REVISTA DE ECONOMÍA Y ESTADÍSTICA
ISSN 0034-8066
INSTITUTO DE ECONOMÍA Y FINANZAS
FACULTAD DE CIENCIAS ECONÓMICAS
UNIVERSIDAD NACIONAL DE CÓRDOBA

OPEN ACCESS

ARTÍCULOS

Modelos mixtos generalizados para el estudio del desempleo en los grandes aglomerados urbanos deArgentina

Fernando García, Margarita Díaz

Revista de Economía y Estadística, Cuarta Época, Vol. 49, No. 1 (2011), pp. 79-98.

http://revistas.unc.edu.ar/index.php/REyE/article/view/6510

La Revista de Economía y Estadística, se edita desde el año 1939. Es una publicación semestral del Instituto de Economía y Finanzas (IEF), Facultad de Ciencias Económicas, Universidad Nacional de Córdoba, Av. Valparaíso s/n, Ciudad Universitaria. X5000HRV, Córdoba, Argentina.
Teléfono: 00 - 54 - 351 - 4437300 interno 253.
Contacto: rev_eco_estad@eco.unc.edu.ar
Dirección web http://revistas.unc.edu.ar/index.php/REyE/index

Revista de Economía y Estadística - Vol. XLIX, N° 1 (2011) - pp. 79-98
Instituto de Economía y Finanzas - Facultad de Ciencias Económicas
Universidad Nacional de Córdoba – Argentina

Modelos mixtos generalizados para el estudio del desempleo en los grandes aglomerados urbanos de Argentina

Fernando García

Instituto de Estadística y Demografía, Facultad de Ciencias Económicas
Universidad Nacional de Córdoba
fgarcia.unc@gmail.com

Margarita Díaz

Instituto de Estadística y Demografía, Facultad de Ciencias Económicas
Universidad Nacional de Córdoba
mdiazlujan@gmail.com

Resumen

El objetivo de este trabajo es identificar los factores de riesgo socio-económicos y demográfico que influyen en la condición de empleo en los principales aglomerados urbanos de Argentina mediante un Modelo Lineal Generalizado Mixto. En este artículo se aplica un modelo logístico que incorpora efectos fijos y aleatorios y se evalúa la evolución promedio marginal inducida por el modelo a través de la marginalización de sus resultados. Para la estimación se utilizaron los datos prevenientes de la Encuesta Permanente de Hogares período 2004-2005. Para el conjunto de datos analizados, la modelación realizada proporcionó similares conclusiones a las arribadas en un estudio previo en el que se trabajó con un enfoque marginal.

Palabras clave: Condición de empleo, Modelo Lineal Generalizado Mixto, Marginalización del Modelo Mixto.

Clasificación JEL: C23, C51, J64.

Abstract

The aim of this analysis is to identify the socio-economic and demographics risk factors that affect the employment status in the main Argentine's urban areas by a Generalized Linear Mixed Models. In this article, we present a technique to estimate a logistic model which has the possibility of include fixed or random effects. In addition, it is possible to make the marginalization of the results in order to evaluate the marginal average evolution induced by the mixed model. The procedure is illustrated by using data from the Permanent Household Survey for the period 2004-2005. The results in this study provided similar findings to the above in a previous study in which we worked with a marginal model.

Keywords: Employment Status, Generalized Linear Mixed Models, Marginalized Mixed Models.

JEL Classification: C23, C51, J64.

I. Introducción

El análisis de datos longitudinales, que surgen al observar un individuo repetidamente en el tiempo, toma en cuenta la correlación entre las respuestas de un mismo sujeto y puede realizarse desde un enfoque promedio poblacional (modelo marginal) o desde los modelos sujeto-específicos. En situaciones donde la variable respuesta de interés es binaria, ambas aproximaciones pueden implementarse desde los Modelos Lineales Generalizados Mixtos (GLMM), los cuales constituyen una combinación natural de dos líneas de modelación, los modelos lineales mixtos y los modelos lineales generalizados (Molenberghs y Verbeke, 2005; Rabe-Hesketh y Skrondal, 2009). En el modelo marginal es posible medir la asociación longitudinal intra- sujeto mediante distintas estrategias, entre ellas usando cocientes de chances entre las respuestas de un mismo individuo. Esto puede lograrse mediante el uso de regresiones logísticas alternadas (Alternating Logistic Regressions, ALR), constituyendo este método, una extensión de las ecuaciones de estimación generalizadas (Carey, Zeger y Diggle, 1993). En el enfoque sujeto-específico (modelo mixto) las respuestas binarias comúnmente se expresan según un modelo logístico donde uno o más parámetros se asocian a componentes aleatorias.

El presente trabajo postula como Objetivo General identificar los factores de riesgo socio-económicos y demográficos que inciden en la precariedad laboral, en los principales aglomerados urbanos del país. En particular, procura describir perfiles de los individuos que permanecieron ocupados durante todo el período de observación y de aquellos que, en al menos una medición, fueron clasificados como desocupados. En términos más generales se pretende difundir la aplicación de estas metodologías estadísticas en el campo de las Ciencias Sociales y de brindar información adecuada para el diseño de políticas públicas y privadas que contribuyan a atenuar el flagelo de la desocupación.

Específicamente, se aplica un modelo logístico mixto en el que se introduce una ordenada al origen aleatoria para captar el efecto sujeto-específico (Fahrmeir y Tutz, 2001, Diggle *et al*, 2002). Se trabaja con la información de la Encuesta Permanente de Hogares (EPH) en la que cada hogar permanece cuatro trimestres en el panel.[1] Como los hogares son medidos repetidamente a través del tiempo, las series de datos son longitudinales, no resultando apropiados los enfoques que suponen independencia estocástica entre los datos.

