CRITERIOS DE INFORMACIÓN Y PREDICTIVOS PARA LA SELECCIÓN DE UN MODELO LINEAL MIXTO*

García María del Carmen[*]
Noelia Castellana*; Cecilia Rapelli*
Liliana Koegel*; Mara Catalano*

Resumen. En los estudios longitudinales las unidades experimentales se observan repetidamente en varias ocasiones. Una herramienta importante para el análisis de este tipo de datos son los modelos lineales mixtos que permiten modelar las múltiples mediciones de la variable respuesta en función de las covariables y la correlación entre las mismas. La construcción de este tipo de modelos comprende la elección de covariables, la determinación del número de efectos aleatorios y fijos, y la especificación de la estructura de correlación del error aleatorio. Para la selección del modelo "óptimo" existe una amplia gama de criterios de información y predictivos. Los valores que proporcionan los paquetes estadísticos usan para calcularlos la formulación marginal del modelo, priorizando la inferencia sobre los parámetros poblacionales. Sin embargo, algunos autores argumentan que los efectos aleatorios individuales a menudo son de interés e introducen las versiones condicionales de los mismos. El propósito de este trabajo es presentar algunos de estos criterios tanto en su versión marginal como condicional e ilustrar su uso con datos provenientes de la Encuesta Permanente de Hogares. En la aplicación se observa que el desempeño de los mismos es disímil en términos de su comportamiento de elección.

Palabras clave: Datos longitudinales; Criterios marginales y condicionales; Akaike condicional

[*] Docentes- Investigadoras integrantes del Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística. Facultad de Ciencias Económicas y Estadística, Universidad Nacional de Rosario
Contacto: mgarcia@fcecon.unr.edu.ar

INFORMATION AND PREDICTIVE CRITERIA FOR SELECTION OF MIXED LINEAR MODEL.

Abstract. In longitudinal studies the experimental units are observed repeatedly in several occasions. Linear mixed models are an important tool for analyzing this type of data because allows modeling separately the multiple response variable measurements (as a function of the covariates) and the correlation between them. The model-building process includes: the selection of covariates, the definition of the number of random and fixed effects, and the specification of the random error correlation structure. For the selection of the "optimal" model a wide range of information and predictive criteria are available. The statistical packages use the marginal model to calculate them, prioritizing the inference about population parameters. However, some authors argue that individual random effects are often of interest and introduce the conditional versions of them. The purpose of this paper is to introduce some of these criteria considering both the conditional and marginal version and illustrate its use with data from the “Argentina Household Permanent Survey“. In the application it is observed that the performance of the criteria is dissimilar in terms of their choice behavior.

Keywords: Longitudinal data; Marginal and conditional criteria; Conditional Akaike

Original recibido el 15-05-2014.

Aceptado para su publicación el 11-08-2014.

1. Introducción

En los estudios longitudinales las unidades experimentales se observan repetidamente en varias ocasiones. Los modelos lineales mixtos permiten analizar este tipo de datos, modelando, por un lado, la evolución de la respuesta promedio en función del tiempo y otras covariables mediante efectos fijos (estructura media) y, por otro lado, la variabilidad entre las medidas repetidas dentro y entre sujetos por medio del error y los efectos aleatorios (estructura de covariancia). Una etapa importante de la modelación es la determinación de los efectos aleatorios y fijos, como así también de las estructuras de covariancias entre e intra unidad, debido a que una mala especificación del modelo puede tener un impacto considerable sobre las propiedades asintóticas de los estimadores y, por consiguiente, en las inferencias que se realizan sobre ellos. La selección del modelo "óptimo" entre múltiples modelos candidatos representa un desafío importante para interpretar adecuadamente los datos. Esta selección se puede realizar mediante el uso de los criterios de información de Akaike y Schwarz, la prueba del cociente de verosimilitudes y de los criterios basados en las respuestas estimadas, que son menos utilizados. Todos ellos se calculan usando la formulación marginal del modelo. Para tener en cuenta los efectos aleatorios, que a menudo resultan de interés, se considera la formulación condicional del modelo y las versiones condicionales de los criterios de selección. Recientemente, ha surgido una estadística que se construye utilizando este modelo que se denomina Akaike condicional.

