Garcia, María del Carmen
Servy, Elsa
Instituto de Investigaciones Teóricas y Aplicadas, de la Escuela de Estadística

COMPARACION DE METODOS PARA EL ANALISIS DE DATOS LONGITUDINALES CON VALORES PERDIDOS.

1.- INTODUCCION

En los años recientes ha habido gran interés en el estudio de datos longitudinales. Estos datos están compuestos por mediciones repetidas (realizadas en ocasiones sucesivas) de una variable sobre una misma unidad. Cuando la variable respuesta es aproximadamente Gaussiana se desarrollaron métodos, para el análisis de los mismos, que son muy aceptados.

Para el análisis de los datos longitudinales no Gaussianos, debido a la falta de distribuciones multivariadas, no se pueden usar métodos basados en la función de verosimilitud. Liang y Zeger (1986) usan una extensión de los modelos lineales generalizados (MLG) que no requiere la especificación completa de la distribución conjunta de las respuestas. Ellos suponen que las distribuciones marginales de la variable respuesta en cada tiempo siguen un modelo lineal generalizado (McCullagh y Nelder, 1983) y proponen modelos de asociación entre las observaciones de cada sujeto a través de una estructura de correlación de "trabajo".

Se suele denominar al método de Liang y Zeger ecuaciones de estimación generalizadas (GEE) y pueden interpretarse como una extensión del método de "quasi"-verosimilitud, creado para datos no normales por Wedderburn (1974).

Es probable que los datos longitudinales posean datos perdidos. En escenarios multivariados donde los faltantes ocurren sobre más de una variable los valores perdidos son a menudo una porción sustancial del conjunto completo.

El proceso que conduce a la pérdida de información se denomina mecanismo de pérdida. De acuerdo a que la probabilidad de no respuesta sea independiente o no de las respuestas observadas se dice que el mecanismo de datos perdidos es "completamente al azar" o "al azar". En caso contrario se dice que es "no aleatorio".

Los métodos usuales para el análisis de datos con información faltante en escenarios multivariados son aplicables cuando los valores están perdidos completamente al azar en

cuyo caso el mecanismo de pérdida se puede ignorar.

Pero, generalmente, no se presentan pérdidas de este tipo, sino que el mecanismo que produce las faltas está relacionado con los valores pasados y/o futuros de la misma variable que produjo la pérdida.
Ante esta situación se pueden adoptar diferentes estrategias para tratar los datos perdidos: ignorarlos, imputarlos o utilizar métodos de análisis modificados para adaptarlos a la estructura de los datos.

Según el tipo de estructura, estos métodos pueden ser más o menos ventajosos. Si se ignoran las pérdidas y el análisis se lleva a cabo únicamente con los datos que presentan todas las observaciones completas, la muestra se reduce de tamaño y los estimadores de los parámetros no poseen, en general, buenas propiedades. Cuando el modelo de datos faltantes es complejo, buscar esquemas de imputación que preserven aspectos importantes de la distribución conjunta puede ser dificultoso.

Los estimadores obtenidos utilizando métodos de estimación modificados, que ponderan por la inversa de la probabilidad de no respuesta (IPCW), poseen propiedades que se mantienen para muestras grandes (asintóticas).

En este trabajo se presenta un estudio preliminar realizado para un conjunto de datos longitudinales no Gaussianos con valores perdidos, para comparar el comportamiento de varios estimadores de los parámetros, utilizando diferentes enfoques, para muestras de tamaño pequeño. Los métodos que se utilizan son: estimación directa usando las ecuaciones de estimación generalizadas (mediante el procedimiento GENMOD de SAS); imputación de los valores perdidos y luego estimación mediante el GENMOD y, por último, IPCW.

En la sección 2 se realiza una breve reseña sobre las propuestas existentes en la bibliografía para el tratamiento de datos faltantes. Se presenta, además, el método que se utiliza para estimar los parámetros. En la sección 3 se describe el estudio de simulación realizado. La aplicación y discusión de los resultados hallados se presenta en la sección 4 y por último, en la sección 5 las consideraciones finales.

2.- CONSIDERACIONES METODOLÓGICAS

En este trabajo se presentan métodos de inferencia estadística para conjuntos de datos multivariados con valores perdidos en la variable respuesta.

Cuando se utiliza un modelo para describir un conjunto de observaciones, es importante identificar el mecanismo de pérdida que sigue la variable respuesta.

