**Guillamet Chargue, Cecilia**

**Rapelli, Cecilia**

**Garcia, María del Carmen**

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

# USO DEL VARIOGRAMA PARA LA SELECCIÓN DE MODELOS DE COVARIANCIA EN LOS MODELOS MIXTOS

**Resumen:**

Los modelos lineales mixtos son apropiados para el análisis de datos longitudinales ya que permiten considerar las distintas fuentes de variación presentes en ese tipo de datos. La modelación adecuada de la covariancia no solamente es útil para la interpretación de la variación aleatoria de los datos, sino que es esencial para obtener inferencias válidas de la estructura media, la cual es de interés primario. Los procedimientos usuales para identificar un modelo para la covariancia requieren que los datos sean balanceados. Cuando los datos son no balanceados la identificación de un modelo se vuelve dificultosa. El variograma es una herramienta ampliamente utilizada en la estadística espacial y adaptada para los datos longitudinales que permite describir la asociación entre las medidas repetidas aún cuando los datos son no balanceados. Una vez seleccionado un modelo para la covariancia, el variograma se puede utilizar como herramienta de diagnóstico para corroborar si el modelo seleccionado es apropiado En este trabajo se presenta el uso del variograma como herramienta exploratoria y de diagnóstico. Se ilustra su aplicación mediante un conjunto de datos referidos a un estudio de pérdida de peso en mujeres.

Palabras claves: Datos longitudinales. Modelos lineales mixtos. Variograma

**Abstract:**

Mixed linear models are suitable for modeling longitudinal data because the model allows considering three qualitatively different sources of random variation. The adequate modeling of the covariance is not only useful for the interpretation of the random variability of the data, but also it is essential for obtaining valid inferences on the mean structure, which is of prime interest. Equally spaced observation periods are required for the usual methods of detecting a model for the covariance. The selection of a suitable model is more complicated when there are unequally spaced observation periods. The variogram has been widely used in spatial statistics to represent the covariance structure in geostatistical data. For longitudinal data, the variogram describes the association among repeated values and allows estimation with irregular observation periods. After a candidate model for the covariance is selected, the variograma can be used as a diagnostic tool for assessing the adequacy of the selected model. In this paper, the variogram is used as a descriptive and diagnostic tool. For its application we used a set of data related to a weight loss study in women.

Keywords: Longitudinal data. Linear mixed models. Variogram

## 1. Introducción

Los estudios longitudinales son frecuentes en una amplia variedad de disciplinas. Éstos están conformados por mediciones repetidas de una variable de interés realizadas a una misma unidad a través del tiempo y permiten caracterizar el cambio en la respuesta y los factores que influencian el cambio.

Una característica distintiva de los datos longitudinales es que las mediciones sobre una misma unidad están correlacionadas positivamente, la cual se debe modelar correctamente para obtener inferencias válidas.

Los modelos lineales mixtos son adecuados para el análisis de datos longitudinales. Los mismos permiten la inclusión de covariables mediante efectos fijos mientras que los efectos aleatorios del mismo reflejan las múltiples fuentes de heterogeneidad y/o correlación entre y dentro de las unidades.

Una etapa fundamental del proceso de construcción del modelo es la selección de la estructura de covariancia. Existen diferentes herramientas para identificar modelos para la misma. Sin embargo, la mayoría sólo se puede utilizar cuando las ocasiones de medición son las mismas para todas las unidades. Cuando esto no ocurre**,** la identificación se puede realizar mediante el variograma, una herramienta alternativa ampliamente utilizada en la estadística espacial y adaptada para datos longitudinales.

El variograma es una herramienta descriptiva que permite determinar qué componentes de variación es conveniente incluir en el modelo. Para ello se grafican los valores de una función, el variograma, versus las diferencias entre ocasiones de medición.

Una vez seleccionado un modelo para la covariancia, el variograma resulta una herramienta de diagnóstico para corroborar si el modelo seleccionado es correcto.

Este trabajo presenta el uso del variograma en datos longitudinales, tanto en la etapa exploratoria como confirmatoria. Para ilustrar su uso se utiliza un conjunto de datos referidos a un estudio de pérdida de peso en mujeres, extraídos del libro “Modeling Longitudinal Data” (Weiss, 2005).

## 2. Datos longitudinales y modelos mixtos

Los datos longitudinales están conformados por mediciones repetidas de una variable de interés realizadas a una misma unidad o individuo. Un rasgo característico de los estudios longitudinales es que las mediciones se realizan a través del tiempo. El objetivo principal en este tipo de estudios es caracterizar el cambio en la respuesta a través del tiempo y los factores que influencian el cambio.

Las mediciones repetidas obtenidas de un mismo individuo en las diferentes ocasiones están correlacionadas y dicha correlación se debe tener en cuenta en el análisis para obtener inferencias válidas.

En los datos longitudinales se pueden distinguir dos fuentes de variabilidad que tienen impacto sobre la correlación entre las mediciones repetidas de una unidad: la heterogeneidad entre unidades y la variabilidad intra unidad. Esta última se puede producir por la variabilidad de la variable respuesta inherente a cada unidad o por errores de medición.

Los modelos lineales mixtos son útiles para modelar datos longitudinales debido a su flexibilidad para representar las múltiples fuentes de variación y correlación. En estos modelos la respuesta media se expresa como una combinación de características poblacionales (efectos fijos), que se asumen comunes a todos los individuos, y un conjunto de efectos específicos que son únicos para cada individuo (efectos aleatorios).

Sea la variable aleatoria $\mathrm{Y_{ij}}$ la respuesta de interés para el individuo $i$-ésimo medido en la ocasión $j(t_{ij}),\ i = 1, \dots, m,\ j = 1, \dots, n_i$, y sea $\mathbf{Y_i}$ un vector de dimensión $(n_i \times 1)$ de todas las mediciones repetidas de la $i$-ésima unidad, $\mathbf{Y_i} = \left(\mathrm{Y_{i1}}, \mathrm{Y_{i2}}, \dots, \mathrm{Y_{in_i}}\right)'$.

