Un modelo para corregir el error de medida en el análisis de la movilidad de la renta

Jesús Pérez Mayo[1]
Miguel Ángel Fajardo Caldera
Departamento de Economía Aplicada. Universidad de Extremadura

Resumen

Este trabajo considera un modelo de Markov mixto latente para corregir el error de medida que afecta a los datos de renta y, por tanto, al análisis de su movilidad. De esta manera, el estudio muestra, en primer lugar, los resultados observados de la transición y, más tarde, el modelo proporciona la transición latente, mejorando la fiabilidad del análisis. Los datos analizados corresponden a los microdatos de la Encuesta Continua de Presupuestos Familiares *para el año 1995.*

Palabras clave: *renta de las economías domésticas, proceso de Markov, datos de panel, España.*
Clasificación JEL: *C23, D10, D31.*

Abstract

This work considers a latent mixed Markov model to correct measurement error in tax data and therefore in the analysis of income mobility. The study first shows the results from the transition and, later, the model provides the latent transition, thus enhancing the reliability of the analysis. The data analysed were micro-data taken from the Continuous Household Budget Survey for the year 1995.

Keywords*: household income, Markov' process, panel data, Spain.*
JEL Classification*: C23, D10, D31.*

1. Introducción

A pesar de que la mayoría de los modelos que analizan la movilidad de la renta suponen que los datos de renta son fiables, dichos datos presentan un nivel bajo de fiabilidad debido a la existencia del error de medida, causado por la falta de sinceridad al declarar la cifra real de renta. Además, es bien conocido que el error de medida provoca una estimación por exceso de la movilidad, ya que se consideran movimientos que no ocurren realmente. Por lo tanto, se hace necesario encontrar un modelo que corrija dicho error.

2. Metodología

2.1 El modelo de clases latentes

Generalmente, se supone que la variable objeto del estudio se ha medido sin error. Sin embargo, puesto que tal supuesto no es realista en gran parte de las situaciones, es impor-

---

[1] Una versión preliminar de este trabajo fue presentada en el seminario «Fighting Poverty and Inequality through Tax-benefit Reform: Empirical Approaches» celebrado en Barcelona en 2000 y, más tarde, publicada en *Estudios de Economía Aplicada* con el título «Correcting Classification Error in Income Mobility».

tante poder considerar el error de medida a la hora de especificar los modelos estadísticos. Este problema ha provocado que se haya propuesto una familia de modelos llamados modelos de estructuras latentes, basados en el supuesto de la independencia local. Esto es, se supone que las variables observadas, utilizadas para medir la variable no observada objeto del estudio, son independientes entre sí para un valor dado de la variable latente.

Los modelos de estructuras latentes presentan distintas modelizaciones dependiendo del tipo de variables latentes y observadas que tengamos (Bartholomew, 1987; Heinen, 1993). Algunas variables manifiestas continuas se utilizan como indicadores para una o más variables latentes continuas en el análisis factorial. En los modelos de rasgo latente, usualmente se supone una variable latente detrás de un conjunto de indicadores categóricos. En el caso que nos ocupa, donde todas las variables son categóricas, el modelo de estructuras latentes se conoce como modelo de clases latentes (Lazarsfeld, 1950; Lazarsfeld y Henry, 1968; Goodman, 1974; Haberman, 1979) (véase Esquema 1).

ESQUEMA 1
MODELOS DE ESTRUCTURAS LATENTES

	Latente	
Observada	Continua	Categórica
Continua	Análisis factorial	Modelo de perfil latente
Categórica	Modelo de rasgo latente	Modelo de clases latentes

Un modelo de clases latentes, por tanto, se compone de un conjunto de variables cuyos valores se observan directamente y una variable latente no observable directamente.

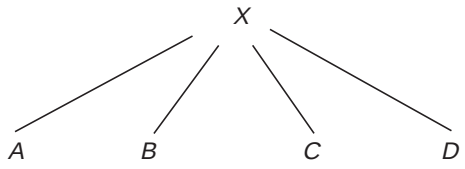
Supongamos un modelo de clases latentes con una variable latente X y cuatro indicadores o variables observadas A, B, C y D. Además X^* es el número de clases latentes y A^*, B^*, C^* y D^*, las categorías de las variables A, B, C y D, respectivamente. La representación gráfica de dicho modelo aparece en la Figura 1 y su ecuación básica es:

$$\pi_{abcd} = \sum_{x=1}^{X^*} \pi_{xabcd}\,, \qquad [1]$$

donde

$$\pi_{xabcd} = \pi_x \pi_{abcd|x} = \pi_x \pi_{a|x} \pi_{b|x} \pi_{c|x} \pi_{d|x} \qquad [2]$$

FIGURA 1
UN MODELO DE CLASES LATENTES



La formulación anterior del modelo de clases latentes es la clásica debida a Lazarsfeld (1950), donde π_{xabcd} es la probabilidad conjunta de todas las variables (manifiestas y latente), π_x la probabilidad de pertenecer a la clase latente x y los distintos términos $\pi_{i|x}$, las pro-

babilidades condicionadas de estar en la categoría *i* de las respectivas variables, dada la pertenencia a la clase latente *x*.

Por tanto, los parámetros del modelo de clases latentes son las probabilidades condicionadas $\pi_{a|x}$, $\pi_{b|x}$, $\pi_{c|x}$, $\pi_{d|x}$ y las probabilidades de las clases latentes π_x, que estarán sometidas a las siguientes restricciones:

$$\sum_{a=1}^{A^*}\pi_{a|x} = \sum_{b=1}^{B^*}\pi_{b|x} = \sum_{c=1}^{C^*}\pi_{c|x} = \sum_{d=1}^{D^*}\pi_{d|x} = 1$$

$$\sum_{x=1}^{X^*}\pi_x = 1 \qquad [3]$$

De nuevo, en la ecuación [1] se manifiesta la hipótesis de independencia local, ya que la población se divide en *X** clases exhaustivas y mutuamente excluyentes por lo que la probabilidad conjunta de las variables observadas se obtiene sumando sobre las clases latentes.

