



“Modelos mixtos para datos longitudinales: uso del variograma para la selección de un modelo para la covariancia”

Tesina

Universidad Nacional de Rosario
Facultad de Ciencias Económicas y Estadística
Licenciatura en Estadística

Alumna: Cecilia Guillamet Chargue
Directora: Lic. Cecilia Rapelli
Año: 2015

Agradecimientos

A mis padres y mi hermano, por su apoyo incondicional en toda mi vida.

A mis amigos y compañeros de la facultad, por compartir todos estos años.

A Ceci, por su tiempo, dedicación y paciencia.

A los profesores de la carrera, por todo lo enseñado durante este tiempo.

A mis amigos de la vida, por estar siempre presentes.

Índice

1. INTRODUCCIÓN	1
2. DATOS LONGITUDINALES	2
3. MODELO LINEAL MIXTO	3
3.1 Técnicas exploratorias	8
3.1.1 Estructura media.....	8
3.1.2 Estructura de covariancia	9
3.1.3 Variograma	11
3.2 Estimación de los parámetros	22
3.2.1 Máxima verosimilitud	22
3.2.2 Máxima verosimilitud restringida	23
3.3 Inferencia sobre los parámetros del modelo estimado.....	24
3.4 Criterios para la selección de modelos de covariancia	26
3.5 Predicción de los efectos aleatorios.	28
3.6 Análisis de residuos para modelos lineales mixtos	29
3.6.1 Tipos de residuos	29
3.6.2 Evaluación de los supuestos del modelo	31
4. APPLICACIÓN.....	33
4.1 Caso 1: Datos simulados.....	33
4.2 Caso 2: Análisis de datos referidos al estudio de pérdida de peso en mujeres	40
5. CONCLUSIONES	51
6. BIBLIOGRAFÍA.....	53
7. ANEXO	55

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 o individuo 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.

Un paso fundamental en el proceso de construcción del modelo es la selección de la estructura de covariancia. Existen distintas herramientas para identificar modelos para la correlación en datos longitudinales. Sin embargo, la mayoría sólo puede utilizarse cuando las ocasiones de medición son las mismas para todos los individuos. Cuando esto no ocurre la identificación de una estructura de covariancia se vuelve un proceso complicado. Una herramienta alternativa, ampliamente utilizada en la estadística espacial y adaptada para datos longitudinales, que permite describir la asociación entre las mediciones repetidas y se construye fácilmente es el *variograma*.

El variograma es una herramienta descriptiva que sirve para determinar qué componentes de variación se incluirán en el modelo. Para ello se grafican los valores del variograma versus las diferencias entre ocasiones de medición.

Una vez seleccionado un modelo para la covariancia, el variograma puede servir como herramienta de diagnóstico para corroborar si el modelo seleccionado es correcto.

En esta tesina se realiza una presentación del uso del variograma en datos longitudinales.

Para ilustrar su uso se utilizan datos simulados a partir de distintos modelos de correlación y un conjunto de datos longitudinales referidos a un estudio de pérdida de peso en mujeres, extraídos del libro "Modeling Longitudinal Data" (Weiss, 2005).

2. Datos longitudinales

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 de un mismo individuo permiten capturar el cambio intra individuo con una gran precisión dado que cada individuo actúa como su propio control. La evaluación de los cambios intra individuo a través del tiempo sólo puede lograrse con los estudios longitudinales.

Las mediciones repetidas obtenidas de un mismo individuo en las diferentes ocasiones están correlacionadas, y dicha correlación debe ser tenida en cuenta en el análisis para obtener inferencias válidas. Los datos longitudinales también presentan un orden temporal, la primera medición dentro de un agrupamiento está necesariamente antes de la segunda, y así sucesivamente.

En los datos longitudinales existen tres fuentes de variación que tienen impacto sobre la correlación entre las mediciones repetidas de una unidad:

- I. **Variación entre unidades:** Representa la variación existente en la respuesta debida a la heterogeneidad entre las unidades, es decir, la propensión natural de los individuos a responder. Por lo tanto, las mediciones de un mismo individuo serán más similares que las mediciones entre individuos.
- II. **Variación inherente a cada unidad (variación intra unidades):** Representa la variabilidad inherente a la variable respuesta a través del tiempo. Describe cómo las observaciones de una misma unidad varían respecto de su trayectoria media. Esta fuente de variación produce correlaciones seriales entre las mediciones repetidas debido a que mediciones más cercanas serán más similares que las más alejadas.
- III. **Errores de medición (variación intra unidades):** representa la variación aleatoria proveniente del proceso de medición o del muestreo.

Cuando todas las unidades tienen el mismo número de mediciones repetidas obtenidas en un conjunto común de ocasiones de medición se dice que el estudio es *balanceado*. Por el contrario, un estudio *no balanceado* ocurre si la secuencia de ocasiones de medición no es común para todos los individuos.

3. Modelo lineal mixto

Los modelos lineales mixtos son útiles para modelar datos longitudinales debido a su flexibilidad para representar las múltiples fuentes de variación y correlación, y para manejar datos incompletos y no balanceados que son comunes en los datos longitudinales.

En los modelos mixtos la respuesta media se modela como una combinación de características poblacionales, que se asumen comunes a todos los individuos, y un

conjunto de efectos específicos que son únicos para cada individuo. Los primeros se denominan *efectos fijos*, mientras que los últimos *efectos aleatorios*. El término mixto se utiliza en este contexto para denotar que el modelo contiene tanto efectos fijos como aleatorios.

Sea la variable aleatoria 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, $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{in_i})'$.

Si las ocasiones de medición son comunes para todos los individuos es posible omitir el subíndice i de t_{ij} .

La expresión del modelo lineal mixto es,

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i, \quad (3.1)$$

donde,

\mathbf{X}_i es una matriz de diseño de ($n_i \times p$) que caracteriza la parte sistemática del modelo.

$\boldsymbol{\beta}$ es un vector de dimensión ($p \times 1$) de parámetros denominados efectos fijos.

\mathbf{Z}_i es una matriz de diseño de dimensión ($n_i \times k$) que caracteriza la parte aleatoria del modelo.

\mathbf{b}_i es un vector de dimensión ($k \times 1$) de efectos aleatorios que representa las diferencias entre individuos.

\mathbf{e}_i es un vector de dimensión ($n_i \times 1$) que representa las desviaciones introducidas por las fuentes de variación intra individuo.

Los supuestos que se realizan sobre el modelo son:

- $\mathbf{e}_i \stackrel{ind}{\sim} N_{n_i}(0, \mathbf{R}_i)$, donde \mathbf{R}_i es una matriz de covariancia de dimensión ($n_i \times n_i$) que caracteriza la variación intra individuo.

- $\mathbf{b}_i \stackrel{iid}{\sim} N_k(0, \mathbf{D}_i)$, donde \mathbf{D}_i es una matriz de covariancias de dimensión ($k \times k$), que representa la variación entre unidades debido a que las trayectorias individuales difieren.
- \mathbf{b}_i y \mathbf{e}_i son independientes.

En los modelos lineales mixtos es posible distinguir entre las distribuciones marginales y condicionales. Para una unidad particular (una realización particular de los efectos aleatorios) se plantea la distribución condicional de $\mathbf{Y}_i/\mathbf{b}_i$, cuya media es la media específica de la unidad,

$$E(\mathbf{Y}_i/\mathbf{b}_i) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i. \quad (3.2)$$

Y su variancia es,

$$Var(\mathbf{Y}_i/\mathbf{b}_i) = \mathbf{R}_i. \quad (3.3)$$

Mientras que, la media de la distribución marginal, que surge promediando sobre la distribución de los efectos aleatorios, es,

$$E(\mathbf{Y}_i) = \mathbf{X}_i\boldsymbol{\beta}. \quad (3.4)$$

Y la variancia es,

$$Var(\mathbf{Y}_i) = \boldsymbol{\Sigma}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}'_i + \mathbf{R}_i, \quad (3.5)$$

El vector $\boldsymbol{\beta}$ de efectos fijos se supone que es el mismo para todos los sujetos y tiene interpretación promedio poblacional. En contraste, el vector \mathbf{b}_i contiene coeficientes específicos de la unidad. Cuando se combinan ambos se obtiene el perfil medio de cada individuo.