II. Marco Teórico

II.1. Modelos sujeto-específicos

Los modelos sujeto-específicos se caracterizan por la inclusión de coeficientes que son específicos para cada sujeto. Condicional sobre estos, las respuestas se suponen independientes. Los efectos sujeto-específicos pueden ser tratados como efectos fijos, aleatorios o condicionándolos a ellos, siendo la más popular la aproximación de efectos aleatorios (Molenberghs y Verbeke, 2005). La idea básica de este modelo es que existe una heterogeneidad natural a través de los individuos en sus coeficientes de regresión, pudiendo dicha variabilidad ser representada por una distribución de probabilidad. Este modelo es más útil cuando el objetivo del estudio es hacer inferencias acerca de los individuos más que del promedio poblacional y describir los efectos de las covariables sobre los cambios en la respuesta para un individuo específico a través del tiempo (Fitzmaurice y Verbeke, 2009).

1. Con el fin de modelar el desempleo a partir del mismo conjunto de datos, se estimó previamente un modelo marginal aplicando el enfoque ALR. Para modelar la asociación entre las respuestas de un mismo individuo se utilizó la opción no estructurada. García F. "Aplicación de Modelos Estadísticos para datos longitudinales binarios: el caso del desempleo en los grandes aglomerados urbanos de la Argentina en el período 2004-2005.", Tesis Magíster en Estadística Aplicada.

Existen varias formas de introducir aleatoriedad en los parámetros del modelo. Stiratelli, Laird y Ware (1984) (citados por Molenberghs y Verbeke, 2005) suponen que el vector de parámetros se distribuye normalmente. Breslow y Clayton (1993) desarrollaron esta idea en los Modelos Lineales Generalizados Mixtos (GLMM), que constituyen uno de los modelos de efectos aleatorios para datos discretos utilizados con mayor frecuencia. Dichos modelos pueden ser ajustados maximizando la verosimilitud marginal, obtenida integrando sobre los efectos aleatorios, a través de la siguiente expresión:

$$L(\boldsymbol{\beta},\mathbf{D},\phi)=\prod_{i=1}^{K} f_i(y_i \mid \boldsymbol{\beta},\mathbf{D},\phi)=\prod_{i=1}^{K}\int\prod_{i=1}^{T_i} f_{it}(y_{it} \mid \mathbf{b}_i,\boldsymbol{\beta},\phi) f(\mathbf{b}_i \mid \mathbf{D}) d\mathbf{b}_i)$$

donde:

$y_i = (y_{i1}. y_{i2}, ...y_i T_i)'$ es el vector de respuestas del sujeto i

$\boldsymbol{\beta}$ es un vector de parámetros fijos de regresión,

$\mathbf{b}_i$ es un vector de coeficientes sujeto-específico, expresados como variables aleatorias distribuidas $N(\mathbf{0},\mathbf{D})$,

ϕ es un parámetro de escala.

Debido a que no están disponibles expresiones analíticas para resolver (1), son necesarias aproximaciones numéricas. Entre ellas, los métodos más ampliamente utilizados son los basados en la aproximación de los datos, estimaciones cuasi-verosímiles penalizadas (PQL) y cuasi-verosímiles marginales (MQL), y los basados en la aproximación de la integral en si misma (cuadratura Gaussiana y la adaptativa Gaussiana) (Molenberghs y Verbeke, 2005; Rabe-Hesketh y Skrondal, 2009).

II.1.1. Modelo Logístico con ordenada al origen aleatoria

Este modelo es un caso especial de la familia de modelos sujeto-específico usado para datos binarios. Sea $\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{iTi})'$ el vector de respuestas del sujeto i donde y_{it} denota la respuesta en el momento t : $i=1,2,.....K$; $t=1,2,....T_i$ y sea $\mathbf{x}_i=(\mathbf{x}_{i1},...,\mathbf{x}_{iTi})'$ la matriz $T_i \; x \; p$ conteniendo los valores de las covariables medidas en los T_i tiempos sobre el sujeto i-ésimo.

El modelo es:

$$\begin{aligned} y_{it} \mid b_i &\sim Bernoulli(\pi_{it}), \\ \log\left(\frac{\pi_{it}}{1-\pi_{it}}\right) &= \mathbf{x}_{it}'\boldsymbol{\beta} + b_i, \end{aligned} \qquad (2)$$

donde:
$\pi_{it}=P(y_{it}=1/b_i)$,
x_{it} es un vector de covariables asociadas,
β es un vector de parámetros fijos de regresión y
b_i es una variable aleatoria asociada a la ordenada al origen que se supone normalmente distribuida con media 0 y varianza τ^2.

Se trata de un GLMM con función enlace logit y con una estructura de efectos aleatorios simplificada a ordenada al origen aleatoria. El grado de heterogeneidad en las respuestas a través de los sujetos, no atribuible a las covariables, es captado por τ^2 (Diggle *et al*, 2002). Asimismo, la ordenada al origen aleatoria puede ser pensada como el efecto combinado de covariables sujeto-específicas omitidas que provocan que algunos individuos sean más propensos a estar desocupados que otros. Resulta atractivo modelar esta heterogeneidad no observada de la misma forma que la heterogeneidad observada a través de la simple adición de una ordenada al origen aleatoria en el predictor lineal (Rabe-Hesketh y Skrondal, 2008).

Debido a que el ajuste del modelo está basado en principios de máxima verosimilitud, la inferencia acerca de los parámetros es obtenida aplicando la teoría clásica. Suponiendo que el modelo ajustado es apropiado, los estimadores tienen distribución normal asintótica con matriz de covarianzas igual a la inversa de la matriz de información de Fisher. En este caso pueden utilizarse tanto las pruebas de Wald, score como de razón de verosimilitud.

II.1.2. Marginalización del Modelo Mixto

En un modelo logístico con ordenada al origen aleatoria, los coeficientes de regresión en (2) necesitan ser interpretados condicionalmente sobre los efectos aleatorios b_i (interpretación sujeto-específica). En este modelo, las esperanzas condicionales están dadas por:

$$\pi_{it} = E(y_{it} \mid b_i) = \frac{\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + b_i)}{1+\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + b_i)}. \qquad (3)$$

El modelo supone que todas las esperanzas condicionales satisfacen un modelo logístico, pero con diferentes ordenadas al origen $\beta_0 + b_i$ para cada individuo.

