REGRESIÓN LOGÍSTICA CON INTERCEPTOS ALEATORIOS. APLICACIÓN A DATOS DE PANEL

Dr(c) Ramiro Coa Clemente *
✉ clementecoa@gmail.com

RESUMEN

En este artículo se presenta sucintamente el modelo lineal generalizado de efectos mixtos, un modelo de mucha utilidad para abordar el análisis estadístico en profundidad, en diferentes campos. Un caso particular de este modelo es el denominado regresión logística con interceptos aleatorios, un modelo alternativo para el análisis de datos de panel. Se ilustra su aplicación en el ámbito de la nutrición. El propósito es determinar si fumar durante el embarazo afecta o no el bajo peso al nacer. Los resultados sugieren un efecto muy significativo del consumo de tabaco durante el embarazo sobre el bajo peso al nacer.

PALABRAS CLAVE

Efectos mixtos, Interceptos aleatorios, Datos de panel

ABSTRACT

This article succinctly presents the generalized linear model of mixed effects, a very useful model to address in-depth statistical analysis in different fields. A particular case of this model is the so-called logistic regression with random intercepts, an alternative model for the analysis of panel data. Its application in the field of nutrition is illustrated. The purpose is to determine whether or not smoking during pregnancy affects low birth weight. The results suggest a very significant effect of tobacco use during pregnancy on low birth weight.

KEYWORDS

Mixed Effects, Random Intercepts, Panel Data

1. EL MODELO LINEAL GENERALIZADO DE EFECTOS MIXTOS

Un Modelo Lineal de Generalizado de Efectos Mixtos (MLGEM) tiene la siguiente forma general:

$$g[E(Y/X,u)] = X\beta + Zu$$

donde Y es un vector de respuestas de dimensión $n*1$ con función de distribución de probabilidad F, X es matriz de covariables $n*p$ asociado al vector de efectos fijos β, Z es una matriz de covariables $n*q$ asociado al vector de efectos aleatorios u, β es vector de efectos fijos $p*1$, u es vector de efectos aleatorios $q*1$, $\eta=X\beta+Zu$ es el predictor lineal, $g(.)$ es la función de enlace para la cual se supone que existe su función inversa $g^{-1}(.)$ de modo que $E(Y/X,\ u) = g^{-1}(X\beta+Zu)=h(\eta)=\mu$, donde μ es el vector de medias de dimensión $n*1$. Al considerar varias definiciones para $g(.)$ y F se tiene una amplia variedad de modelos,

* Ex Director de Investigación de la Unidad de Análisis y Política Social (UDAPSO)

entre los cuales se encuentra el modelo de regresión logística con *interceptos aleatorios*. Generalmente se asume que el vector de efectos aleatorios u tiene una distribución normal multivariada con media 0 y matriz de varianzas-covarianzas Σ de dimensión $q*q$, es decir, $u \sim N_q(0,\Sigma)$. Los efectos aleatorios no son estimados directamente, estos son caracterizados por sus varianzas, denominados comúnmente componentes de varianza. Estos componentes de varianza son elementos de la matriz de varianzas-covarianzas $G = Var(u)$.

El MLGEM permite modelar la correlación dentro de un *cluster* o conglomerado. Esto es, los sujetos dentro de un mismo *cluster* podrían estar correlacionados producto de un intercepto aleatorio compartido, producto de una pendiente aleatoria compartida, o como consecuencia de ambas situaciones.

Cuando se tiene datos *clusterizados*, no es conveniente considerar el total de las n observaciones al mismo tiempo, por el contrario, es ventajoso organizar el modelo mixto como una serie de M *clusters* independientes. La formulación apropiada del modelo es:

$$g[E(Y_j / X_j, u_j)] = X_j \beta + Z_j u_j$$

donde $j=1,\ldots,M$ y el *cluster* j consiste de n_j observaciones. El vector de respuestas Y_j es de dimensión n_j*1 e incluye todas las observaciones correspondientes al j-ésimo *cluster*. Lo mismo para las matrices X_j, Z_j y el vector u_j. Nuevamente se asume que el vector de efectos aleatorios u_j está distribuido normalmente con media 0 y matriz de varianzas-covarianzas Σ de dimensión $q*q$, es decir $u_j \sim N_q(0,\Sigma)$. Este modelo es el propuesto por Laird y Ware (1982) y ofrece dos ventajas importantes. Primero, se puede especificar los términos de los efectos aleatorios con facilidad. Si los *clusters* son escuelas, se puede especificar simplemente un efecto aleatorio al nivel de la escuela. Segundo, el modelo se puede generalizar fácilmente a más de un conjunto de efectos aleatorios. Por ejemplo, si las clases están anidadas dentro de escuelas, el modelo puede ser generalizado para incluir efectos aleatorios a nivel de escuelas y a nivel de clases dentro de escuelas.