La modelación adecuada de la matriz de covariancia (3.5) no solamente es útil para la interpretación de la variación aleatoria de los datos, sino que es esencial para obtener inferencias válidas de los parámetros de la estructura media, lo cual es de interés primario.

Usualmente la estructura de la $Var(\mathbf{e}_i) = \mathbf{R}_i$ se supone diagonal $\mathbf{R}_i = \sigma^2 \mathbf{I}_{n_i}$, siendo \mathbf{I}_{n_i} la matriz identidad de dimensión n_i . Sin embargo, este modelo para la covariancia puede ser poco realista cuando predominan las causas biológicas en la variabilidad intra individuo, puesto que producen correlaciones seriales entre las mediciones repetidas. Para contemplar esta situación se puede relajar el supuesto de independencia condicional permitiendo estructuras más generales para \mathbf{R}_i .

Existen muchos modelos en la literatura para representar la variabilidad intra individuo. Diggle, Liang y Zeger (1994) proponen una descomposición de \mathbf{e}_i de manera que refleje las dos fuentes de variación intra unidad,

$$\mathbf{e}_i = \mathbf{e}_{(1)i} + \mathbf{e}_{(2)i}, \quad i = 1, \dots, m, \quad (3.6)$$

donde:

$\mathbf{e}_{(1)i}$ es un vector de dimensión $(n_i \times 1)$ de los errores de medición, que refleja la variación debida al proceso de medición,

$\mathbf{e}_{(2)i}$ es un vector de dimensión $(n_i \times 1)$ que caracteriza la variabilidad biológica intra individuo.

El modelo resultante es,

$$\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \mathbf{e}_{(1)i} + \mathbf{e}_{(2)i}, \quad i = 1, \dots, m, \quad (3.7)$$

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.

Los supuestos que se realizan son:

- $\mathbf{e}_{(1)i} \stackrel{iid}{\sim} N_{n_i}(0, \sigma^2 \mathbf{I}_{n_i})$, donde σ^2 es la variancia debida a los errores de medición e \mathbf{I}_{n_i} la matriz identidad de dimensión n_i .
- $\mathbf{e}_{(2)i} \stackrel{iid}{\sim} N_{n_i}(0, \tau^2 \mathbf{H}_i)$, donde τ^2 es la variancia de la componente de correlación serial y \mathbf{H}_i es una matriz de correlaciones de dimensión $(n_i \times n_i)$.
- $\mathbf{b}_i \stackrel{iid}{\sim} N_k(0, \mathbf{D}_i)$.
- $\mathbf{b}_1, \dots, \mathbf{b}_m, \mathbf{e}_{(1)1}, \dots, \mathbf{e}_{(1)m}, \mathbf{e}_{(2)1}, \dots, \mathbf{e}_{(2)m}$ son independientes.

Se asume que el elemento $h_{ijj'}$ de \mathbf{H}_i (nótese que el primer subíndice corresponde al individuo y los otros dos a las posiciones dentro de la matriz) se modela como $h_{ijj'} = g(u_{ijj'})$, siendo $u_{ijj'} = |t_{ij} - t_{ij'}|$, para una función decreciente $g(\cdot)$ 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.

Frecuentemente se utilizan dos funciones para representar la correlación serial, la exponencial y la gaussiana definidas como $g(u_{ijj'}) = \exp(-\phi u_{ijj'})$ y $g(u_{ijj'}) = \exp(-\phi u_{ijj'}^2)$, respectivamente, $\phi > 0$.

Si bien es posible que las tres fuentes de variación contribuyan a la variabilidad de los datos longitudinales, una puede ser más dominante que otra, 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.

3.1 Técnicas exploratorias

3.1.1 Estructura media

Todo análisis estadístico debe estar precedido por un análisis exploratorio simple que permita conocer las características sobresalientes de los datos para luego incluir lo observado en un modelo. Los métodos exploratorios utilizados para identificar la parte media del modelo se presentan a continuación.

Gráficos de perfiles individuales

El gráfico de perfiles individuales es una representación de las respuestas versus el tiempo. Se pueden graficar sólo los puntos o unirlos con segmentos. Estos últimos resultan más informativos dado que permiten identificar las mediciones repetidas sobre el mismo individuo. Este tipo de gráfico no siempre es muy útil para visualizar la forma de la respuesta media, sin embargo, resulta útil para detectar si la variabilidad de los datos cambia a través del tiempo e identificar outliers.

Gráficos de perfiles promedios

El gráfico de perfiles promedios resulta más informativo que el anterior dado que permite identificar la tendencia de la respuesta media en el tiempo. Cuando las ocasiones de medición son las mismas para todos los individuos, éste resulta en un gráfico de la respuesta media versus la ocasión de medición con los puntos sucesivos unidos por líneas rectas. Se puede realizar un perfil promedio general o por grupos (sexo, tratamiento, etc.).

Cuando los datos longitudinales son no balanceados o están irregularmente espaciados, para realizarlo es conveniente utilizar métodos de regresión semi paramétricos para

identificar la tendencia de la respuesta media en el tiempo. Estos métodos muestran la relación entre la respuesta media y el tiempo haciendo supuestos mínimos acerca de la forma de la misma. Uno de estos métodos es la técnica Loess. El estimador Loess se calcula en un intervalo, denominado “ventana”, centrado en un tiempo t_0 . En cada ventana se ajusta una función de regresión usando una técnica de regresión robusta que asigna un mayor peso a las observaciones más cercanas al centro que a las más alejadas. La estimación Loess de la media en el momento t_0 es simplemente el valor predicho en el tiempo t_0 de la función de regresión ajustada. La curva Loess se obtiene moviendo la ventana de ancho fijo desde la primera ocasión hasta la última, y repitiendo el proceso a través de todos los tiempos.

Esta técnica de suavizado requiere la especificación de un parámetro que controla el suavizado de la curva ajustada. La elección del mismo involucra cuestiones de sesgo y precisión, por lo cual es necesario buscar un equilibrio entre ambos. Un suavizado excesivo disminuye la variancia de la estimación de la tendencia media pero puede introducir sesgo. Un suavizado insuficiente no presentará sesgo pero se obtendrá una estimación muy variable de la tendencia de la respuesta media.

3.1.2 Estructura de covariancia

Se han desarrollado técnicas gráficas para datos balanceados que ayudan a detectar la estructura de correlación de un conjunto de unidades con mediciones repetidas. Las técnicas más utilizadas son:

Gráfico de Draftman

Consiste en un conjunto de gráficos de dispersión cuyo objetivo es identificar si el patrón de variación tiene características sistemáticas. Para cada par de ocasiones de medición t_j y $t_{j'}$, se grafican los pares de valores observados estandarizados $(Z_{ij}, Z_{ij'})$ para todos los individuos donde,

$$Z_{ij} = \frac{Y_{ij} - \bar{Y}_{.j}}{\sqrt{S_{.j}^2}} \quad i = 1, \dots, m \quad j = 1, \dots, n,$$

siendo, $\bar{Y}_{.j} = \frac{1}{m} \sum_{i=1}^m Y_{ij}$, $S_{.j}^2 = \frac{1}{m-1} \sum_{i=1}^m (Y_{ij} - \bar{Y}_{.j})^2$.