La ecuación [1] se puede expresar mediante un modelo log-lineal (Haberman, 1979). El concepto de independencia local provoca que dicho modelo sea {*AX, BX, CX, DX*}, cuya expresión es

$$\log m_{xabcd} = u_0 + u_x + u_a + u_b + u_b + u_c + u_d + u_{xa} + u_{xc} + u_{xd} \qquad [4]$$

donde $m_{xabcd} = N\pi_{xabcd}$

La ecuación anterior, además de la media general y los términos de una variable, contiene sólo los términos de interacción entre la variable latente *X* y las variables manifiestas. Como las variables manifiestas son independientes entre sí dada la clase latente, no aparecen los términos de interacción entre algunas variables observadas.

Finalmente, es posible relacionar los parámetros de las ecuaciones [1] y [4] según un modelo logit (Haberman, 1979).

$$\pi_{a|k} = \frac{\exp(u_a + u_{xa})}{\sum_a \exp(u_a + u_{xa})} \qquad [5]$$

2.2. El modelo de Markov latente

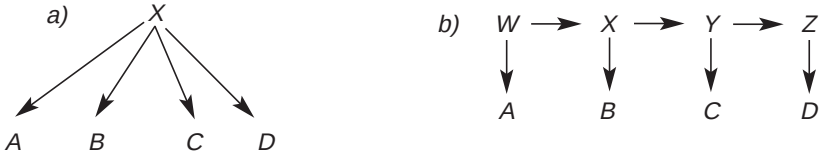
El análisis de una tabla de movilidad presenta la peculiaridad de que se estudia la misma variable en, al menos, dos momentos de tiempo distintos. La consideración de las variables latentes en este contexto puede deberse a distintas causas. En primer lugar, quizás la variable cuya movilidad o cambio se estudia es una variable no observable (por ejemplo, la calidad de vida o el bienestar). Por tanto, es necesario utilizar una o más variables observadas como indicadores de las latentes. Un fenómeno que también implica la introducción de las variables latentes es el error de medida. Es decir, se analiza el cambio de una variable observada, cambio formado por un componente real y otro espurio debido a los errores de repuesta. Finalmente, las variables latentes ayudan a reconocer la heterogeneidad de la población respecto de la movilidad. En este caso, la variable latente divide la población en grupos homogéneos para el cambio, esto es, con matrices de movilidad comunes.

Si se combina el conocido modelo de Markov con el modelo de clases latentes expresado en la ecuación [1], se obtiene un modelo que puede usarse para analizar el cambio, pero donde las categorías ocupadas en los distintos momentos de tiempo pueden estar medidas erróneamente. Este modelo, originalmente propuesto por Wiggins (1973), se llama modelo de Markov latente. Autores como Poulsen (1982), Van de Pol y De Leeuw (1986) y Van de Pol y Langeheine (1990) han contribuido a su aplicación práctica.

Como se ha indicado en la introducción, el error de medida atenúa las relaciones entre las variables. Esto significa que la relación entre dos variables observadas sujetas a algún tipo de error de medida será generalmente más débil que la relación real. Cuando se analiza la movilidad, este hecho implica que la fuerza de las relaciones entre las categorías realmente ocupadas en dos momentos de tiempo será estimada por defecto, o, en otras palabras, la magnitud de la movilidad será estimada por defecto cuando las categorías observadas sufran algún error. Cuando se produce este hecho, las transiciones observadas son, de hecho, una mezcla de la movilidad real y cambio espurio resultante del error de medida (Van de Pol y De Leeuw, 1986; Hagenaars, 1992). El modelo de Markov latente hace posible separar ambos efectos.

Como ejemplos de los modelos latentes para analizar el cambio, se presentan los dos modelos de la Figura 2.

FIGURA 2
MODELOS DINÁMICOS DE VARIABLES LATENTES



El modelo a) refleja aquella situación donde todo el cambio observado se debe al error de medida puesto que sólo existe una variable latente, es decir, no existe cambio latente o real. Por otro lado, en el modelo b) se observa una estructura donde se produce un cambio latente analizado mediante las relaciones entre las variables no observadas *W*, *X*, *Y* y *Z* medidas por las variables indicadores *A*, *B*, *C* y *D*. Además, tenemos que cada variable manifiesta está relacionada con una y sólo una variable latente y cada una de éstas se asocia sólo con la variable latente que recoge el estado real en el período anterior.

Este último modelo es la representación gráfica de un modelo latente de Markov (Wiggins, 1973; Poulsen, 1982) con un único indicador por ocasión. Dicho modelo presenta problemas de identificabilidad que se pueden resolver introduciendo al menos un indicador más o imponiendo restricciones adicionales a las relaciones del modelo. Algunas restricciones posibles son la estacionariedad de la matriz de transición, la independencia de los valores observados en diferentes momentos dados los latentes, la dependencia del error de medida sólo respecto del valor latente actual y no de los previos, entre otras.

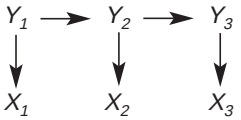
Para simplificar la exposición, supongamos una única variable indicador por ocasión. Supongamos *t* valores consecutivos X_1, X_2, ..., X_t de una misma variable discreta *X* observada con *X** categorías. Además, se supone un vector *p* de probabilidades iniciales y un conjunto de matrices de transición T_t para cada período. Al ser un modelo de Markov de

primer orden, cada matriz de transición T recoge las probabilidades condicionadas de pertenecer a la categoría j de la variable X_t dada la pertenencia al estado j de la variable X_{t-1}.

Además, se introduce una variable discreta latente Y presente en cada uno de los períodos. No es necesario que el número Y^* de estados latentes coincida con la cantidad X^* de categorías observadas.

La estructura de las relaciones entre las distintas variables latentes y observadas para tres momentos de tiempo sigue la asociación representada en la Figura 3.