En este trabajo se ilustra el uso de estos criterios de información y predictivos en la selección de un modelo que explica el ingreso individual en función de covariables. Los datos se obtuvieron a partir de la información suministrada por la Encuesta Permanente de Hogares, relevada por el Instituto Nacional de Estadística y Censos (INDEC), para diferentes aglomerados en el período 2005-2006.

2. Modelos lineales mixtos

El modelo lineal mixto se puede expresar como:

$$\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\varepsilon}_i \qquad (2.1)$$

siendo,

$\mathbf{Y}_i = Y_{i1}, Y_{i2}, \ldots Y_{in_i}'$, i=1,... N, el vector $(n_i \times 1)$ de medidas repetidas del i-ésimo individuo,

$\mathbf{X}_i$ y $\mathbf{Z}_i$ son matrices conocidas de dimensión $(n_i \times p)$ y $(n_i \times k)$ respectivamente,

$\mathbf{b}_i$ es un vector aleatorio de dimensión (k x 1), cuyas componentes se denominan efectos aleatorios,

$\boldsymbol{\varepsilon}_i$ es un vector $(n_i \times 1)$ que contiene los errores aleatorios (intra-sujeto),

$\boldsymbol{\beta}_i$ un vector $(p \times 1)$ de parámetros.

El número total de observaciones es n y las componentes aleatorias tienen las siguientes distribuciones

$\mathbf{b_i} \overset{iid}{\sim} \mathbf{N_K(0,G)}$, $\boldsymbol{\varepsilon_i} \overset{ind}{\sim} \mathbf{N_{n_i}}(0,\mathbf{R_i})$

Los vectores $\mathbf{b}_i$ y $\boldsymbol{\varepsilon}_i$ son estadísticamente independientes.

Bajo el modelo (2.1) para una realización particular de los efectos aleatorios se obtiene la distribución condicional ($\mathbf{Y}_i$ / $\mathbf{b}_i$), con media $E(\mathbf{Y}_i / \mathbf{b}_i) = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i$, y matriz de covariancias $\mathrm{Var}(\mathbf{Y}_i / \mathbf{b}_i) = \mathbf{R}_i$. Marginalmente, el vector $\mathbf{Y}_i$ está distribuido con media**,** $E(\mathbf{Y}_i) = \mathbf{X}_i \boldsymbol{\beta}$, y matriz de covariancias, cuyos elementos se sintetizan en un vector $\boldsymbol{\theta}$ de parámetros de covariancia, $\mathrm{Var}(\mathbf{Y}_i) = \mathbf{Z}_i \mathbf{G} \mathbf{Z}_i' + \mathbf{R}_i = \mathbf{V}_i(\theta)$.

La estimación de los parámetros $\boldsymbol{\beta}$ y $\boldsymbol{\theta}$ se realiza minimizando la función objetivo, menos dos veces el logaritmo de la función de verosimilitud o de verosimilitud restringida (-2logL), mediante el algoritmo de Newton-Raphson. Condicional a $\mathbf{G}$ y $\mathbf{V}_i(\boldsymbol{\theta})$ el estimador máximo verosímil de $\boldsymbol{\beta}$ es:

$$\hat{\boldsymbol{\beta}} = \left[\sum_{i=1}^{N} \mathbf{X}_i' \left[\hat{\mathbf{V}}_i (\boldsymbol{\theta}) \right]^{-1} \mathbf{X}_i \right]^{-1} \sum_{i=1}^{N} \mathbf{X}_i' \left[\hat{\mathbf{V}}_i (\boldsymbol{\theta}) \right]^{-1} \mathbf{Y}_i$$

y la predicción del vector de efectos aleatorios:

$$\hat{\mathbf{b}} = \hat{\mathbf{G}} \mathbf{Z}_i' \left[\hat{\mathbf{V}}_i (\boldsymbol{\theta}) \right]^{-1} (\mathbf{Y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}})$$

El método de máxima verosimilitud restringida se utiliza para la estimación de los parámetros de covariancia y el de máxima verosimilitud para realizar inferencias respecto a los efectos fijos.