La expresión del modelo lineal mixto es,

$$\mathbf{Y_i} = \mathbf{X_i}\boldsymbol{\beta} + \mathbf{Z_i}\mathbf{b_i} + \mathbf{e}_{(1)\mathrm{i}} + \mathbf{e}_{(2)\mathrm{i}}\,, \qquad i = 1, \dots, m\,, \tag{1}$$

siendo, $\mathbf{X_i}$ una matriz de diseño de $(n_i \times p)$ que caracteriza la parte sistemática del modelo, $\boldsymbol{\beta}$ un vector de dimensión $(p \times 1)$ de parámetros denominados efectos fijos, $\mathbf{Z_i}$ una matriz de diseño de dimensión $(n_i \times k)$ que caracteriza la parte aleatoria del modelo, $\mathbf{b_i}$ un vector de dimensión $(k \times 1)$ de efectos aleatorios que representa la variabilidad entre individuos, $\mathbf{e}_{(1)\mathrm{i}}$ un vector de dimensión $(n_i \times 1)$ de los errores de medición, que refleja la variación debida al proceso de medición, $\mathbf{e}_{(2)\mathrm{i}}$ un vector de dimensión $(n_i \times 1)$ que caracteriza la variabilidad biológica intra individuo.

Los supuestos que se realizan son:

- $\mathbf{e}_{(1)\mathrm{i}} \sim N_{n_i}(0, \sigma^2\mathbf{I}_{\mathrm{ni}})$, donde $\sigma^2$ es la variancia debida a los errores de medición e $\mathbf{I}_{\mathrm{n_i}}$ la matriz identidad de dimensión $n_i$.
- $\mathbf{e}_{(2)\mathrm{i}} \sim N_{n_i}(0, \tau^2\mathbf{H_i})$, donde $\tau^2$ es la variancia de la componente de correlación serial y $\mathbf{H_i}$ es una matriz de correlaciones de dimensión $(n_i \times n_i)$.
- $\mathbf{b_i} \sim N_k(0, \mathbf{D_i})$.
- $\mathbf{b_1}, \dots, \mathbf{b_m}, \mathbf{e}_{(1)1}, \dots, \mathbf{e}_{(1)\mathrm{m}}, \mathbf{e}_{(2)1}, \dots, \mathbf{e}_{(2)\mathrm{m}}$ son independientes.

Se asume que $h_{ijj'}$ un elemento de $\mathbf{H_i}$ (donde el primer subíndice corresponde al individuo y los otros dos a las posiciones dentro de la matriz) se modela como $h_{ijj'} = g\left(u_{ijj'}\right)$, siendo $u_{ijj'} = \left|t_{ij} - t_{ij'}\right|$, para una función decreciente $g(.)$ con $g(0) = 1$. Esto significa que la correlación entre $e_{(2)ij}$ y $e_{(2)ij'}$ sólo depende del intervalo de tiempo, o rezago, entre las mediciones $Y_{ij}$ e $Y_{ij'}$, y decrece a medida que dicho intervalo aumenta.

Si bien es posible que los tres elementos contribuyan a la variabilidad de los datos longitudinales, uno puede ser más dominante que otro, y no es necesario incluir todas las componentes estocásticas en el modelo. Los modelos que incluyen varios efectos aleatorios, correlación serial y error de medición pueden tener problemas de estimación (Diggle et al., 1994), por lo cual no siempre es posible modelar por separado las dos fuentes de variación intra unidad.

En este modelo se pueden distinguir las distribuciones marginal y condicional con los siguientes parámetros:

| Marginal | Condicional |
|---|---|
| $E(\mathbf{Y}_\mathrm{i}) = \mathbf{X}_\mathrm{i}\boldsymbol{\beta}$ | $E(\mathbf{Y}_\mathrm{i}/\mathbf{b}_\mathrm{i}) = \mathbf{X}_\mathrm{i}\boldsymbol{\beta} + \mathbf{Z}_\mathrm{i}\mathbf{b}_\mathrm{i}$ |
| $Var(\mathbf{Y}_\mathrm{i}) = \mathbf{Z}_\mathrm{i}\mathbf{D}\mathbf{Z}_\mathrm{i}' + \mathbf{R}_\mathrm{i} = \mathbf{Z}_\mathrm{i}\mathbf{D}\mathbf{Z}_\mathrm{i}' + \tau^2\mathbf{H}_\mathrm{i} + \sigma^2\mathbf{I}_{\mathrm{n_i}} = \boldsymbol{\Sigma}_\mathrm{i}$ | $Var(\mathbf{Y}_\mathrm{i}/\mathbf{b}_\mathrm{i}) = \mathbf{R}_\mathrm{i} = \tau^2\mathbf{H}_\mathrm{i} + \sigma^2\mathbf{I}_{\mathrm{n_i}}$ |

Los parámetros que caracterizan la media o parte sistemática y la variabilidad o parte aleatoria del modelo mixto se pueden estimar utilizando los métodos de máxima verosimilitud y de máxima verosimilitud restringida.

### 3. Variograma

La mayoría de las técnicas utilizadas para seleccionar una estructura de covariancia requieren que el conjunto de datos sea balanceado. Una función que describe la asociación entre mediciones repetidas y se construye fácilmente con datos no balanceados es el variograma. El mismo se utiliza tanto en la etapa exploratoria como en la confirmatoria de selección de la estructura de covariancia.

Históricamente, el variograma ha sido utilizado en la estadística espacial para representar la estructura de covariancia de datos geoestadísticos. A diferencia de los datos espaciales, que son a dos dimensiones, los datos longitudinales tienen una sola dimensión, el tiempo.

En el contexto de datos longitudinales, la función $\gamma\left(u_{ijj'}\right)$ se denomina variograma y se define como,

$$\gamma\left(u_{ijj'}\right) = \frac{1}{2}E\left[\left(r_{ij} - r_{ij'}\right)^2\right], \tag{3}$$

donde $u_{ijj'} = \left|t_{ij} - t_{ij'}\right|, \quad i = 1, \ldots, m \quad j, j' = 1, \ldots, n_i, \ j \neq j'$.