FIGURA 3
UN MODELO LATENTE DE MARKOV PARA TRES MOMENTOS DE TIEMPO



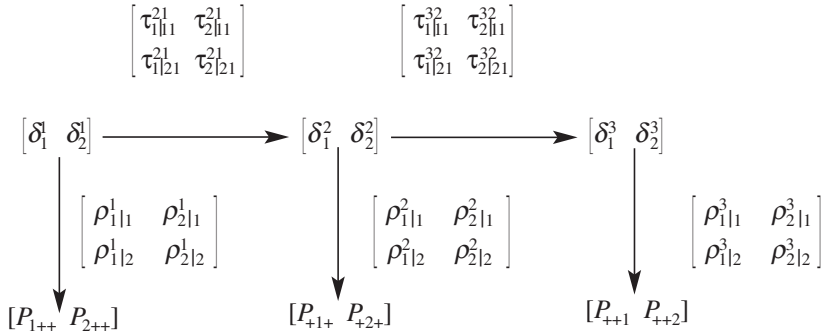
En consecuencia, el vector que recoge la distribución de probabilidad de la variable observada X, p_t, depende de la distribución de probabilidad de la variable latente Y, el vector δ_t, y de la matriz R_t de las probabilidades condicionadas de respuesta o de fiabilidad $\rho_{x|y}$. Estas últimas son las probabilidades de pertenecer a una categoría de la variable X dado un estado latente Y.

Por tanto, podemos expresar la distribución observada de probabilidad en el período t como

$$p_t = \delta_t R_t \qquad [6]$$

En el diagrama siguiente (véase Figura 4), aparecen los parámetros que se estiman si se considera que las variables latentes y observadas son dicotómicas. Se puede observar que el cambio se realiza en la parte latente (estructural) mediante las matrices de transición y que las variables latentes se reflejan en las variables manifiestas a través de las matrices de fiabilidad.

FIGURA 4
RELACIONES Y PARÁMETROS QUE ESTIMAR EN UN MODELO LATENTE DE MARKOV DICOTÓMICO EN TRES PERÍODOS



Por tanto, los parámetros que hay que estimar a partir de los datos observados son:

- La probabilidad inicial δ_{y1} de pertenecer a cada una de las Y^* clases latentes en el momento inicial.
- Las probabilidades condicionadas de respuesta $\rho_{x_i|y_i}$ de estar en cada una de las categorías x_t observadas dado el estado latente y_t en el momento t.
- Las probabilidades de transición latente $\tau_{y_t|y_{t-1}}$ de pasar de cada clase latente y_{t-1} en el momento t-1 a la clase latente y_t en el momento t.

La probabilidad conjunta de pertenecer a una celda determinada de la distribución según este modelo sería[2]:

$$\pi_{y_1y_2y_3x_1x_2x_3} = \delta_{y_1}\rho_{x_1|y_1}\ \tau_{y_2|y_1}\rho_{x_2|y_2}\tau_{y_3|y_2}\rho_{x_3|y_3} \qquad [7]$$

Sin embargo, sólo se conocen los datos observados $\rho_{x_1x_2x_3}$ y, por tanto, nos enfrentamos al problema de estimar estados no observables. Es decir, sólo las probabilidades de las variables observadas tienen una contrapartida empírica para inferir estadísticamente. No obstante, colapsando sobre las variables latentes la expresión [7], podemos obtener una ecuación que relacione la probabilidad de las variables observadas con el producto de probabilidades condicionadas respecto a las latentes:

$$\pi_{x_1x_2x_3} = \sum_{y_1=1}^{y_1^*}\sum_{y_2=1}^{y_2^*}\sum_{y_3=1}^{y_3^*}\ \delta_{y_1}\rho_{x_1|y_1}\ \tau_{y_2|y_1}\rho_{x_2|y_2}\tau_{y_3|y_2}\rho_{x_3|y_3} \qquad [8]$$

Por consiguiente, con una muestra aleatoria de N individuos en el panel, podemos suponer que el modelo latente de Markov sigue una distribución multinomial paramétrica, cuyo parámetro de probabilidad viene dado por la ecuación [8].

Puesto que cada parámetro representa una probabilidad, es necesario imponer estas restricciones:

$$\sum_{w_1=1}^{W^*}\delta_{w_1}=1\quad \sum_{\mathrm{i}}\ \rho^t_{i|w_t}=1\quad \sum_{w_{t+1}}\ \tau^{t+1t}_{w_{t+1}|w_t}=1 \qquad [9]$$

Además, se supone que los errores de clasificación son independientes y, por tanto, los estados medidos son independientes condicionados a los latentes y la probabilidad de estar en una categoría observada concreta depende sólo de la correspondiente latente.

2.3. Estimación mediante el algoritmo EM

A pesar de ser también un modelo loglineal, la determinación de las estimaciones máximo-verosímiles de los parámetros de este modelo es más complicada que en el caso donde se observan todas las variables. Se utilizan distintos métodos de estimación, entre los cua-

---

[2] Se ha utilizado la notación propuesta por los autores que han trabajado sobre estos modelos. Creemos que una letra distinta para cada tipo de probabilidad facilita la comprensión del modelo. Siguiendo la notación usada en el trabajo hasta ahora, la ecuación [7] sería

$$\pi_{y_1y_2y_3x_1x_2x_3} = \pi_{y_1}\pi_{x_1|y_1}\pi_{y_2|y_1}\pi_{x_2|y_2}\pi_{y_3|y_2}\pi_{x_3|y_3}$$

les los más conocidos son el algoritmo de Newton-Raphson y el algoritmo EM (Dempster, Laird y Rubin, 1977).

El último es preferible, ya que como el algoritmo IPF, es sencillo tanto en la teoría como en el cálculo. Además, generalmente los valores iniciales elegidos aleatoriamente son suficientes para llegar a una solución. Presenta frente al Newton-Raphson el inconveniente de necesitar más iteraciones para llegar a la solución. Sin embargo, puesto que cada iteración del algoritmo EM es más rápida, este inconveniente no es tan importante.