La clave para ajustar modelos mixtos cae en la estimación de los componentes de varianza. Existen muchos métodos para tal estimación, uno de ellos es el de máxima verosimilitud (MV). Si $f(Y_j, u_j)$ representa la función de distribución conjunta de Y_j y u_j, la distribución marginal de Y_j es dada por

$$f(Y_j) = \int_{\mathbb{R}^q} f(Y_j, u_j)\, du_j$$

A partir de esta distribución marginal se puede deducir la función de verosimilitud para el *cluster* j, la cual queda expresada como

$$L_j(\beta,\Sigma) = \frac{1}{(2\pi)^{\frac{q}{2}}|\Sigma|^{\frac{1}{2}}} \int_{\mathbb{R}^q} e^{\left\{\ln f(Y_j/u_j) - \frac{u_j'\Sigma^{-1}u_j}{2}\right\}}\, du_j.$$

Como se supuso que los M *clusters* son independientes, la función de verosimilitud total para el vector de respuestas Y es dada por

$$L(\beta,\Sigma) = \prod_{j=1}^{M} L_j(\beta,\Sigma).$$

Para aproximar la integral que aparece en $L_j(\beta,\Sigma)$ se recurre a métodos numéricos. Primero, se hace un cambio de variable para

transformar la integral multivariable en un conjunto anidado de integrales univariables; segundo, cada integral univariable puede entonces ser evaluada usando la cuadratura Gauss-Hermite.

2. REGRESIÓN LOGÍSTICA CON INTERCEPTOS ALEATORIOS

Un caso particular del modelo lineal generalizado de efectos mixtos es el modelo de regresión logística con interceptos aleatorios. El modelo es expresado como:

$$P(Y_{ij} = 1)/X_{ij}, u_j) = H(X_{ij}\beta + u_j) = \mathrm{H}(\eta_{ij})$$

donde el efecto aleatorio u_j es una variable unidimensional con distribución normal, es decir, $u_j \sim N(0, \sigma^2)$. Se asume que $u_1,\ldots,u_M$ son independientes.

Para precisar algunas ideas, consideremos el siguiente ejemplo. Asumamos que la variable respuesta representa la ocurrencia o no de cáncer de pulmón y la variable explicativa es la condición de fumador, un factor de riesgo muy importante. Supongamos también que la muestra consiste de M submuestras conducidas en diferentes departamentos del país. Sea j el subíndice asociado al departamento e i a la persona. Entonces la variable binaria Y_{ij} representa la presencia o ausencia de cáncer de pulmón ($Y_{ij} = 1$ = *con cáncer*; $Y_{ij} = 0$ = *sin cáncer*) y X_{ij} representa la condición de fumador de la *i-ésima* persona en el *j-ésimo* departamento ($X_{ij} = 1$ = *fumador*; $X_{ij} = 0$ = *no fumador*). Sea n_j el número de personas encuestadas en el *j-ésimo* departamento. La regresión logística estándar aplicada al conjunto de datos $\{Y_{ij}, X_{ij}\}, j=1,\ldots,M, i=1,\ldots,n_j$, implícitamente asume que la incidencia de cáncer de pulmón es constante para todos los departamentos. Claramente este supuesto puede ser incorrecto porque los departamentos pueden tener diferentes condiciones ambientales, diferentes campañas contra el tabaco, diferentes tradiciones, diferentes políticas de salud y diferente población por edad, entre otros. Estos factores pueden conducir a diferentes incidencias de cáncer entre los departamentos. Por tanto, al asumir que esta incidencia es la misma se puede obtener conclusiones incorrectas con relación al efecto de fumar. No cabe duda que es más coherente y realista asumir que los interceptos difieran de un departamento a otro, por lo que un modelo más apropiado es el expresado anteriormente.

La función de verosimilitud – y consecuentemente la función *log-verosimilitud* – es un caso particular de la anterior función de verosimilitud correspondiente al MLGEM. La función *log-verosimilitud* para la regresión logística con interceptos aleatorios queda expresado como

$$l = l(\beta, \sigma^2)$$

$$= -\frac{M}{2} ln(2\pi\sigma^2) + \beta \sum_{j=1}^{M}\sum_{i=1}^{n_j} Y_{ij}X_{ij}$$

$$+ \sum_{j=1}^{M} ln\left[\int e^{u_j \sum_{i=1}^{n_j} Y_{ij} - \sum_{i=1}^{n_j} ln\left(1+e^{X_{ij}\beta+u_j}\right) - \frac{u_j^2}{2\sigma^2}}\, du_j\right]$$

