplified latent Markov model is

$$p_{w_1 a_1 w_2 a_2} = p_{w_1} p_{a_1|w_1} p_{w_2|w_1} p_{a_2|w_2} \quad \textbf{(4)}$$

Nevertheless, when there are four or more periods, another way to get identifiability is by assuming stationarity, that is, constant changes.

$$p_{a_1|w_1} = p_{a_2|w_2} \quad \textbf{(5)}$$

Now, we are going to express the model in the usual way, that is, a different notation is used for every probability.

$$P_{ij} = \sum_{w_1=1}^{W^*} \sum_{w_2=1}^{W^*} d^1_{w_1} r^1_{i|w_1} t^{21}_{w_2|w_1} r^2_{j|w_2} \quad \textbf{(6)}$$

where:

- $d^1_{w_1}$ represents the initial proportion of the w_1-th latent class,
- $r^1_{i|w_1}$, the probability of staying in the manifest category i given the w_1-th latent class in the first period,
- $t^{21}_{w_2|w_1}$, the latent transition probability, that is, the probability of change from the latent class w_1 in the first period to the w_2-th class on the second, and
- $r^2_{j|w_2}$, the probability of staying in the manifest category j given the w_2-th latent class in the second period.

As each parameter represents a probability, we must impose these restrictions:

$$\sum_{w_1=1}^{W^*} d_{w_1} = 1; \quad \sum_i r^t_{i|w_t} = 1; \quad \sum_{w_{t+1}} t^{t+1t}_{w_{t+1}|w_t} = 1 \quad \textbf{(7)}$$

Besides, we assumed that the classification errors are independent and, therefore, the measured states are conditionally independent on the latent ones and the probability of staying in a given manifest category depends only on the corresponding latent one.

2.3 Estimation by the EM algorithm

The EM algorithm (Dempster, Laird and Rubin, 1977) is an iterative algorithm used to estimate model parameters when some data are missing. In our case, the latent state W_t are missing for all individuals.

This algorithm is divided in two separated steps per cycle: an E(xpectation) step and a M(aximization) step. In the former, we obtain auxiliary estimates for the missing data using the incomplete data and the parameters from the previous EM iteration. The iterations continue until convergence is reached.

If we assumed a multinomial scheme sampling, the following incomplete data log-likelihood must be maximised.

$$\log L_{(p)} = \sum_{a_1 a_2} n_{a_1 a_2} \log \sum_{w_1 w_2} \hat{p}_{w_1 w_2 a_1 a_2} \qquad \textbf{(8)}$$

In the equation 8, the estimated probabilities are collapsed over the latent dimensions.

In the E step the expectation of the complete data log-likelihood given the observed data and the parameters from the previous iteration is computed

$$\log L^{*}_{(p)} = \sum_{w_1 w_2 a_1 a_2} \hat{n}_{w_1 a_1 w_2 a_2} \log \hat{p}_{w_1 w_2 a_1 a_2} \qquad \textbf{(9)}$$

The E step involves estimating the unobserved frequencies including the latent variables,

$$\hat{n}_{w_1 a_1 w_2 a_2} = n_{a_1 a_2} \hat{p}_{w_1 w_2 | a_1 a_2} \qquad \textbf{(10)}$$

where $\hat{p}_{w_1 w_2 | a_1 a_2}$ is the estimated probability of the missing data given the observed data, evaluated using the estimated probabilities from the previous iteration.

In the M step, the complete data log-likelihood is maximised to improve parameter estimates. The new estimates for the estimated probabilities $\hat{p}_{w_1 w_2 a_1 a_2}$ are used in another E step to get new estimates for the frequencies of the complete table.

The iterations continue until the convergence is reached.

2.4 Testing

We may use the likelihood ratio chi-square statistic L^2 and the Pearson chi-square X^2 to compare the observed frequencies with the estimates of the expected frequencies, both models with the N-1-P degrees of freedom, where N is the number of cells and P the number of free parameters.

$$L^2 = -2 \sum_{ij} n_{ij} \log \left[\frac{N \hat{p}_{ij}}{n_{ij}} \right]$$

$$c^2 = \sum_{ij} \frac{\left(n_{ij} - N \hat{p}_{ij} \right)}{N \hat{p}_{ij}} \qquad \textbf{(11)}$$

If nested models are compared, it must be used the likelihood ratio difference $DL^2 = L(M_2) - L(M_1)$, that follows a χ^2 distribution with as many degrees of freedom as the difference between those of the both models.

Finally, if non-nested models are compared and as the large number of individuals may cause that small differences may be statistically significant, the Bayesian Information Criterion (BIC; Raftery 1986) may be used.

$$BIC = L^2 - \log Ndf \qquad \textbf{(12)}$$

3. EMPIRICAL ANALYSIS

3.1 Dataset

The dataset used is the Spanish Household Panel Survey (Encuesta Continua de Presupuestos Familiares; ECPF) is a panel survey done by the Spanish Statistical Office (INE) since 1985 and every quarter an eight of the sample is renewed. Therefore, the maximum interview period of a household is two years (eight quarters). This survey has been mainly designed to provide information on household budgets and to build the weights for the computation of retail price indices.

