Servy, Elsa
Hachuel, Leticia
Boggio, Gabriela
Cuesta, Cristina
Giordani, Natalia
Méndez, Fernanda
Instituto de Investigaciones Teóricas y Aplicadas en Estadística. Facultad de Ciencias Económicas y Estadística. Universidad Nacional de Rosario.

MODELOS MARGINALES PARA EL ESTUDIO DE LA DESOCUPACION[1]

1. INTRODUCCIÓN

Los datos que se analizan en este trabajo provienen de la muestra de la Encuesta Permanente de Hogares del Gran Buenos Aires, (EPH; GBA). Las fechas de esas encuestas son Mayo/Octubre de 1996, Mayo/Octubre de 1997.

Estos datos ya fueron analizados por este equipo de trabajo (Servy et al, 1999) mediante un estudio de cortes transversales. Dicho estudio tenía por objetivo explicar la desocupación mediante variables demográficas y socio-económicas a través del ajuste de un modelo de tipo "logit" en cada una de las ondas. Los coeficientes de las variables explicativas se calcularon independientemente para cada onda de la EPH. Sin embargo, los datos analizados en ondas diferentes no son probabilísticamente independientes, pues debido a la estructura de la muestra de la EPH, para dos ondas sucesivas cualesquiera, el 75% de las respuestas a cada pregunta de la EPH, son contestadas por los mismos individuos en las dos ocasiones. Sólo el 25% de la muestra se renueva de una onda para la siguiente, según el esquema rotativo de la EPH.

En los análisis transversales no se efectuaron tests para comparar los valores de coeficientes análogos obtenidos en diferentes ondas. Toda comparación entre ondas se realizó por inspección visual de las diferencias observadas. Sin embargo, hay métodos, de aparición reciente, que permiten un mejor aprovechamiento de los datos. Ellos permiten estimar los coeficientes de regresión de cada onda, haciendo uso, no sólo del conjunto de encuestas de la propia onda, sino del conjunto de las cuatro ondas. Además, es posible realizar tests estadísticos sobre la igualdad de coeficientes análogos a lo largo de las ondas.

Entre las metodologías que permiten una mejor utilización de los datos, merecen destacarse dos: la metodología de Grizzle, Starmmer y Koch (GSK) (1977), basada en los mínimos cuadrados ponderados, y la más reciente creada por Zeger y Liang (1986), que introducen las Ecuaciones de Estimación Generalizadas (GEE) como método de estimación.

La primera sólo admite como variables explicativas de las probabilidades de desocupación a variables de tipo categórico, que no varían con los tiempos sucesivos. Este método fue utilizado en el análisis de las mismas cuatro ondas de la EPH, por este mismo equipo de trabajo, en 1999. En ese trabajo el método GSK se adaptó para adecuarlo a la estructura compleja de la EPH.

[1] Trabajo realizado en el marco del Proyecto de Investigación y Desarrollo (Nº19/E069): "Estudio de datos categóricos medidos a través del tiempo". Secretaría de Ciencia y Tecnología de la UNR.

El método GEE, que es el método básico utilizado en este trabajo, sirve para el análisis de repuestas de tipo normal, Poisson, Binomial o Gamma. El método admite datos incompletos y las variables explicativas pueden ser categóricas o continuas y pueden variar con el tiempo o no.

La metodología ligada a las Ecuaciones de Estimación Generalizadas expresa una función conocida de la esperanza marginal de la variable dependiente como una función lineal de una o más variables independientes. Puede incluirse dentro de estas últimas a la variable que representa al tiempo y sus productos con las explicativas, de modo que, condicionalmente a los tiempos, las funciones de regresión pueden resultar diferentes de un tiempo a otro. A través de los coeficientes asociados con la variable tiempo y sus interacciones es posible realizar tests para detectar la influencia del tiempo sobre la relación entre la variable dependiente y las explicativas.

El método GEE supone que la variancia es una función conocida de la media y provee estimadores consistentes de los coeficientes de regresión y de sus variancias bajo suposiciones débiles sobre las correlaciones entre las respuestas de un mismo sujeto. La estructura de covariancia intra-sujeto se trata como un parámetro “nuissance”. Frente a la ignorancia (total o parcial) sobre dicho parámetro se lo reemplaza por una matriz impuesta por el investigador, que se denomina matriz de correlación de trabajo, y que no debe coincidir necesariamente con la verdadera.

El Apéndice 1 presenta una descripción de la metodología GEE. En la Sección 2 se describen los datos que se analizan en este trabajo; el detalle de las variables explicativas y sus codificaciones figura en el Apéndice 2. Se presenta, además, el modelo en su expresión general y comentarios sobre la interpretación de los coeficientes. En la Sección 3 se presentan los resultados hallados y por último los comentarios sobre el estudio realizado.

2. MATERIAL Y MÉTODOS

Los datos que se analizan son los correspondientes a las ondas de 1996 (Mayo y Octubre) y 1997 (Mayo y Octubre) de la EPH para el aglomerado Gran Buenos Aires.

El esquema de rotación de la encuesta impone que en cada onda salga de ella una submuestra, constituida por el 25% de los hogares, los que son reemplazados por un número equivalente de hogares elegidos en forma independiente, de modo que la muestra después de cuatro ondas es sustituida en su totalidad. Por lo tanto, para las cuatro ondas de 1996 y 1997, se pueden obtener paneles independientes de diferentes duraciones, medidos en distintas ocasiones, como lo indica el esquema siguiente:

Tabla 1: Esquema de rotaciones para el período mayo 96/octubre 97

Rotación	*Onda (t)*			
	Mayo 96 (1)	*Octubre 96 (2)*	*Mayo 97 (3)*	*Octubre 97 (4)*
1	X	X	X	X
2	X	X	X	-
3	-	X	X	X
4	X	X	-	-
5	-	-	X	X
6	X	-	-	-
7	-	-	-	X

Se toman en cuenta, entonces, las encuestas de aquellos individuos que fueron interrogados una, dos, tres o cuatro veces en el período bajo estudio.

En el análisis se consideran las encuestas que no poseen faltas en ningunas de las variables involucradas en el análisis. De esta manera el tamaño de la muestra depende de las variables incluidas en el modelo.

En el modelo se incluyen las mismas variables utilizadas en los análisis de los cortes transversales, a las que ahora se agregan las que identifican las ondas y algunos productos (interacciones) con las anteriores, particularmente de aquellas covariables que en el análisis transversal mostraron un comportamiento diferente de onda a onda.

Las variables reflejan los efectos de las ondas, el sexo, la edad, la escolaridad, el nivel de ingreso, la rama de la actividad y el tamaño de la empresa en que trabaja el individuo. La edad, escolaridad y sexo se consideran constantes a través del tiempo (se fija su valor en el registro de la primera onda) porque sus valores son permanentes o cambian poco a través del periodo bajo estudio. Las definiciones de las variables y sus codificaciones se muestran en el Apéndice 2.