El algoritmo EM es un procedimiento iterativo y cada iteración está compuesta por dos pasos. En el paso *E*speranza se calculan todos los valores esperados dados los valores observados y los «actuales» parámetros del modelo. En el paso *M*aximización, se maximiza la función de verosimilitud de todos los datos a partir de los valores esperados calculados en el paso anterior. Esto implica el cálculo de estimaciones actualizadas de los parámetros del modelo como si no faltaran datos, es decir, se utilizan las estimaciones $\hat{n}_{xabcd}$ como si fueran frecuencias observadas. Para hacerlo, se utilizan los mismos procedimientos en la obtención de las estimaciones máximo-verosímiles de un modelo loglineal normal: Newton-Raphson e IPF. Las estimaciones obtenidas se utilizan en un nuevo paso *E*speranza para lograr nuevas estimaciones para las frecuencias de la tabla completa. Las iteraciones continúan hasta que se alcanza la convergencia.

En primer lugar, para poder aplicar el algoritmo, buscaremos los estadísticos suficientes de los datos completos para los parámetros del modelo. Partimos del logaritmo de la función de verosimilitud según la expresión [7].

$$\sum_{y_1y_2y_3}\sum_{x_1x_2x_3} n_{y_1y_2y_3x_1x_2x_3}\log\pi_{y_1y_2y_3x_1x_2x_3} = \sum_{y_1} n_{y_1\cdots}\log\delta_{y_1} + \sum_{x_1,y_1} n_{y_1\cdots x_1\cdots}\log\rho_{x_1|y_1} +$$
$$+ \sum_{y_1,y_2} n_{y_1y_2\cdots}\log\tau_{y_2|y_1} + \sum_{x_2,y_2} n_{\cdot y_2\cdots x_2\cdots}\log\rho_{x_2|y_2} + \sum_{y_2,y_3} n_{\cdot y_2y_3\cdots}\log\tau_{y_3|y_2} + \sum_{x_3,y_3} n_{\cdots y_3\cdots x_3\cdots}\log\rho_{x_3|y_3}$$

Como para que el modelo sea identificable se impone la restricción de que las probabilidades de respuesta sean iguales en cada momento, podemos escribir la expresión anterior como:

$$\sum_{y_1y_2y_3}\sum_{x_1x_2x_3} n_{y_1y_2y_3x_1x_2x_3}\log\pi_{y_1y_2y_3x_1x_2x_3} =$$
$$= \sum_{y_1} n_{y_1\cdots}\log\delta_{y_1} + \sum_{x,y}(n_{y\cdots x\cdots} + n_{\cdot y\cdots x\cdots})\log\rho_{x|y} + \sum_{y_1,y_2} n_{y_1y_2\cdots}\log\tau_{y_2|y_1} + \sum_{y_2,y_3} n_{\cdot y_2y_3\cdots}\log\tau_{y_3|y_2} \quad [10]$$

A partir de esta expresión es posible obtener los estadísticos suficientes para los parámetros.

Una vez hecho esto, ya puede aplicarse el algoritmo EM. En primer lugar, se expone el paso E[3].

$$E\left[n_{y_1\cdots}\,|\,\hat{\pi}(p)\right] = \sum_{x_1x_2x_3} n_{x_1x_2x_3}\,\hat{\pi}_{y_1|x_1x_2x_3}(p)$$

$$E\left[n_{y\cdots x\cdots} + n_{\cdot y\cdots x\cdots} + n_{\cdots y\cdots x}\,|\,\hat{\pi}(p)\right] = \sum_{x_2,x_3} n_{x_1x_2x_3}\,\hat{\pi}_{y_1|x_1x_2x_3}(p) + \sum_{x_2,x_3} n_{x_1x_2x_3}\,\hat{\pi}_{y_2|x_1x_2x_3}(p) +$$

3 El término (*p*) que aparece junto a las distintas probabilidades refleja la iteración p-ésima del algoritmo.

$$+ \sum_{x_2 x_3} n_{x_1 x_2 x_3} \hat{\pi}_{y_3|x_1 x_2 x_3}(p)$$

$$E\left[n_{y_1 y_2}.... \mid \hat{\pi}(p)\right] = \sum_{x_1, x_2, x_3} n_{x_1 x_2 x_3} \hat{\pi}_{y_1 y_2|x_1 x_2 x_3}(p) \quad E\left[n_{.y_2 y_3}.... \mid \hat{\pi}(p)\right] = \sum_{x_1, x_2, x_3} n_{x_1 x_2 x_3} \hat{\pi}_{y_2 y_3|x_1 x_2 x_3}(p),$$

donde

$$\hat{\pi}_{y_1|x_1 x_2 x_3}(p) = \frac{\hat{\pi}_{y_1 \cdot\cdot x_1 x_2 x_3}(p)}{\hat{\pi}_{\cdot\cdot\cdot x_1 x_2 x_3}(p)},$$

$$\hat{\pi}_{y_2|x_1 x_2 x_3}(p) = \frac{\hat{\pi}_{\cdot y_2 \cdot x_1 x_2 x_3}(p)}{\hat{\pi}_{\cdot\cdot\cdot x_1 x_2 x_3}(p)},$$

$$\hat{\pi}_{y_3|x_1 x_2 x_3}(p) = \frac{\hat{\pi}_{\cdot\cdot y_3 x_1 x_2 x_3}(p)}{\hat{\pi}_{\cdot\cdot\cdot x_1 x_2 x_3}(p)},$$

$$\hat{\pi}_{y_1 y_2|x_1 x_2 x_3}(p) = \frac{\hat{\pi}_{y_1 y_2 \cdot x_1 x_2 x_3}(p)}{\hat{\pi}_{\cdot\cdot\cdot x_1 x_2 x_3}(p)} \quad \text{y}$$

$$\hat{\pi}_{y_2 y_3|x_1 x_2 x_3}(p) = \frac{\hat{\pi}_{\cdot y_2 y_3 x_1 x_2 x_3}(p)}{\hat{\pi}_{\cdot\cdot\cdot x_1 x_2 x_3}(p)}$$

Una vez obtenidas las estimaciones de los parámetros del modelo en el paso *p*-ésimo, se determinan las probabilidades estimadas que maximizan la verosimilitud en la etapa *M*aximización.