Since the author wants to propose some models for complete panels, it is not possible to use the whole sample. Thus, we built a complete panel by selecting those households that are interviewed during the four quarters of 1995 (the last year we have data): 1689 households.

The ECPF collects information about demographic features of household, individuals' characteristics, income and expenditures. The main limitation of the survey is the temporal limitation and the lack of identification of households.

3.2 Analysis

The variable used in this study is the quartile where the household belongs in Spain every quarter. Therefore, we have not a quarter of the sample in every category.

The transitions between the income categories are analysed, but the measurements are not assumed completely reliable.

Table 1. Results for the manifest models

Model	L^2	Df	BIC
Non stationary	1123.8064	216	-481.4822
Stationary	1145.7914	240	-637.8627
Source: LEM output			

The first step of the study is to observe the goodness-of-fit of the manifest models: non-stationary and stationary Markov model. Recall that the existence of large samples recommends us to use the BIC statistic. By using this statistic, you should choose the lower one and a negative value means that this model is preferred to the saturated one. In this case, the least value corresponds to the stationary Markov model and so, the latent transition probabilities are assumed time-homogeneous.

Besides, this last assumption guarantees the identifiability of all the parameters and the model that has been used follows the form given in equation 6.

3.2.1 Testing

In the following table the test results for the models appears: a model without error, a model with heterogeneous measurement error, a model with homogeneous measurement error and a model with correlated measurement errors.

Table 2. Results for the models for latent income mobility

Model	**L^2**	**df**	**BIC**
No error	1145.7914	240	-637.8627
Heterogeneous error	425.9545	192	-1000.9687
Homogeneous error	483.1642	228	-1211.3071
Correlated measurement error	7419.7232	240	5636.0692

Source: LEM output

The model with heterogeneous error should be preferred to the manifest one, indicating that the fit of a stationary Markov model can be improved if the measurement errors in the manifest states are taken into account.

Since the heterogeneous error model has more parameters than the homogeneous error model we assume that the measurement error is equal across periods. This model is the best due to the fact that it shows the least BIC value.

Finally, to consider that the measurement errors at the different periods are correlated, a model is specified with a direct effect of A_{t-1} on A_t. The value of the BIC statistic suggests that the errors are not correlated.

3.2.2 Parameters

The estimation process produces:

1) four initial latent proportions,
2) one reliability matrix,
3) one transition matrix.

The standard errors of the estimated values appear below the estimates between brackets. Table 3 shows the values reported for the initial proportion of those households belonging to each quartile.

Table 3. Initial latent proportions

	1	2	3	4
d_1^w	0.2225	0.2250	0.2625	0.2900
	(0.0135)	(0.0119)	(0.0123)	(0.0142)

Source: LEM output

In this table it is possible to observe how the latent states are distributed in a very similar way than the manifest ones. Every latent class is nearly occupied by a quarter of the sample.

Table 4 reports the reliability matrix of estimates.

Table 4. Reliability matrix

		Latent		
Manifest	**1**	**2**	**3**	**4**
1	0.0049 (0.0030)	0.0028 (0.0015)	0.8933 (0.0114)	0.0434 (0.0096)
2	0.0000 (0.0000)	0.0000 (0.0000)	0.1043 (0.0113)	0.7692 (0.0158)
3	0.8483 (0.0157)	0.0495 (0.0101)	0.0000 (0.0000)	0.1858 (0.0148)
4	0.1468 (0.0154)	0.9478 (0.0102)	0.0024 (0.0013)	0.0016 (0.0022)

Source: LEM output

It is possible to observe that the codes of the latent class are not the same as those of the manifest ones, for instance, the first quartile is not an index for the first class. Nevertheless, we can see that the model corrects the measurement error. It shows a high probability of correct classification for every class: 0.8483, 0.9478, 0.8933 and 0.7692.

By combining the information collected in Tables 3 and 4, we can say that the 26.25% of households are low-income households, the 29% low-medium income households, the 22.25% medium-high households and the 22.5% high-income households.

We can observe the dynamic of income in the transition matrix.