El modelo marginal elegido para describir la desocupación es de tipo "logit". Es decir, si π_{it} es la probabilidad de desocupación para el sujeto *i*-ésimo en la onda *t*-ésima, un modelo posible e,s

$$log\left(\frac{\pi_{it}}{1-\pi_{it}}\right) = \beta_0 + \sum_{h=1}^{3}{}^{(onda)} \beta_h X_h + \beta_4 X_4 + \sum_{h=5}^{6}{}^{(sexo)} \beta_h X_h + \sum_{h=7}^{11}{}^{(edad)} \beta_h X_h + \sum_{h=12}^{13}{}^{(escolaridad)} \beta_h X_h + \sum_{h=14}^{19}{}^{(ingreso)} \beta_h X_h + \sum_{h=20}^{23}{}^{(rama)} \beta_h X_h + \sum_{h=24}^{25}{}^{(tamaño)} \beta_h X_h + \sum_{h=26}^{28}{}^{(sexo*edad)} \beta_h X_h + \sum{}^{(sexo*onda)} \beta_h X_h . \qquad (2.1)$$

Si en el modelo no figuran los términos que incluyen a la variable "onda" (efectos directos e interacciones) se está haciendo el supuesto de que una misma regresión logística describe las probabilidades de desocupación en cualquier onda del periodo 96/97.

Si sólo se incluyen los efectos directos del tiempo, se supone que los niveles de desocupación pueden variar con las ondas, pero no varía la influencia que cada variable explicativa ejerce sobre la probabilidad de desocupación.

Cuando se agregan las interacciones de la variable onda con las otras variables explicativas, existe la posibilidad de que éstas influyan sobre la probabilidad de desocupación en forma específica, según las diferentes ondas. Esta última afirmación puede ser ilustrada por el caso de las variables sexo y onda. En función de la codificación escogida (véase el Apéndice 2) en presencia de interacción entre sexo y onda los coeficientes que expresan los efectos principales o directos del sexo sobre la desocupación en las diferentes ondas son:

β_4 en la primera onda

$\beta_4 + \beta_{26}$ en la segunda onda

$\beta_4 + \beta_{27}$ en la tercera onda

$\beta_4 + \beta_{28}$ en la cuarta onda.

El test de H_0: $\beta_{26}=0$ sirve para detectar si el efecto del sexo es igual en la segunda que en la primera onda. Análogo significado tienen las hipótesis $\beta_{27}=0$ y $\beta_{28}=0$ con respecto a la tercera y cuarta onda.

El test de la hipótesis H_0: $\beta_{26}=\beta_{27}=\beta_{28}=0$ es un test de homogeneidad del efecto sexo a través de las ondas.

Además de poder interpretar los efectos de los diferentes factores socio-demográficos sobre la probabilidad de desocupación, la ecuación (2.1) permite estimar el "logit" que co-

rresponde a valores específicos de sexo, edad, escolaridad, nivel de ingreso, etc. La estimación del "logit" es c,

$$log \frac{\hat{\pi}}{1-\hat{\pi}} = c \,. \tag{2.2}$$

A partir de allí se puede estimar:

- la probabilidad de desocupación, $\hat{\pi} = \frac{1}{(e^{-c} + 1)}$; (2.3)

- el "odds", $\frac{\hat{\pi}}{1-\hat{\pi}} = e^{c}$, (2.4)

  que se interpreta como chance o riesgo de desocupación;

- la razón de "odds", $\hat{\theta} = \frac{\hat{\pi}_1 / 1-\hat{\pi}_1}{\hat{\pi}_2 / 1-\hat{\pi}_2} = e^{c_1 - c_2}$, (2.5)

  donde $\hat{\pi}_1$ y $\hat{\pi}_2$ son dos probabilidades calculadas para valores específicos de todas las covariables salvo una, por ejemplo rama de actividad, y $\frac{\hat{\pi}_1}{1-\hat{\pi}_1} = e^{c_1}$ y $\frac{\hat{\pi}_2}{1-\hat{\pi}_2} = e^{c_2}$ son los correspondientes "odds" que representan, por ejemplo, las chances de desocupación en la rama manufacturera y la construcción respectivamente.

  La razón de "odds" significa que la chance de desocupación en la rama manufacturera es $\hat{\theta}$ veces la correspondiente a la construcción. Este coeficiente varía entre 0 e ∞. El valor 1 indica que ambas probabilidades son iguales. Si es menor que 1 la rama manufacturera tiene menos probabilidad de desempleo y si es mayor que 1, la situación se invierte. Cuando la razón de "odds" estimada es menor que 1, puede resultar más claro interpretar su inversa, es decir $\frac{1}{e^{c_1 - c_2}} = e^{c_2 - c_1}$.

Por otra parte, si se explicitan c_2 y c_1 en términos del modelo, se encuentre que c_2-c_1= $\hat{\beta}_{14}$. Por lo tanto, $e^{\hat{\beta}_{14}}$ estima la razón de "odds" de desocupación entre ambas ramas de actividad.

Cuando la covariable está presente también en interacción resulta más conveniente su interpretación a través de la diferencia de "logits" c_2-c_1, la cual resulta ser una función de la variable con quien interactúa.

El ajuste de este modelo marginal (2.1) para la probabilidad de desocupación se realiza mediante el uso de la metodología conocida por Ecuación Generalizada de Estimación (GEE) que proporciona estimadores consistentes de los parámetros de regresión y de sus variancias y cuya descripción detallada es presentada en el Apéndice 1.

3. RESULTADOS

3.1. Ajuste e interpretación del modelo

Para el ajuste del modelo (2.1) a los datos de las cuatro ondas disponibles, se eliminaron los registros con valores faltantes en algunas variables, lo cual condujo a un total de 5796 individuos que según el número de ondas en los que fueron entrevistados dio lugar a 10404 mediciones u observaciones.

Los tests de score generalizados, definidos en el Apéndice (véase A.5), para cada efecto del modelo (2.1) se presentan en la Tabla 2.

Tabla 2: Estadísticas score generalizadas del modelo

Efecto	*gl*	*Chi-cuadrado*	*Prob. Asociada*
Sexo	1	0.35	0.5553
Edad	1	33.30	**0.0000**
Edad cuadrado	1	23.06	**0.0000**
Escolaridad	5	6.45	0.2652
Nivel de ingresos	2	272.91	**0.0000**
Rama de actividad	6	85.30	**0.0000**
Tamaño de la empresa	4	13.88	**0.0077**
Onda	3	8.77	**0.0326**
Sexo*edad	1	3.14	**0.0764**
Sexo*edad cuadrado	1	4.44	**0.0352**
Sexo*onda	3	6.49	**0.0900**

Salvo escolaridad, el resto de las variables e interacciones presentan considerable significación estadística ($p<0.10$).

En la Tabla 3 se presentan los estimadores de los parámetros β del modelo ajustado conjuntamente con su desvío estándar y la probabilidad asociada al test de Wald univariado para cada coeficiente (Apéndice 1, véase A.2, A.3, A.4).

Los resultados son consistentes con los arrojados por los tests de score, pero en la desagregación es posible observar por ejemplo que la interacción sexo*onda presenta sólo un coeficiente significativo (al 10%) cuando se compara el efecto del sexo en las ondas de mayo y octubre de 1996.

La interpretación de los coeficientes de los modelos “logit” es comúnmente presentada en términos de razones de “odds” de acuerdo a lo presentado en la Sección anterior. Estas razones de “odds” constituyen una medida aproximada de la magnitud del riesgo de desocupación ante diferentes escenarios definidos por las covariables consideradas.

