Artículo Original
Revista Varianza - Año 2021 Nro. 18
ISSN 2789-3529 versión en línea
Páginas: 1-7

# MODELOS MARGINALES APLICACIÓN A DATOS DE PANEL

# MARGINAL MODELS APPLICATION TO PANEL DATA

Ramiro Coa Clemente[1]
Instituto de Estadística Teórica y Aplicada, Universidad Mayor de San Andrés, La Paz - Bolivia
✉ *clementecoa@gmail.com*

Artículo recibido: 2021-08-15
Artículo aceptado: 2021-09-07

## RESUMEN

El objetivo de este artículo es aplicar el modelo marginal en el análisis de datos tipo panel sobre la situación nutricional de los recién nacidos. Luego de examinar sucintamente los aspectos centrales de los modelos marginales, se revisa brevemente el método de ecuaciones de estimación generalizada (EEG), un método apropiado para la estimación de este tipo de modelos. Con base en un modelo marginal logístico con patrón de correlación intercambiable, se concluye que fumar durante el embarazo y un servicio prenatal inadecuado incrementan significativamente la probabilidad de un nacimiento con bajo peso al nacer.

**Palabras clave:** *Modelo marginal, ecuaciones de estimación generalizada, datos tipo panel*

## ABSTRACT

The aim of this article is to apply the marginal model in the analysis of panel data on the nutritional status of newborns. After a brief review of the central aspects of marginal models, the generalized estimating equations (GEE) method, an appropriate method for estimating this type of model, is briefly reviewed. Based on a logistic marginal model with exchangeable correlation pattern, it is concluded that smoking during pregnancy and inadequate prenatal care significantly increase the probability of a low birth weight.

**Keywords:** *Marginal model, generalized estimating equations, panel data*

## INTRODUCCIÓN

Los modelos lineales generalizados constituyen una clase unificada de modelos para análisis de regresión con respuesta discreta o continua y observaciones independientes. No es posible, sin embargo, la aplicación directa de estos modelos a datos tipo panel debido a la correlación entre las observaciones obtenidas en las mismas unidades. Por tal razón, se consideran extensiones para datos panel. Hay muchas formas de extender los modelos lineales generalizados a fin de tomar en cuenta la correlación entre las observaciones. En este trabajo se revisa y aplica uno de ellos: los modelos marginales.

Los modelos marginales constituyen una metodología para analizar datos de panel cuando la variable respuesta es discreta o continua. Esta metodología no

[1] Ex-Director de Investigación en la Unidad de Análisis y Política Social de Bolivia (UDAPSO). Ex-Director Nacional de la Encuesta de Demografía y Salud, en 1989 y 1998. M.Sc. Estadística, Pontificia Católica de Chile. Candidato a Doctor en Demografía. Docente investigador de la carrera de Estadística, Universidad Mayor de San Andrés. ORCID: 0000-0002-2955-0204

requiere supuestos distribucionales para las observaciones. El método depende solamente del supuesto de cómo la respuesta media está relacionada con las covariables. La evasión de supuestos distribucionales conduce al método de estimación conocido como ecuaciones de estimación generalizada, EEG. El enfoque de EEG es una buena alternativa a la estimación de máxima verosimilitud.

## MÉTODO

### El modelo

Para el análisis de los datos tipo panel sobre la situación nutricional de los recién nacidos se usa el modelo marginal. Según Fitzmaurice et al. (2011), el modelo marginal para datos de panel es especificado en tres partes:

i. La esperanza o media condicional de cada respuesta depende de las covariables a través de una función de enlace $g$ conocida

$$g(\mu_{it}) = \eta_{it} = X'_{it}\beta$$

ii. La varianza condicional de la respuesta en cada ocasión es función de la media

$$Var(Y_{it}/X_{it}) = \phi\, h(\mu_{it})$$

donde $h(\mu_{it})$ es una función-varianza conocida y $\phi$ es un parámetro de escala.

iii. Se asume que existe asociación entre pares de observaciones dentro de los *clusters*. A partir de un modelo para las correlaciones por pares, la correspondiente matriz de varianzas-covarianzas de trabajo se construye como:

$$V_i = A_i^{1/2}\, Corr(Y_i)\, A_i^{1/2}$$

donde $A_i$ es una matriz diagonal con $Var(Y_{it}/X_{it}) = \phi h(\mu_{it})$ a través de su diagonal y $Corr(Y_i)$ representa la matriz de correlación.