3. Criterios para la selección de modelos

El procedimiento denominado “construcción de un modelo” se basa en la comparación de modelos con diferentes estructuras para las componentes aleatoria y sistemática y, mediante el uso de determinadas medidas, se decide cuál es el más adecuado. Aunque no existe unanimidad acerca de la manera de seleccionar el modelo óptimo, frecuentemente se usa la prueba de la razón de verosimilitudes (TRV) o los criterios de información (CI). Los denominados criterios predictivos (CP), que hacen uso de los valores predichos, son herramientas menos utilizadas. A pesar de la amplia utilización de la prueba de la razón de verosimilitudes, su uso tiene ciertas limitaciones pues está definido para comparar modelos anidados y sólo permite comparar dos al mismo tiempo, por lo que se debe proceder jerárquicamente. Por el contrario, los criterios de información y los predictivos son válidos para comparar y seleccionar modelos anidados y no anidados. Además, permiten la comparación simultánea de un conjunto de modelos.

3.1 Criterios de información

Estos criterios, que se utilizan para seleccionar la estructura de covariancia y media del modelo, son funciones del logaritmo de la verosimilitud (logL) y un término de penalidad basado en el número de parámetros del modelo (s). Entre ellos se encuentran los criterios de Akaike (AIC) (Akaike, 1973) y Bayesiano de Schwarz (BIC) (Schwarz, 1978). Cuyas expresiones vienen dadas por:

$$BIC = -2\ \log(\hat{L}) + s\ \log(N)\infty$$

$$AIC = -2\ \log(\hat{L}) + 2\ s$$

Ambos criterios se diferencian por el valor de la penalización. Se prefieren modelos con valores pequeños de ambos criterios.

Para el cálculo de AIC clásico se utiliza la función de verosimilitud marginal. Esta función surge maximizando una aproximación de la verosimilitud, integrando sobre los efectos aleatorios. Por esta razón se lo denomina AIC marginal (AICm). Si bien este criterio se utiliza frecuentemente como herramienta de selección de modelos,

Greven et al. (2010) muestran que es un estimador sesgado de la información de Akaike bajo el modelo marginal y el sesgo depende de los parámetros de covariancia desconocidos. Si los efectos aleatorios individuales resultan de interés Vaida *et al.* (2005) introducen, como una alternativa al anterior, el AIC condicional (AICc):

$$AIC_c = -2\ \log(\hat{L}) + 2\ \rho$$

Aunque su expresión es similar a la del criterio tradicional, la log(L) corresponde al modelo condicional y el término de penalización es $\rho=tr(H_1)$, donde la matriz H_1 se calcula en función de X, Z y de los elementos de la matriz de covariancia de los efectos aleatorios.

Varios autores trabajaron la expresión del AICc proporcionando una fórmula general, sin embargo el valor de penalidad sólo se puede calcular numéricamente. Greven *et al*. (2010) muestran que este cálculo requiere una alta exigencia computacional, sobre todo para grandes conjuntos de datos, lo que causa problemas cuando se desean tomar decisiones entre varios modelos y obtienen una representación analítica de una versión corregida del AICc, implementándola computacionalmente. De la misma manera que para los criterios marginales se prefieren modelos con valor bajo.

3.2 Criterios predictivos

Los criterios predictivos consideran las respuestas observadas y ajustadas. Se presentan a continuación el Coeficiente R^2, el Coeficiente de concordancia (r_c) y la Suma de cuadrados de errores de predicción (PRESS). Los dos primeros fueron adaptados a los modelos mixtos por Vonesh *et al.* (1996), mientras que el último por Allen (1974).