La utilización de la variable respuesta estandarizada ayuda a eliminar la variabilidad de los datos asociada con diferencias en las medias y variancias en el tiempo, permitiendo visualizar más claramente el patrón de correlación.

Función de Autocorrelación

La función de autocorrelación es una representación gráfica de la correlación entre las mediciones repetidas rezagadas,

$$\rho(u_{ijj'}) = \text{Corr}(Y_{ij}, Y_{ij'}) \quad (3.8)$$

versus el rezago $u_{ijj'} = |t_{ij} - t_{ij'}|$.

Esta función describe cómo cambia la correlación cuando aumenta la separación entre las observaciones y es útil para identificar modelos para la correlación.

Gráfico PRISM (Partial Regression on Intervenors Scatterplot Matrix)

El gráfico PRISM está constituido por gráficos de regresión parcial que se arreglan en la parte superior de una matriz de gráficos de dispersión. Muchas estructuras de correlación tienen asociada una estructura de correlación parcial distintiva. Se utiliza como

complemento del gráfico de Draftman cuando se requiere identificar estructuras autorregresivas o de otro tipo de correlación serial. En la diagonal principal se grafican las variables estandarizadas separadas por un rezago, Z_j y $Z_{j+1} j = 1, \dots, n - 1$. Fuera de ella, se realizan gráficos de regresión parcial de las variables rezagadas por una cierta cantidad de tiempos, ajustadas por las respuestas de los tiempos intermedios. De esta forma la segunda diagonal contiene gráficos de regresión parcial de las respuestas separadas dos ocasiones, ajustada por la respuesta en el tiempo intermedio.

En la tercera diagonal se encuentran los gráficos de regresión parcial de las respuestas apartadas tres tiempos, ajustada por la respuesta dos tiempos intermedios. Y así sucesivamente.

3.1.3 Variograma

Los procedimientos vistos hasta ahora para visualizar la estructura de covariancia requieren que los datos sean balanceados. Una función alternativa que describe la asociación entre mediciones repetidas y se construye fácilmente con datos no balanceados es el *variograma*.

La selección de una estructura de covariancia apropiada es un paso no trivial en el proceso de selección del modelo, especialmente cuando éste incluye efectos aleatorios, ya que la variancia puede estar, a menudo, dominada por los efectos aleatorios.

El variograma 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 $\gamma(u_{ijj'})$ se denomina variograma y se define como,

$$\gamma(u_{ijj'}) = \frac{1}{2} E[(r_{ij} - r_{ij'})^2], \quad (3.9)$$

donde $u_{ijj'} = |t_{ij} - t_{ij'}|$, $i = 1, \dots, m$ y $j \neq j' = 1, \dots, n_i$.

La estimación del variograma, $\hat{\gamma}(u_{ijj'})$, se denomina *variograma muestral* y se calcula a partir de los valores $v_{ijj'} = \frac{1}{2}(r_{ij} - r_{ij'})^2$ 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 $v_{ijj'}$ para cada valor de $u_{ijj'}$ y $\hat{\gamma}(u_{ijj'})$ es el promedio de todos los $v_{ijj'}$, correspondientes al mismo $u_{ijj'}$. Cuando los datos son no balanceados se usan métodos de suavizado (por ejemplo, Loess) para estimar el variograma, $\hat{\gamma}(u_{ijj'})$.

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

$$\frac{1}{2N^*} \sum_{i \neq i'}^m \sum_{j=1}^{n_i} \sum_{j'=1}^{n_{i'}} (r_{ij} - r_{i'j'})^2, \quad (3.10)$$

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

El gráfico de $\hat{\gamma}(u_{ijj'})$ 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 (3.9) y, por lo tanto, un único outlier puede afectar la estimación del variograma en varios rezagos $u_{ijj'}$.

El modelo (3.7),

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_{(1)i} + \mathbf{e}_{(2)i}$$

se puede escribir como,

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i,$$

siendo $\boldsymbol{\varepsilon}_i = \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_{(1)i} + \mathbf{e}_{(2)i}$ con variancia,

$$Var(\boldsymbol{\varepsilon}_i) = \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i' + \tau^2 \mathbf{H}_i + \sigma^2 \mathbf{I}_{n_i}.$$

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

$$Var(\boldsymbol{\varepsilon}_i) = v^2 \mathbf{J}_{n_i} + \tau^2 \mathbf{H}_i + \sigma^2 \mathbf{I}_{n_i}, \quad (3.11)$$

donde \mathbf{J}_{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 ε_{ij} tienen variancia constante $v^2 + \tau^2 + \sigma^2$ y que la correlación entre ε_{ij} y $\varepsilon_{ij'}$ correspondientes al mismo individuo resulta,

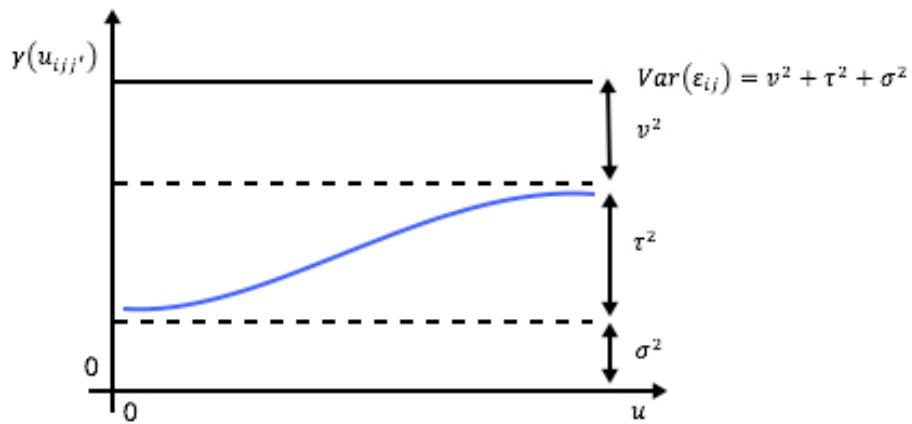
$$Corr(\varepsilon_{ij}, \varepsilon_{ij'}) = \frac{v^2 + \tau^2 g(u_{ijj'})}{v^2 + \tau^2 + \sigma^2}.$$

La expresión del variograma resulta de (3.9),

$$\begin{aligned}
 \gamma(u_{ijj'}) &= \frac{1}{2} E \left[(\varepsilon_{ij} - \varepsilon_{ij'})^2 \right] \\
 &= \frac{1}{2} [Var(\varepsilon_{ij}) + Var(\varepsilon_{ij'}) - 2cov(\varepsilon_{ij}, \varepsilon_{ij'})] \\
 &= \frac{1}{2} [2(v^2 + \tau^2 + \sigma^2) - 2(v^2 + \tau^2 g(u_{ijj'}))] \\
 &= \sigma^2 + \tau^2 [1 - g(u_{ijj'})]
 \end{aligned} \tag{3.12}$$

para $i = 1, \dots, m$ y para $j \neq j' = 1, \dots, n_i$. 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 $\gamma(u_{ijj'}) = \sigma^2 + \tau^2$ cuando $u_{ijj'}$ tiende a infinito. El siguiente gráfico muestra el variograma para el modelo (3.7).

Gráfico 3.1: Variograma correspondiente al modelo (3.7)



A continuación se presentan gráficos del variograma que permiten visualizar diferentes estructuras de covariancia. La comparación del variograma muestral con los casos

presentados permitirá decidir cuáles componentes de variancia deben incluirse en el modelo.