Un proceso de no respuesta se denomina

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

2.-**perdidos al azar** (**MAR:** missing at random) si la pérdida, condicional a las respuestas observadas, es independiente de las respuestas 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 respuestas (observadas o no) y posiblemente de los regresores.

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

2.1. Estrategias para tratar los valores perdidos

Los métodos para tratar los conjuntos de datos con respuestas perdidas se pueden agrupar aproximadamente en las siguientes categorías, no excluyentes:

1.- Procedimientos basados exclusivamente en las *unidades registradas completamente* Cuando algunos de los valores de la variable respuesta no se registran para algunas de las unidades se procede a descartarlas, analizándose sólo las completas. Este procedimiento es generalmente fácil de llevar a cabo y puede ser satisfactorio cuando existen pocos datos perdidos.

El análisis usando los casos completos tiene la ventaja de la simplicidad, pues se pueden aplicar métodos estadísticos estándares sin modificaciones y permite la comparabilidad de las estadísticas univariadas. La desventaja proviene de la potencial pérdida de información al descartar los casos incompletos.

Una inquietud que se presenta es si la selección de las unidades con los datos completos conduce a estimadores sesgados. Bajo el supuesto MCAR los datos completos representan una submuestra aleatoria de los datos originales y por lo tanto, descartar los incompletos no sesga las estimaciones. Sin embargo, en general, las unidades registradas completamente difieren de la muestra original de distintas maneras. En estudios de panel o lon-

gitudinales, por ejemplo, las unidades que se pierden son, a menudo, diferentes de los que permanecen en el estudio y en estos casos el análisis de datos completos puede producir resultados sesgados (Afifi y Elashoff, 1966).

2.- Procedimientos basados en *imputación*

Este procedimiento consiste en completar los valores perdidos y analizar los conjuntos completos resultantes por métodos estadísticos estándares.

Para datos longitudinales existen varios métodos que permiten imputar valores. Algunos de ellos usan sólo información que pertenece a los individuos cuyas respuestas están perdidas, mientras que otros usan los valores de los otros individuos. Los métodos de imputación se pueden dividir en cuatro grupos de acuerdo al tipo de valores que ellos usan para realizar las estimaciones de los valores perdidos: poblacional, preliminar o basal ("baseline"), anterior y anterior y posterior (B/A).

2.1.- Grupo poblacional: no usan información específica de la persona. para estimar el valor de la respuesta faltante en una ocasión. Se usa la media o mediana de esa ocasión (o ese tiempo), a través de todos los individuos.

2.2.- Grupo preliminar o basal ("baseline"): los procedimientos que pertenecen a este grupo usan los valores de las variables registradas al comienzo del estudio. No se usa información longitudinal. Uno de los métodos denominado *media (mediana) de la clase* procede de la siguiente manera: se forman grupos de unidades de acuerdo a la similitud de los valores de sus observaciones preliminares. Dada una unidad que presenta un tiempo perdido, este se sustituye por la media (mediana) de las observaciones preliminares del grupo a que la unidad pertenece.

Otros métodos son el de *regresión* y el *"hot deck".*

2.3.- Grupo anterior: Si una unidad tiene un tiempo faltante estos métodos usan los datos longitudinales de esa unidad, pero sólo hasta la ocasión en que la observación está perdida. Entre los métodos se encuentran:

i) *media o mediana de la unidad*: consiste en calcular la media o la mediana con los valores previos a la falta de esa unidad y reemplazar la respuesta perdida por ese valor.

ii) *adjudicación de la última observación registrada* al valor faltante (*locf*): como lo indica su nombre, el método consiste en extrapolar la última respuesta observada en la unidad con pérdidas, al resto de los tiempos previstos para esa unidad. Un refinamiento de este método sería estimar una tendencia en el tiempo, para una unidad o un conjunto de unidades de un

grupo particular, y extrapolar mediante la tendencia estimada en lugar de utilizar un único valor.

2.4.- Grupo anterior/posterior: estos métodos usan para cada unidad los valores anterior y posterior a la respuesta perdida. Los métodos de este grupo son: *último y posterior* (promedia estos dos valores alrededor del dato perdido para cada persona); *la media/mediana de la unidad* (asigna la media/mediana de todos los valores conocidos de una unidad); *próxima observación siguiente* (asigna al valor perdido el próximo valor conocido).