Tabla 3: Coeficientes estimados, desvío estándar y probabilidad asociada al test de Wald

Parámetro	***Coeficiente***	***Error estándar***	***Prob. Asociada***
Constante	2.2917	0.3695	**0.0000**
Onda			
Octubre 96 (X_1)	-0.0504	0.1011	0.6183
Mayo 97 (X_2)	-0.2069	0.1100	**0.0600**
Octubre 97 (X_3)	-0.3173	0.1056	**0.0027**
Sexo			
Femenino (X_4)	-0.3399	0.5772	0.5559
Edad(X_5)	-0.1321	0.0175	**0.0000**
Edad Cuadrado(X_6)	0.0015	0.0002	**0.0000**
Escolaridad			
Primario completo (X_7)	-0.0744	0.1246	0.5507
Secundario incompleto (X_8)	0.0098	0.1386	0.9439
Secundario completo (X_9)	0.1981	0.1495	0.1850
Superior o universitario incompleto (X_{10})	0.0937	0.1839	0.6105
Superior o universitario completo (X_{11})	-0.1211	0.2300	0.5986
Nivel de Ingreso			
Medio(X_{12})	-0.8029	0.0772	**0.0000**
Alto(X_{13})	-2.1069	0.1389	**0.0000**
Rama de Actividad			
Manufactura(X_{14})	-0.8765	0.1313	**0.0000**
Servicios comerciales(X_{15})	-0.9913	0.1253	**0.0000**
Intermediación Financiera(X_{16})	-0.5779	0.1644	**0.0004**
Administración Pública y Defensa(X_{17})	-1.5425	0.3015	**0.0000**
Instrucción Pública(X_{18})	-1.1643	0.2106	**0.0000**
Otras Actividades de Servicios(X_{19})	-1.1404	0.1434	**0.0000**
Tamaño de la empresa			
2 a 5 personas(X_{20})	-0.2518	0.1000	**0.0118**
6 a 25 personas(X_{21})	-0.2432	0.1182	**0.0396**
26 a 100 personas(X_{22})	-0.2713	0.1280	**0.0341**
101 o más personas(X_{23})	-0.5035	0.1451	**0.0005**
Edad*Sexo(X_{24})	0.0551	0.0313	**0.0783**
Edad Cuadrado*Sexo(X_{25})	-0.0008	0.0004	**0.0379**
Sexo*onda			
Femenino* Octubre 96(X_{26})	-0.2630	0.1579	**0.0957**
Femenino* Mayo 97(X_{27})	0.0860	0.1703	0.6137
Femenino* Octubre 97(X_{28})	0.1803	0.1636	0.2705

Las estimaciones obtenidas fueron:

Rama de actividad

Para una onda dada y para valores fijos del sexo, la edad, la escolaridad, el nivel de ingreso y el tamaño de la empresa se estima que

- la chance de estar desocupado en la rama intermediación financiera es el 44% menor que en la rama de la construcción.
- la chance de estar desocupado en la rama manufacturera es el 58% menor que en la rama de la construcción.
- la chance de estar desocupado en la rama servicios comerciales y de transporte es el 63% menor que en la rama de la construcción.
- la chance de estar desocupado en la rama otras actividades de servicios es el 68% menor que en la rama de la construcción.

- la chance de estar desocupado en la rama instrucción pública y servicios de salud es el 69% menor que en la rama de la construcción.
- la chance de estar desocupado en la rama administración pública y defensa es el 78% menor que en la rama de la construcción.

Tamaño de la empresa

Para una onda dada y para valores fijos del sexo, la edad, la escolaridad, el nivel de ingreso y la rama de actividad se estima que

- la chance de estar desocupado si trabajó en una empresa de 2 a 5 personas es el 22% menor que si trabajó en una empresa unipersonal.
- la chance de estar desocupado si trabajó en una empresa de 6 a 25 personas es el 21% menor que si trabajó en una empresa unipersonal.
- la chance de estar desocupado si trabajó en una empresa de 26 a 100 personas es el 24% menor que si trabajó en una empresa unipersonal.
- la chance de estar desocupado si trabajó en una empresa de 101 o más personas es el 39% que si trabajó en una empresa unipersonal.

Nivel de ingreso

Para una onda dada y para valores fijos del sexo, la edad, la escolaridad, la rama de actividad y el tamaño de la empresa se estima que

- la chance de estar desocupado si se tiene un ingreso medio es el 55% menor que si se tiene un ingreso bajo.
- la chance de estar desocupado si se tiene un ingreso alto es el 88% menor que si se tiene un ingreso bajo.

Sexo

Debido a que existe interacción entre el sexo y la onda, no es pertinente hablar de la influencia que ejerce el sexo sobre la desocupación en general, sino a la influencia que ejerce el sexo sobre la desocupación cuando se cambia de una onda a otra.

Si se considera la diferencia de los "logits" de las probabilidades de desocupación entre mujeres y hombres cuando están fijos los valores de todas las variables, excepto los de sexo y la onda, se obtiene

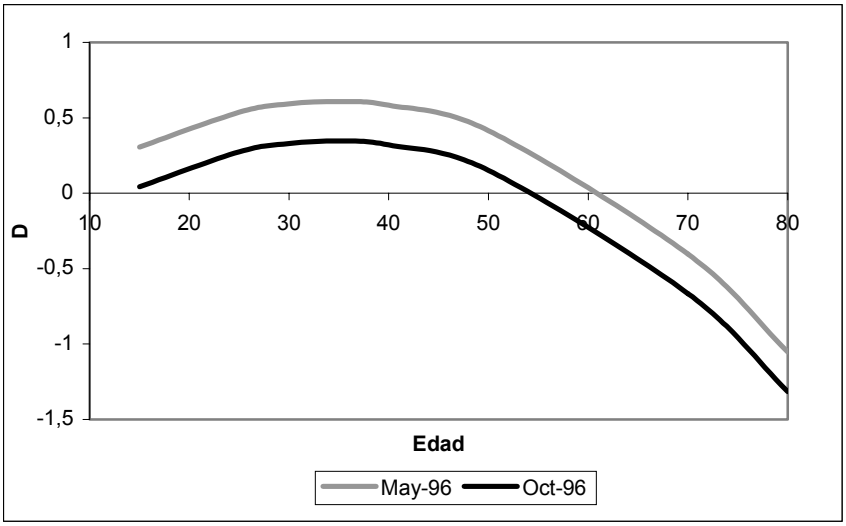
$$\left(log\frac{\pi}{1-\pi}\right)_{fem} - \left(log\frac{\pi}{1-\pi}\right)_{masc} = \beta_4 + \sum_{h=24}^{25}\beta_h X_h + \sum_{h=26}^{28}\beta_h X_h \,.$$

Esta expresión se reduce a $\beta_4 + \sum_{h=24}^{25}\beta_h X_h$ para la onda mayo 96 y a $\beta_4 + \sum_{h=24}^{25}\beta_h X_h + \beta_{26}$ para octubre 96.

Restando estas dos últimas expresiones se obtiene β_{26}. Representa el efecto de la onda octubre 96 comparada con mayo 96 sobre la diferencia entre el "logit" de desocupación de ambos sexos.

