Garcia, María del Carmen
Blaconá, María Teresa

Instituto de Investigaciones Teóricas y Aplicadas, de la Escuela de Estadística

METODOS DE ESTIMACION PARA DATOS LONGITUDINALES CON INFORMACIÓN FALTANTE

1.- INTRODUCCIÓN

Los datos longitudinales en los que cada sujeto o unidad experimental se mide u observa en ocasiones múltiples es probable que posean muchos datos perdidos, debido a que los sujetos no completan el estudio o salen antes de que el mismo finalice.
En los últimos años hubo un gran crecimiento en la literatura de métodos para analizar la falta de respuesta en los estudios longitudinales. Este interés por el estudio de la no respuesta refleja el reconocimiento que los sujetos pueden dejar un estudio longitudinal sobre las bases de características no medidas que pueden estar correlacionadas con los resultados bajo investigación.

En los estudios longitudinales a menudo es de interés estimar la evolución, a través del tiempo, de la media de la variable respuesta como una función de variables explicativas. Pero con frecuencia este tipo de datos contiene mucha información faltante.

El método de estimación de máxima verosimilitud puede proveer inferencias válidas sobre los parámetros, si el modelo para la distribución conjunta de las variables está correctamente especificada y la probabilidad de no respuesta no depende de los valores observados de la variable respuesta. Pero para datos incompletos, el método de verosimilitud puede ser sensible a una mala especificación del modelo, cuando se imputan las pérdidas a partir de la distribución condicional dados los datos observados (Dempster, Laird y Rubin,1977). Además, aún con datos completos si el interés está sobre modelos para la distribución marginal de la respuesta, los modelos paramétricos completos para ciertos datos no normales, que preservan la esperanza marginal de la respuesta dada las covariables, puede ser engorroso y computacionalmente difícil.

Liang y Zeger (1976) propusieron un método de estimación que se utiliza especialmente para variables no gaussianas, denominado ecuaciones de estimación generalizadas (GEE). Ellas proveen soluciones consistentes para los parámetros de interés mientras se especifi-

que correctamente sólo el modelo para las medias marginales de la respuesta en cada ocasión. Su enfoque es una extensión del método de quasi-verosimilitud a un escenario multivariado y provee estimadores mínimo cuadráticos iterativamente reponderados. Como Liang y Zeger (1976) puntualizaron las inferencias con la GEE son válidas sólo bajo el supuesto fuerte que los datos perdidos responden al esquema de pérdida "completamente al azar", esto es, dadas las covariables, el proceso de no respuesta es independiente tanto de los valores observados como de los no observados de la respuesta y de covariables que dependen del tiempo.

El propósito de este trabajo es plantear una discusión introductoria sobre las dificultades que se presentan en el análisis de datos longitudinales, por la posible falta de respuesta en la variable de interés. Además se utiliza para la estimación de los parámetros del modelo una clase de ecuaciones de estimación ponderadas que consideran la probabilidad de no respuesta en un tiempo determinado, dado el pasado y conducen a estimadores consistentes y asintóticamente normales de los parámetros (Robins, Rotnitzky y Zhao, 1995).

Se comparan diferentes métodos utlizados para la estimación de modelos que explican el ingreso individual, planteando un modelo a partir de la teoría económica existente sobre los ingresos individuales. La aplicación se realiza utilizando la información sumistrada por la Encuesta Permanente de Hogares (EPH), relevada por el Instituto Nacional de Estadística y Censos, correspondiente a los años 1998 y 1999.

En la sección 2 se realiza una breve reseña sobre propuestas existente en la bibliografía en el tratamiento de datos faltantes. Se presenta, además, el método que se utiliza para estimar los parámetros. En la sección 3 se definen las variables que se incorporan al modelo para explicar el ingreso y las estimaciones de los modelos Por último, en la sección 4 se discuten los resultados hallados.

2.- CONSIDERACIONES METODOLÓGICAS

En este trabajo se presentan métodos de inferencia estadística para conjuntos de datos multivariados con valores perdidos donde las pérdidas pueden ocurrir en alguna o todas las variables. Tales datos surgen frecuentemente en la práctica, pero las herramientas para tratarlos no son fácilmente disponibles para el analista.

Cuando se encuentran valores perdidos, los investigadores frecuentemente recurren a métodos ad hoc para borrar datos o imputarlos imponiendo que tengan un formato rectangular. Muchos paquetes automáticamente omiten en un análisis de regresión cualquier caso que tiene un valor perdido para cualquier variable. Imputación es un término genérico que

se refiere a la acción de completar un valor perdido con valores estimados. Por ejemplo, en conjuntos de datos multivariados cada valor faltante se puede reemplazar por la media observada para esa variable o por el valor predicho a partir de un modelo de regresión. Después que el conjunto de datos se altera por la imputación (mediante alguno de esos métodos) o por la eliminación, se procede como si los datos omitidos nunca hubiesen sido realmente observados o como si los valores imputados fueran datos reales.