Para estimar los parámetros β y σ^2 se puede usar el siguiente procedimiento iterativo

$$\hat{\beta}_{s+1} = \hat{\beta}_s + H^{-1}\left(\frac{\partial l}{\partial \beta}|_{\beta=\beta_s}\right)$$

$$\hat{\sigma}^2_{s+1} = \frac{1}{M}\sum_{j=1}^{M}\frac{I_{2j}}{I_{1j}}$$

donde

$$\frac{\partial l}{\partial \beta} = \sum_{j=1}^{M}\sum_{i=1}^{n_j} Y_{ij}X_{ij} - \sum_{j=1}^{M}\frac{\mathbf{I_{3j}}}{I_{1j}}$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{M}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{j=1}^{M}\frac{I_{2j}}{I_{1j}}$$

y las tres integrales están definidas como

$$I_{1j} = \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{h_j(\beta;u)}}\, du$$

$$I_{2j} = \int_{-\infty}^{\infty} u^2\, \mathrm{e}^{\mathrm{h_j(\beta;u)}}\, du$$

$$\mathbf{I_{3j}} = \int_{-\infty}^{\infty}\left[\sum_{i=1}^{n_j} X_{ij}\,\frac{e^{\beta' \mathrm{X_{ij}}+u}}{1+e^{\beta' \mathrm{X_{ij}}+u}}\,\mathrm{e}^{\mathrm{h_j(\beta;u)}}\right] du$$

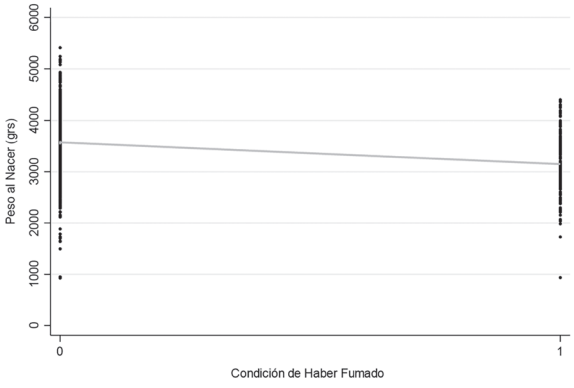
Notar que la integral $\mathbf{I_{3j}}$ es un vector $p*1$ y que H, I_{kj} para $k=1,2,3$, $j=1,\ldots,M$ son calculados en los valores actuales, $\beta = \beta_s$ y $\sigma = \sigma_s$.

3. APLICACIÓN A DATOS DE PANEL

Consideremos los datos de panel donde se tiene 648 *clusters* (mujeres en edad fértil) y en cada *cluster* se observa la condición de peso al nacer para cada uno de tres nacimientos. Luego se tiene un total de 1944 observaciones. La variable respuesta es la condición de bajo peso al nacer. Los nacimientos con bajo peso son los que tuvieron un peso de 2500 gramos o menos al momento de nacer. Una de las variables explicativas consideradas en el análisis es la condición de la madre de haber fumado o no durante cada embarazo. Adicionalmente se incluyeron en el modelo otras variables que, de acuerdo la experiencia, pueden afectar el peso al nacimiento. Estas variables son la edad de la madre, el estado civil de la madre, la educación de la madre, el control prenatal, momento del primer control prenatal, calidad del control prenatal y el sexo del recién nacido. El propósito del análisis es determinar si el haber fumado durante el embarazo tiene un efecto significativo sobre la probabilidad de nacer con bajo peso.

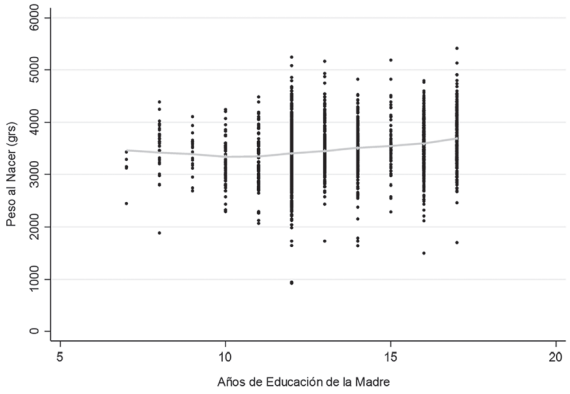
En los siguientes dos gráficos se exhiben las relaciones entre la variable peso al nacer y las variables condición de haber fumado durante el embarazo y educación de la madre. En términos generales se puede apreciar que el peso al nacer de los recién nacidos disminuye cuando la madre fuma durante el embarazo. Al relacionar el peso al nacer con la educación de la madre se observa que desde alrededor de los 12 años de educación comienza a incrementarse suavemente el peso al nacer, sin embargo, previo a este número de años de educación se observa incluso un leve descenso en el peso de los recién nacidos.