Variograma con variación biológica de cada unidad

Uno de los posibles casos que se puede presentar es cuando la variancia del error se debe solamente a la variabilidad biológica inherente a cada individuo, es decir $\varepsilon_i = \mathbf{e}_{(2)i}$, y por lo tanto, $Var(\varepsilon_i) = \tau^2 \mathbf{H}_i$. De aquí surge que

$$Var(\varepsilon_{ij}) = \tau^2,$$

y las correlaciones entre los distintos ε_{ij} están determinadas por la función de autocorrelación,

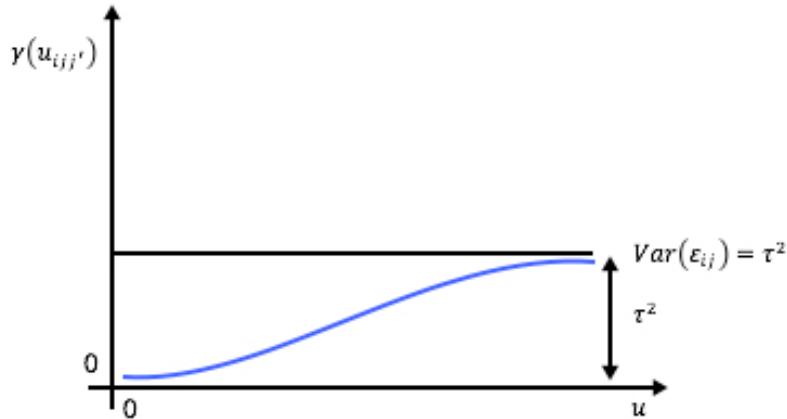
$$Corr(\varepsilon_{ij}, \varepsilon_{ij'}) = g(u_{ijj'}).$$

La expresión del variograma es,

$$\gamma(u_{ijj'}) = \tau^2 [1 - g(u_{ijj'})].$$

Para $u_{ijj'} = 0$ la autocorrelación es uno y $\gamma(0) = 0$. Cuando el rezago aumenta, la autocorrelación se aproxima a cero y $\gamma(u_{ijj'})$ se aproxima a la variancia total, que en este caso resulta ser τ^2 (Gráfico 3.2). De acuerdo a estas características, si el variograma muestral es una curva creciente con $\hat{\gamma}(0) \cong 0$ y valores cercanos a la variancia total cuando $u_{ijj'}$ toma sus valores máximos, sólo será necesario incluir en el modelo la componente de correlación serial, $\mathbf{e}_{(2)i}$.

Gráfico 3.2: Variograma con varianción intra individuo.



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 es,

$$g(u_{ijj'}) = e^{(-\phi u_{ijj'})}, \quad \phi > 0,$$

siendo ϕ un parámetro de correlación y el variograma correspondiente a dicha estructura resulta,

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

Los Gráficos 3.3 y 3.4 presentan el modelo de correlación exponencial y su correspondiente variograma, respectivamente.

Gráfico 3.3: Función de correlación exponencial

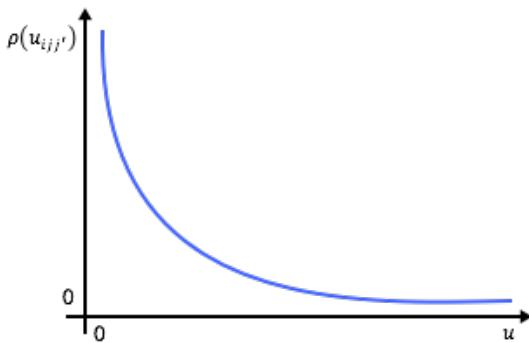
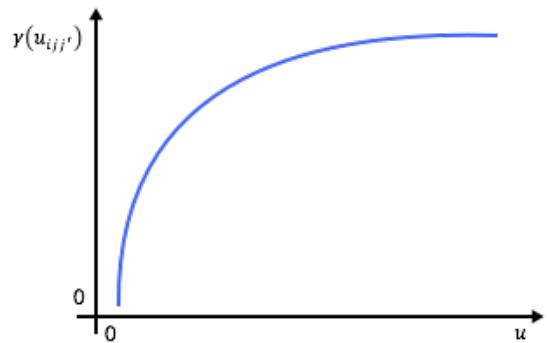


Gráfico 3.4: Variograma correspondiente a la función de correlación exponencial



❖ El modelo de correlación gaussiana es,

$$g(u_{ijj'}) = e^{-\phi(u_{ijj'}^2)}, \quad \phi > 0.$$

El variograma para este tipo de estructura resulta,

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

Ambos se observan en los Gráficos 3.5 y 3.6.

Gráfico 3.5: Función de correlación gaussiana

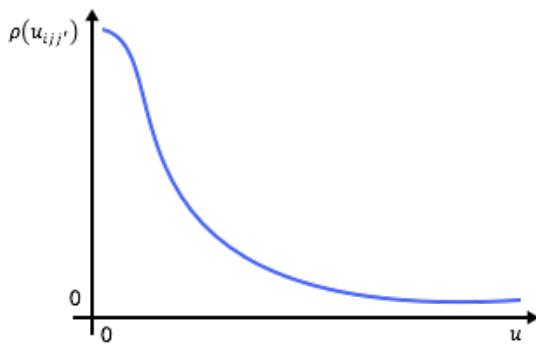
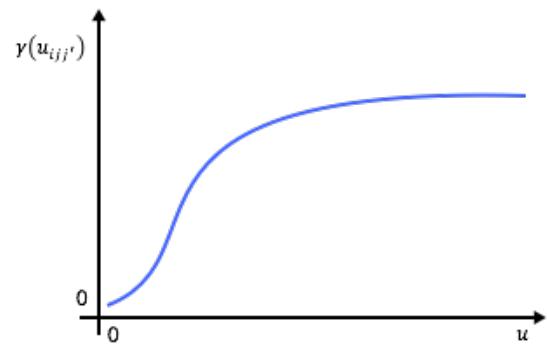


Gráfico 3.6: Variograma correspondiente a la función de correlación gaussiana



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.

Variograma con variación biológica del individuo y errores de medición

Cuando la variancia del error incluye no sólo la variación intra individuo, sino también los errores de medición, $\boldsymbol{\varepsilon}_i = \mathbf{e}_{(1)i} + \mathbf{e}_{(2)i}$, se tiene que $Var(\boldsymbol{\varepsilon}_i) = \sigma^2 \mathbf{I}_{n_i} + \tau^2 \mathbf{H}_i$, y resulta,

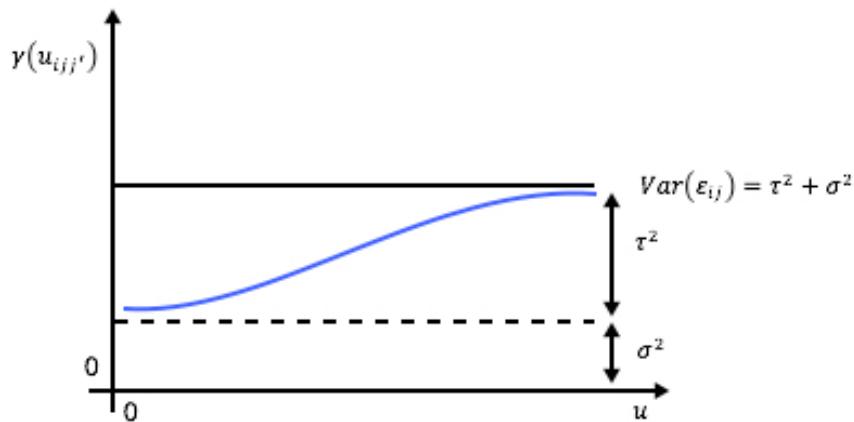
$$Var(\varepsilon_{ij}) = \sigma^2 + \tau^2 \quad \text{y} \quad Cov(\varepsilon_{ij}, \varepsilon_{ij'}) = \tau^2 g(u_{ijj'}).$$

La expresión del correspondiente variograma es,

$$\gamma(u_{ijj'}) = \sigma^2 + \tau^2 [1 - g(u_{ijj'})].$$

A diferencia del caso anterior, $\gamma(0) = \sigma^2 \neq 0$, lo que indica la presencia de errores de medición. Sin embargo, cuando $u_{ijj'}$ tiende a infinito, $\gamma(u_{ijj'})$ tiende a la variancia total, $\tau^2 + \sigma^2$.