Si es de interés obtener la evolución promedio marginal implicada por el modelo mixto, es necesario integrar sobre los efectos aleatorios:

$$E(y_{it}) = E\left[E(y_{it} \mid b_i)\right]. \quad (4)$$

Para ello, son utilizadas técnicas de integración numérica. Sin embargo, es más frecuente el uso de promedios numéricos muestreando un número grande M de efectos aleatorios b_i de su distribución ajustada $N(0, \hat{\tau})$ y estimando $E(y_{it})$ como sigue:

$$E(y_{it}) = \frac{1}{M}\sum_{i=1}^{M} \frac{\exp(\mathbf{x}_{it}^{'}\hat{\boldsymbol{\beta}} + b_i)}{1 + \exp(\mathbf{x}_{it}^{'}\hat{\boldsymbol{\beta}} + b_i)}. \quad (5)$$

Cabe aclarar que este procedimiento no produce estimaciones formales para el vector de parámetros de regresión en el modelo marginal (Molenberghs y Verbeke, 2005).

III. Aplicación: el caso del desempleo en los grandes aglomerados urbanos de Argentina

III.1 Datos

Se analizaron los datos de la EPH, encuesta de relevamiento continuo de propósitos múltiples realizado por el Instituto Nacional de Estadísticas y Censos (INDEC) y por las Direcciones Provinciales de Estadística de Argentina. Con el diseño de muestreo implementado en la EPH a partir del año 2003, se renueva periódicamente el conjunto de hogares a encuestar (panel de respondentes) con un "esquema de rotación" llamado 2-2-2, en el que los hogares permanecen 4 trimestres no consecutivos en el panel (se encuestan a todas las personas del hogar). El método de rotación brinda la posibilidad de realizar estudios a través del tiempo, siguiendo a un respondente desde su ingreso al panel hasta la última vez que es encuestado. De manera específica, se tomó en consideración la base de datos de la EPH para los grandes aglomerados urbanos del país, es decir aquellos que tienen más de 500.000 habitantes: Gran Córdoba, Gran Rosario, Ciudad de Buenos Aires, partidos del Gran Buenos Aires, Gran Mendoza, Gran Tucumán-Tafí Viejo y Mar del Plata-Batán período 2004-2005 (www.indec.mecon.gov.ar).

La variable respuesta fue la condición de empleo del encuestado, la que se obtuvo mediante la segmentación de la población económicamente activa en Ocupados y Desocupados. Para la selección de las variables determinantes de la situación laboral (factores explicativos de la condición de

empleo), se consideraron tres dimensiones básicas: Características personales, Características del hogar y Perfil laboral de las personas (Díaz *et al*, 2005). Con respecto a las características personales se tuvieron en cuenta el Sexo, Edad, Estado Civil, Jefatura de Hogar y Nivel Educacional del encuestado. En cuanto a las características del hogar donde viven las personas se consideraron el Tamaño del Hogar y los indicadores que definen la Condición Necesidades Básicas Insatisfechas (NBI) del Hogar. Para analizar la dimensión perfil laboral se incluyeron indicadores que reflejan la experiencia laboral de las personas como Categoría Ocupacional, Tamaño del Establecimiento y Rama de Actividad.

La mayoría de los factores explicativos se trataron como variables categorizadas. En la tabla 1 se muestra la operacionalización de las variables utilizadas.

Tabla 1
Operacionalización de las variables predictoras

Variable	Categorías
Edad	*Continua*
Sexo	Varón y Mujer
Estado Civil	Sin pareja y Con pareja
Jefatura del Hogar	No jefe y Jefe
Nivel Educacional	Sin instrucción o primario incompleto, Primario completo o secundario o superior incompleto y Superior completo
Condición NBI del Hogar	Hogar pobre y Hogar no pobre
Tamaño del Hogar	Hogares con 1 o 2 personas y Hogares con 3 o más personas
Categoría Ocupacional	Trabajador independiente y Trabajador en relación de dependencia
Rama de Actividad	Construcción, Comercio, Servicios personales, Servicios sociales, Otras ramas e Industria
Aglomerado	Gran La Plata, Gran Rosario, Gran Mendoza, Gran Córdoba, Gran Tucumán - Tafí Viejo, Ciudad de Buenos Aires, Mar del Plata - Batán y Partidos del Gran Buenos Aires
Tamaño del Establecimiento	Hasta 5 empleados, De 6 a 40 empleados y Más de 40 empleados
Tiempo	*Continua*

Fuente: elaboración propia.

Una limitación importante en los datos de la EPH viene dado por el efecto desgranamiento, es decir la pérdida de observaciones en el tiempo, porque hay hogares que dejan de responder a partir de una observación en adelante (dropout). Esta situación podría sesgar las estimaciones en el caso de encontrarse asociado con el fenómeno de estudio. Si la pérdida de datos es completamente aleatoria e independiente, este desgranamiento no debería ocasionar sesgos. Esta es la conclusión a la que se arriba en un estudio realizado con base en información de la EPH en el aglomerado Córdoba.[2] Se considera que se puede extender esa conclusión a todos los aglomerados incluidos en el estudio, por tratarse de datos en los que las muestras se extraen con el mismo diseño y la recolección se realiza con idénticos procedimientos.

En este trabajo se realizaron las depuraciones pertinentes de modo que la información disponible es referida a la población económicamente activa (PEA) con edades comprendidas entre 15 y 65 años de edad. La muestra quedó conformada por 697 personas, quienes respondieron a la encuesta en las cuatro entrevistas planificadas: inicial, a los 3 meses, a los 12 meses y a los 15 meses.