$$E\left[n_{y_1}.... \mid N, \hat{\pi}(p+1)\right] = N\delta_{y_1}(p+1),$$

$$E\left[n_{y_1 \cdot\cdot x_1}.. \mid N, \hat{\pi}(p+1)\right] = N\delta_{y_1}(p+1)\rho_{x|y}(p+1),$$

$$E\left[n_{\cdot y_2 \cdot\cdot x_2}. \mid N, \hat{\pi}(p+1)\right] = N\rho_{x|y}(p+1) \sum_{y_1} \delta_{y_1} \tau_{y_2|y_1}(p+1),$$

$$E\left[n_{\cdot\cdot y_3 \cdot\cdot x_3} \mid N, \hat{\pi}(p+1)\right] = N\rho_{x|y}(p+1) \sum_{y_1, y_2} \delta_{y_1}(p+1)\, \tau_{y_2|y_1}(p+1)\, \tau_{y_3|y_2}(p+1),$$

$$E\left[n_{y_1 \cdot\cdot x_1}.. + n_{\cdot y_2 \cdot\cdot x_2}. + n_{\cdot\cdot y_3 \cdot\cdot x_3} \mid N, \hat{\pi}(p+1)\right] = N\rho_{x|y}(p+1)\,(\delta_{y_1}(p+1) +$$

$$+ \sum_{y_1} \delta_{y_1}(p+1)\tau_{y_2|y_1}(p+1) + \sum_{y_1, y_2} \delta_{y_1}(p+1)\, \tau_{y_2|y_1}(p+1)\, \tau_{y_3|y_2}(p+1)$$

$$E\left[n_{y_1 y_2}.... \mid N, \hat{\pi}(p+1)\right] = N\delta_{y_1}(p+1)\tau_{y_2|y_1}(p+1)$$

$$E\left[n_{\cdot y_2 y_3}... \mid N, \hat{\pi}(p+1)\right] = N\tau_{y_3|y_2}(p+1) \sum_{y_1} \delta_{y_1} \tau_{y_2|y_1}(p+1)$$

Partiendo de las expresiones anteriores, podemos encontrar las estimaciones p+1-ésimas de los parámetros que maximizan la función de verosimilitud dadas las probabilidades de la iteración anterior.

$$\hat{\delta}_{y_1}(p+1) = \frac{\sum_{x_1, x_2, x_3} n_{x_1x_2x_3} \hat{\pi}_{y_1|x_1x_2x_3}(p)}{N} \qquad [11.a]$$

$$\hat{\rho}_{x|y}(p+1) = \frac{\sum_{x_1, x_2, x_3} n_{x_1x_2x_3} \left(\hat{\pi}_{y_1|x_1x_2x_3}(p) + \hat{\pi}_{y_2|x_1x_2x_3}(p) + \hat{\pi}_{y_3|x_1x_2x_3}(p) \right)}{N\rho_{x|y} \left(\delta_{y_1} + \sum_{y_1} \delta_{y_1}\tau_{y_2|y_1} + \sum_{y_1, y_2} \delta_{y_1}\tau_{y_2|y_1}\tau_{y_3|y_2} \right)} \qquad [11.b]$$

$$\hat{\tau}_{y_2|y_1}(p+1) = \frac{\sum_{x_1, x_2, x_3} n_{x_1x_2x_3} \hat{\pi}_{y_1y_2|x_1x_2x_3}(p)}{\sum_{x_1, x_2, x_3} n_{x_1x_2x_3} \hat{\pi}_{y_1|x_1x_2x_3}(p)} \qquad [11.c]$$

$$\hat{\tau}_{y_3|y_2}(p+1) = \frac{\sum_{x_1, x_2, x_3} n_{x_1x_2x_3} \hat{\pi}_{y_2y_3|x_1x_2x_3}(p)}{\sum_{y_1} \sum_{x_1, x_2, x_3} n_{x_1x_2x_3} \hat{\pi}_{y_1y_2|x_1x_2x_3}(p)} \qquad [11.d]$$

En la ecuación [11.b] se comprueba cómo las restricciones de igualdad sobre los parámetros producen una expresión similar a una media ponderada de las probabilidades de respuesta no restringidas.

2.4. Contraste del modelo

La calidad del ajuste de un modelo loglineal concreto puede determinarse con la comparación de las frecuencias observadas, n, con las esperadas estimadas, $\hat{m}$, mediante el contraste de la χ^2 de Pearson y la razón de verosimilitud L^2, cuyas expresiones son las siguientes para un modelo con tres variables, A, B y C.

$$\chi^2 = \sum_a \sum_b \sum_c \frac{(n_{abc} - \hat{m}_{abc})^2}{\hat{m}_{abc}}$$

$$L^2 = 2\sum_a \sum_b \sum_c n_{abc} \ln \frac{n_{abc}}{\hat{m}_{abc}} \qquad [12]$$

Si el modelo es válido para la población, ambos estadísticos siguen asintóticamente una distribución χ^2. Para cada modelo el número de grados de libertad de la distribución se obtiene a partir de la expresión

gl = número de celdas – número de parámetros independientes.

Si algunas frecuencias esperadas estimadas son ceros estructurales o no pueden calcularse algunos parámetros al existir ceros en algunos estadísticos suficientes, Clogg y Eliason (1987) mostraron que la diferencia anterior pasaría a ser

gl = número de celdas sin ceros - número de parámetros estimables.

El estadístico L^2 tiene una ventaja sobre el de Pearson porque puede descomponerse en distintas componentes referidas a diferentes efectos, submodelos o subgrupos. Esta propiedad será muy interesante cuando busquemos un modelo que se ajuste bien y, simultáneamente, sea parsimonioso.

Llegados a este punto, conocemos las herramientas que permiten al investigador determinar en qué medida el modelo propuesto *a priori* se ajusta o no a los datos observados. Sin embargo, el objetivo es encontrar el mejor modelo, aquél que explica las relaciones existentes entre las variables en la población que generan los datos observados. Por tanto, los errores posibles al seleccionar un modelo se producirán cuando éste contenga más parámetros de los necesarios o se excluyan algunos parámetros que forman parte del mejor modelo.