Cuando los casos incompletos comprenden sólo una pequeña fracción borrar un caso puede ser una solución perfectamente razonable. En escenarios multivariados donde los faltantes ocurren sobre más de una variable los datos perdidos son a menudo una porción sustancial del conjunto completo. Por eso, borrarlos puede ser ineficiente, provocando gran cantidad de información descartada. Omitirlos tenderá a introducir sesgo y los casos que queden no serán representativos de la población para la cual se desea realizar la inferencia: la poblaciòn total más que la población de datos sin valores faltantes.
Los métodos de imputación no son menos problemáticos. Imputar por promedios puede preservar las medias muestrales pero distorsiona la estructura de covariancias; imputar por regresión tiende a sobreestimar las correlaciones observadas. Cuando el modelo de datos faltantes es complejo, puede ser dificultoso buscar esquemas de imputación que preserven aspectos importantes de la distribución conjunta. Más aún, si podría ser preservada adecuadamente la distribución conjunta de todas las variables, sería un serio error tratar los datos imputados como si ellos fueran reales.

Cuando se utiliza un modelo para describir un conjunto de datos, es importante identificar el mecanismo de pérdida que siguen los datos, ya que ciertos métodos de estimación no alteran las propiedades de los estimadores obtenidos.

Un proceso de no respuesta se denomina

1.- **completamente al azar** (**MCAR:** missing completely at random) si la pérdida es independiente tanto de los datos observados como de los no observados y de las variables explicativas,

2.-**perdidos al azar** (**MAR:** missing at random) si la pérdida, condicional a los datos observados, es independiente de las medidas no observadas, pero puede depender de las variables explicativas,

3.- un proceso que no es ni MCAR ni MAR se denomina **no aleatorio** (**MNAR:** missing non at random), ya que la pérdida puede depender de las observaciones y posiblemente de los regresores.

En el contexto de la inferencia verosimil y cuando los parámetros que describen el proceso de medida son funcionalmente independientes de los parámetros que describen el proceso de pérdida, un mecanismo se dice ignorable. Entonces MCAR y MAR son ignorables, mientras que un proceso no aleatorio es no ignorable.

Rubin (1976) define MAR en término de un modelo probabilístico para las pérdidas. Sea R la matriz de variables indicadoras cuyos elementos son cero o uno dependiendo si los correspondientes elementos de Y están perdidos u observados. En general, no se debería esperar que la distribución de R no esté relacionada a Y, así se postula un modelo probabilístico para R, el cual depende de Y como también de parámetros desconocidos.

Se necesita suponer que el parámetro del modelo y el mecanismo de pérdida sean distintos. Así, si se cumple este supuesto, los datos son MAR y el mecanismo se dice ignorable.

Siguiendo los argumentos dados por Rubin(1976) y Little y Rubin (1987) se puede mostrar que bajo ignoraribilidad, cuando se realizan inferencias acerca de los parámetros del modelo basadas en la verosimilitud, no se necesita considerar el modelo para R ni los parámetros del mismo.

2.1.- Estimación de modelos con datos faltantes

Se presenta, a continuación, el método de estimación cuando existen datos faltantes, para conjuntos de datos longitudinales gaussianos y no gaussianos, propuesto por Robins, Rotnitzky y Zhao (1995) y la nomenclatura que se utiliza. Este método se denominará método RRZ.

Se considera un estudio realizado sobre un período de tiempo fijo de 1 a n_i. Sea

- $\mathbf{Y}_i=(y_{i1},y_{i1},...,y_{ini})'$ el vector de las respuestas para el sujeto i, i=1,..,K, medidas en los momentos 0,1,... n_i, donde el cero ocurre antes de comenzar el estudio,
- $\mathbf{X}_i = (\mathbf{X}_{i0}^{'},...,\mathbf{X}_{ini}^{'})^{'}$, donde $\mathbf{X}_{it}$ es un vector de variables explicativas para el momento t, asociado con y_{it} y que incluye una constante 1.
- V_{it} , t=0,.., n_i, son las covariables que dependen del tiempo
- $W_{it} = (V_{it}^{'}, y_{it}^{'})^{'}$,t=1,.., n_i, comprende los valores de las variables que dependen del tiempo y las respuestas
- $W_{i0} = (x_i, V_{i0}, x_{i0})$ comprende los valores X_i y los valores base y_{i0} y V_{i0} .

- $\overline{W}_{it} = (W_{i0}^{'}, W_{i1}^{'},..., W_{i(t-1)}^{'})^{'} = ((X_i, V_{i0}, Y_{i0}),...,(V_{i(t-1)}, Y_{i(t-1)}))^{'}$ vector que depende de los datos pasados registrados hasta el momento t pero no incluye momento actual.

Se define R_{it} =1 si el sujeto i se observa en el tiempo t (es decir, y_{it} y V_{it} se observan) y R_{it}=0 en otro caso.