III.2. Procedimientos

En una primera etapa de la modelación se utilizó un modelo de regresión logístico incorporando un vector de efectos aleatorios con dos coeficientes aleatorios, uno para cada individuo y otro para cada aglomerado. Esta decisión obedeció a la necesidad de capturar la asociación existente entre las respuestas de un mismo sujeto y entre las respuestas de los sujetos que viven en un mismo aglomerado respectivamente. La idea subyacente era que existía una heterogeneidad natural a través de los individuos en los distintos aglomerados. El método de estimación utilizado fue el enfoque de cuasi-verosimilitud penalizada implementado en el procedimiento *GLIMMIX* de *SAS*, pese a que suele producir estimaciones sesgadas de los parámetros de regresión en el caso de datos binarios con pocas repeticiones por sujeto. A

2. Se ha realizado un análisis del mecanismo de pérdida de datos de la EPH en el aglomerado Córdoba. Para identificar si la pérdida se produce en forma completamente aleatoria, aleatoria o no aleatoria, fue necesario realizar el análisis del mecanismo de dropout a través de una regresión logística. Los resultados obtenidos permiten concluir que el proceso es completamente aleatorio, ya que se observó una clara independencia del dropout en relación con el estado del hogar en la visita previa (pobre o no pobre), así como con el resto de las covariables incluidas en el análisis. Por lo tanto, es de esperar que los trabajos basados en datos completos no estén afectados por un sesgo. Stanecka N. Modelos para datos longitudinales binarios completos y con información faltante aplicados al estudio de la pobreza en Argentina, Tesis Magíster en Estadística Aplicada.

efectos de la estimación de los parámetros de covarianza del vector de efectos aleatorios se trabajó con el modelo saturado. En el mismo se incluyeron todos los efectos principales para las variables definidas anteriormente, todas las interacciones de las predictoras con el tiempo y algunas interacciones dobles entre predictoras juzgadas pertinentes en esta aplicación tales como *Estado Civil x Sexo, Estado Civil x Jefatura de Hogar, Jefatura de Hogar x Sexo, Nivel de Educación x Sexo y Edad x Sexo.* La variable Edad, medida en escala continua, fue modelada a través de una componente lineal y otra cuadrática a fin de captar la diferencia observada de cambios en la probabilidad de desocupación según la edad del sujeto.

La utilización del procedimiento *GLIMMIX* de *SAS* se hizo con la intención de realizar una primera selección de los efectos principales e interacciones cuya incorporación al modelo podían resultar pertinentes. Las estimaciones obtenidas fueron utilizadas como valores iniciales para los parámetros del modelo en la implementación del procedimiento *NLMIXED* de *SAS* (SAS Institute, 2002-2003) con el que se ajustó el modelo final con el cual se hicieron las interpretaciones. *NLMIXED* es un procedimiento alternativo de SAS, que implementa cuadratura Gaussiana y cuadratura adaptativa Gaussiana como aproximaciones de la integral en la verosimilitud marginal (1). De esta manera, permite obtener estimadores máximo verosímiles de los parámetros del modelo aproximando numéricamente el logaritmo de la verosimilitud. A efectos de determinar los puntos de cuadratura necesarios para la aproximación de las integrales se siguió la sugerencia de Agresti (2002) de ir aumentando secuencialmente el valor de q hasta que los cambios en las estimaciones y sus errores estándar fueran despreciables. Se utilizó la cuadratura adaptativa Gaussiana debido a que es más eficiente y permite reducir el número de puntos de cuadratura en relación a la cuadratura Gaussiana. Para la selección del modelo más adecuado, dado el conjunto de datos disponibles, se evaluaron los criterios AIC, AICC y BIC provistos por *PROC NLMIXED* para varios modelos alternativos. Finalmente se eligió aquel modelo que tuvo valores más pequeños para dichos criterios.

III.3 Resultados

Del análisis de las estimaciones de las componentes de varianza obtenidas al aplicar un modelo mixto con efectos aleatorios de sujeto y aglomerado, supuestos independientes, surge que la variabilidad entre aglomerados es muy baja b_1=0,3992 ($\sqrt{d11}$=0,3899) en relación a la variabilidad entre sujetos b_2=4,4049 ($\sqrt{d22}$=0,5254). Estos resultados indican que

la heterogeneidad de los datos es atribuible más bien a diferencias entre los sujetos que entre aglomerados. Por tal motivo, se decidió trabajar con un sólo efecto aleatorio, incorporando el *Aglomerado* como factor fijo dentro del predictor junto con el resto de covariables. En este caso, la asociación entre las medidas repetidas será modelada a través de la incorporación de un coeficiente aleatorio específico para cada individuo.

A los efectos de la interpretación de los resultados obtenidos se desarrollaron varias estrategias. En primer lugar, se analizaron los cocientes de chances para las categorías significativas de las variables incluidas en el modelo. Adicionalmente fueron calculadas las probabilidades de desempleo según distintas características socio-económicas y demográficas de las personas. Para tal fin, se construyó un perfil de individuo que fue tomado como referencia: varón de 35 años que vive en un hogar con menos de tres integrantes sin NBI (Necesidades Básicas Insatisfechas), pertenece al Gran Buenos Aires, tiene pareja, es jefe de hogar, tiene estudios secundarios completos y trabaja en una empresa industrial que tiene entre 6 y 40 empleados. Para la interpretación de aquellas covariables cuya interacción con el factor Tiempo no resultó significativa, la comparación con el perfil de referencia podría haber sido evaluada en cualquiera de las cuatro mediciones. En todos los casos que se comentan a continuación, las probabilidades de desempleo fueron calculadas en la primera medición.

Los resultados del ajuste del modelo sujeto-específico se presentan en la Tabla 2. De los predictores considerados no resultaron significativos respecto a su contribución en la explicación de las chances de desempleo la *Categoría Ocupacional* y el *Tamaño del Hogar*. Igual suerte siguieron las *interacciones dobles entre predictoras*. En relación al *Tiempo*, sólo resultó significativa la interacción *con el Sexo*.