3.- Procedimientos de *estimación basados en el modelo para los datos perdidos*.

Cuando hay valores perdidos el funcionamiento de los métodos de análisis depende del mecanismo que conduce a los valores perdidos, por lo cual muchos procedimientos se generan definiendo un modelo para las pérdidas y basan las inferencias sobre la verosimilitud de ese modelo, estimando los parámetros por diferentes procedimientos (el de máxima verosimilitud es el más frecuentemente usado).

Para un conjunto de datos longitudinales con respuestas que pueden ser o no Gaussianas, Robins, Rotniztky y Zhao (1995) propusieron una clase de ecuaciones de estimación que admiten valores perdidos. El conjunto de los estimadores producidos por este método de estimación semiparamétrico incluye los estimadores GEE (Liang y Zeger, 1986), entre otros. El método se describe en la próxima sección.

2.2.- Método de estimación para conjuntos de datos con información faltante

Cuando existen datos faltantes en conjuntos de datos longitudinales gaussianos y no gaussianos, Robins, Rotnitzky y Zhao (1995) proponen una extensión del método GEE a datos con pérdidas al azar, ponderando las ecuaciones GEE por la inversa de la probabilidad de no respuesta. Este método se denominará método RRZ (o IPCW).

Se considera un estudio realizado sobre un período de tiempo de 1 a T, que puede incluir una medición preliminar, que será identificada con 0. La nomenclatura que se presenta a continuación es la utilizada por Robins, Rotnitzky y Zhao (1995).

- $\mathbf{Y}_i=(y_{i0},y_{i1},...,y_{iT})'$ es el vector de las respuestas para la unidad i, i=1,..,K, medidas en los momentos 0,1,... T, donde el cero ocurre antes de comenzar el estudio,
- $\mathbf{X}_i = (\mathbf{X}_{i0}^{'},...,\mathbf{X}_{iT}^{'})^{'}$ es un vector de variables explicativas, donde $\mathbf{X}_{it}$ está asociado con y_{it} y que incluye una constante 1.
- $\mathbf{V}_{it}$, t=0,.., T, es un vector de variables que dependen del tiempo. Este vector a veces

puede no existir ya que no siempre interesa registrar estas variables.

- $\mathbf{W}_{it} = (\mathbf{V}_{it}', y_{it})'$, t=1,.., T, comprende el vector de las variables que dependen del tiempo y la respuesta en cada tiempo.
- $\mathbf{W}_{i0} = (\mathbf{X}_i, \mathbf{V_{i0}}, y_{i0})$ comprende los valores $\mathbf{X_i}$ y los valores preliminares y_{i0} y $\mathbf{V}_{i0}$.
- $\bar{\mathbf{W}}_{it} = (\mathbf{W}_{i0}', \mathbf{W}_{i1}', ..., \mathbf{W}_{i(t-1)}')' = ((\mathbf{X_i}, \mathbf{V_{i0}}, y_{i0}), ..., (\mathbf{V}_{i(t-1)}, y_{i(t-1)}))'$ vector que depende de los datos pasados registrados hasta el momento t pero no incluyen el momento actual.

Se define r_{it} =1 si el sujeto i se observa en el tiempo t (es decir, y_{it} y $\mathbf{V}_{it}$ se observan) y r_{it}=0 en otro caso.

El proceso de datos perdidos es perdido al azar (MAR) si satisface

$$P(r_{it} = 1 / r_{i(t-1)} = 1, \bar{\mathbf{W}}_{it}, \mathbf{Y}_i) = P(r_{it} = 1 / r_{i(t-1)} = 1, \bar{\mathbf{W}}_{it}),$$

donde $\bar{\mathbf{W}}_{it}$ contiene variables que dependen de los datos pasados registrados hasta el momento t pero no lo incluye.

Si para cada tiempo el objetivo principal es modelar la esperanza marginal como una función de las variables explicativas, mientras que la correlación entre las respuestas de una misma unidad 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 variables gaussianas y no gaussianas. En un modelo marginal, el efecto de las covariables sobre la respuesta se modela separadamente de la correlación dentro de la unidad. Se supone que dicha estructura tiene una forma arbitraria y se la denomina "covariancia de trabajo".