El *Coeficiente R^2*, similar al coeficiente de determinación de los modelos lineales, provee una medida de la concordancia entre las respuestas observadas y ajustadas, siendo su expresión,

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\sum_{j=1}^{n_i} r_{ij}^2}{\sum_{i=1}^{N}\sum_{j=1}^{n_i} Y_{ij} - \bar{Y}^{\ 2}}$$

donde, $\mathbf{r}_{ij} = Y_{ij} - \hat{Y}_{ij}$ es la diferencia entre la observación y su valor ajustado $\hat{Y}_{ij}$. Se disponen dos versiones de este coeficiente, R^2_m si se utilizan los residuos marginales $r_{ijm} = Y_{ij} - \mathbf{x}'_{ij}\hat{\beta}$ y R^2_c si los residuos utilizados son los condicionales $r_{ijc} = Y_{ij} - \mathbf{x}'_{ij}\hat{\beta} - \mathbf{z}_{ij}\hat{\mathbf{b}}_i$, siendo, $\mathbf{x}'_{ij}$ y $\mathbf{z}'_{ij}$ las filas j-ésimas de las matrices $\mathbf{X_i}$ y $\mathbf{Z_i}$ respectivamente.

El *Coeficiente de concordancia (r_c)* mide la correlación entre los valores observados Y_{ij} y estimados $\hat{Y}_{ij}$.

$$r_c = 1 - \frac{\sum_{i=1}^{N}\sum_{j=1}^{n_i} Y_{ij} - \hat{Y}_{ij}^{\ 2}}{\sum_{i=1}^{N}\sum_{j=1}^{n_i} Y_{ij} - \bar{Y}^{\ 2} + \sum_{i=1}^{N}\sum_{j=1}^{n_i} \hat{Y}_{ij} - \bar{\hat{Y}}^{\ 2} + n \quad \bar{Y} - \bar{\hat{Y}}^{\ 2}}$$

Donde

$$\bar{Y} = \sum_{i=1}^{N}\sum_{j=1}^{n_i}\frac{Y_{ij}}{n}, \quad \bar{\hat{Y}} = \sum_{i=1}^{N}\sum_{j=1}^{n_i}\frac{\hat{Y}_{ij}}{n} \quad \text{y} \quad n = \sum_{i=1}^{N} n_i$$

es el total de observaciones.

Los valores ajustados pueden corresponder a los obtenidos usando tanto el modelo marginal como el condicional, proporcionando el coeficiente de concordancia marginal (r_{cm}) y el condicional (r_{cc}).

Debido a que ambos coeficientes (R^2 Y r_c) aumentan a medida que se sobreparametrizan los modelos, sus valores se pueden ajustar por el número de parámetros, para ambas versiones, de la siguiente manera:

$$R^2_a = 1 - k(1 - R^2)$$
$$r_{c,a} = 1 - k(1 - r_c)$$

siendo, k=N/(N-d) y d el número de parámetros de efectos fijos y covariancias. Los valores de R^2_a y $r_{c,a}$ pueden ser menores que 0. Se prefieren modelos con valores altos de estos criterios.

La *Suma de cuadrados de errores de predicción (PRESS)* se define como la suma de los cuadrados de las diferencias entre los valores observados y predichos, siendo este último obtenido omitiendo las observaciones de a una por vez. Su expresión es,

$$PRESS = \sum_{i=1}^{N} \sum_{j=1}^{n_i} Y_{ij} - \hat{Y}_{ij\,(-ij)}^{\;2}$$

donde, $\hat{Y}_{ij\,(-ij)}$ es el valor ajustado que se obtiene al haber omitido del conjunto de datos esa observación. Los valores de esta estadística difieren si se usan valores ajustados marginales ($PRESS_m$) o condicionales ($PRESS_c$). Los modelos seleccionados deben presentar un valor bajo de este criterio.