Figura 1.
Relación entre el Peso al Nacer y Condición de Fumar



Fuente: Elaboración Propia

Figura 2.
Relación entre el Peso al Nacer y la Educación de la Madre



Fuente: Elaboración Propia

El modelo de regresión logística de *interceptos aleatorios* usado para el análisis de los datos de panel es el siguiente

$$P(Y_{ij} = 1/X_{ij}, u_j) = g^{-1}(u_j + X_{ij}\beta) = \frac{e^{u_j + X_{ij}\beta}}{1 + e^{u_j + X_{ij}\beta}}$$

donde Y_{ij} es la condición de bajo peso al nacer del *i-ésimo* nacimiento para la *j-ésima* madre; X_{ij} es el vector fila de variables explicativas para el *i-ésimo* nacimiento de la *j-ésima* madre; β es el vector de coeficientes de efectos fijos y u_j es el efecto aleatorio de la *i-ésima* madre, que hace que el modelo tenga *interceptos aleatorios*. Los resultados se exhiben en el siguiente cuadro.

Cuadro 1.
Regresión Logística con interceptos aleatorios aplicado a datos de panel

Bajo peso al nacer	Razón de chances	Error estandar robusto	P>|z|	Intervalo de confianza 95%	
Fumó	3,17	1,21	0,003	1,50	6,70
Sexo del recién nacido	0,62	0,20	0,130	0,33	1,15
Edad de la madre	0,95	0,04	0,204	0,88	1,03
Educación de la madre	0,88	0,08	0,178	0,73	1,06
Madre casada	0,49	0,22	0,105	0,20	1,16
Prenatal de calidad intermedia	2,86	1,31	0,022	1,17	7,02
Prenatal de calidad inadecuada	6,19	3,58	0,002	1,99	19,22
Sin control prenatal	0,89	0,78	0,898	0,16	4,97
1er C.prenatal en 2do trimestre	0,65	0,32	0,382	0,24	1,71
1er C.prenatal en 3er trimestre	0,05	0,07	0,034	0,00	0,80
Constante	0,41	0,42	0,382	0,06	3,02

Fuente: Elaboración Propia

Recordemos que el objetivo del ejemplo es principalmente determinar si haber fumado durante el embarazo tiene un efecto significativo sobre la probabilidad de nacer con bajo peso. De los resultados expuestos en el cuadro se puede concluir que el efecto de fumar sobre el bajo peso al nacer es altamente significativo. Cuando la madre fuma, la chance de tener bajo peso al nacer es más de tres veces que cuando la madre no fuma. Por otra parte, como ya se advirtió en el gráfico, la educación de la madre no tiene un efecto significativo sobre el bajo peso al nacer. Sin embargo, la calidad del cuidado prenatal tiene un efecto altamente significativo sobre el bajo peso al nacer. Cuando el cuidado prenatal es de baja calidad, la chance de tener bajo peso al nacer es seis veces más que cuando el cuidado prenatal es adecuado.

4. ALGUNAS CONSIDERACIONES

El modelo lineal generalizado de efectos mixtos se caracteriza por incluir tanto efectos fijos como efectos aleatorios. La introducción de efectos aleatorios permite realizar el análisis de datos con estructura más compleja y, consecuentemente, permite un análisis más próximo de la compleja realidad. En la modelación se puede permitir, por ejemplo, que la probabilidad de nacer con bajo peso varíe de una madre a otra. En cambio, con un modelo lineal generalizado estándar no es posible realizar este tipo de análisis.

Si bien el modelo lineal generalizado de efectos mixtos permite un análisis más profundo de los datos, la maximización de la función *log-verosimilitud* para estimar los coeficientes llega a ser bastante compleja, puesto que involucra la solución de integrales complejas como la integral logística-normal. Para solucionar estas integrales se recurre a métodos de aproximación numérica. Principalmente se recurre al método de cuadratura de Gauss-Hermite.

Un modelo particular y muy importante de la familia de modelos lineales generalizados de efectos mixtos es el modelo de regresión logística con *interceptos aleatorios*. Este modelo es útil para analizar datos de panel.

BIBLIOGRAFÍA

1. Pinheiro, J.C. and Bates, D.M. (2000). Mixed-Effects Models in S and S-PLUS. New York: Springer
2. McCullagh, P. and Nelder, J. A. (1989). Generalized linear models, Second Ed. Chapman and Hall/CRC, London.
3. Laird, N. M. and Ware, J. H. (1982). Random-Effects Models for Longitudinal Data; Biometrics, Vol. 38, No. 4, pp. 963-974