En el proceso lógico de la modelización estadística, se parte de unas hipótesis o supuestos *a priori* que se reflejan en una formulación determinada del modelo. Dichas hipótesis naturalmente deben basarse en las ideas que el investigador tenga sobre las relaciones existentes entre las variables en la población, es decir, es conveniente utilizar los conceptos teóricos relacionados con el problema que intentemos resolver. Estadísticamente, pueden existir cientos de modelos para un solo conjunto de datos que se ajusten con la misma calidad. Si no seguimos la orientación proporcionada por el problema teórico que queremos resolver, es difícil dilucidar qué modelo elegir.

Si los supuestos de partida llevan a un único modelo loglineal no saturado, el proceso es fácil, dado que se limitaría a la aplicación de los estadísticos mostrados en la ecuación [12]. Sin embargo, como hemos expuesto anteriormente, la dificultad comienza a la hora de descubrir cuál es el mejor modelo dentro de una gama.

Si los modelos están anidados jerárquicamente, pueden utilizarse contrastes de L^2 condicionados. Dos modelos están anidados jerárquicamente cuando el modelo restringido contiene sólo un subconjunto de los efectos presentes en el modelo libre.

Entonces, el estadístico L^2 de la razón de verosimilitud contrasta la significatividad de los parámetros libres del modelo libre, dado que el modelo libre es cierto para la población.

Utilizando las mismas variables categóricas A, B y C de la expresión [12], podemos representar el estadístico condicionado $L^2_{r|l}$, donde el subíndice r se refiere al modelo restringido y l al libre como

$$L^2_{r|l} = L^2_r - L^2_l$$

$$= 2\sum_a \sum_b \sum_c n_{abc} \ln \frac{\hat{m}_{abc(l)}}{\hat{m}_{abc(r)}} \qquad [13]$$

Los grados de libertad del contraste vienen definidos por el número de parámetros que se fijan en el modelo restringido, es decir, los grados de libertad del modelo restringido menos los del libre.

Como se puede observar en la expresión [13], el estadístico condicionado $L^2_{r|l}$ puede calcularse como la diferencia de los estadísticos L^2 no condicionados de ambos modelos. Se

confirma el comentario anterior sobre los estadísticos L^2 y χ^2, en el que se comentaba la ventaja del primero por su capacidad de descomponerse y así es posible realizar el estadístico condicionado.

Este estadístico condicionado sigue una distribución χ^2 si el modelo libre es válido y la muestra es grande y la aproximación es buena, incluso en aquellas situaciones, como las muestras pequeñas, en que el contraste no condicionado tiene problemas (Haberman, 1978).

Con el estadístico $L^2_{r|l}$, se contrasta la hipótesis nula de que el modelo restringido es válido para la población frente a la hipótesis alternativa del libre. Por tanto, es diferente el significado de la aceptación y el rechazo respecto del contraste del estadístico L^2 sin condicionar, ya que en este último la comparación se hace con el modelo saturado, no con otro modelo no saturado. Podríamos decir que el estadístico $L^2_{r|l}$ contrasta la validez de las condiciones impuestas al modelo libre para obtener el restringido. A la hora de elegir el mejor modelo, es preferible utilizar los contrastes condicionados entre dos modelos no saturados frente al test no condicionado del modelo restringido contra el modelo saturado.

A partir de la teoría de la información, es posible desarrollar otra forma de seleccionar el modelo más adecuado. El objetivo no es descubrir el modelo verdadero, sino aquél que proporciona mayor información sobre la realidad. Por un lado, las frecuencias esperadas estimadas deben ser parecidas a las observadas y, por otro, el modelo debe ser tan reducido como sea posible.

Los contrastes más conocidos basados en la teoría de la información son el *criterio de información de Akaike* (AIC) (Akaike, 1987) y el *criterio de información bayesiano* (BIC) (Raftery, 1986). El primero, penalizando al modelo según su grado de complejidad, determina hasta qué punto un modelo concreto se desvía de la realidad y su expresión es:

$$AIC = -2\log\ell + 2npar \quad , \qquad [14]$$

donde ℓ representa el valor de la función de verosimilitud y *npar* el número de parámetros desconocidos.

Raftery (1986) desarrolló el AIC dentro del contexto de los modelos loglineales y propuso el BIC que puede calcularse como:

$$BIC = -2\log\ell + (\log N)\, npar \qquad [15]$$

Cuanto menores sean los valores de los criterios, mejor será el modelo porque mayor información contendrá. Ambos criterios pueden calcularse a partir del estadístico L^2 de la siguiente forma

$$\begin{aligned} AIC &= L^2 - 2\, gl, \\ BIC &= L^2 - (\log N)\, gl, \end{aligned} \qquad [16]$$

que, como se puede observar, son mucho más sencillas que las anteriores y consisten en la comparación con los respectivos criterios para el modelo saturado.

Por tanto, y refiriéndonos al BIC al ser el más adecuado en los modelos loglineales, se podrá calcular el criterio BIC a partir del estadístico L^2 no condicionado. Un valor negati-

vo indica que el modelo es preferible al modelo saturado y además, debe elegirse aquel modelo con el menor valor. Este criterio elimina los problemas del ajuste por exceso y por defecto. Un modelo que no se ajuste bien a las frecuencias observadas tendrá un L^2 elevado y, en consecuencia, incrementará el primer término de la diferencia que hará poco probable seleccionarlo. Por otra parte, si un modelo se ajusta muy bien porque posee un gran número de parámetros, al tener una cantidad muy pequeña de grados de libertad, de nuevo, el valor del criterio será muy elevado.

Asimismo, una ventaja del BIC se presenta ante las muestras muy grandes (como en la aplicación empírica que nos ocupa). Un tamaño muestral muy elevado provoca que todos los efectos del modelo saturado sean significativos y, por lo tanto, puede llevar al rechazo de los modelos no saturados si la calidad del ajuste se prueba según los estadísticos usuales χ^2 y L^2.

3. Análisis empírico

3.1. Datos

Los datos utilizados proceden de la *Encuesta Continua de Presupuestos Familiares* (ECPF), encuesta realizada por el Instituto Nacional de Estadística desde 1985 en la que cada trimestre se renueva un octavo de la muestra. Por tanto, un hogar dado es entrevistado como máximo durante dos años (ocho trimestres). Esta encuesta se diseñó principalmente para proporcionar información sobre el consumo de los hogares y construir los pesos para el cálculo del Índice de Precios al Consumo.