Si el proceso de datos perdidos satisface

$$P(R_{it} = 1/R_{i(t-1)} = 1, \overline{W}_{it}, y_i) = P(R_{it} = 1/R_{i(t-1)} = 1, \overline{W}_{it}), \qquad (2.1.1)$$

donde $\overline{W}_{it}$ contiene variables que dependen de los datos pasados registrados hasta el momento t pero no lo incluye, se dice que los datos están perdidos al azar (MAR).

Esto se lee entre todos los sujetos observados al tiempo (t-1)(es decir, aquéllos con

$R_{(t-1)}$=1) la no respuesta al tiempo t no está relacionada con los resultados actual y futuros (y_{it},…$y_{i\ ni}$), condicional al pasado observado $\overline{W}_{it}$. Esto significa que la no respuesta no está relacionada con resultados actuales y futuros dado el pasado $\overline{W}_{it}$.

Si el objetivo principal es modelar la esperanza marginal (respuesta promedio para las observaciones que comparten los mismos valores de las covariables) como una función de las variables explicativas, mientras que la correlación entre las respuestas es de interés secundario, se puede utilizar para el análisis de los datos un modelo marginal. Estos modelos fueron propuestos por Liang y Zeger (1986) para variable gaussiana y no gaussiana.

En un modelo marginal, el efecto de las covariables sobre la respuesta se modela separadamente de la correlación dentro del individuo (unidad). Debido a que el interés está en modelar la esperanza marginal o la respuesta promedio, es natural poner más énfasis en especificar correctamente la estructura media marginal que la estructura de covariancia. Por eso se supone que dicha estructura tiene una forma arbitraria y se la denomina “covariancia de trabajo”.

2.1.1.- El modelo marginal y su estimación

Un modelo marginal tiene los siguientes supuestos:

a.- **La media marginal de y_{it}** está correctamente especificada por

$$\mu_{it} = E(y_{it} / \mathbf{X}_{it}) = h(\mathbf{X}_{it}^{'} \boldsymbol{\beta}),\ i = 1,...,K,\ t = 1,...,n_i \ \text{ o } \ \mathbf{X}_{it}^{'} \boldsymbol{\beta} = g(\mu_{it}), \qquad (2.1.1.1)$$

donde, g es la función de enlace y $\mathbf{X}_{it}$ fue definido anteriormente.

b.- El supuesto de estructura media se complementa con la especificación de la **varianza de y_{it}** como una función de μ_{it},

$$\sigma_{it}^2 = \text{Var}(y_{it}) = v(\mu_{it})\phi \text{, } i = 1,..., K \quad t = 1,..., n_i \qquad (2.1.1.2)$$

donde v, es una función de variancia conocida.

Liang y Zeger (1986) suponen que la distribución marginal de y_{it} pertenece a la familia exponencial. Luego $v(\mu_{it})$ está completamente determinada por el supuesto de esta familia.

c.- Para explicar la dependencia entre las observaciones dentro de las unidades se introducen **covariancias de trabajo**, las que posiblemente dependen de algún vector de parámetros desconocido $\boldsymbol{\alpha}$.

Liang y Zeger (1986) suponen que la estructura de variancia es complementada mediante la especificación de una matriz de correlación de trabajo,

$$\mathbf{R}(\boldsymbol{\alpha})=\{\text{corr}(y_{it}, y_{it'})\}=\{c(\mu_{it},\mu_{it'};\alpha)\},$$

con función conocida c. De esta manera la matriz de covariancia de trabajo $\boldsymbol{\Sigma}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ depende de la variancia de las observaciones (σ_{it}^2) y de $\mathbf{R}(\boldsymbol{\alpha})$.

Dentro del contexto de datos longitudinales, Liang y Zeger (1986) sugieren varias estructuras para $\mathbf{R}(\boldsymbol{\alpha})$. La especificación más simple es la de suponer que las observaciones repetidas no están correlacionadas, es decir, $\mathbf{R}_i(\boldsymbol{\alpha}) = \mathbf{I}$, lo que conduce a las ecuaciones de estimación típicas de las observaciones independientes.

Otra elección, denominada simetría compuesta ("exchangeable"), es la que corresponde al modelo de igual correlación.

Si se dispone de suficiente cantidad de datos y los tiempos de observación son los mismos para todas las unidades ($n_i = n$) se puede dejar a $\mathbf{R}(\boldsymbol{\alpha})$ sin especificar, lo cual conduce a un modelo con estructura de correlación arbitraria.

Si las observaciones repetidas para cada unidad están correlacionadas de forma similar a la de un proceso Gaussiano autorregresivo, se obtiene una estructura autorregresiva.