Es importante resaltar algunas características particulares de estos modelos. Los modelos marginales son una forma muy natural de extender los modelos lineales generalizados para tratar respuestas tipo panel correlacionadas y permiten realizar inferencias sobre medias poblacionales. La palabra marginal se usa para enfatizar que el modelo para la respuesta media en cada ocasión no depende de efectos aleatorios. En estos modelos se asume que las respuestas para los distintos *clusters* son independientes entre sí, pero, las medidas repetidas para el mismo *cluster* no son independientes. No requieren supuestos distribucionales para la respuesta, esto porque no hay una especificación completa de la distribución multivariada conjunta para respuestas discretas. La anulación de supuestos distribucionales conduce al método de estimación conocido como ecuaciones de estimación generalizada, una buena alternativa a la estimación de máxima verosimilitud.

Para el caso particular de una respuesta binaria, la especificación completa del modelo marginal, de acuerdo a Fitzmaurice et al. (2011), es expresada como:

i. $ln(\mu_{it}/(1-\mu_{it})) = \eta_{it} = X'_{it}\beta$

ii. $Var(Y_{it}/X_{it}) = \mu_{it}(1-\mu_{it})$

iii. $ln[\,RC(Y_{it}, Y_{it'})/(X_{it}, X_{it'})] = \alpha_{tt'}$

donde la razón de chances entre dos respuestas para los momentos $t$ y $t'$ es definido como:

$$RC(Y_t, Y_{t'}) = \frac{Pr(Y_t = 1,\ Y_{t'} = 1)Pr(Y_t = 0,\ Y_{t'} = 0)}{Pr(Y_t = 1,\ Y_{t'} = 0)Pr(Y_t = 0,\ Y_{t'} = 1)}$$

## Las ecuaciones de estimación generalizada

Cuando las variables respuesta son discretas, no es posible una especificación conveniente de su distribución mutivariante conjunta. Por este hecho, para los modelos marginales se requiere un método alternativo de estimación, alternativo al de máxima verosimilitud. El enfoque de ecuaciones de estimación generalizado representa esta alternativa. Este método proporciona un enfoque muy general y unificado para analizar respuestas correlacionadas que pueden ser discretas o continuas. La idea esencial detrás del enfoque de EEG es generalizar y extender las habituales ecuaciones de verosimilitud para un modelo lineal generalizado incorporando la matriz de varianzas-covarianzas del vector de respuestas. Inicialmente esta idea fue introducida por Wedderburn (1974). En el artículo de Wedderburn se asume independencia entre las observaciones y que la forma de la función varianza es una función conocida de la media sin la exigencia formal de que ellos se originen a partir de una distribución específica. Este es un supuesto menos fuerte que el proveniente de una distribución específica. En consecuencia, uno es libre de elegir cualquier parametrización de las funciones media y varianza, y aplicarlos en la ecuación de estimación.

El término ecuación de estimación generalizada indica que una ecuación de estimación no es el resultado de una derivación basada en la verosimilitud, es obtenido por la generalización de otra ecuación de estimación. La modificación que se hace para obtener una ecuación de estimación generalizada es introducir componentes de varianza de segundo orden directamente en la ecuación de estimación.

Para datos tipo panel, la ecuación de estimación de máxima quasiverosimilitud (MQV) para un modelo lineal generalizado, según Hardin and Hilbe (2013), es expresada como:

$$\Psi(\beta) = \sum_{i=1}^{n} X_{ji}'\ D_i\ V_i^{-1}\ \left(\frac{Y_i - \mu_i}{a(\phi)}\right) = 0$$

donde:

$$D_i = diag\left(\frac{\partial \mu_{it}}{\partial \eta_{it}}\right) \quad t = 1, \ldots, T_i$$

$$V_i = diag\left[h^{1/2}(\mu_{it})\right]\ I_{T_i * T_i}\ diag\left[h^{1/2}(\mu_{it})\right]$$

Notar que $V_i$ es matriz diagonal, por lo que esta ecuación de estimación trata a las observaciones dentro de cada *cluster* como independientes.