La estimación del variograma, $\hat{\gamma}\left(u_{ijj'}\right)$, se denomina variograma muestral y se calcula a partir de los valores $\mathrm{v}_{ijj'} = \frac{1}{2}\left(r_{ij} - r_{ij'}\right)^2$ siendo $r_{ij} = Y_{ij} - X_i\hat{\beta}_{MCO}$ los residuos mínimos cuadrados ordinarios que surgen de considerar un modelo preliminar para la media, ignorando cualquier dependencia entre las mediciones repetidas.

Si los datos son balanceados habrá más de un valor $\mathrm{v}_{ijj'}$ para cada valor de $u_{ijj'}$ y $\hat{\gamma}\left(u_{ijj'}\right)$ es el promedio de todos los $\mathrm{v}_{ijj'}$, correspondientes al mismo $u_{ijj'}$. Cuando los datos son no balanceados se usan métodos de suavizado (por ejemplo, Loess) para estimar el variograma, $\hat{\gamma}\left(u_{ijj'}\right)$.

La variancia total se estima como (Verbeke et al., 2000),

$$\frac{1}{2N^*}\sum_{i \neq i'}^{m}\sum_{j=1}^{n_i}\sum_{j'=1}^{n_i}\left(r_{ij} - r_{i'j'}\right)^2, \tag{4}$$

donde $N^*$ es el número de términos de la suma.

El gráfico de $\hat{\gamma}\left(u_{ijj'}\right)$ versus $u_{ijj'}$, junto con la variancia total estimada, permite decidir cuáles de las tres componentes estocásticas deben incluirse en el modelo. Si se identifica que una componente de correlación serial se debe incluir en el modelo, el gráfico permite seleccionar una función de covariancia apropiada.

Una desventaja del variograma muestral es que puede resultar muy sensible a valores atípicos. Como se basa en el cuadrado de las diferencias entre pares de residuos cada residuo $r_{ij}$ aparece en $(n_i - 1)$ diferencias al cuadrado en (3) y, por lo tanto, un único outlier puede afectar la estimación del variograma en varios rezagos $u_{ijj'}$.

El modelo (1), $\mathbf{Y_i} = \mathbf{X_i}\boldsymbol{\beta} + \mathbf{Z_i}\mathbf{b_i} + \mathbf{e}_{(1)\mathrm{i}} + \mathbf{e}_{(2)\mathrm{i}}$, se puede escribir como $\mathbf{Y_i} = \mathbf{X_i}\boldsymbol{\beta} + \boldsymbol{\varepsilon_i}$, siendo $\boldsymbol{\varepsilon_i} = \mathbf{Z_i}\mathbf{b_i} + \mathbf{e}_{(1)\mathrm{i}} + \mathbf{e}_{(2)\mathrm{i}}$ con variancia $Var(\boldsymbol{\varepsilon_i}) = \mathbf{Z_i}\mathbf{D}\mathbf{Z_i'} + \tau^2\mathbf{H_i} + \sigma^2\mathbf{I}_{\mathrm{n_i}}$.

Suponiendo que el único efecto aleatorio en el modelo es el correspondiente a la ordenada al origen, la covariancia marginal se reduce a,

$$Var(\boldsymbol{\varepsilon_i}) = v^2\mathbf{J}_{\mathrm{n_i}} + \tau^2\mathbf{H_i} + \sigma^2\mathbf{I}_{\mathrm{n_i}}, \tag{5}$$

donde, $\mathbf{J}_{\mathrm{n_i}}$ es una matriz $(n_i \times n_i)$ de unos y $v^2$ es la variancia debida a la ordenada aleatoria. Esto implica que los errores $\varepsilon_{ij}$ tienen variancia constante $v^2 + \tau^2 + \sigma^2$ y que la correlación entre $\varepsilon_{ij}$ y $\varepsilon_{ij'}$ correspondientes al mismo individuo es,

$$Corr\left(\varepsilon_{ij}, \varepsilon_{ij'}\right) = \frac{v^2 + \tau^2 g\left(u_{ijj'}\right)}{v^2 + \tau^2 + \sigma^2}.$$

La expresión del variograma dada en (3) resulta,

$$\begin{aligned}\gamma\left(u_{ijj'}\right) = \frac{1}{2}E\left[\left(\varepsilon_{ij} - \varepsilon_{ij'}\right)^2\right] &= \frac{1}{2}\left[Var\left(\varepsilon_{ij}\right) + Var\left(\varepsilon_{ij'}\right) - 2cov\left(\varepsilon_{ij}, \varepsilon_{ij'}\right)\right] \\ &= \frac{1}{2}\left[2(v^2 + \tau^2 + \sigma^2) - 2\left(v^2 + \tau^2 g\left(u_{ijj'}\right)\right)\right] \\ &= \sigma^2 + \tau^2\left[1 - g\left(u_{ijj'}\right)\right] \end{aligned} \tag{6}$$

para $i = 1, \ldots, m$ y para $j \neq j' = 1, \ldots, n_i$. Se observa aquí que el variograma es una función creciente de $u_{ijj'}$, dado que la autocorrelación es positiva y decrece a medida que aumenta la separación en el tiempo, y sólo depende de las diferencias en el tiempo $u_{ijj'}$. Cuando $u_{ijj'} = 0$, el variograma resulta $\gamma(0) = \sigma^2$, y converge a $\gamma\left(u_{ijj'}\right) = \sigma^2 + \tau^2$ cuando $u_{ijj'}$ tiende a infinito.

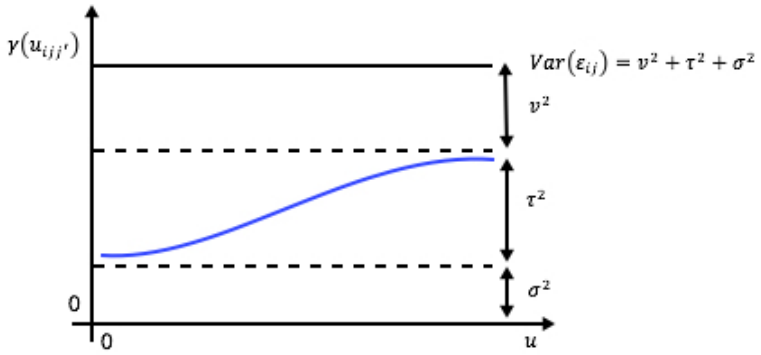
Una característica de los modelos con errores de medición es que $\gamma(0) \neq 0$. Cuando el rezago tiende a infinito, $\gamma\left(u_{ijj'}\right)$ se aproxima a $(\sigma^2 + \tau^2)$, un valor menor que la variancia del error. La diferencia entre la asíntota alcanzada por la línea ajustada y la variancia total corresponde a la variabilidad entre individuos. Un variograma muestral con una curva creciente sugiere la inclusión de una componente de correlación serial. Si además $\hat{\gamma}(0)$ no tiende a

cero estaría indicando que será necesario incluir también errores de medición. Cuando $\hat{\gamma}\left(u_{ijj'}\right)$ no tiende a la variancia total cuando $u_{ijj'}$ aumenta, sino a un valor menor, sugiere la inclusión de una ordenada al origen aleatoria.