El objetivo principal en este tipo de estudios es realizar inferencias acerca de los efectos fijos relacionados a promedios poblacionales. La inclusión de efectos aleatorios es un mecanismo para caracterizar la correlación entre las mediciones para un mismo sujeto pero en general no interesan conclusiones sujeto-específico. Sin embargo, la estimación de la desviación estándar de la ordenada aleatoria del modelo constituye un resumen útil del grado de heterogeneidad de la población bajo estudio (Agresti, 2002). En la aplicación abordada en este trabajo la desviación estándar de la ordenada al origen aleatoria ($\hat{\tau}$ = 6,2203) es significativamente distinta de cero (p<0,0001), sugiriendo que la heterogeneidad en la población es importante,

Tabla 2
Estimaciones de los parámetros del modelo logístico con ordenada al origen aleatoria

Parámetro	Estimación	Error Estándar	valor p	Cociente de chances
Intercepto	-1,432	5,508	0,795	
Estado Civil (Ref: Con pareja)				
Sin pareja	2,481	1,062	0,020	11,9
Edad	-0,649	0,257	0,012	0,52
Edad²	0,007	0,003	0,020	1,01
Nivel de Educación (Ref: Superior completo)				
Sin instrucción - Primario incompleto	5,248	1,847	0,005	190,43
Prim. Comp.-Secundario-Superior incompleto	2,400	1,134	0,035	11
Sexo (Ref: Mujer)				
Varón	0,139	1,21	0,908	
Sexo – Tiempo				
Varón-Tiempo	0,156	0,074	0,036	
Jefatura de Hogar (Ref: Jefe)				
No Jefe	2,550	1,171	0,030	12,8
Aglomerado (Ref. partidos del GBA)				
Gran La Plata	-0,21	1,604	0,896	
Gran Rosario	-0,85	1,494	0,570	
Gran Mendoza	-4,330	1,871	0,021	0,013
Gran Córdoba	-3,480	1,588	0,029	0,031
Gran Tucumán-Tafí Viejo	-5,530	2,319	0,017	0,004
Ciudad de Buenos Aires	-2,091	1,391	0,133	
Gran Mar del Plata – Batán	-4,221	2,426	0,082	
Condición NBI del Hogar (Ref: No pobre)				
Pobre	4,250	0,619	<0,0001	70,1
Rama de Actividad (Ref: Industria)				
Construcción	1,822	1,073	0,09	6,2
Comercio	0,680	0,838	0,417	
Servicios Personales	0,550	0,842	0,514	
Servicios Sociales	-1,481	1,348	0,273	
Otras ramas	4,850	1,236	<0,0001	127,8
Tamaño del Establecimiento (Ref: más de 40 empleados)				
hasta 5 empleados	2,719	0,729	0,000	15,2
entre 6 y 40 empleados	2,408	0,689	0,001	11,1
Tiempo	-0,175	0,068	0,010	0,84
Desvío Estándar intercepto aleatorio	6,220	0,886	<0,0001	

Fuente: elaboración propia.

siendo atribuible la misma a diferencias entre los sujetos. Este resultado es consistente con el obtenido en el primer modelo mixto utilizado (con efectos aleatorios para sujeto y aglomerado). Las estadísticas descriptivas calculadas para las predicciones de los efectos específicos de sujeto ($\hat{b}_i$), sugieren que la magnitud de la varianza de la ordenada al origen aleatoria es influenciada por la existencia de una proporción baja de efectos de sujeto muy atípicos (Tabla 3). Sin embargo, la mayoría de los efectos de sujeto se encuentran en el entorno del cero (valor esperado de la distribución de efectos de sujeto).

Tabla 3
Estadísticas descriptivas para las predicciones de los efectos de sujeto ($\hat{b}_i$)

Variable	Mínimo	Percentil 10	Mediana	Percentil 90	Máximo
	-4,245	-0,228	-0,001	3,539	21,389

Fuente: elaboración propia.

Revisado el conjunto de datos de la EPH nuevamente, se pudo constatar que tales efectos de sujeto atípicos correspondían a individuos desocupados pese a que no presentaban los factores de riesgo esperados según el modelo ajustado. Este hallazgo sugiere que otras características, distintas a las consideradas como predictoras en el modelo ajustado y presentes en estos sujetos, son las que determinan su condición de desempleo. A los efectos de evaluar el impacto de tales observaciones en los resultados obtenidos, se estimó el modelo eliminándolas, obteniendo idénticos resultados a los alcanzados con los datos completos.

Los coeficientes de las covariables incluidas en el modelo se interpretan en términos de cocientes de chances condicionales (el modelo es condicional sobre el efecto aleatorio para cada sujeto). Es decir que constituyen una medida del riesgo de desempleo para una persona determinada, según sea el valor asumido por cada covariable considerada, manteniendo constante el valor de las restantes. Estos coeficientes, de acuerdo a la especificación del modelo, son iguales para todos los sujetos, es decir que su estimación se obtiene promediando sobre todos los individuos de acuerdo a la distribución del efecto aleatorio para sujeto.

Del análisis de los coeficientes estimados surge que la chance estimada de desempleo para un determinado individuo es ciento noventa veces mayor si no tiene instrucción o primario incompleto con respecto a si tuviera estudios superiores completos. Esta relación, disminuye notablemente si tiene estudios secundarios o superiores incompletos en relación a si tuviera es-

tudios superiores completos (cociente de chances igual a 11). Considerando las otras características personales podemos señalar que la chance estimada de desempleo es casi doce veces mayor si el sujeto no tiene pareja a si la tuviera. Idéntico efecto tiene la *Jefatura del Hogar* (chance igual a 12,8), resultando un factor de riesgo de desempleo no ser jefe de hogar. En relación a las características del hogar la chance estimada de desempleo para un individuo determinado es setenta veces mayor si vive en un hogar pobre que si vive en un hogar sin NBI. Con respecto a perfil laboral de las personas, resultaron factores determinantes el *Tamaño del Establecimiento* y la *Rama de Actividad*. Si consideramos el *Tamaño del Establecimiento*, podemos señalar que la chance estimada de desempleo es quince veces mayor si el individuo trabaja en un establecimiento que tiene menos de 5 empleados a si lo hiciera en establecimientos con más de 40 empleados, disminuyendo a once si el individuo trabaja en un establecimiento que tiene entre 6 y 40 empleados. En lo atinente a la *Rama de Actividad,* el riesgo de desempleo se incrementa (seis veces mayor) si el individuo trabaja en la rama de la construcción a si lo hiciera en la industrial. Si consideramos el *Aglomerado* donde vive el sujeto, la chance estimada de desempleo es un 98 % menor si el sujeto vive en el Gran Mendoza a si viviera en el Gran Buenos Aires, en tanto que si la persona vive en el Gran Córdoba el riesgo de desempleo es un 97 % menor. La mayor diferencia se presenta si el individuo vive en el Gran Tucumán, siendo la chance de desempleo un 99 % menor a si viviera en el Gran Buenos Aires.