Para estimar los parámetros no se puede aplicar el método de Máxima Verosimilitud porque no se conoce la distribución conjunta de las observaciones repetidas. En su defecto se usa una versión multivariada de la "quasi"-verosimilitud denominada Ecuaciones de Estimación Generalizadas (GEE). La expresión de estas ecuaciones es

$$S_{\boldsymbol{\beta}}(\boldsymbol{\beta},\boldsymbol{\alpha}) = \sum_{i=1}^{K} \mathbf{X}_i' D_i(\boldsymbol{\beta}) \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\beta},\boldsymbol{\alpha})(y_i - \boldsymbol{\mu}_i) = K^{-1/2} \sum_{i=1}^{K} \mathbf{H}_i(\beta) = 0\,, \qquad (2.1.1.3)$$

siendo,

- $H_i(\boldsymbol{\beta}) = D_i(\boldsymbol{\beta})\boldsymbol{\varepsilon}_i(\boldsymbol{\beta}) = d(X_i,\boldsymbol{\beta})\ \boldsymbol{\varepsilon}_i(\boldsymbol{\beta}) = d(X_i,\boldsymbol{\beta})\ (y_{it} - E(Y_{it}/X_i))$ y $\varepsilon_i(\beta) = (\varepsilon_{i1}(\beta),...,\varepsilon_{ini}(\beta))'$
- $\boldsymbol{\mu}_i = (\mu_{i1},...,\mu_{in_i}) = E(Y_{it}/X_i)$ y μ_{it} t=1,..,n_i, la media marginal de y_{it}, definida como en (2.1.1.1),
- $\boldsymbol{\Sigma}_i(\boldsymbol{\alpha},\boldsymbol{\beta})$ la matriz de covariancias de "trabajo" que explica la dependencia entre observaciones de cada unidad, cuya expresión viene dada por

$$\boldsymbol{\Sigma}_{\mathbf{i}}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \mathbf{V}_i = \mathbf{A}_i^{1/2}\, \mathbf{R}_i(\boldsymbol{\alpha})\, \mathbf{A}_i^{1/2} / \phi\ . \qquad (2.1.1.4)$$

Esta matriz $\mathbf{A}_i$ = diag $\{\phi V(\mu_{it})\}$ = diag $\{\sigma_{i1}^2,..,\sigma_{in_i}^2\}$, es una matriz diagonal que contiene la variancia de las observaciones.

- $D_i(\boldsymbol{\beta})$= diag $(D_{it}(\boldsymbol{\beta}))$, donde $D_{it}(\boldsymbol{\beta}) = \partial h / \partial \eta_{it}$.

La matriz $\Sigma_i(\beta,\alpha)$ se denomina de trabajo pues no se espera que esté correctamente especificada. Será igual a la verdadera matriz de covariancias para las observaciones de un individuo, Cov(Y_i), si R_i (α) es la verdadera matriz de correlación de Y_i.

Se consideran a continuación las ecuaciones de estimación ponderadas para los parámetros del modelo cuando hay datos faltantes.

2.1.1.1.- Ecuaciones de estimación para datos faltantes

Estas ecuaciones son las presentadas en (2.1.1.3), modificadas convenientemente de acuerdo a que las probabilidades de pérdida sean o no conocidas.

a.- Estimación sin datos faltantes

Cuando y_i se observa para todos los sujetos las ecuaciones de estimación que se utilizan son las dadas por (2.1.1.4) y simbolizadas

$$U_{total}(\beta) = K^{-1/2} \sum_{i=1}^{K} H_i(\beta) = 0 \qquad (2.1.1.1.1)$$

Estas ecuaciones poseen una solución para β que es consistente y asintóticamente normal (Liang y Zeger, 1986).

Las inferencias basadas en GEE siguen siendo válidas bajo el supuesto que las pérdidas son completamente al azar.

b.- Estimación con datos faltantes

Cuando las observaciones no son completamente observadas y los faltantes no son MCAR las soluciones a (2.1.1.1.1) basadas sobre los datos observados pueden no ser consistentes ya que los residuos y_{it}-$g_t(X_i,\beta)$ no tienen media cero (Liang y Zeger,1976). Esto es así pues sujetos con R_i=1 pueden representar una muestra sesgada.

Cuando los datos perdidos son MCAR el estimador mínimo cuadrados ponderado de los parámetros está restringido a los resultados observados. Las soluciones obtenidas a partir de las ecuaciones para los datos observados

$$U_{comp}(\beta) = K^{-1/2} \sum_{i=1}^{K} D_i^*(\beta) \varepsilon_i^*(\beta) = 0 \qquad (2.1.1.1.2)$$

tienen las propiedades mencionadas anteriormente y $\varepsilon_i^*(\beta)$ es el vector de residuos observados para el sujeto i.

b1.- Probabilidades de respuesta conocidas

En estudios donde los datos faltantes son por diseño los datos son MAR y las probabilidades de respuesta son funciones conocidas de los datos observados.

Si los individuos (**Y**i) no son completamente observados y las probabilidades de respuesta $P(R_i = 1/W_i)$ son funciones conocidas de Wi se pueden estimar los parámetros utilizando las ecuaciones de estimación

$$\sum_{i=1}^{n}\left[\frac{I(R_i=1)}{\pi_i(1)}d_i(X_i,\beta)\{Y_i - g(X_i;\beta)\}\right]=0 \qquad (2.1.1.1.3)$$

donde, la función $I(R_i=1)=1$ si $R_i=1$ y 0 en otro caso y $\pi_i(1)=P(R_i=1/W_i)$ es la probabilidad de observar el vector completo de datos Wi para el sujeto i, dado W.