Un modelo marginal tiene los siguientes supuestos:

a.- **La media marginal de $\mathbf{y_{it}}$** está especificada, en cada ocasión t, por

$$\boldsymbol{\mu}_{it} = E(y_{it} / \mathbf{X}_{it}) = h_t(\mathbf{X}_{it}' \boldsymbol{\beta}),\ i = 1,...,K,\ t = 1,...,T \quad o \quad \mathbf{X}_{it}' \boldsymbol{\beta} = \boldsymbol{\eta}_{it} = g(\mu_{it}),$$

donde, g es la función de enlace, $\boldsymbol{\beta}$ es un vector (p+1 x 1) de parámetros y $\mathbf{X}_{it}$ fue definido anteriormente. A la expresión $\mathbf{X}_{it}' \boldsymbol{\beta}$ se la denomina el predictor lineal.

b.- El supuesto de estructura media se complementa con la especificación de la **varian-cia de $\mathbf{y_{it}}$** como una función de $\boldsymbol{\mu}_{it}$,

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

donde v, es una función conocida.

Liang y Zeger (1986) suponen que la distribución marginal de y_{it} pertenece a la familia exponencial.

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, **ρ**. La matriz de correlación de trabajo se simboliza,

$$\mathbf{R}(\rho)=(corr(\, y_{it}, y_{it'})), \quad t,t'=1,\ldots,T$$

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

Cuando la respuesta $\mathbf{Y_i}$ se observa para todos los sujetos las ecuaciones de estimación propuestas por Liang y Zeger (1986) se simbolizan

$$U_{total}(\beta) = K^{-1/2} \sum_{i=1}^{K} \mathbf{H}_i(\beta) = K^{-1/2} \sum_{i=1}^{K} \mathbf{D}_i(\beta)\Sigma_i^{-1}(\beta,\rho)(\mathbf{Y_i} - \mu_i) = 0$$

(2.2.1)

siendo, $\mathbf{D}_i(\beta)$= diag $(D_{it}(\beta))$ y $D_{it}(\beta) = \dfrac{\partial h(\mathbf{X}_{it}^{'}\beta)}{\partial \beta}$.

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

Una característica útil del enfoque GEE es que no se necesita especificar correctamente la matriz de correlación de trabajo para obtener un estimador consistente de los parámetros o estimar consistentemente la variancia robusta. Sin embargo, esta propiedad de robustez se mantiene sólo cuando existe una diminuta fracción de datos perdidos o cuando ellos son del tipo MCAR.

Se verá a continuación un caso donde se supone que la forma de la pérdida es MAR. Las probabilidades de respuesta $\lambda_{it} = P(r_{it} = 1 \mid r_{i(t-1)} = 1, \mathbf{W}_{it})$ son funciones conocidas de $\mathbf{W}_{it}$ y de un vector de parámetros desconocidos α. Generalmente se elige la función logística

para modelar $\bar{\lambda}_{it}$. Esto es, se supone que dado $r_{i(t-1)}=1$, r_{it} sigue un modelo logístico, con parámetro α, en función de $\bar{\mathbf{W}}_{it}$.

Se pueden usar procedimientos estándares para investigar la forma funcional de $\bar{\lambda}_{it}(.)$, debido a que la probabilidad de respuesta $P(r_{it}=1/r_{i(t-1)}=1,\bar{\mathbf{W}}_{it})$ depende sólo de lo observado $\bar{\mathbf{W}}_{it}$.

El estimador máximo verosímil $\hat{\alpha}$ de α es aquél que maximiza

$$L(\alpha)=\prod_i L_i(\alpha)=\prod_i\prod_t (\bar{\lambda}_{it}(\alpha)^{r_{it}}(1-\bar{\lambda}_{it}(\alpha))^{1-r_{it}})^{r_{i(t-1)}}.$$

(2.2.2)

Se define $\bar{\pi}_{it}(\alpha)=\bar{\lambda}_{i1}$ x....x $\bar{\lambda}_{it}$. Sea

$$\Delta_i(\alpha)=\begin{bmatrix} \bar{\pi}_{i1}(\alpha)^{-1}r_{i1} & 0 & \dots & 0 \\ 0 & \bar{\pi}_{i2}(\alpha)^{-1}r_{i2} & \dots & 0 \\ . & . & \dots & . \\ 0 & 0 & \dots & \bar{\pi}_{iT}(\alpha)^{-1}r_{iT} \end{bmatrix}$$

una matriz diagonal TxT con elementos $\Delta_{it}(\alpha)=\bar{\pi}_{it}(\alpha)^{-1}r_{it}$.