Jerarquizando el efecto de los factores determinantes del desempleo de acuerdo a su importancia, debemos destacar en primer lugar el Nivel de Educación (sin instrucción o con primario incompleto). En segundo lugar la condición de vivir en un hogar con NBI (pobre), seguido por trabajar o haber trabajado en un establecimiento pequeño o mediano, no ser jefe de hogar, no tener pareja y trabajar o haber trabajado en la rama de la construcción.

III.3.1. Marginalización del Modelo Mixto

Como fue señalado en la Sección 2.1.2, para evaluar la evolución promedio marginal inducida por el modelo mixto es necesario integrar sobre los efectos aleatorios. De los dos procedimientos mencionados en dicha sección, seguiremos el último, es decir usaremos promedios numéricos, para lo cual se generaron 2000 realizaciones del coeficiente aleatorio b_i tomados de una distribución normal con media cero y varianza $(\hat{\tau}^2) = 38{,}69$, siendo esta última la estimación de la varianza de la ordenada al origen

aleatoria en el modelo finalmente ajustado. Para cada una de las 2000 realizaciones del coeficiente aleatorio, la esperanza condicional fue calculada utilizando la expresión (3), reemplazando los parámetros de regresión $\boldsymbol{\beta}$ por sus estimaciones (Tabla 2). Una estimación de (4) fue obtenida promediando las 2.000 esperanzas condicionales calculadas.

$$E(\hat{Y}_{it}) = \frac{1}{2000}\sum_{i=1}^{2000}\frac{\exp(\mathbf{x}_{it}^{'}\hat{\boldsymbol{\beta}} + b_i)}{1+\exp(\mathbf{x}_{it}^{'}\hat{\boldsymbol{\beta}} + b_i)}$$

A efectos de los cálculos anteriores se consideró el sujeto de referencia definido originalmente. Adicionalmente se fueron evaluando los cambios en la probabilidad de desempleo para las distintas características socioeconómicas y demográficas de las personas. El comportamiento de la variable *Edad* puede ser estudiado analizando la probabilidad de desempleo para el perfil de referencia según distintas edades. Del mismo surge una disminución en la probabilidad de desempleo hasta alrededor de los 45 años, momento a partir del cual comienza nuevamente a aumentar. Esta disminución en el primer tramo de edades puede estar vinculado a una mayor educación, experiencia en el trabajo, mayores contactos personales. etc. No obstante esta tendencia se revierte a medida que la persona tiene mayor edad. Una posible explicación para el aumento en la probabilidad de desempleo postula que la depreciación del capital humano comienza a ser más veloz que la capacidad de absorción del mismo. En el caso de la variable *Estado Civil* se observa claramente el mayor riesgo de desempleo para un individuo sin pareja. En relación a la variable Sexo, si bien el efecto principal no resultó significativo, la interacción *Sexo x Tiempo* sí lo es. El estudio de la evolución de las probabilidades de desempleo en las cuatro mediciones para el sujeto de referencia, agregado un nuevo perfil similar al anterior pero de *Sexo* mujer, muestra que las probabilidades de desempleo para las mujeres en relación a los varones no permanecen constantes en las cuatro mediciones, resultando mayor el riesgo de desocupación de los varones en relación a las mujeres sobre el final del estudio. Con respecto al *Nivel de Educación*, podemos señalar que la probabilidad de desempleo disminuye a medida que aumenta el nivel de instrucción, presentando mayor riesgo aquellos individuos sin instrucción o con primario incompleto. Similares conclusiones fueron obtenidas en el modelo marginal. En relación a la *Jefatura de Hogar*, el no ser jefe de hogar es un factor de riesgo de desempleo importante, no advirtiéndose una mayor incidencia en las mujeres con respecto a los varones, como fue detectado en el modelo marginal. En el caso de la *Condición NBI del Hogar,* pertenecer a un hogar con NBI

(pobre) aumenta significativamente la probabilidad de desempleo, constituyendo por tanto un factor de riesgo importante.

Considerando los indicadores del perfil laboral podemos mencionar que tanto el *Tamaño del Establecimiento* como la *Rama de Actividad* son factores de riesgo de desempleo. En el primer caso podemos señalar que el riesgo de desempleo se incrementa a medida que disminuye el número de empleados en el establecimiento. En lo que respecta a la *Rama de Actividad,* el mayor riesgo de desempleo se observa para aquellas personas que trabajan o trabajaron en la rama de la construcción y en otras ramas en relación a los que lo hicieron en la industria.

Finalmente, resta mencionar como factor determinante del desempleo el *Aglomerado* donde vive la persona. En este sentido, las probabilidades de desempleo de los individuos que viven en el Gran Córdoba, el Gran Mendoza y el Gran Tucumán-Tafí Viejo son sensiblemente menores en relación a los que viven en el Gran Buenos Aires presentando los individuos del Gran Tucumán-Tafí Viejo la menor probabilidad de desempleo. El análisis anterior puede efectuarse también calculando las probabilidades de desempleo según distintas características socio-económicas y demográficas de las personas. Estas probabilidades se muestran en la siguiente tabla para el momento inicial.