Gráfico 3.7: Variograma con correlación serial y errores de medición.



Por lo tanto, un variograma muestral en el que se observa una curva creciente con $\hat{\gamma}(0) \neq 0$ y con asíntota igual a la variancia total sugiere la inclusión en el modelo de dos componentes para caracterizar las dos fuentes de variación intra individuo en el modelo.

Variograma con variancia intra individuo, errores de medición y efectos aleatorios

En algunas situaciones el modelo podría incluir las tres componentes de error, es decir, la variancia del error incluye la variabilidad intra individuo, la variabilidad entre individuos y los errores de medición (Gráfico 3.1). Suponiendo efectos aleatorios sólo para la ordenada al origen, $\boldsymbol{\varepsilon}_i = \mathbf{Z}_i b_{0i} + \mathbf{e}_{(1)i} + \mathbf{e}_{(2)i}$. Por lo tanto se tiene que $Var(\boldsymbol{\varepsilon}_i) = \nu^2 \mathbf{J}_{n_i} + \sigma^2 \mathbf{I}_{n_i} + \tau^2 \mathbf{H}_i$.

En consecuencia,

$$Var(\varepsilon_{ij}) = \nu^2 + \sigma^2 + \tau^2 \quad \text{y} \quad Cov(\varepsilon_{ij}, \varepsilon_{ij'}) = \nu^2 + \tau^2 g(u_{ijj'}),$$

y el correspondiente variograma resulta $\gamma(u_{ijj'}) = \sigma^2 + \tau^2 [1 - g(u_{ijj'})]$ (Gráfico 3.1).

Una característica de los modelos con errores de medición es que $\gamma(0) \neq 0$. Cuando el rezago tiende a infinito, $\gamma(u_{ijj'})$ se aproxima a $(\sigma^2 + \tau^2)$, un valor menor que la variancia del error. La diferencia entre la asíntota alcanzada por la linea 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 $\hat{\gamma}(u_{ijj'})$ 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 difícil 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.

Variograma con efectos aleatorios y errores de medición

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.

En esta situación se supone que el error está compuesto por los efectos aleatorios y los errores de medición, $\boldsymbol{\varepsilon}_i = \mathbf{Z}_i \mathbf{b}_i + \mathbf{e}_{(1)i}$. Por lo tanto, suponiendo sólo ordenada al origen aleatoria, $Var(\boldsymbol{\varepsilon}_i) = v^2 \mathbf{J}_{n_i} + \sigma^2 \mathbf{I}_{n_i}$, resulta,

$$Var(\varepsilon_{ij}) = v^2 + \sigma^2 \quad \text{y} \quad Cov(\varepsilon_{ij}, \varepsilon_{ij'}) = v^2,$$

y la correlación entre dos mediciones cualesquiera de la misma unidad es $g(u_{ijj'}) = \frac{v^2}{v^2 + \sigma^2}$. Este modelo de correlación, que supone correlación constante, se denomina *simetría compuesta*.

El variograma resulta $\gamma(u_{ijj'}) = \sigma^2$, lo cual implica que la curva ajustada es una función constante para todo $u_{ijj'}$, con pendiente igual a cero. Así, el variograma muestral que presente una curva ajustada con pendiente aproximadamente igual a cero, $\hat{\gamma}(0)$ distinto de cero y $\hat{\gamma}(u_{ijj'})$ no tiende a la variancia total sugiere un modelo con errores de medición y efectos aleatorios.

Variograma con efectos aleatorios y correlación serial

En este caso la variancia del error incluye tanto la variabilidad entre los individuos como la variación intra individuo. Suponiendo efectos aleatorios sólo para la ordenada al origen, $\boldsymbol{\varepsilon}_i = \mathbf{Z}_i \mathbf{b}_{0i} + \mathbf{e}_{(2)i}$, se tiene que $Var(\boldsymbol{\varepsilon}_i) = v^2 \mathbf{J}_{n_i} + \tau^2 \mathbf{H}_i$ y resulta,

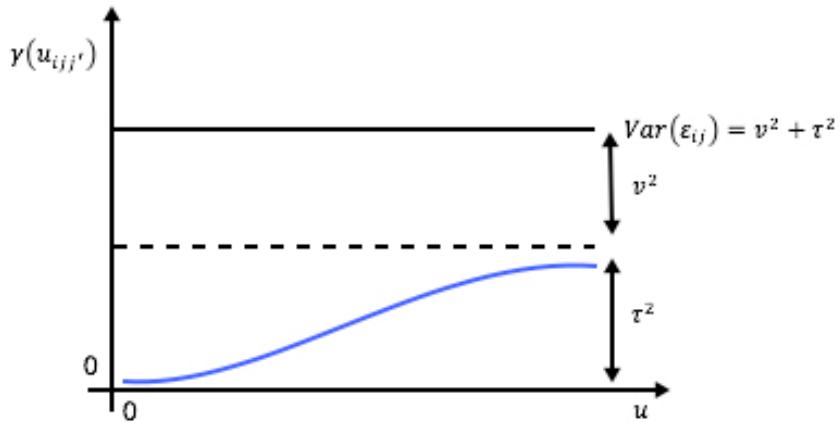
$$Var(\varepsilon_{ij}) = \nu^2 + \tau^2 \quad \text{y} \quad Cov(\varepsilon_{ij}, \varepsilon_{ij'}) = \nu^2 + \tau^2 g(u_{ijj'}).$$

La expresión del correspondiente variograma es,

$$\gamma(u_{ijj'}) = \tau^2 [1 - g(u_{ijj'})].$$

Para $u_{ijj'} = 0$ el variograma $\gamma(u_{ijj'}) = 0$. Cuando $u_{ijj'}$ tiende a infinito, $\gamma(u_{ijj'})$ tiende a τ^2 , un valor menor que la variancia total.

Gráfico 3.8: Variograma con correlación serial y ordenada al origen aleatoria.



Por lo tanto, un variograma muestral en el que se observa una curva creciente con $\hat{\gamma}(0) \cong 0$ y con asíntota menor a la variancia total sugiere la inclusión en el modelo de una única componente estocástica que caracterice ambas fuentes de variación intra individuo, la cual introduce correlación serial, como así también de una ordenada al origen aleatoria.

3.2 Estimación de los parámetros

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*.

3.2.1 Máxima verosimilitud

Bajo el supuesto que los vectores \mathbf{Y}_i se distribuyen normales $\mathbf{Y}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}; \boldsymbol{\Sigma}_i)$, $i = 1, \dots, m$, y son independientes entre sí, la expresión de la función de densidad conjunta es,

$$f(\mathbf{Y}) = \prod_{i=1}^m f_i(\mathbf{Y}_i) = \prod_{i=1}^m (2\pi)^{-n_i/2} |\boldsymbol{\Sigma}_i|^{-1/2} \exp \left\{ -\frac{1}{2} [(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})' \boldsymbol{\Sigma}_i^{-1} (\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})] \right\}, \quad (3.13)$$

donde,

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_m \end{pmatrix}.$$

Los valores de los parámetros desconocidos $\boldsymbol{\beta}$ y $\boldsymbol{\omega}$, donde el vector $\boldsymbol{\omega}$ contiene los parámetros que caracterizan a la matriz $\boldsymbol{\Sigma}_i$, que maximizan la expresión (3.13) son los estimadores máximos verosímiles.