Además, para cada onda en particular el efecto del sexo depende de la edad del individuo. La Figura 1 muestra las diferencias entre los "logits" de desocupación para ambos sexos en las ondas con diferencias significativas (mayo 96, octubre 96).

Figura 1: Representación gráfica de la diferencia entre el "logit" de desocupación para sexo femenino y masculino (D)



De acuerdo a esta figura la desocupación de las mujeres supera a la de los hombres hasta los 62 años de edad para mayo 96, y hasta los 55 años para octubre 96, onda en la que las diferencias entre sexos cambian de signo, es decir, la desocupación de los hombres es superior a la de las mujeres. En cambio, en las ondas del año 1997 el análisis muestra que la desocupación tiene un comportamiento similar a mayo 96.

3.2. Diagnóstico

En el análisis de respuestas binarias correlacionadas utilizando la metodología GEE resulta problemático evaluar si el modelo es adecuado debido a que no existe una verosimilitud y los residuos para cada individuo están correlacionados.

Por tal razón se realiza un simple análisis de diagnóstico comparando las probabilidades de desocupación predichas con las proporciones observadas de desocupación. Las probabilidades predichas para cada sujeto se obtienen valorizando el modelo ajustado en los valores de las covariables correspondientes a cada individuo. En base a intervalos de estas probabilidades predichas se construye la distribución de frecuencias de los individuos. En la submuestra de individuos cuyas probabilidades estimadas caen en un intervalo dado, se observa la verdadera proporción de desocupados. Si el modelo es bueno es de esperar que esta proporción caiga dentro del intervalo de probabilidades predichas. La Tabla 4 incluye esa comparación en forma general y desagregada por onda. Expresa, por ejemplo, que en mayo 96 hubo 771 individuos cuyos valores predichos estaban entre 0.01 y 0.05. Para esos 771 individuos la frecuencia relativa observada de desocupación fue 0.0376. En general se visualiza que las observadas caen dentro del intervalo de los valores predichos. En las últimas categorías de las probabilidades predichas, las frecuencias observadas de desocupados no están incluidas en la categoría correspondiente, probablemente, porque las muestras en ellas son pequeñas.

Tabla 4: Distribución de frecuencias de los valores estimados por el modelo*

Prob. Predicha	***General***	***Mayo 96***	***Octubre 96***	***Mayo 97***	***Octubre 97***
Menor a 0.01	101(0.0000)	13 (0.0000)	20 (0.0000)	25 (0.0000)	43 (0.0000)
[0.01; 0.05)	3018 (0.0288)	771 (0.0376)	601 (0.0266)	637 (0.0173)	1009 (0.0307)
[0.05; 0.10)	2656 (0.0614)	690 (0.0623)	550 (0.0618)	609 (0.0624)	807 (0.0595)
[0.10; 0.15)	1751 (0.1314)	502 (0.1215)	332 (0.1295)	377 (0.1379)	540 (0.13709)
[0.15; 0.20)	1170 (0.1624)	359 (0.1727)	262 (0.1450)	243 (0.1564)	306 (0.1699)
[0.20; 0.25)	691 (0.2127)	249 (0.2329)	119 (0.2101)	121 (0.1983)	202 (0.1980)
[0.25; 0.30)	418 (0.2967)	129 (0.2558)	76 (0.3553)	88 (0.2273)	125 (0.3520)
[0.30; 0.35)	277 (0.3321)	96 (0.2500)	55 (0.7273)	63 (0.4286)	63 (0.3651)
[0.35; 0.40)	173 (0.4220)	80 (0.4125)	28 (0.5000)	28 (0.3929)	37 (0.4054)
[0.40; 0.45)	80 (0.5125)	47 (0.5106)	11 (0.6364)	11 (0.4545)	11 (0.4545)
[0.45; 0.50)	34 (0.4706)	14 (0.6429)	5 (0.2000)	6 (0.5000)	9 (0.3333)
[0.50; 0.55)	23 (0.3913)	6 (0.6667)	5 (0.2000)	5 (0.4000)	79 (0.2857)
[0.55; 0.60)	8 (0.2500)	3 (0.3333)	3 (0.3333)	2 (0.0000)	0
[0.60; 0.65)	4 (0.0000)	3 (0.0000)	1 (0.0000)	0	0
Total	10404 (0.1128)	2962 (0.1286)	2068 (0.1088)	2215 (0.1043)	3159 (0.1067)

*Los números entre paréntesis muestran la proporción real de desocupación observada para las muestras correspondientes a cada intervalo de valores predichos.

Otra medida que se suele emplear para evaluar el modelo es la correlación, R, entre la respuesta binaria y las probabilidades predichas por el modelo. Para tales modelos, R es un índice crudo del poder predictivo, aunque no tiene propiedades tan buenas como en el caso de trabajar con variables normales. Por ejemplo, no se garantiza que R sea no decreciente a medida que el modelo es más complejo. Además, como cualquier otra medida de asociación, su valor puede depender fuertemente del rango de los valores observados de las variables explicativas. Sin embargo, es muy útil para comparar los ajustes de diferentes modelos ajustados al mismo conjunto de datos.

Si bien R resulta informativo cuando se intenta elegir entre varios modelos, se presentan en la Tabla 5 las correlaciones entre las respuestas observadas y las probabilidades estimadas para el modelo presentado anteriormente con fines ilustrativos.

Tabla 5: Correlación entre el valor de la variable respuesta observado y el predicho por el modelo

Grupo	***R***
General	0.30861
Mayo 96	0.31048
Octubre 96	0.30848
Mayo 97	0.30520
Octubre 97	0.30453

4. CONCLUSIONES

El análisis realizado muestra que los efectos de las variables bajo estudio sobre la desocupación no dependen de la onda particular que se considere, excepto en el caso de la variable sexo donde se observa que octubre de 1996 es un período más favorable a la mujer que el resto de los períodos bajo estudio. Las otras ondas tienen un comportamiento homogéneo con respecto a la diferencia en la desocupación del hombre y la mujer. Además, para cualesquiera de las ondas el signo de la diferencia entre los sexos depende de la edad de los trabajadores. Hasta los 55 (octubre 1996) o 60 años (resto del período), la desocupación

en las mujeres es mayor que en los hombres. A partir de allí la relación se invierte, probablemente debido al hecho de que la mujer abandona más tempranamente la población económicamente activa. Este cambio de signo de la diferencia entre los sexos explica la falta de significación del coeficiente β_4 (efecto directo del sexo), ya que él mide el efecto promedio de esa diferencia.

Con respecto a la influencia de la rama de actividad, tamaño de la empresa y nivel de ingreso del individuo, se puede enunciar lo siguiente sin diferenciación por onda.

En cuanto a los efectos de la rama de actividad sobre la desocupación, la construcción es la presenta mayor probabilidad y en el otro extremo está la administración pública que es la que acusa menor desocupación. La intermediación financiera y la industria manufacturera tienen altas probabilidades de desocupación, y en menor medida los servicios comerciales y de transporte y la instrucción pública.

En cuanto al tamaño de las empresas las más vulnerables a la desocupación son las unipersonales y en general, los mayores tamaños están asociados con menores probabilidades de desocupación.

El nivel de ingreso es un factor que tiene influencia sobre la probabilidad de desocupación. Los niveles superiores presentan menores probabilidades de desocupación.