Muchas veces resulta dificultosa la identificación de la función de correlación, principalmente cuando existen efectos aleatorios. Verbeke y Molenberghs (2000) sugieren que incluir la correlación serial, si el variograma indica que está presente, es más importante que especificar correctamente la función de correlación, dado que ignorar la presencia de la misma conduce a inferencias incorrectas sobre los coeficientes de regresión y a estimaciones ineficientes de los parámetros.

El siguiente gráfico muestra el variograma para el modelo (1).

Gráfico 1: Variograma correspondiente al modelo (1).



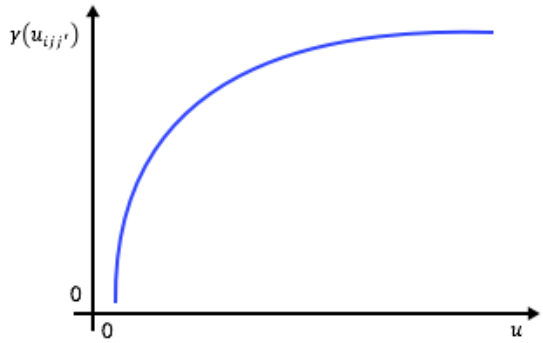

Cuando existe variación biológica intra unidad, la forma que presenta el variograma es distinta según la estructura de correlación que caracterice a los datos. Las estructuras más utilizadas para modelar la correlación serial son la exponencial y la gaussiana. La diferencia entre estas estructuras es la forma en que la correlación, entre los términos de error, decrece a medida que aumenta la distancia en el tiempo entre las observaciones.

El modelo de correlación exponencial $g\left(u_{ijj'}\right) = e^{\left(-\phi u_{ijj'}\right)}$, $\phi > 0$, siendo $\phi$ un parámetro de correlación y el variograma correspondiente resulta,

$$\gamma\left(u_{ijj'}\right) = \tau^2\left(1 - e^{-\phi u_{ijj'}}\right).$$

El Gráfico 2 presenta el variograma para esa estructura.

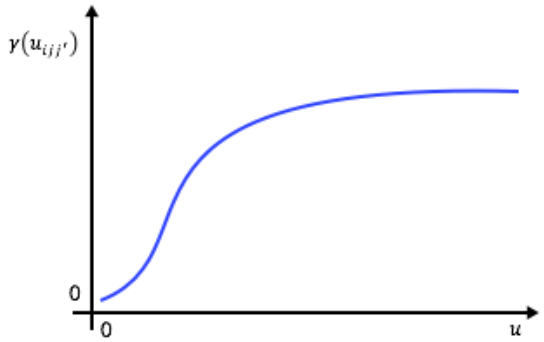
Gráfico 2: Variograma correspondiente a la función de correlación exponencial.



Para el modelo de correlación gaussiana $g\left(u_{ijj'}\right) = e^{\left(-\phi\left(u_{ijj'}^2\right)\right)}$, $\phi > 0$, el variograma es,

$$\gamma\left(u_{ijj'}\right) = \tau^2\left(1 - e^{-\phi\left(u_{ijj'}^2\right)}\right).$$

Gráfico 3: Variograma correspondiente a la función de correlación gaussiana.




La diferencia más importante entre las dos funciones de correlación, que permite la identificación de las mismas, es su comportamiento en la cercanía de $u_{ijj'} = 0$, es decir, la forma en la cual el variograma aumenta en los primeros rezagos y la forma en la que el mismo tiende a la variancia total.

A pesar de que la correlación serial parece un rasgo característico de los modelos de datos longitudinales, en algunas situaciones podría estar dominada por la combinación de efectos aleatorios y errores de medición. La curva no paramétrica ajustada en el variograma muestral podría tener pendiente cero, lo cual indica que una estructura de covariancia que incorpora correlación serial es innecesaria en el modelo.

Como se mencionó anteriormente, el variograma, se puede utilizar también como herramienta de diagnóstico para verificar si el modelo seleccionado para la covariancia es apropiado, utilizando los residuos Cholesky.

Dada una estimación de la matriz de covariancia, $\widehat{\mathbf{\Sigma}}_\mathrm{i}$, la descomposición de Cholesky crea una matriz triangular inferior, $\mathbf{L}_\mathrm{i}$, tal que $\widehat{\mathbf{\Sigma}}_\mathrm{i} = \mathbf{L}_\mathrm{i}\mathbf{L}_\mathrm{i}'$. Se utiliza la matriz $\mathbf{L}_\mathrm{i}$ o, más específicamente $\mathbf{L}_\mathrm{i}^{-1}$, para transformar los residuos, $\mathbf{r}_\mathrm{i}^* = \mathbf{L}_\mathrm{i}^{-1}\mathbf{r}_\mathrm{i}$, de modo que se obtiene un conjunto de residuos transformados aproximadamente no correlacionados y con variancia unitaria, denominados residuos Cholesky.

Cuando el variograma se construye usando los residuos de Cholesky marginales $\left(r_{ij}^*\right)$ se tiene que,

$$\gamma\left(u_{ijj'}\right) = \frac{1}{2}Var\left(r_{ij}^*\right) + \frac{1}{2}Var\left(r_{ij'}^*\right) - Cov\left(r_{ij}^*; r_{ij'}^*\right) \cong \frac{1}{2}(1) + \frac{1}{2}(1) - 0 = 1.$$

Entonces, en un modelo correctamente especificado para la matriz de covariancia, el gráfico del variograma de estos residuos versus el tiempo transcurrido entre las correspondientes observaciones, debería fluctuar aleatoriamente alrededor de una línea centrada en uno y no mostrar ningún patrón sistemático.