Estas ecuaciones usan los datos sólo para los sujetos con información completa, es decir, sujetos con $R_i=1$, para los cuales se dispone $\pi_i(1)$. En (2.1.1.1.3) cada sujeto con información completa aporta un término igual a su contribución a las ecuaciones de estimación U_{total} (2.1.1.1.1) ponderada por la probabilidad inversa de observar el vector de resultados completo Y_i dado W_i.

Bajo condiciones de regularidad, la solución a (2.1.1.1.3) es consistente y asintóticamente normal, pues la clave para la consistencia se debe a que la funciones de estimación definidas es insesgada. Las inferencias basadas en las soluciones a (2.1.1.1.3) no hacen uso efectivo de los datos disponibles. Por ejemplo, los sujetos con información incompleta no contribuyen a (2.1.1.1.3) y así, no se usan los valores observados de y_{it} y V_{it}.

Para incrementar la eficiencia con la cual se estima β*, se extiende la clase de ecuaciones de estimación (2.1.1.1.3) a

$$\sum_{i=1}^{n}\left[\frac{I(R_i=1)}{\pi_i(1)}d(X_i)\{Y_i - g(X_i;\beta)\}+A_i\right]=0 \qquad (2.1.1.1.4)$$

con

$$A_i=\sum_{r\neq 1}\left[I(R_i=r)-\frac{I(R_i=1)}{\pi_i(1)}\pi_i(r)\right]\phi_r(W_{(r)i}) \qquad (2.1.1.1.5)$$

y $\pi_i(r)=P(R_i=r/W_i)$.

Aquí $\phi_r(W_{(r)i})$ es una función de un vector px1 arbitraria de los datos $W_{(r)i}$ que son observados cuando Ri=r y Ai es función sólo de los datos.

En particular en las ecuaciones (2.1.1.1.4) se usan los datos observados para todos los sujetos del estudio, incluyendo sus valores registrados de V_{it}.

La inclusión de A_i mejora la eficiencia de un escenario en el cual las probabilidades de no respuesta son funciones conocidas de los datos (posiblemente no observados)

b2.- Probabilidades de respuesta desconocidas

Si bien las probabilidades de respuesta son desconocidas, las probabilidades de respuesta condicionas a cada tiempo t (λ_{it}) se suponen seguir un modelo paramétrico.

No se hace restricciones sobre la distribución conjunta de W_i pero se supone un modelo para la probabilidad de no respuesta, es decir sobre la ley de R_i dado W_i. El modelo que se sugiere considera que $\mathbf{R}_i$ es un vector de variables binarias, posiblemente correlacionadas, que toman valores en el conjunto $\{r=(r_1,...,r_T)': r_i=1 \text{ ó } 0,\ 1\le t \le n_i\}$. Sea $\overline{R}_{it} = (R_{i1},...,R_{i(t-1)})'$ y por conveniencia se define $\overline{R}_{i1} = 1$. La distribución condicional de $\mathbf{R}_i$ dado $\mathbf{W}_i$ es

$$P(R_i = r \mid W_i) = \prod_{t=1} P(R_{it} = 1 \mid \overline{R}_{it}, W_i)^{r_i} P(R_{it} = 0 \mid \overline{R}_{it}, W_i)^{t-r_i},$$

donde, $P(R_i = 1 \mid \overline{R}_{it}, W_i)$, t=1,...., n_i, tiene un modelo paramétrico conocido con vector qx1 de parámetros ρ. Si $\lambda_{it} = P(\mathbf{R}_i = 1 \mid \overline{R}_{it}, \mathbf{W}_i)$ se supone que $\lambda_{it} = \lambda_{it}(\rho)$ satisface

$$\text{logit}\ \lambda_{it}(\rho) = h_i(\overline{R}_{it}, W_i; \rho) \qquad (2.1.1.1.6)$$

donde $h_t(.,.;.)$, t=1,...,T son funciones conocidas.

Para parametrizar λ_{it} se puede elegir otra función como por ejemplo la probit.

En lo que sigue $W_{(r)i}$ es el subvector de $\mathbf{W}_i$ cuando $\mathbf{R}_i = \mathbf{r}$ (Si r=(1,1,0,..,0) entonces $W_{(r)i} = (W'_{i0}, W'_{i1}, W'_{i2})'$).

Cuando las probabilidades de respuestas son desconocidas pero las probabilidades de respuesta condicional en cada tiempo t, λ_{it}, se suponen que siguen el modelo (2.1.1.1.6) con un vector qx1 de parámetros desconocidos ρ. Es conveniente definir

$$\pi_i(r;\rho) = \prod_{t=1} (\lambda_i(\rho))^{r_i} (1-\lambda_{it}(\rho))^{1-r_t}$$

Si un estimador consistente de ρ fuese disponible, se reemplaza $\pi_i(r)$ en (2.1.1.1.3) por $\pi_i(r;\rho)$ y luego se resuelve obteniendo el estimador consistente de β.