Table 5. Latent transitions

	t+1			
t	**Low**	**Low-medium**	**Medium-high**	**High**
Low	0.9531 (0.0095)	0.0469 (0.0095)	0.0000 (0.0000)	0.0000 (0.0000)
Low-medium	0.0250 (0.0086)	0.8873 (0.0141)	0.0849 (0.0115)	0.0028 (0.0024)
Medium-high	0.0065 (0.0036)	0.0730 (0.0132)	0.8710 (0.0178)	0.0495 (0.0119)
High	0.0000 (0.0000)	0.0031 (0.0026)	0.0084 (0.0093)	0.9885 (0.0095)

Source: LEM output

This table shows a high degree of expected immobility because the lower probability in the main diagonal is 87.10% (for medium-high households). If this matrix is compared with the corresponding one to the stationary Markov manifest model (table 6), it is possible to observe that the measurement error causes mobility. Mobility is higher in the last model.

This immobility may be caused by the categorisation of the income: there are only four groups. However, if deciles were used, the table and the number of parameters to estimate would be very large.

Table 6. Observed transitions

	t+1			
t	1	2	3	4
1	0.8090	0.1553	0.0157	0.0085
2	0.1595	0.6237	0.1959	0.0092
3	0.0242	0.2034	0.6144	0.1476
4	0.0073	0.0176	0.1740	0.8347

Source: LEM output

Besides, these results do not coincide with those from other studies about income mobility in Spain as Cantó-Sánchez (1998). These differences may be caused by the short accounting period (only one year), the chosen year or the number of categories (it is expected to have more changes if there are more categories).

Therefore, there some hints to improve this study and, for the moment, this paper has to be considered as a first approximation to measurement error and a test of Latent Markov model as a very good tool to correct measurement error in income mobility.

4. REMARKS

We have analysed in this paper the income mobility in Spain during 1995 and we have used the household as unit of analysis.

Although the income is usually assumed to be measured without measurement error, it is not a realistic assumption. Thus, a model to correct this error, a Latent Markov model, is proposed and tested.

In this study, we have observed that the reliability of the data is very high, that is, the households are well classified and the errors are low. Nevertheless, the comparison between manifest and latent stationary models reveals that the measurement error causes more mobility than the actual situation. The latent transition matrix shows a high degree of immobility.

We think that this paper presents an alternative model to analyse mobility and it must be improve to test if the results are consistent with different years, a large sample period and more manifest categories.

Gratitude

An anonymous referee's comments that have improved this paper are aknowledged by the authors.

REFERENCES

BARTHOLOMEW, D.J. (1987): *Latent variables models and factor analysis*. Griffin London.

CANTÓ-SÁNCHEZ, O.(1998): *Income mobility in Spain: How much is there*? FEDEA Working Papers EEE17, Madrid

DEMPSTER, A.P.; LAIRD, N.M.; RUBIN, D.B. (1977): "Maximum likelihood estimation from incomplete data via the EM algorithm" *Journal of the Royal Statistical Society* B 39: 1-38.

GHELLINI, G.; PANNUZZI, N.; TARQUINI, S. (1995): *A latent Marko model for poverty análisis: the case of the GSOEP*.CEPS Document 11, Differdange (Luxembourg).

GOODMAN, L.A. (1974): "Exploratory latent structure analysis using both identifiable and unidentifiable models". *Biometrika* 61: 215-231.

HAGENAARS, J.A. (1990): *Categorical longitudinal data. Log-linear Panel, Trend, and Cohort Analysis*. Sage Newbury Park.

HABERMAN, S.J. (1979): *Analysis of qualitative data vol. 2 New developments*. Academic Press New York.

HEINEN, A. (1993): *Discrete latent variables models*. Tilburg University Press. Tilburg University (Netherlands)

LANGEHEINE, R.; VAN DE POL, F. (1994): "Discrete-Time Mixed Markov latent class models". In: DALE, A.; DAVIES, R.B. (ed): *Analyzing social and political change*. Sage Publications, London Thousand Oaks New Delhi.

LAZARSFELD, P.F.; HENRY, N.W. (1968): *Latent structure analysis*. Houghton Mifflin Boston.

POULSEN, C.A. (1982): *Latent structure analysis with choice modelling*. Aarhus School of Business Aministration and Economics, Aarhus (Denmark)

RAFTERY, A.E. (1986): "Choosing models for cross- classifications". *American Sociological Review* 51: 145-146.

VAN DE POL, F.; DE LEEUW, J. (1986): "A latent Markov model to correct for measurement error". *Sociological Methods and Research* 15: 188-141.

VAN DE POL, F.; LANGEHEINE, R. (1990): "Mixed Markov latent class models". In CLOGG, C.C. (ed.) *Sociological Methodology* 1990. Basil Blackwell Oxford.

VERMUNT, J.K.; LANGEHEINE, R.; BOCKENHOLT, U. (1995): *Discrete-time discrete-state latent Markow models with time-.constant and time-varying covariates.* WORC Paper 95.06.013/7. Tilburg University (Netherlands).

VERMUNT, J.K. (1997): *Log-linear Models for Event Histories*. Sage Publications, London Thousand Oaks New Delhi.

WIGGINS, L.M. (1973): *Panel analysis*, Elsevier, Amsterdam.