Informalmente la idea de Robins et al. (1995) expresa que si $\boldsymbol{\pi}_{it}(\boldsymbol{\alpha})$ indica la probabilidad que la unidad i permanezca hasta el tiempo t, dada la historia de respuestas observadas del sujeto $y_{i1},\dots,y_{i(t-1)}$ y cualquier información relevante de las covariables, luego la observación y_{it} de esta unidad es representativa de todas las unidades con respuestas y covariables comparables que habrían sido observados si no hubiesen salido del estudio. De esta manera para restablecer el insesgamiento de los estimadores GEE para los datos completos se necesita ponderar la contribución de y_{it} por la inversa de $\bar{\pi}_{it}(\alpha)$.

Esto conduce a extender las ecuaciones de estimación, produciendo las ecuaciones de estimación modificadas, que ponderan las GEE por la inversa de la probabilidad de respuesta,

$$U(\boldsymbol{\beta},\boldsymbol{\alpha}) = K^{-1/2}\sum_{i=1}^{K} U_i(\boldsymbol{\beta},\boldsymbol{\alpha}) = 0\ ,$$

(2.2.3)

Siendo $U_i(\boldsymbol{\beta},\hat{\boldsymbol{\alpha}}) = \mathbf{D}_i(\boldsymbol{\beta})\Delta_i(\hat{\boldsymbol{\alpha}})\boldsymbol{\varepsilon}_i(\boldsymbol{\beta})$. (2.2.4)

$$\boldsymbol{\varepsilon}_i(\boldsymbol{\beta}) = \mathbf{Y}_i - h(\mathbf{X}_i.\boldsymbol{\beta})$$

$$\mathbf{D}_i(\boldsymbol{\beta}) = \text{diag}\left[\frac{\partial h(\mathbf{X}_{it}'\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\Sigma^{-1}(\boldsymbol{\beta},\boldsymbol{\rho})\right],\ \text{siendo}\ \eta_{it} = \mathbf{X}_{it}'\boldsymbol{\beta}\ ,\ \ \mu_{it} = h(\eta_{it})$$

y $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta},\boldsymbol{\rho})$ la matriz de covariancias de trabajo.

Multiplicar $\boldsymbol{\varepsilon}_i(\boldsymbol{\beta})$ por $\Delta_i(\boldsymbol{\alpha})$ es equivalente a ponderar cada componente de los $\boldsymbol{\varepsilon}_i(\boldsymbol{\beta})$ observados por $\pi_{it}(\boldsymbol{\alpha})^{-1}$.

Las ecuaciones (2.2.3) poseen una solución $\hat{\beta}$, que es consistente y asintóticamente normal para $\boldsymbol{\beta}$. En las ecuaciones (2.2.3) se usan las respuestas observadas para todos los sujetos del estudio. La estimación de β se realiza obteniendo α de (2.2.2) y luego resolviendo (2.2.3).

3. ESTUDIO DE SIMULACIÓN

Este estudio se efectúa para varios conjuntos de datos longitudinales, que poseen una distribución gamma multivariada con estructura de correlación entre las medidas repetidas que responde a un proceso autorregresivo de primer orden. Si bien, para las variables gamma no existe una densidad conjunta multivariada, se dice que un vector aleatorio tiene distribución gamma multivariada si cada una de las funciones de densidad marginales es una gamma univariada (Bustos, 2000).

La diferencia entre los conjuntos generados reside en el grado de la correlación entre las medidas repetidas (baja y alta), el número de mediciones repetidas por unidad (T) y el tamaño de la muestra (o número de unidades) (K).

3.1.- Generación de las gammas

El algoritmo construido para obtener variables con distribución gamma multivariada genera muestras de observaciones a partir de un vector con distribución normal multivariada

y utiliza la función inversa de la función de distribución acumulada de la gamma (Bustos, 2000).

Se considera que cada elemento del vector aleatorio (que representa la respuesta multivariada de cada unidad), es decir, cada y_{it}, pertenece a una gamma con parámetros (μ_{it}, ν_{it}), t=1, 2, …, T, es decir, $G(\mu_{it}, \nu_{it})$, siendo $E(y_{it})= \mu_{it}$ y $Var(y_{it})= \mu^2_{it} / \nu_{it}$.