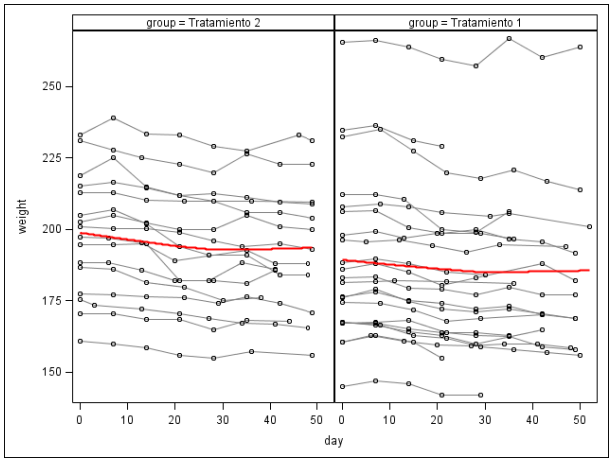
## 4. Resultados

Los datos que se utilizan para la aplicación, extraídos del libro “Modeling Longitudinal Data” (Weiss, 2005), consisten en los pesos, en libras, de 38 mujeres inscriptas en un estudio de pérdida de peso. Al comienzo del estudio, luego de la primera medición, las mujeres fueron asignadas aleatoriamente a dos grupos que difieren en el tratamiento recibido. Una vez por semana se registró el peso de las mujeres. Como las pacientes no concurrían a la cita en las mismas fechas se registró, además del peso, el día en el que se realizó la medición. Los días en que se realizaron las mediciones no fueron los mismos para todas las mujeres, haciendo de éste un conjunto de datos no balanceados.

En el gráfico de perfiles individuales y promedio por grupo (Gráfico 4) se muestra la evolución del peso de las mujeres a través del tiempo para ambos tratamientos. En el mismo se observa que el peso presenta una leve disminución, sin embargo para muchas mujeres permanece constante. Al comienzo del estudio (día cero) se observa mucha variabilidad en los pesos de las mujeres, en ambos grupos, dando un indicio de que un efecto aleatorio en la ordenada al origen debería ser tenido en cuenta en el análisis. Por el contario, un efecto aleatorio en la pendiente no es necesario, dado que las pendientes entre las participantes parecen ser similares.

Resulta importante destacar que la paciente 26, asignada al tratamiento 1, presenta durante todo el periodo de estudio, pesos mayores que las demás pacientes, por lo cual podría ser considerada una unidad atípica.

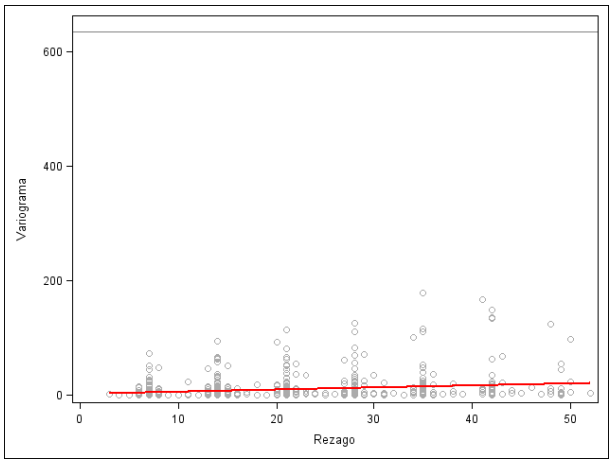
Gráfico 4: Gráfico de perfiles individuales y de perfiles promedio por tratamiento.




La construcción del modelo lineal mixto comienza con la especificación de un modelo adecuado para la covariancia. Para guiar esta selección se utiliza el variograma muestral como herramienta exploratoria.

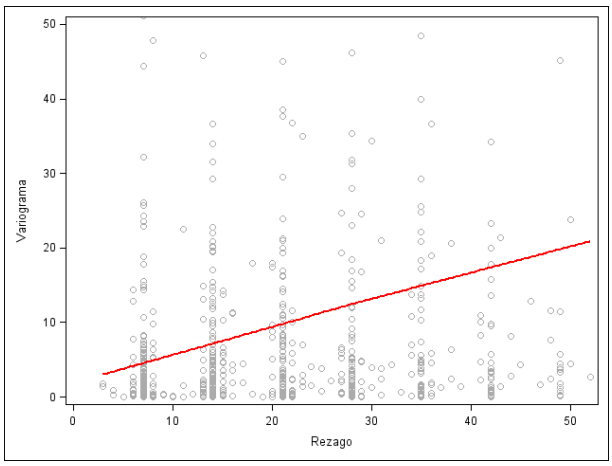
En el Gráfico 5 se observa que resultaría apropiado ajustar un modelo con un efecto aleatorio en la ordenada, dado que la curva ajustada dista en gran magnitud de la variancia total, lo que está en correspondencia con lo observado en el gráfico de perfiles individuales por grupo (Gráfico 4). La curva ajustada presenta una pendiente distinta de cero, sugiriendo que es necesario incluir una componente de correlación serial en el modelo.

Gráfico 5: Variograma muestral.




La identificación de la función de correlación en base al Gráfico 5 resulta dificultosa. Visualmente se asemeja a una línea recta que aumenta lentamente por lo que el patrón observado se correspondería con una función exponencial, con parámetro cercano a cero. Se puede apreciar mejor la forma de la misma en el Gráfico 6, que se focaliza en la curva ajustada.

Gráfico 6: Variograma muestral focalizado.




Por otro lado, en los Gráfico 5 y 6, se puede apreciar que la curva ajustada tiende a cero para rezagos chicos, indicando que la variación debida los errores de medición es despreciable, y que sería apropiado modelar ambas fuentes de variación intra individuo de forma conjunta.