Tabla 4
Probabilidades de desempleo según distintas características socio-económicas y demográficas de las personas (Modelo Mixto Marginalizado)

Probabilidad de desempleo	**Inicial**
Sujeto de referencia	0,046
Construcción	0,081
Sin pareja	0,097
No jefe	0,099
Sin instrucción	0,107
Pobre	0,152

Fuente: elaboración propia.

El mismo análisis fue realizado para un sujeto típico o promedio (b_i=0) obteniendo conclusiones similares a las anteriores.

IV. Discusión y conclusiones

El análisis de los datos proporcionados por la EPH presenta desafíos metodológicos desde el punto de vista estadístico, debido a la correlación existente entre las respuestas de un mismo respondente a diferentes tiempos. Esta asociación es producto de considerar posibles efectos de individuos. Desde el ingreso de un respondente al panel, este es encuestado en cuatro oportunidades entre las que transcurren 15 meses. En cada momento de tiempo, se registran variables que proveen información en torno a distintas temáticas, entre las cuales se encuentra la situación laboral, que permite abordar estudios de la desocupación a nivel país. Comúnmente este tipo de estudios se realiza transversalmente, es decir para algún momento de tiempo dado. No obstante, la característica temporal de la encuesta brinda también la posibilidad de abordar estudios longitudinales.

Debido a que el estudio de la desocupación se hizo a través de una respuesta binaria (Ocupado/Desocupado), en este trabajo se utilizaron modelos generalizados para respuesta dicotómica. Para contemplar la correlación en las series de datos longitudinales, se utilizaron extensiones de estos modelos lineales generalizados para datos binarios (Fahrmeir y Tutz, 2001). El uso del enlace logit fue preferido debido a que la probabilidad de desempleo en promedio asumió valores cercanos a 0,10 en este conjunto de datos, por lo que funciones de enlaces más simples podían proveer estimaciones fuera del límite esperado. Además, la interpretación de los parámetros en términos de cocientes de chances resultó apropiada para evaluar el efecto de los factores de interés sobre el desempleo.

Se ajustó un modelo logístico con ordenada al origen aleatoria para medidas repetidas binarias, que incorpora vía el coeficiente aleatorio la correlación entre las respuestas de un mismo sujeto, al tiempo que permite modelar la dependencia de la respuesta en términos de las variables explicatorias. Ignorar la correlación entre las respuestas podría provocar que se detecten erróneamente efectos no significativos de las variables predictoras, ya que los errores estándares son subestimados. También se intentó modelar la correlación entre las respuestas de los individuos que viven en un mismo aglomerado mediante la incorporación de una componente aleatoria específica para cada aglomerado. Debido a que la variabilidad entre aglomerados fue muy baja en comparación con la variabilidad entre sujetos, el modelo finalmente usado para la interpretación del fenómeno en estudio fue el modelo logístico con ordenada al origen aleatoria debido al efecto de sujeto.

Este modelo se ajustó maximizando la verosimilitud marginal obtenida integrando sobre los efectos aleatorios. Para ello, se utilizó la cuadratura adaptativa Gaussiana como método de aproximación numérica. Se ajustaron y compararon modelos de regresión logística con efecto aleatorio de sujeto distintos respecto a la estructura de medias. Estos modelos incluyeron el efecto del tiempo, una o más covariables y la interacción de estas con el tiempo.

Los resultados obtenidos mediante la aplicación de estos modelos al conjunto de datos de la EPH muestran que las estimaciones de los efectos vía modelos mixtos pueden son mayores que las obtenidas mediante los modelos marginales. La discrepancia fue importante debido a la considerable heterogeneidad existente entre sujetos, que se traduce en una varianza de la ordenada al origen aleatoria muy alta ($\hat{\tau}^2$ =38,69). Los parámetros en ambos modelos tienen interpretaciones diferentes. Por ello, la elección del modelo a estimar se debe hacer en función del tipo de inferencia que se quiere efectuar. Los modelos marginales son más apropiados cuando se desean obtener inferencias acerca de promedios poblacionales. En este caso, el enfoque ALR permite obtener estimaciones eficientes y consistentes de los parámetros de regresión aunque la estructura de asociación no esté correctamente especificada. Sin embargo, si se desea obtener una interpretación específica para cada sujeto, es decir controlando por el efecto del individuo, el modelo a elegir es el mixto. En este caso, los errores estándares serán mayores, dado que el espacio de inferencia es más amplio. La gran variabilidad observada entre sujetos hace que la magnitud de los errores estándares obtenidos en el modelo mixto sean mayores que en el modelo marginal. En cuanto a la consistencia en la inferencia acerca de los parámetros, el enfoque marginal sólo requiere que la relación entre el valor esperado y las covariables esté correctamente especificada. El modelo mixto, en cambio, exige que no sólo la función de enlace se especifique correctamente, sino también la distribución de probabilidad supuesta para los efectos aleatorios, así como que dichos efectos sean independientes de las covariables. Sin embargo, Lee y Nelder (2004) sugieren que el enfoque sujeto-específico es más "rico", ya que a partir de un modelo mixto es posible estudiar una relación marginal (marginalización del modelo mixto), pero no es posible estudiar una relación sujeto-específico a partir de un modelo marginal.

En la estrategia de análisis seguida en este trabajo se estimó un modelo mixto que permitió estimar los cocientes de chances para las categorías significativas de las variables incluidas en el modelo (ver Tabla 2). Adicio-

nalmente, a efectos de evaluar la evolución promedio se marginalizó el modelo mixto a partir de una muestra de 2000 realizaciones del efecto aleatorio. A partir de ello, fueron obtenidas las probabilidades de desempleo según distintas características socio-económicas y demográficas de las personas. Para tal fin, se construyó un perfil de individuo que fue tomado como referencia (ver sección 3.3) y que permitió evaluar los cambios en probabilidad de desempleo según las distintas características socio-económicas y demográficas de las personas (ver Tabla 4).