Liang and Zeger (1986) propusieron una EEG que es una modificación de la ecuación de estimación de MQV. La modificación consiste en remplazar la matriz identidad con una matriz de correlación más general

$$\mathrm{V_i} = \mathrm{diag}\left[\mathrm{h}^{1/2}(\mu_{it})\right]\ \mathrm{R}(\alpha)_{\mathrm{T_i * T_i}}\ \mathrm{diag}\left[\mathrm{h}^{1/2}(\mu_{it})\right.$$

Adicionalmente, también se tiene una ecuación de estimación para los parámetros auxiliares α, la cual se la expresa como (Hardin and Hilbe, 2013).

$$\Psi(\alpha) = \sum_{\mathrm{i=1}}^{\mathrm{n}} \left(\frac{\partial \varepsilon_i}{\partial \alpha}\right)' \mathrm{H_i^{-1}}(\mathrm{W_i} - \varepsilon_i) = 0_{\mathrm{q*1}}$$

donde $q = \binom{T_i}{2}$, $W_i$ y $\varepsilon_i$ son vectores de dimensión $q*1$, $H_i$ es una matriz diagonal $q*q$, definidos como:

$$W_i = \left[r_{i1}r_{i2}, r_{i1}r_{i3}, \ldots, r_{iT_i-1}r_{iT_i}\right]'$$

$$H_i = Diag\ [Var\ (W_{it})]$$

$$\varepsilon_i = E(W_i)$$

y $r_{it}$ es el *it-ésimo* residuo de Pearson.

Combinando las ecuaciones de estimación para los parámetros de la regresión y para los parámetros auxiliares, la completa EEG para modelos marginales está dada por:

$$\Psi(\beta,\alpha)=\begin{bmatrix}\sum_{i=1}^{n} X_{ji}'\ D_i\ V_i^{-1}\ \left(\frac{Y_i-\mu_i}{a(\phi)}\right)\\ \sum_{i=1}^{n}\left(\frac{d\varepsilon_i}{d\alpha}\right)'\ H_i^{-1}(W_i-\varepsilon_i)\end{bmatrix}$$

donde:

$$V_i=\mathrm{D}[h^{1/2}(\mu_{it})]\ \mathrm{R}(\alpha)\ \mathrm{D}[h^{1/2}(\mu_{it})]$$

En cada paso primero se estima R($\alpha$) y luego se lo usa para estimar $\beta$. Se declara convergencia cuando el cambio en las estimaciones de los parámetros es menor a algún valor o vector de valores pre-definidos, o cuando el cambio en la suma de los cuadrados de las devianzas es inferior a un valor pre-determinado.

## Estructuras para la matriz de correlación

Hardin and Hilbe (2013) plantean varias estructuras estándar para la estimación de la correlación dentro de los *clusters*, algunas de esas estructuras son:

## Correlación intercambiable

Como una extensión simple a la estructura independiente se puede lanzar la hipótesis de que las observaciones dentro de un panel tienen una correlación común. En este caso, la matriz de correlación tiene la siguiente estructura

$$R_{tt'}=\begin{cases}1 & si\ \ t=t'\\ \alpha & si\ \ t\neq t'\end{cases}$$

## Correlación autorregresiva

Si las observaciones repetidas dentro de los paneles tienen un orden natural puede ser más razonable asumir una dependencia del tiempo para la asociación. En este caso, la matriz de correlación es la estructura de correlación autoregresiva de orden *k*. En particular, para un AR(1) se tiene:

$$R_{tt'}=\begin{cases}1 & si\ \ t=t'\\ \alpha^{|t-t'|} & si\ \ t\neq t';\end{cases}$$

## Correlación estacionaria

Como una alternativa a la hipótesis de autocorrelación, se puede postular que existen las correlaciones para algún número pequeño de unidades de tiempo. En esta hipótesis se especifica una diferencia de tiempo máxima para la cual las observaciones pueden estar correlacionadas de modo que la matriz de correlación esté acotada. En este caso, α es un vector de correlaciones de hasta *k* rezagos y la matriz de correlación puede ser descrita como:

$$R_{tt'}=\begin{cases}\alpha_{|t-t'|} & si\ \ |t-t'|\leq k\\ 0 & en\ otro\ caso\end{cases}$$

# APLICACIÓN

El propósito de la aplicación es determinar si el hecho de fumar cigarrillo durante el embarazo y la calidad del control prenatal afectan significativamente la probabilidad de que el nacimiento resulte con bajo peso al nacer. Otras variables incluidas en el análisis son la edad, educación y estado civil de la madre, además del momento en el que se realiza la atención prenatal y el sexo del

nacido. Se cuenta con un total 648 madres y 3 nacimientos por cada madre.