El procedimiento de Bustos (2000), permite generar T variables y_{it} con distribución gamma multivariada

$$\Gamma[(\Gamma_{i1}, \Gamma_{i1}),(\Gamma_{i2}, \Gamma_{i2}),...,(\Gamma_{iT}, \Gamma_{iT})] \qquad (3.1.1)$$

con matriz de correlación

$$\mathbf{R}(\Gamma) = \begin{bmatrix} 1 & \Gamma & \Gamma^2 \ldots & \Gamma^{T-1} \\ \Gamma & 1 & \Gamma \ldots. & \Gamma^{T-2} \\ . & . & . & . \\ \Gamma^{T-1} & \Gamma^{T-2} & \Gamma^{T-3} \ldots.. & 1 \end{bmatrix},$$

siendo, la relación entre los parámetros de $G(\mu_{it}, \nu_{it})$ y $G(\theta_{it}, \gamma_{it})$ la siguiente

$\theta_{it} = \mu_{it} . \gamma_{it}$ y $\nu_{it} = \theta_{it}$. (3.1.2)

Para ello se procede así:

- Generar el vector **¡Error! No se pueden crear objetos modificando códigos de campo.** a partir de una distribución $N(\mathbf{0}, \mathbf{R}(\rho))$
- Calcular **¡Error! No se pueden crear objetos modificando códigos de campo.** , t=1, 2,...,T (3.1.3) donde $\Gamma(\Gamma_{it})$ es la función de distribución acumulada de la N(0,1), valorizada en ξ_{it} y F^{-1} es la inversa de la función de distribución acumulada de una gamma con parámetros $(\alpha_t,1)$. El vector $\mathbf{Z}_i$ que tiene componentes $\{z_{it}$ t=1,2,...,T} posee distribuciones marginales $\Gamma(\Gamma_{i1},1),\Gamma(\Gamma_{i2},1),...,\Gamma(\Gamma_{iT},1)$ y matriz de variancias y covariancias $\mathbf{R}(\Gamma)$.
- Calcular $\mathbf{Y}_i= \mathbf{Z}_i / \boldsymbol{\gamma}_i$ con $\boldsymbol{\gamma}_i = (\Gamma_{i1}, \Gamma_{i2},..., \Gamma_{iT})$ (3.1.4)

3.2.- Descripción del estudio de Montecarlo

En el estudio de Montecarlo, se supone que μ_{it} depende de las variables di y t y que el predictor lineal es de la forma $\mathbf{X}_{it}'\boldsymbol{\beta} = \beta_0 + \beta_1 d_i + \beta_2 t$, i=1,…,K, t=1,…,T, siendo $\mathbf{X}_{it}' = \{1, d_i, t\}$. Los parámetros involucrados en el mismo se fijan en β_o=0.3, β_1 =0.6 y $\boldsymbol{\beta}_2$ =0.032. Los valores de T son 4 y 6 (es decir, se simulan dos conjuntos de observaciones que tienen el primera 4 mediciones repetidas y el otro, 6) y la variable d_i toma los valores 0 ó 1.

1. *Generación de los datos*

Se simulan 8 conjuntos de datos, que corresponden a las combinaciones de K (10 y 20) unidades, para cada una de las cuales se registran T (4 ó 6) mediciones y la correlación entre dos observaciones consecutivas ρ (0.1 y 0.6). La matriz de correlación entre las observaciones de una unidad responde a un proceso autorregresivo de orden 1 (AR(1)) con coeficiente ρ.

1.1. Cada conjunto simulado se obtiene por el proceso que se describe a continuación. La descripción corresponde a la generación del vector de observaciones de la i-ésima unidad que se registra en 4 tiempos $(y_{i1}, y_{i2}, y_{i3}, y_{i4})$. El procedimiento para 6 tiempos (T=6) no se presenta pues es similar.

1.1.1 Se genera un vector de variables normales multivariadas $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3}, \varepsilon_{i4})'$

```
ERROR: rangecheck
OFFENDING COMMAND: .buildcmap

STACK:

-dictionary-
/WinCharSetFFFF-V2TT3491565Ct
/CMap
-dictionary-
/WinCharSetFFFF-V2TT3491565Ct
```