De acuerdo a lo observado en los gráficos, se postula un modelo lineal con un efecto aleatorio para la ordenada al origen y modelando la variabilidad intra individuo conjuntamente. Además se considera la covariable tratamiento asignado, resultando el modelo,

$$Y_{ij} = (\beta_0 + \beta_{01} d_i + b_{0i}) + (\beta_1 + \beta_{11} d_i) t_{ij} + e_{ij} \qquad i = 1, \dots, 38 \quad j = 1, \dots, n_i, \tag{7}$$

siendo,

$Y_{ij}$ el peso de la $i$-ésima mujer en la $j$-ésima ocasión de medición.

$t_{ij}$ la fecha en que se realiza la $j$-ésima medición de la participante $i$.

$d_i$ una variable indicadora que vale uno para las mujeres asignadas al tratamiento uno y cero para las mujeres asignadas al tratamiento dos.

Los supuestos del modelo son:

- $\mathrm{b}_{0\mathrm{i}} \sim N(0, v^2)$
- $\mathbf{e}_{\mathrm{i}} \sim N(0, \tau^2 \mathbf{H}_{\mathrm{i}})$
- $\mathrm{b}_{0\mathrm{i}}$ y $\mathbf{e}_{\mathrm{i}} = \left(\mathbf{e}_{\mathrm{i}1}, \ldots, \mathbf{e}_{\mathrm{in_i}}\right)'$ son independientes.

El Gráfico 5 no permite distinguir claramente la estructura de correlación serial apropiada para representar la variabilidad biológica intra individuo, por lo cual se consideraron dos funciones de correlación serial (exponencial y gaussiana) y una de independencia. Se ajustan tres modelos usando cada una de las estructuras. Para evaluar cuál de ellos es más conveniente utilizar se construye, como elemento diagnóstico, el variograma con residuos Cholesky y se calculan los criterios de información.

Tabla 1: Criterios de información para las distintas estructuras de covariancia.

| | **AIC** | **BIC** |
|---|---|---|
| **Modelo con estructura exponencial** | 1471,8 | 1476,7 |
| **Modelo con estructura gaussiana** | 1481,0 | 1485,9 |
| **Modelo sin correlación serial** | 1538,9 | 1542,1 |

El modelo que supone errores independientes presenta valores de AIC y BIC mucho más grandes que los modelos de correlación serial, en correspondencia con lo observado en los variogramas. La estructura exponencial es la que muestra un mejor ajuste dado que presenta los menores valores de AIC y BIC.

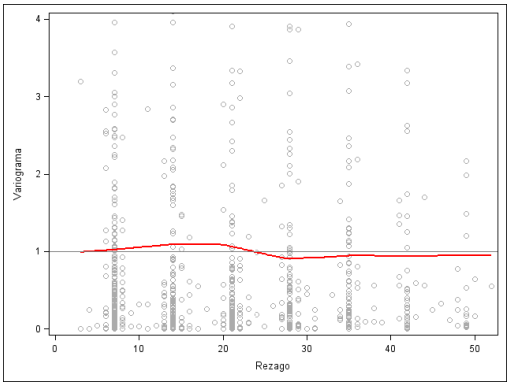
El Gráfico 7 muestra que las dos estructuras de correlación serial proveen una buena caracterización de la correlación entre las medidas repetidas, puesto que el variograma fluctúa aleatoriamente alrededor de uno, pero no son concluyentes respecto de cuál es la mejor. En contraste, en el variograma de los residuos Cholesky correspondientes al modelo que supone errores independientes se observa claramente la especificación errónea de la estructura de covariancias puesto que la curva ajustada no fluctúa alrededor de uno, sino que la misma muestra una forma consistente con el modelo de correlación exponencial.

Tanto las diferencias entre los criterios de información como entre los variogramas de los residuos transformados correspondientes a las dos funciones de correlación serial son muy pequeñas, con lo cual la elección de la misma no influye en gran medida sobre las estimaciones e inferencias que se realicen. Sin embargo, dado que los valores de AIC y BIC son menores para la estructura exponencial se selecciona esta estructura para modelar la corre-

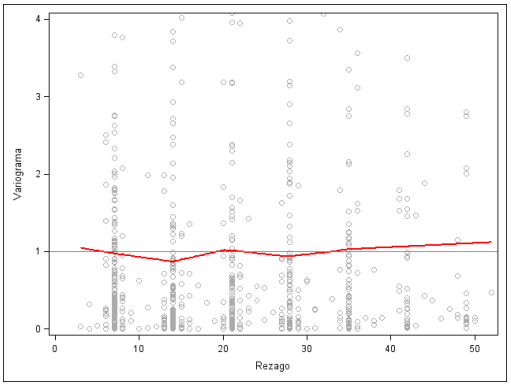
lación serial.

Gráfico 7: Variogramas de los residuos Cholesky para los distintos modelos de correlación.

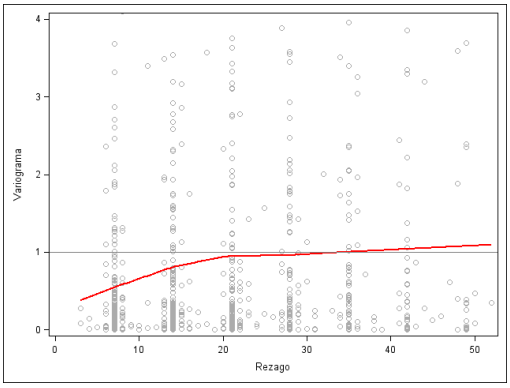
(a) Exponencial




(b) Gaussiana




(c) Independencia

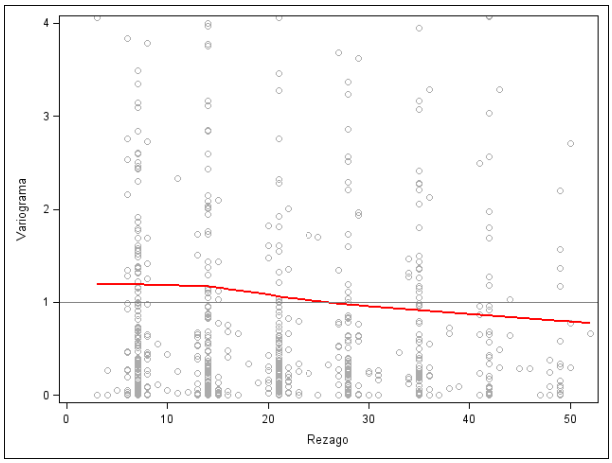



Para corroborar si es necesario incluir un efecto aleatorio para la ordenada al origen se plantea un modelo que no incluya el mismo, se realiza un gráfico del variograma con residuos Cholesky y se calculan los valores de AIC y BIC para probar la bondad del ajuste de ambos modelos.