Si se desconoce $\boldsymbol{\omega}$, los estimadores se calculan utilizando algoritmos iterativos. El estimador $\hat{\boldsymbol{\beta}}$ se puede expresar como,

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^m \mathbf{X}_i' \hat{\boldsymbol{\Sigma}}_i^{-1} \mathbf{X}_i \right)^{-1} \sum_{i=1}^m \mathbf{X}_i' \hat{\boldsymbol{\Sigma}}_i^{-1} \mathbf{Y}_i, \quad (3.14)$$

donde, $\hat{\boldsymbol{\Sigma}}_i$ es la matriz de covariancia para \mathbf{Y}_i con los elementos de $\boldsymbol{\omega}$ reemplazados por sus estimadores.

Como las matrices Σ_i son reemplazadas por $\widehat{\Sigma}_i$ en la expresión de $\widehat{\beta}$, no se puede calcular exactamente la matriz de covariancia para $\widehat{\beta}$ y en conclusión no se conoce la distribución muestral exacta de este estimador. Sin embargo, si se supone que el número de unidades es grande, $m \rightarrow \infty$, se puede demostrar que $\widehat{\beta} \sim N_p(\beta; V(\widehat{\beta}))$. La matriz de covariancia para $\widehat{\beta}$ se puede estimar por,

$$\widehat{V}(\widehat{\beta}) = \left(\sum_{i=1}^m \mathbf{X}'_i \widehat{\Sigma}_i^{-1} \mathbf{X}_i \right)^{-1}. \quad (3.15)$$

Es importante tener presente que los errores estándares de $\widehat{\beta}$ y las inferencias realizadas bajo este supuesto son aproximados.

3.2.2 Máxima verosimilitud restringida

Cuando el número de unidades experimentales, m , no es demasiado grande se presenta un problema con las estimaciones de los parámetros que caracterizan a la estructura de covariancia. Aunque los estimadores máximo verosímiles de los parámetros β son insesgados, los estimadores de ω son sesgados.

Para disminuir el sesgo se pueden estimar los parámetros mediante el método de máxima verosimilitud restringida (REML) que maximiza la siguiente log verosimilitud,

$$-\frac{1}{2} \sum_{i=1}^m \log |\Sigma_i| - \frac{1}{2} \sum_{i=1}^m (\mathbf{Y}_i - \mathbf{X}_i \widehat{\beta})' \Sigma_i^{-1} (\mathbf{Y}_i - \mathbf{X}_i \widehat{\beta}) - \frac{1}{2} \log \left| \sum_{i=1}^m \mathbf{X}'_i \Sigma_i^{-1} \mathbf{X}_i \right|. \quad (3.16)$$

Cuando la verosimilitud residual dada en (3.16) es maximizada, se obtiene un estimador de ω que ha sido corregido en el caso que β ya haya sido estimado.

El estimador de β resultante de maximizar la expresión anterior es igual a la de maximizar la función de verosimilitud (3.13), es decir:

$$\hat{\beta} = \left(\sum_{i=1}^m \mathbf{X}'_i \hat{\Sigma}_i^{-1} \mathbf{X}_i \right)^{-1} \sum_{i=1}^m \mathbf{X}'_i \hat{\Sigma}_i^{-1} \mathbf{Y}_i,$$

donde, $\hat{\Sigma}_i$ es la matriz de covariancia para \mathbf{Y}_i con el estimador para ω hallado maximizando (3.16) conjuntamente con β .

Cuando el tamaño de muestra es pequeño se recomienda utilizar máxima verosimilitud restringida para estimar ω dado que el sesgo es menor que el del estimador de máxima verosimilitud.

Para realizar pruebas sobre los parámetros del vector β se aconseja utilizar el método de máxima verosimilitud, si en cambio el interés está centrado en las estimaciones de los parámetros de covariancia ω , se recomienda utilizar los estimadores máximo verosímiles restringidos ya que ellos tienden a ser menos sesgados.

3.3 Inferencia sobre los parámetros del modelo estimado.

Para realizar pruebas de hipótesis sobre los parámetros del modelo se expresan las mismas como funciones lineales de los elementos de β , $\mathbf{L}'\beta$, siendo \mathbf{L}' una matriz de dimensión ($s \times p$) y rango s .

La distribución muestral del estimador de $\mathbf{L}'\beta$ es,

$$\mathbf{L}'\hat{\beta} \sim N_r(\mathbf{L}'\beta, \mathbf{L}'\hat{\Sigma}(\hat{\beta})\mathbf{L}).$$

Para probar las hipótesis existen varios métodos basados en la función de verosimilitud. Asintóticamente, las pruebas son equivalentes, pero para muestras no demasiado grandes los resultados pueden ser bastante diferentes. Se presentan a continuación dos pruebas, el test de Wald y el test de razón de verosimilitud.

Las hipótesis a contrastar en el test de Wald se expresan como,

$$H_0) \mathbf{L}'\boldsymbol{\beta} = \mathbf{0}$$

$$H_1) \mathbf{L}'\boldsymbol{\beta} \neq \mathbf{0},$$

donde, $\mathbf{0}$ es un vector de dimensión ($s \times 1$) con todos sus elementos iguales a cero.

La estadística de prueba correspondiente es,

$$W = (\mathbf{L}\hat{\boldsymbol{\beta}})'(\mathbf{L}'\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}})\mathbf{L})^{-1}(\mathbf{L}'\hat{\boldsymbol{\beta}}),$$

con, $\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}})$ dada por (3.15). W tiene aproximadamente una distribución chi-cuadrado con s grados de libertad. Por lo tanto, se rechaza la hipótesis nula si $W > \chi_{s,1-\alpha}^2$.

Este test no es demasiado confiable cuando el tamaño de muestra es pequeño, dado que es muy liberal.

Otra forma de probar las hipótesis es mediante la prueba de la razón de verosimilitud. Si bien ambos test se basan en la teoría de las muestras grandes, este enfoque es más confiable que el test de Wald cuando el tamaño muestral no es demasiado grande.

La prueba de la razón de verosimilitud para las hipótesis,

$$H_0) \mathbf{L}'\boldsymbol{\beta} = \mathbf{0}$$

$$H_1) \mathbf{L}'\boldsymbol{\beta} \neq \mathbf{0},$$

se obtiene comparando las log verosimilitud de dos modelos, el modelo "completo", en el que no se realizan restricciones, y el modelo "reducido", que surge de incorporar la hipótesis nula ($\mathbf{L}'\boldsymbol{\beta} = \mathbf{0}$) al modelo, siendo el modelo reducido un caso especial de modelo completo.

La verosimilitud del modelo completo es,

$$L_{completo} = \prod_{i=1}^m (2\pi)^{-n_i/2} |\Sigma_i|^{-1/2} \exp \left\{ -\frac{1}{2} [(\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta})' \Sigma_i^{-1} (\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta})] \right\}. \quad (3.17)$$

Para el modelo reducido, $L_{reducido}$ tiene la misma forma que (3.17) con la diferencia que el vector de parámetros $\boldsymbol{\beta}$ está restringido a tener la forma de la hipótesis nula.