Dado que en este trabajo se quieren proponer algunos modelos para paneles completos, no es posible utilizar la muestra completa. Así, se construirá un panel completo seleccionando aquellos hogares entrevistados durante los cuatro trimestres de 1995 (el último año para el que los autores tenían datos): 1.689 hogares.

La ECPF recoge información sobre algunas características demográficas de los hogares, características de los individuos, la renta y los gastos. Las principales restricciones de la encuesta son la limitación temporal y la falta de identificación de los hogares que no permite un enlace de las muestras.

3.2. Análisis

La variable utilizada es la cuartila a la que el hogar pertenece cada trimestre. Dicha cuartila no ha sido calculada por los autores, sino que se ha tomado la asignación realizada por el INE al realizar la encuesta. Se analizan las transiciones entre las categorías de la renta sin suponer las medidas totalmente fiables. El primer paso consiste en observar la calidad del ajuste de los modelos observados: el modelo de Markov no estacionario y el estacionario. Recordemos que el tamaño muestral recomienda utilizar el estadístico BIC. Utilizando éste, se debería elegir el menor y un valor negativo significa que ese modelo se prefiere al saturado (véase Tabla 1).

TABLA 1
RESULTADOS PARA LOS MODELOS OBSERVADOS

Modelo	L^2	gl	BIC
No estacionario	1123.8064	216	–481.4822
Estacionario	1145.7914	240	–637.8627

FUENTE: Elaboración propia.

En este caso, el menor valor corresponde al modelo de Markov estacionario y, por eso, las probabilidades de transición latentes se supondrán homogéneas en el tiempo. Además, este supuesto garantiza la identificabilidad de todos los parámetros.

3.2.1. Selección del modelo

En la Tabla 2 se recogen los resultados del contraste para los siguientes modelos: un modelo sin error, un modelo con errores de medida heterogéneos, un modelo con errores de medida homogéneos y un modelo con errores de medida correlacionados.

TABLA 2
MODELOS PARA LA MOVILIDAD LATENTE

Modelo	L^2	gl	BIC
Sin error	1145.7914	240	–637.8627
Errores heterogéneos	425.9545	192	–1000.9687
Errores homogéneos	483.1642	228	–1211.3071
Errores de medida correlacionados	7419.7232	240	5636.0692

FUENTE: Elaboración propia.

Se prefiere el modelo con errores heterogéneos al manifiesto, indicando que el ajuste de un modelo de Markov estacionario podría mejorarse si se consideraran los errores de medida en los estados observados. Puesto que el modelo con errores heterogéneos tiene más parámetros que el modelo con errores homogéneos, se supondrá en lo sucesivo que los errores de medida son iguales en todos los períodos. Este modelo es el mejor ya que muestra el menor valor del estadístico BIC.

Finalmente, para considerar que los errores en diferentes períodos estén correlacionados, se especifica un modelo con un efecto directo de A_{t-1} sobre A_t. El valor del estadístico BIC sugiere que los efectos no están relacionados.

3.2.2. Parámetros

El proceso de estimación produce:
1) Cuatro proporciones latentes iniciales.
2) Una matriz de fiabilidad.
3) Una matriz de transición.

La Tabla 3 muestra los valores para las proporciones iniciales de aquellos hogares que pertenecen a cada cuartila. Las desviaciones típicas de los valores estimados aparecen bajo las estimaciones entre paréntesis.

TABLA 3
PROPORCIONES LATENTES INICIALES

	1	2	3	4
δ_1^w	0,2225 (0,0135)	0,2250 (0,0119)	0,2625 (0,0123)	0,2900 (0,0142)

FUENTE: Elaboración propia.

En esta tabla se puede observar cómo los estados latentes se distribuyen de una manera muy parecida a los observados. Cada clase está ocupada aproximadamente por un cuarto de la muestra.

La Tabla 4 recoge la matriz de fiabilidad de las estimaciones. La codificación de las clases latentes no es la misma que la de las categorías observadas, ya que , por ejemplo, la primera cuartila no es un índice para la primera clase. No obstante, es posible reconocer que el modelo recoge la influencia del error de medida. Muestra una probabilidad muy alta de estar ante una clasificación correcta para cada clase: 0,8493, 0,9478, 0,8933 y 0,7692. Sólo en lo referente a la última clase, tenemos un nivel de error mayor, cercano al 25 por 100. En el resto, se estima que la calidad de la clasificación se aproxima al 90 por 100.

TABLA 4
MATRIZ DE FIABILIDAD

	Latente			
Observada	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)

FUENTE: Elaboración propia.

Combinando la información recogida en las Tablas 3 y 4, podemos decir que el 26,25 por 100 de los hogares son hogares de baja renta, el 29 por 100 de renta media-baja, el 22,25 por 100 de renta media-alta y el 22,5 por 100 de renta alta. Observamos cómo los mayores errores corresponden a las clases de renta media-baja y media-alta, mientras que los extremos poseen una alta probabilidad de clasificación correcta.

La dinámica de la renta aparece recogida en la matriz de transición (véase Tabla 5).

TABLA 5
MATRIZ DE TRANSICIÓN LATENTE

	t+1			
t	Baja	Baja-media	Media-alta	Alta
Baja	0,9531 (0,0095)	0,0469 (0,0095)	0,0000 (0,0000)	0,0000 (0,0000)
Media-baja	0,0250 (0,0086)	0,8873 (0,0141)	0,0849 (0,0115)	0,0028 (0,0024)
Media-alta	0,0065 (0,0036)	0,0730 (0,0132)	0,8710 (0,0178)	0,0495 (0,0119)
Alta	0,0000 (0,0000)	0,0031 (0,0026)	0,0084 (0,0093)	0,9885 (0,0095)

FUENTE: Elaboración propia.

Esta tabla muestra un grado de inmovilidad esperada alto porque la menor probabilidad en la diagonal principal es 87,10 por 100 (para los hogares de renta media-alta). Si esta matriz se compara con la correspondiente al modelo de Markov observado, es posible ver que los errores de medida causan movilidad, ya que la movilidad es mayor en el segundo.