La escolaridad no parece influir sobre la probabilidad de desocupación. Esta aparente paradoja se debe a que esta conclusión, como todas las anteriores están referidas a escenarios donde todas la variables, bajo estudio, excepto las específicamente comentadas, se consideran fijas.

REFERENCIAS

AGRESTI, A. (1989): A survey of models for repeated ordered categorical response data. Statistics in Medicine, 8, 1209-1224.

AGRESTI, A. (1996): An Introduction to Categorical Data Analysis. Wiley Series in Probability and statistics. John Wiley & Sons, INC.

AGRESTI, A. (1990): Categorical Data Analysis. John Wiley & Sons.

ALLISON, P. (1999): Logistic Regresion Using the SAS System: Theory and Application. SAS Institute.

BINDER, D. (1982): On the Variances of Asymptotically Normal Estimators from Complex Surveys.

COX, D. R. (1983): Some remarks on overdispersion. Biometrika, 70, 269-274.

DAVIDIAN, M. (1998): Applied Longitudinal Data Analysis. Lecture Notes.

DIGGLE, P.; LIANG, K.; ZEGER, S. (1994): Analysis of Longitudinal Data. Oxford Statistical Science Series 13.

Encuesta Permanente de Hogares. Manual de instrucciones. INDEC.

LIANG, K.; ZEGER, S. (1986): Longitudinal data analysis using generalized linear models. Biometrika, 73, 1, 31-22.

ROTNITZKY, A.; JEWELL, N. (1990): Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. Biometrika, 77, 3, 485-497.

SAS Institute Inc., Cary, NC, USA. (1997) SAS/STAT Software: Changes and enhancements through release 6.12.

SAS Institute Inc. SAS Online (TM). Version 7-1. Copyright 1999, SAS Institute Inc.

SERVY, E., HACHUEL, L., BOGGIO, G., CUESTA, C, LEONE, G. (1999): Modelos estadísticos para el estudio de la desocupación. Parte I: modelos para cortes transversales

SERVY, E., HACHUEL, L., BOGGIO, G., CUESTA, C, LEONE, G. (2000): Adaptación de la metodología GSK a muestras de panel rotativas para modelos de respuesta binaria repetida.

STOKES, M., SAS Institute Inc.; Cary, NC; La Vange, L., Quintiles Inc., RTP, NC (1998): Applications of GEE Methodology Using the SAS System. SAS Chapter Workshop, Research Triangle Park.

STOKES, M.; DAVIS, C.; KOCH, G. (1995): Categorical Data Analysis Using the SAS System. SAS Institute INC.

AGRADECIMIENTOS

Se agradece al INDEC la provisión de las bases de datos y en especial a la Estadística Clyde Charre de Trabuchi por su asesoramiento en relación a la interpretación de las mismas.

APÉNDICE 1

ECUACIÓN DE ESTIMACIÓN GENERALIZADA

La ecuación de estimación generalizada (GEE) (Liang, K. y Zeger, S., 1986) modela una función conocida de la esperanza marginal de la variable dependiente como una función lineal de una o más variables explicativas. El modelo describe cómo los promedios a través de la población de respuestas en diferentes puntos del tiempo se relacionan con el tiempo y con las covariables adicionales, que pueden ser tanto discretas como continuas, dependientes o no del tiempo.

La metodología GEE posee la propiedad de estimar consistentemente los coeficientes de regresión y sus variancias bajo suposiciones débiles sobre la correlación entre las observaciones de los sujetos. Evita la necesidad de distribuciones multivariadas asumiendo una forma funcional para la distribución marginal en cada punto del tiempo.

Notación y estimación

Para una fácil notación, considérese la situación en la cual se obtienen medidas repetidas para cada uno de n sujetos en t puntos del tiempo (t podría diferir de sujeto a sujeto y entonces denominarse t_i, sin embargo para simplificar la notación se lo llama t). Aunque esta notación es más natural para los estudios longitudinales, donde t denota el número de ocasiones bajo las cuales se obtienen las mediciones dependientes para cada sujeto, también se usa para el caso general de respuestas correlacionadas, donde cada individuo es un "cluster" con a lo sumo t unidades experimentales.

Sea entonces y_{ij} la respuesta para el sujeto i en el tiempo j, para i=1,..., n y j=1,..., t. Se asume que los vectores de datos $\mathbf{y}_i=(y_{i1}, ..., y_{it})'$ son independientes a través de las unidades individuales. Aunque se puede usar la metodología GEE para respuestas distribuidas de diversas formas (por ejemplo, gamma y normal), se restringe la atención a situaciones en las cuales y_{ij} es una respuesta binaria (0,1). Sea $\mathbf{x}_{ij}=(x_{ij1}, ..., x_{ijp})'$ el vector px1 de variables expli-

cativas (covariables) asociadas con y_{ij}. Si todas las covariables son independientes del tiempo, $x_{i1}=x_{i2}=...=x_{it}$. Si no se obtuvieron observaciones para el tiempo *j*, y_{ij} y x_{ij} serán observaciones faltantes.

El *primer paso* de la GEE es relacionar la respuesta marginal $\mu_{ij}=E(y_{ij})$ (para respuestas binarias $\mu_{ij}=\pi_{ij}$) con una combinación lineal de las covariables: $h(\mu_{ij})=\mathbf{x}_{ij}'\boldsymbol{\beta}$, donde $\boldsymbol{\beta}_{px1}=(\beta_1, ..., \beta_p)'$ es el vector de parámetros desconocidos que caracteriza cómo la distribución de respuesta del corte transversal depende de las variables explicativas y *h* es una función conocida. La función *h* de uso más frecuente para respuestas dicotómicas es la función "logit": $h(\pi_{ij})=log(\pi_{ij}/(1-\pi_{ij}))$.

El *segundo paso* es describir la variancia de y_{ij} como una función de la media: $var(y_{ij})=v(\mu_{ij})\phi$, donde *v* es una función de variancia conocida y ϕ es un parámetro de escala posiblemente desconocido. Al especificar la variancia de y_{ij}, se modela la variancia dentro y entre unidades. La forma en que la variancia está relacionada con la media depende del tipo de datos. Por ejemplo, para respuestas binarias, $v(\mu_{ij})=v(\pi_{ij})=\pi_{ij}(1-\pi_{ij})\phi$; luego se espera que ϕ sea igual a 1.

El *tercer paso* es modelar la correlación $\mathbf{R}_i(\boldsymbol{\alpha})_{txt}$ entre las observaciones para cada $\mathbf{y}_i$. El elemento (*j, j'*) de $\mathbf{R}_i(\boldsymbol{\alpha})$ es la correlación conocida, en forma hipotética, entre y_{ij} e $y_{ij'}$, es decir, el modelo describe la correlación entre pares de observaciones del mismo vector de datos. Además, representa toda la variación que podría deberse a la correlación que surge por la naturaleza de recolección de datos, ya que las observaciones sobre la misma unidad suelen ser más semejantes que las de diferentes unidades.