La estimación de β se realiza obteniendo ρ de (2.1.1.1.6)y luego resolviendo (2.1.1.1.3). Sin embargo, se pueden estimar conjuntamente β y ρ a partir de las soluciones de un conjunto de p+q ecuaciones de estimación,

$$\sum U_i(\beta, \rho, d, \phi) = 0 \qquad (2.1.1.1.7)$$

donde

$$U_i(\boldsymbol{\beta}, \boldsymbol{\rho}, d, \boldsymbol{\phi}) = \frac{I(R_i = 1)}{\boldsymbol{\pi}_i(1;\boldsymbol{\rho})} \begin{bmatrix} d^{(1)}(X_i;\boldsymbol{\beta}) \\ d^{(2)}(X_i;\boldsymbol{\beta}) \end{bmatrix} \{Y_i - g(X_i, \boldsymbol{\beta})\} - \begin{bmatrix} A_i^{(1)}(\boldsymbol{\rho}) \\ A_i^{(2)}(\boldsymbol{\rho}) \end{bmatrix}. \qquad (2.1.1.1.8)$$

3.- APLICACIÓN

Se especifican y estiman modelos con el fin de poder explicar los ingresos laborales individuales, usando datos registrados por la Encuesta Permanente de Hogares (EPH), en presencia de datos faltantes.

Aún hoy existen frecuentes discusiones sobre los factores que intervienen en la formación del ingreso y los métodos con que se los estima. Sin embargo, en general, los trabajos econométricos recientes sobre estos temas y otros análogos realizados en nuestro país, han adoptado la teoría del capital humano (Becker, 1983). Una vertiente teóricamente afín de la investigación empírica, mucho menos explorada en la Argentina, se ha centrado en la explicación económica de los comportamientos del ingreso a través del tiempo.

En este estudio se desea modelar el ingreso de la fuente laboral de los varones jefe de hogar registrado en 4 períodos consecutivos, en función de ciertas variables de interés.

Se decide trabajar con el ingreso laboral neto como variable respuesta, debido a que es la variable registrada en la EPH.

Las variables usadas como explicativas, y que se registran en la EPH, son las mismas se utilizaron en trabajos realizados anteriormente por Blaconá et al (2002, 2001), a saber:

- **edadv:** variable continua calculada a partir de la fecha de nacimiento declarada por el encuestado. Esta variable se toma como *proxy* de la experiencia de la persona.

- **edad2v:** variable edad elevada al cuadrado. Se incluye para captar el comportamiento no lineal del retorno por año de actividad, que se supone creciente hasta una cierta edad y luego puede comenzar a decrecer lentamente.
- **medescov:** años de escolaridad declarados por el encuestado. Típica variable del capital humano que mide el retorno por cada año adicional de educación del individuo.

3.1.-Características de la muestra en estudio

Para realizar el estudio empírico se selecciona el aglomerado urbanos: de Rosario.

La muestra en estudio está compuesta por los varones jefe de hogar que permanecieron en la encuesta durante 4 ondas consecutivas, es decir entrevistados desde la primera onda de 1998 hasta la segunda onda de 1999. Estas muestras no incluyen aquellos individuos que tiene ingreso cero o es inactivo o tiene 65 o más años de edad. Se eliminan los varones con ingreso cero porque su inclusión apareja problemas en la estimación. Además se presentan en un número pequeño de casos y no producen cambios importantes en los resultados.

3.2.-El modelo

Para analizar este conjunto de datos se modela la respuesta media marginal de los ingresos en función de las variables mencionadas anteriormente. Para ello, se considera el predictor lineal de la forma

$$\eta_i = \beta_o + \beta_1 \,\text{edadv} + \beta_2 \,\text{edad2v} * \beta_3 \,\text{medes cov}$$

Las observaciones seleccionadas en la muestra se utilizan para estimar los parámetros del predictor lineal mediante la metodología de las ecuaciones de estimación generalizadas (GEE) (Liang y Zeger, 1986; Zeger y Liang, 1986), usando en la matriz de covariancias como matriz de correlación la de simetría compuesta e independencia.

El procedimiento utilizado en este trabajo fue el siguiente:

1.- Se estima el modelo propuesto para el conjunto completo de datos, utilizando el procedimiento GENMOD de SAS para dos matrices de covariancia de trabajo distintas.