4. Resultados

En este trabajo se realiza una aplicación de los criterios presentados utilizando datos provenientes de la Encuesta Permanente de Hogares (EPH), relevada por el Instituto Nacional de Estadística y Censos (INDEC). La información relativa a los diferentes aglomerados que surge mediante esa encuesta ofrece importantes ventajas para el análisis empírico de variables registradas en la misma. Se considera la modelación de la variable ingreso de 102 individuos de tres aglomerados diferentes (Rosario, Córdoba y Gran Buenos Aires) que permanecieron en la encuesta en el período comprendido entre el primer trimestre de 2005 hasta el cuarto trimestre de 2006. En este período, los individuos fueron relevados en 4 oportunidades identificadas como "onda" y se registró el ingreso de los mismos en cada una de ellas. Además, se consideraron las siguientes variables: nivel máximo de educación alcanzado (primaria completa, secundaria completa, universitario completo), sexo (masculino, femenino) y edad. La representación gráfica de la evolución promedio de los ingresos individuales durante el período para cada uno de los tres aglomerados se presenta en el Gráfico 1.

De acuerdo a la tendencia evidenciada a partir de los datos (Gráfico 1), se propuso un modelo mixto cuadrático con tres efectos aleatorios y matriz de covariancias no estructurada para estos efectos, cuya expresión es,

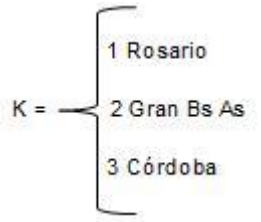
$$Y_{ij} = \beta_{0i} + \beta_{1i}\, t_j + \beta_{2i}\, t_j^2 + \varepsilon_{ij}\,, \quad (4.1)$$

donde,

$$\beta_{0i} = \beta_{0k} + \beta_{0k1} N_{1i} + \beta_{0k2} N_{2i} + \beta_{0k3} S_i + \beta_{0k4} E_i + b_{0i}$$

$$\beta_{1i} = \beta_{1k} + \beta_{1k1} N_{1i} + \beta_{1k2} N_{2i} + \beta_{1k3} S_i + \beta_{1k4} E_i + b_{1i}$$

$$\beta_{2i} = \beta_{2k} + \beta_{2k1} N_{1i} + \beta_{2k2} N_{2i} + \beta_{2k3} S_i + \beta_{2k4} E_i + b_{2i}$$



siendo,

Y_{ij}: logaritmo natural del ingreso del i-ésimo individuo en la onda j-ésima i=1,2,..,102; j=1,2,3,4

N_{1i}: = 1 si el individuo i-ésimo tiene nivel primario completo; = 0 en otro caso

N_{2i}: = 1 si el individuo i-ésimo tiene nivel secundario completo; = 0 en otro caso

S_i: sexo del i-ésimo individuo (1 masculino; 0 femenino)

E_i: edad del i-ésimo individuo

t_j: j-ésima ocasión de medición del ingreso correspondiente a la onda j, j=1,2,3,4

ε_{ij} : error aleatorio intra individuo. El vector que contiene estos errores tiene distribución $\boldsymbol{\varepsilon}_i \overset{ind}{\sim} N_{n_i}(\mathbf{0}, \mathbf{R}_i)$

$\mathbf{b}_i = \begin{pmatrix} b_{0i} \\ b_{1i} \\ b_{2i} \end{pmatrix}$: vector de efectos aleatorios del i-ésimo individuo, con distribución $\mathbf{b}_i \overset{iid}{\sim} N_k(\mathbf{0}, \mathbf{G})$

b_{0i} efecto aleatorio asociado a la ordenada (β_{0i})