La estadística de prueba del test de verosimilitud se define como,

$$G^2 = -2\log\hat{L}_{reducido} + 2\log\hat{L}_{completo}. \quad (3.18)$$

Cuando el tamaño muestral es grande, la estadística G^2 tiene una distribución χ^2 con grados de libertad igual a la diferencia entre el número de parámetros del modelo completo y el número de parámetros del modelo reducido.

3.4 Criterios para la selección de modelos de covariancia

En los modelos para datos longitudinales es importante modelar correctamente tanto la estructura de la media como la estructura de covariancias. Dado que la selección de la estructura media depende de la correcta especificación de la estructura de covariancia, se debe seleccionar un modelo apropiado para ésta última.

Se debe reconocer que la elección del modelo para la covariancia depende del modelo asumido para la media, por lo tanto, el mismo debería ser un modelo maximal, de modo que minimice cualquier mala especificación del modelo para la media.

Con datos longitudinales balanceados y un número pequeño de covariables discretas, la elección del modelo maximal es relativamente sencilla, dado que es posible elegir como modelo maximal uno que incluya el tiempo y todos los efectos principales de las covariables, además de las interacciones de mayor orden. Sin embargo, cuando hay muchas covariables, la elección del modelo maximal es más difícil, por lo tanto sólo se deberán incluir covariables de tratamiento o aquellas que se consideren realmente importantes para el estudio. El modelo maximal sólo incluirá los efectos principales y las interacciones con el tiempo.

Si los modelos de covariancia están anidados, es decir, uno es un caso particular del otro, se puede utilizar el test de cociente de verosimilitud para seleccionar el más adecuado.

El mismo se construye comparando las log-verosimilitud maximizadas REML, $\hat{L}_{completo}$ y $\hat{L}_{reducido}$, para el modelo completo y reducido, respectivamente. El uso de máxima verosimilitud restringida, en vez de máxima verosimilitud, es preferible dado que disminuye el sesgo, en la estimación de la covariancia.

La estadística para el test de cociente de verosimilitud resulta,

$$G^2 = -2\log\hat{L}_{reducido} + 2\log\hat{L}_{completo}. \quad (3.19)$$

La estadística G^2 se compara con una distribución chi-cuadrado con grados de libertad igual a la diferencia entre el número de parámetros de covariancia del modelo completo y del modelo reducido.

Puede ser de interés comparar modelos no anidados, para lo cual se utilizan los criterios de información de Akaike (AIC) y Schwarz (BIC). Ambos criterios penalizan las log verosimilitudes por el número de parámetros de los modelos, aunque las dos penalizaciones son diferentes.

El criterio de información de Akaike (AIC) penaliza la log verosimilitud de acuerdo al número de parámetros del modelo. Se selecciona el modelo que minimiza

$$AIC = -2\log\hat{L} + 2q,$$

donde, \hat{L} es la verosimilitud del modelo maximizada y q es el número de parámetros desconocidos del modelo, los parámetros incluidos en β y ω ($q = \dim(\beta) + \dim(\omega)$).

El criterio de información bayesiano de Schwarz (BIC) es similar al anterior excepto en el valor de la penalización, cuya fórmula es,

$$BIC = -2\log\hat{L} + q\log(m).$$

Al igual que para el AIC, se selecciona el modelo que presenta un menor valor de BIC.

Este criterio tiende a favorecer la selección de modelos más parcios y no es recomendable su uso para elegir estructuras de covariancia.

3.5 Predicción de los efectos aleatorios.

En muchas aplicaciones el interés se centra en caracterizar el comportamiento individual, es decir, estimar o predecir la trayectoria individual a través del tiempo.

Los modelos lineales mixtos son modelos específicos para la unidad en el sentido que los modelos de regresión individuales se caracterizan por tener media

$$E(\mathbf{Y}_i/\mathbf{b}_i) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i.$$

De manera que si se desea caracterizar el comportamiento individual se tendrían que conocer los valores de $\boldsymbol{\beta}$ y \mathbf{b}_i , estimar $\boldsymbol{\beta}$ y predecir el efecto específico de la unidad i , \mathbf{b}_i .

Se puede demostrar que el mejor predictor lineal insesgado (BLUP) de \mathbf{b}_i es su esperanza condicional, dado el vector de respuestas,

$$E(\mathbf{b}_i/\mathbf{Y}_i) = \mathbf{D}\mathbf{Z}'_i\boldsymbol{\Sigma}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})$$

donde $\boldsymbol{\Sigma}_i = Var(\mathbf{Y}_i) = \mathbf{Z}_i\mathbf{D}\mathbf{Z}'_i + \mathbf{R}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}'_i + \tau^2\mathbf{H}_i + \sigma^2\mathbf{I}_{n_i}$.

Si \mathbf{D} y \mathbf{R}_i son desconocidas se reemplazan por sus estimadores REML y se obtiene,

$$\hat{\mathbf{b}}_i = \widehat{\mathbf{D}}\mathbf{Z}'_i\widehat{\boldsymbol{\Sigma}}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}). \quad (3.20)$$

El predictor (3.20) se denomina mejor predictor lineal insesgado empírico (o EBLUP, por su nombre en inglés "empirical best linear unbiased predictor") o estimador empírico de Bayes.

Se puede obtener el perfil de respuesta predicho de la unidad i como,

$$\widehat{\mathbf{Y}}_i = \mathbf{X}_i \widehat{\boldsymbol{\beta}} + \mathbf{Z}_i \widehat{\mathbf{b}}_i. \quad (3.21)$$

La expresión anterior (3.21) puede escribirse como,

$$\widehat{\mathbf{Y}}_i = (\widehat{\mathbf{R}}_i \widehat{\Sigma}_i^{-1}) \mathbf{X}_i \widehat{\boldsymbol{\beta}} + (\mathbf{I}_n - \widehat{\mathbf{R}}_i \widehat{\Sigma}_i^{-1}) \mathbf{Y}_i.$$

Es decir, el perfil de respuesta predicho para la unidad i se puede expresar como un promedio ponderado de la estimación del perfil de respuesta medio para el promedio poblacional y el perfil de respuesta observado de la unidad i .

3.6 Análisis de residuos para modelos lineales mixtos

El análisis de datos longitudinales no está completo sin el examen de los residuos. Éstos se utilizan para evaluar la adecuación del modelo ajustado y para detectar la presencia de valores atípicos o extremos.

Los métodos de diagnóstico para los modelos lineales clásicos están bien establecidos. En contraste, los diagnósticos para modelos mixtos son más difíciles de realizar y de interpretar dado que el modelo es más complejo debido a la presencia de efectos aleatorios y diferentes estructuras de covariancia.

3.6.1 Tipos de residuos

En los modelos mixtos se distinguen los residuos marginales, los condicionales y los EBLUP, que predicen los errores marginales y condicionales, y los efectos aleatorios respectivamente.

Los *residuos marginales* se definen como la diferencia entre el valor observado y la media marginal estimada,

$$\mathbf{r}_{mi} = \mathbf{Y}_i - \widehat{E}(\mathbf{Y}_i) = \mathbf{Y}_i - \mathbf{X}_i \widehat{\boldsymbol{\beta}},$$

y estiman los errores marginales,

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

Los *residuos condicionales* se definen como la diferencia entre el valor observado y el valor predicho o media condicional,

$$\mathbf{r}_{ci} = \mathbf{Y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}} - \mathbf{Z}_i \hat{\mathbf{b}}_i,$$

y estiman los errores condicionales,

$$\mathbf{e}_i = \mathbf{Y}_i - E(\mathbf{Y}_i/\mathbf{b}_i) = \mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta} - \mathbf{Z}_i \mathbf{b}_i.$$

En un modelo sin efectos aleatorios, los dos conjuntos de residuos coinciden.