2.- Se realiza un estudio de simulación produciendo, en el conjunto de datos completos, pérdidas del tipo MCAR y MAR. Se realizan 100 repeticiones del proceso para cada ley de pérdida y se estiman los parámetros del conjunto incompleto usando:
a.- el procedimiento GENMOD de SAS para dos matrices de covariancia de trabajo distintas: independencia (I) y simetría compuesta (CS). Este procedimiento utiliza como matriz de

covariancias de trabajo la matriz de covariancias empírica calculada con los "casos disponibles", es decir usa todos los valores no perdidos. Al estimar los parámetros la contribución de cada unidad se calcula omitiendo los elementos que corresponden a los datos faltantes; b.- el método de RRZ para las dos matrices de covariancia de trabajo mencionadas modela la probabilidad de respuesta mediante un modelo logístico saturado e introduce en la GEE la inversa de esa probabilidad de pérdida como pesos.

Durante el proceso de simulación se estiman los coeficientes de regresión y los errores estándares de los mismos para cada caso, utilizando los procedimientos mencionados. Con estos 100 valores se calcula el promedio y la variancia empírica de los estimadores Estos valores se comparan con los obtenidos para los datos completos.

Las tablas 3.2.1 y 3.2.2 muestran el promedio de los parámetros estimados para cada simulación y la variancia empírica. Los valores calculados se comparan con los reales.

En la primera tabla se observa que el uso de ambos métodos no produce sesgo en las estimaciones. Este resultado era esperado debido a que la métodologia GEE produce estimaciones válidas cuando las pérdidas son completamente al azar.

En la tabla 3.2.2 se presentan los resultados cuando las pérdidas son MAR. Cuando se usa el procedimiento GENMOD aparece un pequeño sesgo en las estimaciones. Con el procedimiento ponderado el sesgo se elimina, aunque el error estandar del estimador será algo más alto que el verdadero valor. La razón se podría deber a que por defecto el GENMOD asume que los pesos son conocidos, cuando en realidad se estiman a través de los datos.

Tabla 3.2.1 Estimación de los parámetros del conjunto de datos completos e incompleto con distintos porcentaje de perdidas del tipo MCAR, para el aglomerado Rosario

PARAMETRO	MODELO COMPLE	METODO		PORCENTAJE DE PERDIDA			
				6%	10%	20%	4%(t3) 6%(t4)
ORDENADA	3.0612	GENMOD	BCS	3.05099	3.04490	3.01205	3.03851
	0.8788		SEBC	0.88120	0.88314	0.89006	0.88439
			BI	3.05238	3.04205	3.01334	3.04208
			SEBI	0.88081	0.88150	0.89426	0.88473

		RRZ	BCS	3.05069	3.00542	2.98622	3.05485
			SEBC	0.90028	0.91554	1.00172	0.91634
			BI	3.05069	3.00542	2.98622	3.05485
			SEBI	0.90028	0.91554	1.00172	0.91634
EDADV	0.1068 0.0507	GENMOD	BCS	0.10720	0.10747	0.10893	0.10780
			SEBC	0.05076	0.05080	0.05103	0.05095
			BI	0.10723	0.10772	0.10888	0.10766
			SEBI	0.05073	0.05074	0.05127	0.05099
		RRZ	BCS	0.10773	0.11017	0.11040	0.10709
			SEBC	0.05159	0.05249	0.05626	0.05265
			BI	0.10773	0.11017	0.11040	0.10709
			SEBI	0.05159	0.05249	0.05626	0.05265
EDADV2	-0.011 0.0006	GENMOD	BCS	-.00106	-.00107	-.00108	-.00107
			SEBC	0.00065	0.00065	0.00065	0.00065
			BI	-.00107	-.00107	-.00108	-.00107
			SEBI	0.00065	0.00065	0.00065	0.00065
		RRZ	BCS	-.00108	-.00111	-.00111	-.00107
			SEBC	0.00065	0.00066	0.00071	0.00067
			BI	-.00108	-.00111	-.00111	-.00107
			SEBI	0.00065	0.00066	0.00071	0.00067
MEDESCOV	0.0961 0.0170	GENMOD	BCS	0.09585	0.09602	0.09584	0.09605
			SEBC	0.01703	0.01705	0.01710	0.01709
			BI	0.09581	0.09600	0.09599	0.09608
			SEBI	0.01704	0.01708	0.01717	0.01708
		RRZ	BCS	0.09525	0.09516	0.09554	0.09583
			SEB	0.01780	0.01817	0.01928	0.01798
			BI	0.09525	0.09516	0.09554	0.09583
			SEBI	0.01780	0.01817	0.01928	0.01798

Referencias BCS: Promedio Beta con Simetría compuesta. BI: Promedio Beta con Independencia.SEBCS: Variancia empírica de los coeficientes de regresión con Simetría compuesta.SEBI: Variancia empírica de los coeficientes de regresión con Independencia

Tabla 3.2.2 Estimación de los parámetros del conjunto de datos completos e incompleto con distintos porcentaje de perdidas del tipo MAR, para el aglomerado Rosario

PARAMETRo	MODELO COMPLetO	METODO		PORCENTAJE DE PERDIDA		
				6%	20%	30%
ORDENADA	3.0612	GENMOD	BCS	3.1080	3.1194	3.1557
	0.8788		SEBC	0.8775	0.8741	0.8722