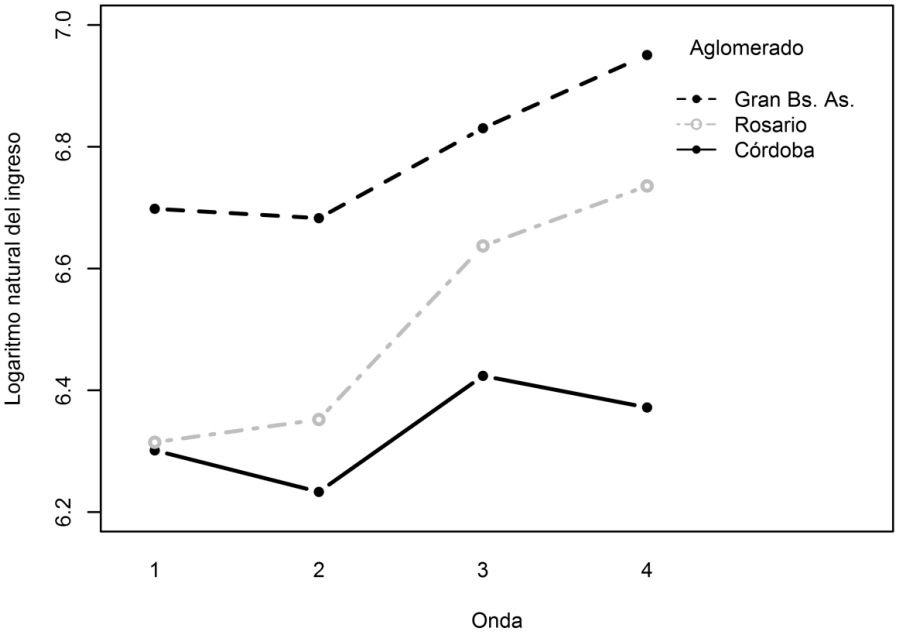
b_{1i} efecto aleatorio asociado al efecto lineal (β_{1i}) denominado efecto aleatorio onda,

b_{2i} efecto aleatorio asociado al efecto cuadrático (β_{2i}) denominado efecto aleatorio onda^2,

$$\mathbf{G} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{bmatrix}$$

La selección del modelo final, es decir, un modelo que aproxime el verdadero mecanismo que generan los datos se realiza a partir de un conjunto de modelos candidatos.

Gráfico 1. Perfiles promedio por aglomerado



Fuente: Elaboración propia con datos del INDEC.

En base a un estudio exploratorio previo se considera como modelo completo al Modelo (4.1), que incluye efectos aleatorios para la ordenada, onda y onda cuadrado y como efectos fijos a aglomerado, nivel, sexo, edad, las interacciones entre aglomerado y sexo y entre aglomerado y nivel. Los modelos reducidos surgen de eliminar algunos de los efectos anteriores. La selección entre esos modelos se realiza usando los criterios mencionados.

El ajuste de las curvas se realiza con el paquete estadístico SAS mientras que los criterios predictivos se calculan utilizando el procedimiento IML de SAS (SAS Institute Inc., 2004) y el paquete estadístico R.

El proceso de selección se efectúa en varios pasos. En el primer paso se selecciona la matriz de covariancia intra individuo. Para ello se plantean diferentes estructuras (Tabla 1) y mediante los criterios de información se selecciona la más adecuada.

La siguiente tabla (Tabla 2*)* presenta los valores de los criterios de información obtenidos al utilizar diferentes estructuras para la matriz R del modelo completo*.*

Tabla 1. Diferentes estructuras de covariancias

Estructura	**Componentes**
Independencia	$\mathrm{Var}(\varepsilon_{ij}) = \sigma^2 \quad \mathrm{Cov}(\varepsilon_{ij}, \varepsilon_{ij'}) = 0$
Diagonal	$\mathrm{Var}(\varepsilon_{ij}) = \sigma_j^2 \quad \mathrm{Cov}(\varepsilon_{ij}, \varepsilon_{ij'}) = 0$
Simetría compuesta	$\mathrm{Var}(\varepsilon_{ij}) = \sigma^2 \quad \mathrm{Cov}(\varepsilon_{ij}, \varepsilon_{ij'}) = \rho$
Simetría compuesta heterogéna	$\mathrm{Var}(\varepsilon_{ij}) = \sigma_j^2 \quad \mathrm{Cov}(\varepsilon_{ij}, \varepsilon_{ij'}) = \rho \sigma_j \sigma_{j'}$
Exponencial	$\mathrm{Var}(\varepsilon_{ij}) = \sigma^2 \quad \mathrm{Cov}(\varepsilon_{ij}, \varepsilon_{ij'}) = \rho^{d_{jj'}} \; d_{jj'} = \lvert t_{ij} - t_{ij'} \rvert$

Fuente: Elaboración propia

Según los valores marginales de los criterios de información se elige el modelo 1 (M1). Se descartan los modelos M2 y M4 porque presentan problemas de convergencia. Para todos los modelos el Akaike condicional es de la misma magnitud, AICc =43683,1.