El modelo con efecto aleatorio en la ordenada al origen presenta un valor de AIC de 1471,8 y de BIC 1476,7, resultando mejor que el modelo sin efecto aleatorio (AIC=1484,9 y BIC=1488,1). Lo cual está en correspondencia con lo observado en el Gráfico 5.

En el Gráfico 8 se observa que el variograma no fluctúa aleatoriamente alrededor de la línea centrada en uno, sugiriendo que este modelo no está correctamente especificado.

Gráfico 8: Variograma de los residuos Cholesky para un modelo sin efecto aleatorio en la ordenada.




En este estudio interesa determinar si el tratamiento al que fueron sometidas las participantes influye en el cambio del peso a través del tiempo. Se prueba la significación del efecto tratamiento mediante el test de razón de verosimilitud.

El valor de la estadística resulta $G^2 = 1{,}6$ ($p = 0{,}4493$) por lo que se concluye que no hay evidencia muestral que sugiera que el peso medio de las participantes difiere debido a los tratamientos a través de todo el periodo del estudio.

En base a los resultados obtenidos se postula un modelo sin el efecto tratamiento, resultando,

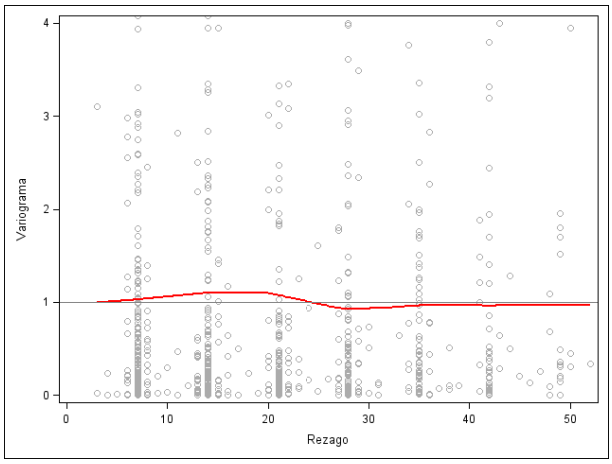
$$\hat{Y}_{ij} = 192{,}85 + \hat{b}_{0i} - 0{,}1567 t_{ij} \qquad i = 1, \ldots, 38 \quad j = 1, \ldots, n_i.$$

La variancia estimada del efecto aleatorio es $\widehat{\mathrm{D}} = \hat{v}^2 = 627{,}44$ y las estimaciones de los parámetros de la matriz de covariancia intra individuo, $\hat{Var}(\mathbf{e}_\mathrm{i}) = \hat{\tau}^2 \widehat{\mathrm{H}}_\mathrm{i}$, con elementos de $\widehat{\mathrm{H}}_\mathrm{i} = e^{\left(-\hat{\phi} u_{ijj'}\right)}$, $\hat{\tau}^2 = 11{,}2497$ y $\hat{\phi} = \frac{1}{15{,}2049} = 0{,}0658$.

De acuerdo al valor de $\hat{\phi}$ la correlación para mediciones separadas por un día resulta de 0,9363.

Para completar el análisis de los datos es necesario realizar un análisis de residuos para evaluar la adecuación del modelo ajustado. Para verificar si el modelo seleccionado para la covariancia resulta apropiado, se realiza el variograma de los residuos de Cholesky marginales del modelo seleccionado. En el Gráfico 9 se puede observar que la curva ajustada fluctúa alrededor del uno, lo cual indica que el modelo para la covariancia está correctamente especificado.

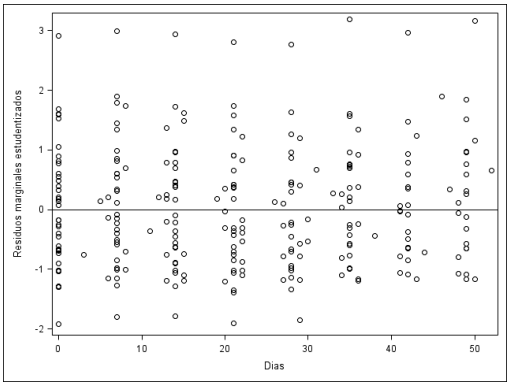
Gráfico 9: Variograma de los residuos Cholesky.



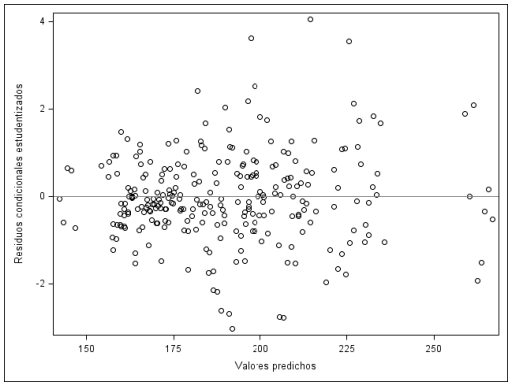

El Gráfico 10 (a) de los residuos marginales ($\mathbf{r}_{\mathbf{mi}} = \mathbf{Y}_\mathbf{i} - \hat{E}(\mathbf{Y}_\mathbf{i}) = \mathbf{Y}_\mathbf{i} - \mathbf{X}_\mathbf{i}\widehat{\boldsymbol{\beta}}$) versus los días permite evaluar la estructura media del modelo. Dado que en el gráfico los puntos se distribuyen de forma aleatoria alrededor de una media constante igual a cero y no presentan ningún patrón sistemático se puede concluir que el modelo para la respuesta media está correctamente especificado. Resulta de interés destacar que los puntos que se encuentran más alejados del resto son los pertenecientes a la participante número 26, la cual se había advertido como un posible outlier.

Gráfico 10: Gráfico de los residuos.