Dado que $Y_{it}$ es una respuesta binaria que toma valores de 1 (bajo peso al nacer) y 0 (no bajo peso), interesa relacionar los cambios en la $E(Y_{it} / X_{it})$ con los cambios en las covariables. El modelo marginal especificado es un modelo de regresión logística con un patrón de asociación intercambiable. Los resultados se presentan en el siguiente cuadro.

**Cuadro No. 1**
**Regresión logística con patrón de correlación intercambiable**

| *Bajo Peso al nacer* | *Razón de chances* | *Error estándar robusto* | *P > z* | *Intervalo de confianza del 95%* | |
|---|---|---|---|---|---|
| *Fumó* | *2,77* | *0,95* | *0,003* | *1,41* | *5,43* |
| *Sexo del nacido* | *0,67* | *0,19* | *0,149* | *0,39* | *1,15* |
| *Edad de la madre* | *0,95* | *0,04* | *0,201* | *0,88* | *1,03* |
| *Educación de la madre* | *0,90* | *0,07* | *0,220* | *0,77* | *1,06* |
| *Madre casada* | *0,55* | *0,22* | *0,127* | *0,25* | *1,19* |
| *Prenatal de calidad intermedia* | *2,50* | *0,93* | *0,014* | *1,20* | *5,19* |
| *Prenatal de calidad inadecuada* | *5,49* | *2,66* | *0,000* | *2,13* | *14,17* |
| *Sin control prenatal* | *0,83* | *0,59* | *0,797* | *0,20* | *3,38* |
| *1er C. prenatal en 2do trimestre* | *0,68* | *0,27* | *0,331* | *0,31* | *1,49* |
| *1er C. prenatal en 3er trimestre* | *0,07* | *0,07* | *0,014* | *0,01* | *0,58* |
| *Constante* | *0,60* | *0,53* | *0,568* | *0,11* | *3,42* |

Fuente: Elaboración propia

Claramente el efecto de fumar sobre el bajo peso al nacer es altamente significativo (valor - p = 0,003). Cuando la madre fuma, la chance de tener bajo peso al nacer es 2,8 veces más que cuando la madre no fuma. La educación de la madre no tiene un efecto significativo (valor - p = 0,220) sobre la probabilidad de nacer con bajo peso, sin embargo, el control prenatal inadecuado tiene un efecto altamente significativo sobre la probabilidad de nacer con bajo peso (valor - p = 0,000).

## DISCUSIÓN

Muchos estudios se han realizado a fin de determinar el efecto de fumar durante el embarazo sobre el peso al nacer, entre los cuales se encuentran los trabajos de Alonso et al. (2005) y Caraballoso (1999). En ambos trabajos se concluye que fumar durante el periodo de gestación afecta negativamente el peso del recién nacido. En el trabajo de Alonso et al. (2005) se aplican dos modelos, por una parte, un modelo de regresión lineal para explicar el peso de los recién nacidos en función de la condición de fumar de la madre y de su pareja, además de la edad gestacional y, por otra parte, se aplica una regresión logística con las mismas variables explicativas. En el estudio de Caraballoso (1999) también se aplicó un modelo de regresión logística con una serie de variables explicativas, entre las que se consideró, además del hábito de fumar durante el embarazo, variables como las vinculadas al alcoholismo, el deseo del embarazo y los antecedentes de hijos con bajo peso al nacer.

Hay, sin embargo, dos diferencias metodológicas entre los dos trabajos citados y el presente estudio. Primero, en este estudio se usan datos de panel, los cuales son apropiados para un control más efectivo de los otros factores que podrían afectar el peso al nacer, como el consumo de alcohol durante el embarazo y el estado nutricional de la madre. Segundo, el uso de un modelo marginal, un modelo apropiado para el análisis de datos tipo panel.