Se debe asumir que $\mathbf{R}_i(\boldsymbol{\alpha})$ es conocida excepto por el número fijo de parámetros $\boldsymbol{\alpha}$, que deben ser estimados de los datos. Aunque esta matriz de correlación puede diferir de sujeto a sujeto, comúnmente se usa una matriz de correlación de trabajo $\mathbf{R}(\boldsymbol{\alpha})$ que representa aproximadamente la dependencia promedio entre las observaciones repetidas de los sujetos. Por simplicidad en la escritura, se considera $\mathbf{R}(\boldsymbol{\alpha})=\mathbf{R}$.

Con respuestas no normales, la correlación entre las respuestas de los sujetos podría depender de los valores medios y, por lo tanto, de $\mathbf{x}_{ij}'\boldsymbol{\beta}$. Pero, como se pretende capturar la variancia y la correlación de todos los factores con un modelo simple, se considera a **R** como una matriz de trabajo y no la que representa necesariamente la verdad; más aún, al no haber un método formal para chequear la elección realizada.

El método GEE provee estimadores consistentes de los coeficientes de regresión y de sus variancias, aún cuando las especificaciones de la estructura de la matriz de covariancias no sean ciertas. Sin embargo, la eficiencia de los estimadores crece cuando la elección de **R** se acerca a la correlación verdadera. Cuando el número de sujetos es grande, no se pierde eficiencia como consecuencia de una elección incorrecta de **R**.

Se han sugerido muchas posibilidades para la estructura de correlación de trabajo. Primero, cuando el número de sujetos es grande en relación al número de observaciones por sujeto, la influencia de la correlación es lo suficientemente pequeña como para que los coeficientes de regresión mínimo cuadrático ordinarios sean bastante eficientes. Sin embargo, las correlaciones entre medidas repetidas tendrían un efecto sustancial sobre las variancias estimadas de los parámetros y por lo tanto deben tomarse en cuenta para hacer inferencias correctas. El modelo de trabajo de *independencia*, con **R** igual a la matriz identidad **R=I,** adopta la suposición que las observaciones repetidas para un sujeto son independientes:

$$\mathbf{R} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix}.$$

En este caso, aplicar la GEE es lo mismo que ajustar el modelo de regresión usual para datos independientes, y puede ser resuelto usando paquetes de software estándares.

El modelo de correlación de trabajo *constante* ("exchangeable") asume que la correlación entre dos observaciones cualesquiera de la misma unidad es fija, es decir, $R_{jj'}=\rho$, para $j \neq j'$. Esta estructura de correlación podría ser escrita en términos de un solo parámetro de correlación simple $0<\rho<1$ ($\alpha=\rho$):

$$\mathbf{R}=\begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \dots & \dots & \dots & \dots \\ \rho & \dots & \rho & 1 \end{pmatrix}.$$

En principio, este modelo se usaría con datos balanceados, idealmente balanceados con datos perdidos, y datos desbalanceados donde las distintas unidades se han medido en puntos diferentes del tiempo. Aunque la suposición de correlación constante entre dos medidas repetidas cualesquiera parecería difícil de verificar en un estudio longitudinal, puede ser razonable en situaciones en las cuales las medidas repetidas no se obtienen sobre el tiempo.

Se permite un número arbitrario de observaciones por sujeto tanto para una estructura de correlación de trabajo de independencia como constante.

Cuando la matriz de correlación no presenta ninguna especificación, hay $t(t-1)/2$ parámetros para ser estimados. En esta situación se obtienen los estimadores más eficientes, pero sólo es útil de aplicar cuando el número de mediciones repetidas es pequeño. Luego, si y_{ij} e $y_{ij'}$, $j, j'=1, \dots, t$, son dos observaciones de la misma unidad donde todas las unidades son observadas en los mismos t tiempos, y $\rho_{jj'}$ representa la correlación entre y_{ij} e $y_{ij'}$, resulta $\rho_{jj}=1$ si $j=j'$ y es $-1 \leq \rho_{jj'} \leq 1$ si $j \neq j'$; la matriz de correlación de trabajo *no estructurada* es:

$$\mathbf{R}=\begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1t} \\ \rho_{21} & 1 & \dots & \rho_{2t} \\ \dots & \dots & \dots & \dots \\ \rho_{t1} & \dots & \rho_{t,t-1} & 1 \end{pmatrix},$$

donde $\rho_{jk}=\rho_{kj}$ para todo j, k ($\alpha=(\rho_{12}, \dots, \rho_{1t, \dots,} \rho_{t,t-1})$).

La estructura de correlación *autorregresiva AR(1)* supone que la correlación entre observaciones decrece con el tiempo: $R_{j,j+d}=\rho^d$ $d=0, 1, 2, \dots, t_{i-j}$, con $-1<\rho<1$ ($\alpha=\rho$). La matriz de correlación de trabajo es

$$\mathbf{R}=\begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{t-1} \\ \rho & 1 & \rho & \dots & \rho^{t-2} \\ \dots & \dots & \dots & \dots & \dots \\ \rho^{t-1} & \dots & \rho^2 & \rho & 1 \end{pmatrix}.$$

En principio este modelo podría usarse en cualquier situación; sin embargo para datos desbalanceados observados en diferentes puntos del tiempo podría no tener sentido.

Finalmente, el modelo de correlación de trabajo *m-dependiente* supone

$$\mathbf{R}_{j,j+d} = \begin{cases} 1, d = 0 \\ \rho_d, d = 1,2,...,m \\ 0, d > m \end{cases}.$$

Por ejemplo, cuando m=1 sólo las observaciones adyacentes en el tiempo están correlacionadas por la misma cantidad $-1<\rho<1$ ($\alpha=\rho$). En principio, este modelo también podría ser usado en cualquier situación, pero para datos desbalanceados medidos en puntos de tiempo diferentes podría no tener sentido.

$$\mathbf{R}=\begin{pmatrix} 1 & \rho & 0 & ... & 0 \\ \rho & 1 & \rho & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & ... & 0 & \rho & 1 \end{pmatrix}.$$

En este momento, se cuenta con un modelo para la variancia de y_{ij} (del segundo paso $var(y_{ij})=v(\mu_{ij})\phi$) y con otro para la correlación entre individuos ($\mathbf{R}_i(\boldsymbol{\alpha})$). A partir de la combinación de ellos surge un modelo para la matriz de covariancias del individuo i-ésimo, mejor dicho, del vector de datos $\boldsymbol{y}_i$: $\mathbf{V}_i(\boldsymbol{\alpha})=\phi\mathbf{A}_i^{1/2}\mathbf{R}_i(\boldsymbol{\alpha})\mathbf{A}_i^{1/2}$, donde $\mathbf{A}_i$ es la matriz diagonal txt con $v(\mu_{ij})$ como elemento diagonal j-ésimo para la unidad i.

El *cuarto paso* de la GEE es estimar el vector de parámetros $\boldsymbol{\beta}$ y su matriz de covariancias.

Dado que las consideraciones hechas no son suficientes para especificar completamente la distribución de probabilidad multivariada apropiada, no es posible apelar al principio de máxima verosimilitud para desarrollar una estructura de estimación y prueba.