Este grado de inmovilidad puede estar causado por la categorización de la renta: sólo se toman cuatro grupos. Es de esperar que si se consideran más grupos, la movilidad será mayor. Sin embargo, si se utilizaran decilas, la tabla y la cantidad de parámetros que estimar serían muy grandes y, puesto que el objeto de este análisis empírico es ilustrar el modelo propuesto, los autores han decidido optar por un número menor de categorías.

Además, debemos decir que esos resultados no coinciden con los de otros trabajos sobre movilidad en España como el de Cantó-Sánchez (1998). Creemos que estas discrepancias pueden deberse a la categorización tomada o a la longitud del período considerado (sólo un año). No obstante, pensamos que esta metodología es útil para describir la influencia de los errores de medida en la movilidad y así, poder corregirla. Es preciso comentar que los resultados coinciden con los expuestos por Pena (1996), aunque éstos se refieren al período 1987-1988. Por lo tanto, existen varias líneas para mejorar este trabajo y, por el momento, debe ser considerado como una primera aproximación al problema del error de medida y una prueba de que el modelo de Markov latente puede ser una herramienta muy útil para corregirlo.

4. Conclusiones

En este trabajo se ha analizado la movilidad de la renta en España durante 1995. Aunque la renta se supone que se mide sin error de medida, éste no es un supuesto realista. Por consiguiente, se propone y contrasta un modelo para corregir este error, el modelo de Markov latente.

Se ha observado que la fiabilidad de la clasificación realizada por el INE es muy alta. No obstante, la comparación entre la transición latente y la observada revela que los errores de medida causan más movilidad de la realmente existente, ya que la matriz de transición latente muestra un grado alto de inmovilidad.

Creemos que este trabajo presenta un modelo alternativo para estudiar la movilidad y que debe ser mejorado para contrastar si los resultados son consistentes al tomar años diferentes, un mayor período muestral o más categorías observadas.

Referencias bibliográficas

[1] AKAIKE, H. (1987): «Factor analysis and AIC». *Psychometrika*, 52, 317-332.

[2] BARTHOLOMEW, D.J. (1987): *Latent Variables Models and Factor Analysis*. Londres: Griffin.

[3] CANTÓ-SÁNCHEZ, O. (1998): *Income Mobility in Spain: How Much is There? Documento de trabajo* FEDEA EEE 17, Madrid

[4] CLOGG, C.C. y ELIASON, S.R. (1987): «Some Common Problems in Log-linear Analysis». *Sociological Methods and Research*, 16, 8-14.

[5] DEMPSTER, A.P., LAIRD, N.M. y RUBIN, D.B. (1977): «Maximum Likelihood Estimation from Incomplete Data Via the EM Algorithm». *Journal of the Royal Statistical Society B* 39:1-38.

[6] GHELLINI, G., PANNUZZI, N. y TARQUINI, S. (1995): *A Latent Markov Model for Poverty Analysis: The Case of the GSOEP*. Documento de trabajo CEPS 11, Differdange (Luxemburgo).

[7] GOODMAN, L.A. (1974): «Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models». *Biometrika* 61: 215-231.

[8] HABERMAN, S.J. (1978): *Analysis of Qualitative Data, Vol. 1, Introduction Topics.* Nueva York: Academic Press.

[9] HABERMAN, S.J. (1979): *Analysis of Qualitative Data, Vol. 2, New Developments*. Nueva York: Academic Press.

[10] HAGENAARS, J.A. (1990): *Categorical Longitudinal Data. Log-linear Panel, Trend, and Cohort Analysis*. Londres: Sage Publications

[11] HAGENAARS, J.A. (1992): «Exemplifying Longitudinal Log-linear Analysis with Latent Variables». En: VAN DER HEIJDEN, P.G.M., JANSEN, W., FRANCIS, B. y SEEBER, G.U.H. (eds.), *Statistical modelling*, 105-120, Amsterdam: Elsevier Science Publishers.

[12] HEINEN, A. (1993): *Discrete Latent Variables Models*. Tilburg University (Netherlands): Tilburg University Press.

[13] LANGEHEINE, R. y VAN DE POL, F. (1994): «Discrete-Time Mixed Markov latent class models». En: DALE, A. y DAVIES, R.B. (ed) *Analyzing Social and Political Change*. Londres: Sage Publications.

[14] LAZARSFELD, P.F. (1950): «The Logical and Mathematical Foundation of Latent Structure Analysis». En: STOUFFER, S.A. *et al.* (eds.), *Measurement and Prediction*, 362-472, Princeton: Princeton University Press.

[15] LAZARSFELD, P.F. y HENRY, N.W. (1968): *Latent Structure Analysis.* Boston: Houghton Mifflin.

[16] PENA, J.B. (1996): «Análisis dinámico de la distribución personal de la renta: la movilidad de la distribución personal». En: PENA, J.B. (director), *Distribución Personal de la Renta en España*, 933-980, Madrid: Ediciones Pirámide.
[17] POULSEN, C.A. (1982): *Latent Structure Analysis with Choice Modelling*. Aarhus School of Business Administration and Economics, Aarhus (Dinamarca)
[18] RAFTERY, A.E. (1986): «Choosing Models for Cross-classifications». *American Sociological Review* 51:145-146.
[19] VAN DE POL, F. y DE LEEUW, J. (1986): «A Latent Markov Model to Correct for Measurement Error». *Sociological Methods and Research* 15: 188-141.
[20] VAN DE POL, F. y LANGEHEINE, R. (1990): «Mixed Markov Latent Class Models». En: CLOGG, C.C. (ed.) *Sociological Methodology 1990*. Oxford: Basil Blackwell.
[21] VERMUNT, J.K., LANGEHEINE, R. y BÖCKENHOLT, U. (1995): *Discrete-time Discrete-State Latent Markov Models with Time-constant and Time-varying Covariates*. WORC Paper 95.06.013/7. Tilburg University (Países Bajos).
[22] VERMUNT, J.K. (1997): *Log-linear Models for Event Histories*. Londres: Sage Publications.
[23] WIGGINS, L.M. (1973): *Panel analysis*. Amsterdam: Elsevier.