Si bien existen diferencias metodológicas para abordar un mismo objetivo – explicar el efecto de consumo de tabaco sobre el peso al nacer - es claro que las conclusiones a las que se arriban en los distintos trabajos son similares. Este y los otros estudios muestran claramente el efecto significativo del consumo de tabaco durante el embarazo sobre el bajo peso al nacer. Sin duda que estos hallazgos son útiles para fines de políticas públicas vinculadas a la salud y nutrición de los infantes.

## CONCLUSIÓN

El modelo marginal es bastante flexible en el sentido que no es necesario especificar una distribución de probabilidad conjunta para las respuestas. Permite hacer inferencias sobre las medias poblacionales, pero, requiere de un método apropiado de estimación de los parámetros, el denominado método de ecuaciones de estimación generalizada. Este método está basado en el concepto de ecuaciones de estimación y proporciona un enfoque muy general y unificado para analizar respuestas correlacionadas, una característica muy frecuente en los datos tipo panel.

Su aplicación permitió evidenciar estadísticamente que fumar durante el embarazo y un control prenatal inadecuado están asociados con un bajo peso al nacer, luego de controlar el efecto de otras variables como la edad y la educación de la madre. Será importante ver, en un posterior trabajo de investigación, si el uso de otros métodos de análisis de datos de panel conduce a las mismas conclusiones.

## REFERENCIAS BIBLIOGRÁFICAS

Agresti, A. (2002). Categorical data análisis (2 ed.). John Wiley & Sons Inc.

Alonso, A., Cano, J., Girón, A., Yep, G. y Sánchez, M. (2005). Peso al nacimiento y tabaquismo familiar. Asociación Española de Pediatria, vol. 63, No. 2, pp. 116-119.

AndreB, H., Golsch, K. and Schmidt, A. (2013). Applied panel data analysis for economic and social surveys. Springer.

Baltagi, B. (2013). Econometric analysis of panel data (5 ed.). John Wiley & Sons.

Banerjee, M. and Frees, E. (1997). Influence diagnostics for linear longitudinal models. Journal of the American Statistical Association vol. 92, pp. 999-1005.

Bickel, P. and Doksum, K. (1977). Mathematical Statistics. Holden-Day.

Bijleveld, C. and Van der Kamp, L.(1998). Longitudinal data analysis: Designs, models and methods. Sage Publications.

Caraballoso, M. (1999). Bajo peso al nacer y tabaquismo. Revista Cubana de Salud Pública vol. 25, No. 1.

Diggle, P., Heagerty, P., Liang, K. and Zeger, S. (2002). Analysis of longitudinal data (2 ed.). Oxford University Press.

Fitzmaurice, G., Laird, N. and Ware, J. (2011). Applied longitudinal análisis (2 ed.). John Wiley & Sons.

Frees, E. (2004). Longitudinal and Panel Data: Analysis and applications for the social sciences. Cambridge University Press.

Gosho, M., Hamada, C. and Yoshimura, I. (2011). Criterion for the selection of a working correlation structure in the generalized estimating equation: Approach for longitudinal balanced data. Communications in Statistics-Theory and Methods No. 40(21), pp. 3839-3856.

Gosho, M. (2014). Criteria to select a working correlation structure for the generalized estimating equations method in SAS. Journal of Statistical Software, vol. 57.

Hardin, J. and Hilbe, J. (2013). Generalized estimating equations (2 ed.). Chapman & Hall.

Hsiao, Cheng. (2003). Analysis of panel data (2 ed.). Cambridge University Press.

Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. Biometrika No. 1, vol. 73, pp. 13-22.

McCullagh, P. and Nelder, J. (1989). Generalized linear models (2 ed.). Chapman & Hall.

Skrondal, A. and Rabe-Hesketh, S. (2004). Generalized latent variable modeling: Multilevel, longitudinal and structural equation models. Chapman & Hall.

Wedderburn, R. (1974). Quasi-likelihood functions, generalized linear models and the gauss-newton method. Biometrika No 61, vol. 3, pp. 439.

Yan, J. and Fine, J. (2004). Estimating equations for association structures. Statistics in Medicine, vol. 23, pp. 859-880.