Una aproximación natural para ajustar el modelo $h(\mu_{ij})=\mathbf{x}_{ij}'\boldsymbol{\beta}$, inspirada en las ecuaciones de estimación de los modelos lineales generalizados (GLM), es resolver una ecuación de estimación consistente de p ecuaciones que es función lineal de los desvíos $\boldsymbol{y}_i-\boldsymbol{\mu}_i$, y que pondera estos desvíos usando la inversa de la matriz de covariancias de trabajo asumida para $\boldsymbol{y}_i$. La ecuación de estimación generalizada es, entonces,

$$U(\beta) = \sum_{i=1}^{n}\left(\frac{\partial\mu_i}{\partial\beta}\right)'[\mathbf{V}_i(\alpha)]^{-1}(\boldsymbol{y}_i - \mu_i) = \mathbf{0}_p, \qquad \text{(A.1)}$$

donde $\boldsymbol{\mu}_i=(\mu_{i1}, ..., \mu_{it})'$ y $\mathbf{0}_p$ es el vector (0, ..., 0)' de dimensión px1.

De aquí surge que si todas las observaciones fueran independientes ($\mathbf{R}=\mathbf{I}$) esta ecuación se reduce a las ecuaciones de estimación bajo independencia. Para cada i,

$$U_i(\beta) = \left(\frac{\partial\mu_i}{\partial\beta}\right)'[\mathbf{V}_i(\alpha)]^{-1}(\boldsymbol{y}_i - \mu_i)$$

es equivalente a las funciones derivadas de la cuasi-verosimilitud, excepto que los $\mathbf{V}_i$ además de ser funciones de $\boldsymbol{\beta}$ son también funciones de $\boldsymbol{\alpha}$.

La estimación para $\boldsymbol{\beta}$ es la solución de la ecuación de estimación (A.1) que se resuelve a partir de un algoritmo numérico. El método iterativo que se sugiere para estimar $\boldsymbol{\beta}$ es:

$$\hat{\beta}^{k+1} = \hat{\beta}^{k} - \left(\sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \hat{\beta}} \right)' \left[\tilde{\mathbf{V}}_i(\hat{\alpha}) \right]^{-1} \left(\frac{\partial \mu_i}{\partial \hat{\beta}} \right) \right)^{-1} \left(\sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \hat{\beta}} \right)' \left[\tilde{\mathbf{V}}_i(\hat{\alpha}) \right]^{-1} (y_i - \mu_i) \right) \qquad \text{(A.2)}$$

donde $\tilde{\mathbf{V}}_i(\hat{\alpha}) = \mathbf{V}_i\left(\beta, \hat{\alpha}(\beta, \hat{\phi})\right)$.

Surge claramente que para estimar $\boldsymbol{\beta}$ se requiere un estimador de $\boldsymbol{\alpha}$. Una forma, propuesta originalmente por Liang y Zeger (1986), es basar la estimación sobre funciones apropiadas de los desvíos $\boldsymbol{y}_i - \hat{\mu}_i$. Para iniciar el proceso se ajusta el modelo sobre los *n* individuos, asumiendo independencia entre todas las observaciones, y se obtienen los estimadores usando las técnicas de GLM. Luego, se usan esos estimadores para formar los desvíos y estimar $\boldsymbol{\alpha}$.

Concretamente, para una iteración dada, sea

$$\hat{e}_{ij} = \frac{y_{ij} - \hat{\mu}_{ij}}{\sqrt{v(\hat{\mu}_{ij})}}$$

la desviación correspondiente a la observación *j* del sujeto *i* dividido por el estimador de su desvío estándar. Luego el parámetro de dispersión se estima por

$$\hat{\phi} = \frac{1}{(\mathrm{N} - p)} \sum_{i=1}^{n} \sum_{j=1}^{t} \frac{\{y_{ij} - \hat{\mu}_{ij}\}^2}{v(\hat{\mu}_{ij})} = \frac{1}{(\mathrm{N} - p)} \sum_{i=1}^{n} \sum_{j=1}^{t} \hat{e}_{ij}^2 ,$$

donde N es el número total de observaciones.

El estimador de $\boldsymbol{\alpha}$ depende de la elección de $\mathbf{R}_i(\boldsymbol{\alpha})$. Si **R** corresponde a la suposición de correlación constante, el parámetro simple ρ se estima por

$$\hat{\rho} = \frac{1}{\left[\left(\sum_{i=1}^{n} t_i (t_i - 1) \right) - p \right] \hat{\phi}} \sum_{i=1}^{n} \sum_{j \neq k} \hat{e}_{ij} \hat{e}_{i,k} ;$$

y si **R** corresponde a la suposición de correlación no estructurada, ρ_{jk} se estima por

$$\hat{\rho}_{jk} = \frac{1}{(n - p)\hat{\phi}} \sum_{i=1}^{n} \hat{e}_{ij} \hat{e}_{ik} .$$

Si en cambio, se supone la estructura de correlación autorregresiva,

$$\hat{\rho} = \frac{1}{\left[\left(\sum_{i=1}^{n} (t_i - 1) \right) - p \right] \hat{\phi}} \sum_{i=1}^{n} \sum_{j \leq t_i - 1} \hat{e}_{ij} \hat{e}_{i,j+1} ;$$

y si la estructura de correlación supuesta es la m-dependiente

$$\hat{\rho}_d = \frac{1}{\left[\left(\sum_{i=1}^{n} (t_i - d \right) - p \right] \hat{\phi}} \sum_{i=1}^{n} \sum_{j \leq t_i - d} \hat{e}_{ij} \hat{e}_{i,j+d} .$$

Sólo para el caso de una estructura de correlación AR(1) es necesario calcular $\hat{\phi}$ para determinar el estimador de $\boldsymbol{\beta}$ y de su variancia.

En resumen, el esquema de estimación sería:

1. Calcular un estimador inicial de $\boldsymbol{\beta}$, asumiendo que todas las observaciones son independientes, a través de la aplicación del método de mínimos cuadrados reponderados usado en los GLM.
2. Calcular la correlación de trabajo $\mathbf{R}_i(\boldsymbol{\alpha})$ basada en los residuos estandardizados y en la estructura asumida de $\mathbf{R}_i(\boldsymbol{\alpha})$.
3. Calcular un estimador de la covariancia $\mathbf{V}_i(\boldsymbol{\alpha})=\phi\mathbf{A}_i^{1/2}\mathbf{R}_i(\boldsymbol{\alpha})\mathbf{A}_i^{1/2}$.
4. Actualizar el valor de $\boldsymbol{\beta}$ usando la extensión de mínimos cuadrados reponderados presentada en (A.2).
5. Repetir los pasos hasta conseguir convergencia.

Es importante notar que al no utilizar el método de máxima verosimilitud para estimar los parámetros del modelo, no es posible obtener criterios que permitan comparar distintas matrices de correlación de trabajo para determinar cuál suposición es más apropiada.