El segundo paso consiste en la determinación de los parámetros del modelo que contendrán efectos aleatorios y posteriormente, si corresponde, la elección del tipo de estructura para la matriz de covariancias de los mismos. Se parte del modelo M1 seleccionado en el paso anterior, que posee efectos aleatorios en todos sus parámetros, y se plantean modelos anidados obtenidos al eliminar

algunos efectos aleatorios. Para comparar el ajuste de los mismos se usan los criterios de información de Akaike marginal y condicional.

Tabla 2. **Criterios de información marginales para la selección de matriz de covariancias del modelo.**

Modelo	R	AICm	BICm
M1	Independencia	841,4	859,8
M2	Diagonal	Nc*	Nc*
M3	Simetría Compuesta	843,4	864,4
M4	Simetría Compuesta heterogénea	Nc”	Nc*
M5	Markov	843,0	864,0

Fuente: Elaboración propia con datos del INDEC.

Nota: (*) Nc indica que el procedimiento no converge.

Tabla 3. **Criterios de Akaike marginal y condicional para la selección de los efectos aleatorios**

Modelo	Efectos aleatorios	AICm	AICc
M1	Ordenada-onda-onda^2	841,4	43683,1
M6	Ordenada	839,6	672,2
M7	onda-onda^2	861,8	13276,5
M8	Onda	903,5	780,6
M9	Ordenada-onda	839,1	3322,6

Fuente: Elaboración propia con datos del INDEC.

Los resultados de la Tabla 3 conducen a seleccionar un modelo con efecto aleatorio sólo en la ordenada (M6). Cabe mencionar que, a diferencia del AICc, el AICm no permite diferenciar entre los modelos M6 y M9. Se considera al modelo M6, que posee sólo un

efecto aleatorio en la ordenada y una estructura de independencia para la matriz R_i, como el modelo completo a partir del cual se seleccionan los efectos fijos.

En los dos pasos anteriores, el modelo especificado incluye todas las covariables de interés. En el último paso se plantean modelos anidados en el modelo M6 que surgen de omitir algunas covariables. La comparación entre estos modelos se realiza mediante los coeficientes R^2 y de concordancia, la estadística PRESS y las dos versiones de AIC, siguiendo la sugerencia de Greven et al. (2010).

La tabla 4 muestra un resumen de algunos de los modelos evaluados en el trabajo. Los valores más pequeños del AICc corresponden a los modelos 12 y 14, siendo muy similares, resultando difícil la elección de uno de ellos. El valor del AICm más chico lleva a optar por el modelo 12. Con la estadística PRESS ocurre una situación similar, su versión marginal sugiere optar por el modelo 12 y su versión condicional por los modelos 12 y 14.

Los valores condicionales de R^2 y r_c (R_c^2 y r_{cc}) son tan similares que no permiten tomar una decisión sobre el modelo a elegir. Si se consideran los valores marginales, R_m^2 selecciona el modelo 12 y r_{cm} opta por lo modelos 10 ó 12.

En resumen, los diversos criterios conducen a conclusiones contradictorias acerca de cuál es el modelo a seleccionar. La mayoría de los criterios, sobre todo las versiones marginales, indican que el modelo 12 se debe elegir, sin embargo las versiones condicionales de AICc y PRESS seleccionaron el modelo 14, que tiene menos parámetros.

5. Consideraciones finales

En este trabajo se utilizan criterios que permiten seleccionar un modelo entre varios candidatos. El criterio de Akaike condicional se usa conjuntamente con los criterios de información tradicionales y los criterios predictivos, condicionales y marginales, con el fin de evaluar su desempeño.

Los valores de R2, rc y PRESS permiten comparar modelos independientemente que estén anidados o no, eligiendo el modelo que tenga el de mayor magnitud. Tradicionalmente para la selección de las estructuras de covariancia de los errores y los

efectos aleatorios a incluir en el modelo se utiliza el AICm. A partir de los trabajos de Vaida *el al*. (2005) y Greven *et al.* (2010) surge la recomendación de considerar la versión condicional del AIC si los efectos aleatorios resultan de interés.