			BI	3.1080	3.1194	3.1557
			SEBI	0.8776	0.8749	0.8793
		RRZ	BCS	3.0639	3.0679	3.0724
			SEBC	0.8779	0.8806	0.8815
			BI	3.0650	3.0681	3.0739
			SEBI	0.8779	0.8860	0.8815
EDADV	0.1068 0.0507	GENMOD	BCS	0.1041	0.1030	0.1018
			SEBCS	0.0506	0.0504	0.0503
			BI	0.1041	0.1037	0.1018
			SEBI	0.0506	0.0505	0.05051
		RRZ	BCS	0.1065	0.1064	0.1061
			SEBCS	0.0508	0.0510	0.0511
			BI	0.1065	0.1064	0.1061
			SEBI	0.0508	0.0510	0.0511
EDADV2	-0.011 0.0006	GENMOD	BCS	-.00106	-.0010	-0.0010
			SEBCS	0.0006	0.0006	0.0006
			BI	-.00106	-.0010	-0.0010
			SEBI	0.0006	0.0006	0.0006
		RRZ	BCS	-.00103	-.0011	-0.0010
			SEBCS	0.00065	0.0007	0.0007
			BI	-.00103	-.0011	-0.0010
			SEBI	0.00065	0.0007	0.0007
MEDESCOV	0.0961 0.0170	GENMOD	BCS	0.09605	0.0957	0.0953
			SEBCS	0.01702	0.0170	0.0170
			BI	0.09606	0.0957	0.0953
			SEBI	0.01702	0.0170	0.0170
		RRZ	BCS	0.09605	0.0960	0.0960
			SEBCS	0.01712	0.0172	0.0174
			BI	0.09605	0.0961	0.0961
			SEBI	0.01712	0.0172	0.0174

Referencias BCS: Promedio Beta con Simetría compuesta BI: Promedio Beta con Independencia SEBCS: Variancia empírica de los coeficientes de regresión con Simetría compuesta SEBI: Variancia empírica de los coeficientes de regresión con Independencia

4.- CONSIDERACIONES FINALES

En este trabajo se presentan métodos de inferencia estadística para conjuntos de datos multivariados con valores perdidos donde las pérdidas pueden ocurrir en alguna o todas las variables.

Cuando se encuentran valores perdidos, los investigadores frecuentemente borran esos datos o los imputan. En escenarios multivariados, donde los datos perdidos son a menudo

una porción sustancial del conjunto completo, ambos procedimientos son peligrosos. Borrarlos puede ser ineficiente, provocando gran cantidad de información descartada, mientras que los métodos de imputación pueden no. preservar aspectos importantes de la distribución conjunta.

Los métodos de estimación convencionales sólo permiten pérdidas MCAR. Robins, Rotnitzky y Zhao (1995) propusieron un método de estimación que permite datos perdidos del tipo MAR.

Se consideran dos enfoques para conocer el comportamiento de los estimadores de un modelo marginal, para un conjunto de datos longitudinales no gaussianos cuando hay datos perdidos: la GEE y el de RRZ.

Para realizar el estudio se considera el ingreso de la fuente laboral de los varones jefe de hogar registrado en 4 períodos consecutivos, que se desea modelar en función de ciertas variables de interés. Se generan pérdidas de tipo MCAR y MAR a estos datos y se evalúan los estimadores resultantes al aplicar ambos métodos.

Cuando las pérdidas son MCAR se obtienen resultados similares, con ambos métodos.

Cuando los datos son MAR el enfoque de RRZ requiere que la ley de pérdida sea modelada y que los pesos correspondientes a la inversa de la probabilidad de no respuesta sea incluída en la GEE. Este método producirá coeficientes insesgados de los parámetros, pero los estimadores del error estándar son sesgados de los verdaderos.

En resumen cuando se estima usando la GEE se debe considerar no sólo el modelo para la media marginal de las observaciones, sino también considerar de qué tipo son los datos perdidos.

5.-BIBLIOGRAFIA

Becker, Gary (1983): *El Capital Humano*, Alianza Universidad, Madrid.

Dempster, A.P., Laird, N.M. y Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series B, 39,*1-38**.**

Liang, K. Y. y Zeger, S. L. (1986). Longitudinal data anlysis using the generalized lineal model. *Biometrika,73,*13-22.

Little, R. J. A. y Rubin, D.B.(1987). S*tatistical analysis with missing data.* New York: Wiley.

Robins, J. M. Y Rotniztky. A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association, 90,* 122-129.

Robins, J. M., Rotniztky. A. y Zhao, L. P. (1994). Estimation of regression coefficients when a regressor is not always observed. *Journal of the American Statistical Association, 89,*846-866.

---------- , (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association, 90,* 106-121.

Rotniztky, A., Robins, J.M. y Scharfstein,D. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association, 93,* 1321-1339.

Rubin, D.B.(1976). Inference and missing data.*Biometrika,63*, 581-592.