La matriz de covariancias estimada de $\hat{\boldsymbol{\beta}}$, por su parte, es

$$\hat{\mathbf{V}}_\beta = \left(\sum_{i=1}^{n} \left(\frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \boldsymbol{\beta}} \right)' [\mathbf{V}_i(\hat{\boldsymbol{\alpha}})]^{-1} \left(\frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \boldsymbol{\beta}} \right) \right)^{-1}.$$

Si la suposición sobre las correlaciones es incorrecta la validez de las inferencias sobre los β podría estar comprometida. Una solución a este problema es modificar la matriz de covariancias estimada $\hat{\mathbf{V}}_\beta$ para que una incorrecta elección de **R** no ocasione problemas. La versión modificada de $\hat{\mathbf{V}}_\beta$ es

$$\hat{\mathbf{V}}_\beta^R = \hat{\mathbf{V}}_\beta \mathbf{M} \hat{\mathbf{V}}_\beta, \tag{A.3}$$

donde

$$\mathbf{M} = \sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)' [\mathbf{V}_i(\hat{\boldsymbol{\alpha}}_i)]^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i)(\boldsymbol{y}_i - \boldsymbol{\mu}_i)' [\mathbf{V}_i(\hat{\boldsymbol{\alpha}}_i)]^{-1} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right).$$

Para *n* grande $\hat{\mathbf{V}}_\beta{}^R$ dará siempre un estimador razonable de la verdadera matriz de covariancia muestral de $\hat{\boldsymbol{\beta}}$, más allá de que la elección de **R** sea incorrecta. La consistencia de $\hat{\boldsymbol{\beta}}$ y de $\hat{\mathbf{V}}_\beta{}^R$ depende solamente de la especificación correcta de la media, no de la elección correcta de **R**. Sin embargo, no ocurre lo mismo con $\hat{\mathbf{V}}_\beta$.

$\hat{\mathbf{V}}_\beta$ es llamada matriz de covariancias estimada basada en el modelo, mientras que $\hat{\mathbf{V}}_\beta{}^R$ se denomina frecuentemente matriz de covariancias estimada empírica o robusta, ya que es "robusta" a que **R** sea incorrecta.

La decisión de usar la estimación basada en el modelo $\hat{\mathbf{V}}_\beta$ o la estimación robusta $\hat{\mathbf{V}}_\beta{}^R$ es personal. No existe consenso sobre cual se prefiere en muestras finitas. Si son muy diferentes, algunos lo toman como una indicación de que la suposición original es incorrecta. Por otro lado, si uno o más de los vectores $\boldsymbol{y}_i$ contiene valores inusuales sería suficiente para restar validez al estimador $\hat{\mathbf{V}}_\beta{}^R$. Por ello no hay una regla, y no se recomienda cual usar.

Distribución muestral y test de hipótesis

Para obtener una aproximación de la distribución muestral del estimador de $\boldsymbol{\beta}$ obtenido por el método GEE se debe apelar a la teoría de las muestras grandes. El término "muestras grandes" se refiere al número de unidades experimentales *n* (sujetos).

Para *n* grande, el estimador GEE para $\boldsymbol{\beta}$ satisface (Liang, K. y Zeger, S., 1986)

$$\hat{\boldsymbol{\beta}} \sim \mathrm{N}\left\{\boldsymbol{\beta}, \left(\sum_{i=1}^{n}\left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)' [\mathbf{V}_i(\boldsymbol{\alpha})]^{-1}\left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)\right)^{-1}\right\}.$$

Para probar una hipótesis nula de la forma H_0: $\mathbf{L}\boldsymbol{\beta}=\mathbf{d}$, donde **L** es una matriz *r*x*p*, se puede utilizar el procedimiento del test de Wald. Para muestras grandes

$$\mathbf{L}\hat{\boldsymbol{\beta}} \sim \mathrm{N}\left(\mathbf{L}\boldsymbol{\beta}, \mathbf{L}\hat{\mathbf{V}}_\beta\mathbf{L}'\right).$$

Luego, para un **L** general la estadística del test χ^2 de Wald es

$$\left(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{d}\right)'\left(\mathbf{L}\hat{\mathbf{V}}_\beta\mathbf{L}'\right)^{-1}\left(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{d}\right) \qquad \text{(A.4)}$$

y se compara con el apropiado valor crítico χ^2 con grados de libertad igual al número *r* de filas de **L**.

Por otra parte, Boss (1992) y Rotnitzky y Jewell (1990) describieron una generalización del test score aplicable para probar esa hipótesis.

Sean $\tilde{\boldsymbol{\beta}}$ los parámetros de regresión resultantes de resolver las GEE bajo las restricciones del modelo $\mathbf{L}\boldsymbol{\beta}=\mathbf{0}$, y sea $U(\tilde{\boldsymbol{\beta}})$ la ecuación de estimación generalizada evaluada en $\tilde{\boldsymbol{\beta}}$.

La estadística score generalizada es

$$\mathbf{T} = U(\tilde{\boldsymbol{\beta}})'\hat{\mathbf{V}}_\beta^{-1}\mathbf{L}'(\mathbf{L}\hat{\mathbf{V}}_\beta^{R}\mathbf{L}')^{-1}\mathbf{L}\hat{\mathbf{V}}_\beta^{-1}U(\tilde{\boldsymbol{\beta}}) \qquad \text{(A.5)}$$

El valor de la probabilidad asociada a la estadística **T** se calcula en base a la distribución χ^2 con *r* grados de libertad.

APÉNDICE 2

Codificaciones utilizadas en los modelos.

- *Onda:* X_1: 1 si es octubre 96, 0 en otro caso; X_2: 1 si es mayo 97, 0 en otro caso; X_3: 1 si es octubre 97, 0 en otro caso.
- *Sexo:* X_4: 1 si es de sexo femenino, 0 en otro caso.
- *Edad:* X_5: variable continua medida en años; X_6: $X_5{}^2$ edad al cuadrado.

- *Escolaridad:* X_7: 1 si posee primaria completa, 0 en otro caso; X_8: 1 si posee secundaria incompleta, 0 en otro caso; X_9: 1 si posee secundaria completa, 0 en otro caso; X_{10}: 1 si posee nivel superior o universitario incompleto, 0 en otro caso; X_{11}: 1 si posee nivel superior o universitario completo, 0 en otro caso.
- *Niel de ingreso:* X_{12}: 1 si es medio, 0 en otro caso; X_{13}: 1 si es alto, 0 en otro caso.
- *Rama de actividad:* X_{14}: 1 si pertenece a la industria manufacturera, 0 en otro caso; X_{15}: 1 si pertenece al sector servicios comerciales y de transporte, 0 en otro caso; X_{16}: 1 si pertenece al sector de intermediación financiera, 0 en otro caso; X_{17}: 1 si pertenece a la administración pública o defensa, 0 en otro caso; X_{18}: 1 si pertenece a la instrucción pública o servicios de salud, 0 en otro caso; X_{19}: 1 si pertenece a otras actividades de servicios, 0 en otro caso.
- *Tamaño de la empresa:* X_{20}: 1 si la empresa tiene de 2 a 5 personas, 0 en otro caso; X_{21}: 1 si la empresa tiene de 6 a 25 personas, 0 en otro caso; X_{22}: 1 si la empresa tiene de 26 a 100 personas, 0 en otro caso; X_{23}: 1 si la empresa tiene 101 o más personas, 0 en otro caso.
- *Interacción entre Sexo y Edad*: $X_{24}= X_4 * X_5$; $X_{25}= X_4 * X_5^2$
- *Interacción entre Onda y Sexo*: $X_{26}= X_1 * X_4$; $X_{27}= X_2 * X_4$; $X_{28}= X_3 * X_4$.

  Debe aclararse que para cada persona entrevistada, la cantidad de observaciones asociada con ella es igual al número de veces en que respondió a la encuesta.