Tabla 4. Criterios para la selección de efectos fijos

Modelo	Efectos fijos	AICm	AICc	PRESSm	R^2_m	R^2_c	r_{cm}	r_{cc}
M6	Completo	839,6	672,15	218,29	0,282	0,748	0,477	0,849
M10	Sin edad	809,2	668,22	217,01	0,287	0,749	0,479	0,849
M11	Sin edad, aglomerado y sexo	793,0	653,53	245,72	0,189	0,758	0,330	0,852
M12	Sin edad y aglomerado	778,5	651,86	211,31	0,306	0,756	0,478	0,851
M13	Sin aglomerado	809,5	655,93	213,39	0,301	0,755	0,475	0,851
M14	Sin edad, aglomerado y nivel	794,4	651,85	260,89	0,138	0,759	0,253	0,853

Fuente: Elaboración propia con datos del INDEC

Para mostrar el comportamiento de los diferentes criterios de información y predictivos se utilizaron datos de la EPH elaborada por el INDEC con el fin de modelar el ingreso de un individuo (en términos de logaritmo).

En la aplicación se observa que, para la selección de la matriz de covariancia de los errores, los criterios de información marginales utilizados (AICm y BICm) conducen a la elección del mismo modelo. El comportamiento del AICc en este sentido no se puede evaluar debido a que su magnitud fue la misma para todos los modelos.

Con respecto a la selección de efectos aleatorios, el criterio de información AICc resulta de gran utilidad comparado con el AICm,

ya que la magnitud del AICm no permite discriminar entre la elección de algunos modelos.

Para la selección de los efectos fijos (o las covariables a incluir en el modelo) la versión condicional del AIC no permite diferenciar entre los modelos propuestos tan claramente como la versión marginal. Considerando los criterios predictivos, las versiones condicionales de R^2, r_c y PRESS se ven poco afectados por la sub o la sobreparametrización del modelo, en cambio, las versiones marginales sí se encuentran afectadas, siendo de mayor utilidad a la hora de seleccionar un modelo.

La aplicación considerada no permite evaluar si algún criterio es consistentemente mejor que los otros en términos de su comportamiento de selección.

Como sugerencia para una continuidad del presente trabajo se podrían utilizar simulaciones para indagar el funcionamiento del AICc y de las versiones condicionales de los criterios predictivos en la selección de los efectos fijos.

Referencias biobliográficas

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. En: B. N. Petrov y F. Csaki (Eds.). *Second International Symposium on Information Theory* (pp. 267–281). Budapest: Akademiai Kiado.

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method of prediction. *Technometrics,* 16, 125–127.

Greven, S. y Kneib, T. (2010). On the behavior of marginal and conditional AIC in linear mixed models. *Biometrika,* 97, 773–789.

Hodges, J. S. y Sargent, D. J. (2001). Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika* 88, 367–379.

Lin, L. I.-K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics,* 45, 255–268.

Orelien, J.G. y Edwards, L.K. (2007). Fixed-effect variable selection in linear mixed models using R^2 statistics. *Computational Statistics and Data Analysis,* 52**,** 1896–1907.

Pinheiro, J. y Bates, D. (2004). *Mixed-Effects Models in S and SPLUS.* New York: Springer.

R Development Core Team (2006). R*: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. Recuperado de http://www.R-project.org.

SAS Institute Inc. 2004. SAS/STAT Software: Versión 9.1. Cary, NC: SAS Institute Inc.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics,* 6, 461–464.

Vaida, F. y Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika,* 92, 351–370.

Vonesh, E. F., Chinchilli, V. M.y Pu, K. (1996). Goodness-of-fit in generalized nonlinear mixed-effects models. *Biometrics*, 52, 575–587.

Wang, J. y Schaalje, G. B. (2009). Model selection for linear mixed models using predictive criteria. Communications in Statistics. *Simulation and Computation,* 38, 788–801.