De dicho análisis surge que la pertenencia del individuo a un hogar con necesidades básicas insatisfechas triplica la probabilidad de desempleo. En cuanto a las características personales se observa un mayor riesgo de desempleo para los individuos sin instrucción, no jefes de hogar y sin pareja, factores que prácticamente duplican la probabilidad. En cuanto a la Edad, se advierte una disminución en la probabilidad de desempleo hasta alrededor de los 45 años, momento a partir del cual comienza nuevamente a aumentar. El estudio de la evolución de las probabilidades de desempleo en las cuatro mediciones para el sujeto de referencia, agregado un nuevo perfil similar al anterior pero de Sexo mujer, muestra que las probabilidades de desempleo para las mujeres en relación a los varones no permanecen constantes en las cuatro mediciones, resultando mayor el riesgo de desocupación de los varones en relación a las mujeres sobre el final del estudio.

Considerando los indicadores del perfil laboral podemos mencionar que tanto el *Tamaño del Establecimiento* como la *Rama de Actividad* son factores de riesgo de desempleo. En lo que respecta a la *Rama de Actividad*, el mayor riesgo de desempleo se observa para aquellas personas que trabajan o trabajaron en la rama de la construcción y en otras ramas en relación a los que lo hicieron en la industria.

Finalmente, resta mencionar como factor determinante del desempleo el *Aglomerado* donde vive la persona. En este sentido, las probabilidades de desempleo de los individuos que viven en el Gran Córdoba, el Gran Mendoza y el Gran Tucumán-Tafí Viejo son sensiblemente menores en relación a los que viven en el Gran Buenos Aires presentando los individuos del Gran Tucumán-Tafí Viejo la menor probabilidad de desempleo.

La experiencia de ajustes de modelos para estos datos de la EPH hace factible proponer esta estrategia de análisis recomendable de ser implementada en otros estudios longitudinales de los factores de riesgo socio-econó-

mico y demográfico determinantes del desempleo cuando este sea medido a través de datos binarios en paneles. El uso de los modelos generalizados mixtos representa un importante aporte metodológico para las Ciencias Sociales, donde es muy común trabajar con variables categorizadas y con efectos muy variables de sujetos. Su aplicación a datos de la EPH para estudiar el desempleo aporta valiosa y adecuada información para el diseño de políticas públicas y privadas que contribuyan a atenuar el flagelo de la desocupación.

V. Referencias

Agresti, A. (2002). *Categorical Data Analysis* (2nd.), New York: John Wiley and Sons.

Breslow, N.E. y Clayton, D.G. (1993) "Aproximate inference in generalized linear mixed models". *Journal of the American Statistical Association*, nº 88, pp 9-25.

Carey, V.C., Zeger, S.L.and Diggle, P.J. (1993) "Modelling multivariate binary data with alternating logistic regressions''. *Biometrika*, nº 80, pp 517-526.

Margarita Díaz, Fernando Ferrero, Cecilia Díaz, Patricia Caro y María Inés Stimolo, (2005). "Análisis del desempleo urbano a través de un estudio comparativo de métodos de clasificación," ***Revista de Economía y Estadística,*** Universidad Nacional de Córdoba, Facultad de Ciencias Económicas, Instituto de Economía y Finanzas, vol. XLIII (2), pp 61-85.

Diggle, P.J., Heagerty, P.J., Liang, K.Y. y Zeger, S.L. (2002). *Analysis of Longitudinal Data*, (2nd ed), Oxford: Oxford University Press.

Fahrmeir, L. y Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models* (2nd ed), New York: Springer-Verlag.

Fitzmaurice, G y Verbeke, G. (2009). "Parametric modeling of longitudinal data: Introduction and overview". En Fitzmaurice, G., Davidian, M., Verbeke, G. y Molenberghs, G. (eds) "Longitudinal Data analysis". *Handbooks of Modern Statistical Methods*, pp.31-41 Chapman & Hall/CRC, New York.

García, F. (2007). "Aplicación de Modelos Estadísticos para datos longitudinales binarios: el caso del desempleo en los grandes aglomerados urbanos de Argentina en el período 2004-2005". Tesis Magíster Estadística Aplicada. Universidad Nacional de Córdoba.

INDEC (2003) "La nueva Encuesta Permanente de Hogares de Argentina". Documento de trabajo.

McCullagh, P. y Nelder, J.A. (1989). *Generalized Linear Models* (2nd ed.), London: Chapman and Hall.

Molenberghs, G y Verbeke, G. (2005). *Model for Discrete Longitudinal Data*. New York. Springer.

Lee, Y. y Nelder, J. (2004), "Conditional and Marginal Models: Another view". *Statistical Science*, nº 19, pp. 219-238.

Liang, K.Y. y Zeger, S.L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models", *Biometrika*, nº 73, pp 13-22.

Rabe-Hesketh, S. y Skrondal, A. (2008) "Multilevel and Longitudinal Modeling Using Stata". Texas, StataCorp.

Rabe-Hesketh, S. y Skrondal, A. (2009). "Generalized linear mixed-effects models". En Fitzmaurice, G., Davidian, M., Verbeke, G. y Molenberghs, G. (eds) "Longitudinal Data analysis". *Handbooks of Modern Statistical Methods*, pp.79-106 Chapman & Hall/CRC, New York.

SAS Institute, Inc. 2002-2003. SAS/STAT User's Guide, Version 9.1.3 Carey, NC, USA.

Stanecka, Nancy (2009). "Modelos para datos longitudinales binarios completos y con información faltante aplicados al estudio de la pobreza en Argentina". Tesis Magíster Estadística Aplicada. Universidad Nacional de Córdoba.

Observatorio de Política

Esta sección incluye artículos que discuten en forma rigurosa, pero no técnica, temas corrientes de política económica que son de interés por su vinculación al mundo real, aún cuando la literatura económica no los haya todavía incorporado definitivamente y artículos que presentan contenidos teóricos o resultados empíricos con implicancias de política relevantes. Esta sección procura acercar a los investigadores académicos con los formuladores de política aportando, respectivamente unos y otros, desarrollos teórico-conceptuales y empíricos importantes y claridad e información sobre las prioridades de política. Los artículos enviados a para esta Sección no están sujetos a los procedimientos normales de referato de la Revista.