(a) marginales estudentizados versus los días



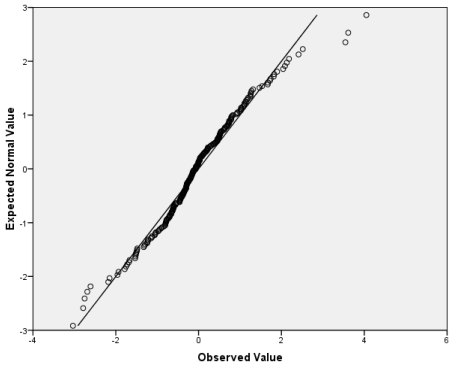

(b) residuos condicionales estudentizados vs. los valores predichos




En el Gráfico 10 (b) de los residuos condicionales estudentizados ($\mathbf{r}_{\mathbf{ci}}^{\boldsymbol{estud}} = \frac{\mathbf{r}_{\mathbf{ci}}}{\sqrt{\widehat{Var}(\mathbf{r}_{\mathbf{ci}})}}$, $\mathbf{r}_{\mathbf{ci}} = \mathbf{Y}_\mathbf{i} - \mathbf{X}_\mathbf{i}\widehat{\boldsymbol{\beta}} - \mathbf{Z}_\mathbf{i}\hat{\mathbf{b}}_\mathbf{i}$) versus los valores predichos se observa que los residuos fluctúan aleatoriamente alrededor de cero con un rango de variación constante, indicando que no se evidencia heterocedasticidad de variancias.

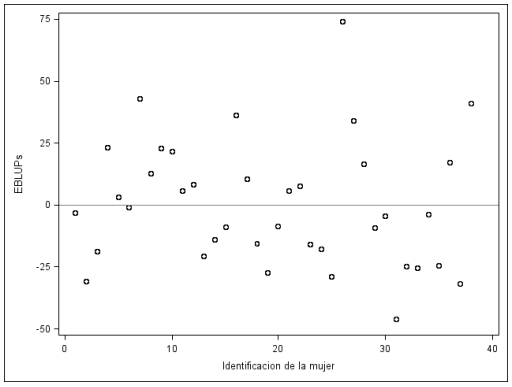
El supuesto de normalidad de los errores intra individuos se evalúa a través de un gráfico probabilístico normal de los residuos condicionales estudentizados. En el Gráfico 11 se observa que los residuos presentan una distribución aproximadamente normal con un leve alejamiento de la normalidad de las colas.

Gráfico 11: Gráfico probabilísticos normal de los residuos condicionales estudentizados.




Se evalúa la existencia de individuos atípicos utilizando el gráfico de los EBLUP ($\mathbf{Z}_i\hat{\mathbf{b}}_i$) versus el número de unidad. De acuerdo al Gráfico 12 se puede observar que la participante número 26 es un posible individuo atípico, dado que se aleja considerablemente del resto.

Gráfico 12: Gráfico de los EBLUP vs. el número de unidad.




## 5. Consideraciones finales

Los modelos lineales mixtos se utilizan para modelar los datos longitudinales debido a su flexibilidad para representar las múltiples fuentes de variación y correlación y para manejar datos incompletos y no balanceados.

Una etapa fundamental en el proceso de construcción del modelo es la selección de la estructura de covariancia. El variograma muestral se presenta como la única herramienta disponible para la identificación de la misma cuando los datos son no balanceados.

En este trabajo se ilustra el uso del variograma en la etapa exploratoria, para la identificación de las componentes estocásticas que se deben incluir en el modelo, así como para la identificación del modelo de correlación serial. También se ilustra su uso en el análisis de residuos como herramienta diagnóstica para evaluar el ajuste del modelo de covariancia seleccionado.

En la aplicación, el variograma muestral permitió identificar claramente qué componentes estocásticas son convenientes incluir en el modelo para caracterizar de forma adecuada la variabilidad de los datos. Sin embargo, la selección de la función de correlación serial no fue

evidente. El uso del variograma como herramienta de diagnóstico permite evaluar de forma sencilla y concluyente el modelo de covariancia seleccionado.

Resulta necesario realizar posteriores análisis para evaluar si la identificación de una unidad atípica afecta el ajuste del modelo y las conclusiones obtenidas.

## REFERENCIAS BIBLIOGRÁFICAS


Akaike, H. (1973). *Information theory and an extension of the maximum likelihood principle.* In Petrov, B. N., Csáki, F., eds Second International Symposium on Information Theory. Budapest: AkadémiaiKiadó, p 267-281.

Dawson, K. S.; Gennings, C.; Carter, W. H. (1997). *Two graphical techniques useful in detective correlation structure in repeated meausresdata.*The American Statistician, vol. 51, 275-283.

Diggle, P. J. (1988): *An approach to the analysis of repeated measures*. Biometrics, 45, 959-971.

Diggle, P. J.; Heagerty, P. J.; Liang, K. Y.; Zeger, S. L. (1994): *Analysis of Longitudinal Data*. Oxford University Press, New York.

Fitzmaurice, G. M.; Laird, N. M. y Ware J. H. (2004): *Applied Longitudinal Analysis*. J. Wiley & Sons.

Littell, R.C.; Milliken, G.A.; Stroup, W.W.; Wolfinger, R.D.(1996) *SAS® System for Mixed Models*. Cary, NC: SAS Institute Inc.

Nobre, J. S.; Singer, J. M. (2007). *Residual analysis for linear mixed models.* Biometrical Journal, vol. 49, 6, 863-875.

Patetta, M. (2002). *Longitudinal Data Analysis with Discrete and Continuous Responses course notes.*SAS Institute Inc., Cary, NC 27513, USA.

Pinheiro, J. C.; Bates, D. M. (2000): *Mixed-Effects Models in S and S-plus*. Springer-Verlag.

Verbeke, G; Molenberghs, G. (2000): *Linear Mixed Models for Longitudinal Data*. Springer-Verlag, New York.

Weiss, R. E. (2005): *Modeling Longitudinal Data.* Springer.

Zimmerman, D. L. (2000). *Viewing the correlation structure of Longitudinal data through a PRISM.* The American Statistician, vol. 54, 310-318.