El BLUP de $\mathbf{Z}_i \mathbf{b}_i$, $\mathbf{Z}_i \hat{\mathbf{b}}_i$, que predice los efectos aleatorios $\mathbf{Z}_i \mathbf{b}_i = E(\mathbf{Y}_i/\mathbf{b}_i) - E(\mathbf{Y}_i)$, refleja la diferencia entre la respuesta predicha para el i -ésimo individuo y el promedio poblacional.

Los residuos marginales y condicionales ordinarios no resultan convenientes para verificar los supuestos del modelo ya que están correlacionados y sus variancias no son necesariamente constantes, en consecuencia, los gráficos se verán considerablemente influenciados. Para evitar este inconveniente se proponen distintas estandarizaciones.

- Residuos estudiantizados

Para eliminar el efecto de las variancias distintas se proponen los residuos estudiantizados. Los mismos se construyen dividiendo el residuo ordinario por una estimación de su desvío estándar,

$$r_i^{estud} = \frac{r_i}{\sqrt{Var(r_i)}}.$$

El cálculo de un residuo estudiantizado también depende de si la observación correspondiente al residuo en cuestión está incluida en la estimación del desvío estándar

o no. Si esta estimación es independiente de la unidad i en estudio el proceso se denomina “estudentización externa” y se logra excluyendo la i -ésima unidad para calcular la estimación del error estándar. En cambio, si la observación contribuye al cálculo del error estándar el residuo se llama “internamente estudentizado”.

- Residuos Cholesky

Otra opción para estandarizar los residuos ordinarios es usar la transformación de Cholesky. Dada una estimación de la matriz de covariancia, $\widehat{\Sigma}_i$, la descomposición de Cholesky crea una matriz triangular inferior, L_i , tal que $\widehat{\Sigma}_i = L_i L_i'$. Se utiliza la matriz L_i o, más específicamente L_i^{-1} , para transformar los residuos,

$$r_i^* = L_i^{-1} r_i, \quad (3.22)$$

de modo que se obtiene un conjunto de residuos transformados aproximadamente no correlacionados y con variancia unitaria.

Estas transformaciones se pueden aplicar tanto a los residuos marginales como a los residuos condicionales.

3.6.2 Evaluación de los supuestos del modelo

Las técnicas gráficas para evaluar el cumplimiento de los supuestos del modelo consideran los distintos tipos de residuos que se definen para el modelo lineal mixto.

Cada tipo de residuo es útil para evaluar algún supuesto. Los residuos condicionales, son útiles para chequear la normalidad de los errores intra individuos, la presencia de heterocedasticidad e identificar outliers; mientras que los residuos marginales se utilizan para evaluar el modelo para la respuesta media y la estructura de covariancia intra

individuo. Los BLUP se utilizan para evaluar los supuestos distribucionales de los efectos aleatorios y para identificar individuos atípicos.

Técnicas gráficas utilizando residuos

- El diagrama de dispersión de los residuos marginales versus el tiempo permite examinar la estructura media del modelo. En un modelo correctamente especificado, se espera que los residuos varíen aleatoriamente alrededor de una media constante igual a cero, sin exhibir patrones sistemáticos.
- La normalidad de los errores intra individuo se puede evaluar a través de un gráfico probabilístico normal de los residuos condicionales. En el mismo, si los residuos se alejan de una recta, el supuesto de normalidad no es sostenible.
- El gráfico de los residuos condicionales versus los valores predichos permite chequear la homocedasticidad del error condicional.
- Con un gráfico probabilístico normal de los EBLUPs, se verifica la normalidad de los efectos aleatorios.
- La presencia de observaciones atípicas se puede visualizar mediante un gráfico de los residuos condicionales versus el número de observaciones.
- A través de un gráfico de los EBLUPs versus el número de unidad se pueden identificar individuos atípicos.

Variograma

El *variograma*, se puede utilizar como herramienta de diagnóstico para verificar si el modelo seleccionado para la covariancia es apropiado. Cuando el variograma se construye usando los residuos de Cholesky marginales (r_{ij}^*) se tiene que,

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

Entonces, en un modelo correctamente especificado para la matriz de covariancia, el gráfico del variograma de los residuos Cholesky 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. Aplicación

Para ilustrar el uso del variograma se utilizan datos simulados a partir de varios modelos de covariancia conocidos y un conjunto de datos longitudinales referidos a un estudio de pérdida de peso en mujeres, extraídos del libro “Modeling Longitudinal Data” (Weiss, 2005). Se utilizan los procedimientos IML y MIXED del programa estadístico SAS® (Littell et al., 1996).

4.1 Caso 1: Datos simulados

Con el objetivo de desarrollar la intuición acerca de la apariencia del variograma muestral se generan conjuntos de datos a partir de un modelo conocido para la media y considerando distintos escenarios para la estructura de covariancias:

- Datos balanceados y no balanceados. Para el primer caso se establecen siete ocasiones de medición equiespaciadas. Para el caso no balanceado, las ocasiones de medición se generan aleatoriamente considerando que las distancias entre las ocasiones de medición tienen distribución exponencial con media uno.

- Distintas componentes de variancia. Un modelo con correlación serial solamente, uno con correlación serial y error de medición, uno con correlación serial y ordenada al origen aleatoria y un modelo que considera las tres fuentes de variación.
- Dos funciones de correlación: exponencial y gaussiana.

Se plantea para la parte media un modelo lineal en el tiempo. La expresión del mismo es,

$$Y_{ij} = 5 + 2,5t_{ij} + \varepsilon_{ij} \quad i = 1, \dots, 100 \quad j = 1, \dots, n_i,$$

donde t_{ij} es el tiempo de medición para la j -ésima medida de la unidad i .

De acuerdo a las diferentes situaciones que se desean evaluar el error aleatorio tiene la forma,

- $\varepsilon_{ij} = e_{(2)ij}$ (Correlación serial solamente)
- $\varepsilon_{ij} = e_{(1)ij} + e_{(2)ij}$ (Correlación serial y errores de medición)
- $\varepsilon_{ij} = b_{0i} + e_{(2)ij}$ (Correlación serial y ordenada al origen aleatoria)
- $\varepsilon_{ij} = e_{(1)ij} + e_{(2)ij} + b_{0i}$ (Correlación serial, errores de medición y ordenada al origen aleatoria)

La generación de los efectos aleatorios y los errores se realizó bajo el supuesto que

$b_{0i} \stackrel{ind}{\sim} N(0, \nu^2)$, $\mathbf{e}_{(1)i} \stackrel{ind}{\sim} N_{n_i}(0, \sigma^2 \mathbf{I}_{n_i})$ y $\mathbf{e}_{(2)i} \stackrel{ind}{\sim} N_{n_i}(0, \tau^2 \mathbf{H}_i)$, donde $\nu^2 = 1$, $\sigma^2 = 0,7$ y $\tau^2 = 1,3$, b_{0i} , $\mathbf{e}_{(1)i}$ y $\mathbf{e}_{(2)i}$ son independientes.

Además, para cada función de correlación considerada para modelar los elementos de \mathbf{H}_i se utilizan tres valores del parámetro ϕ ,

- $\phi = 0,1; 0,5$ y $0,9$ para la función exponencial
- $\phi = 0,1; 0,5$ y $0,9$ para la función gaussiana

Estos valores se corresponden con correlaciones de 0,9; 0,6 y 0,4 respectivamente entre observaciones rezagadas una unidad de tiempo ($u_{ijj'} = 1$).

Para cada conjunto de datos simulados primero se estima el siguiente modelo,

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \varepsilon_{ij} \quad i = 1, \dots, 100 \quad j = 1, \dots, n_i,$$

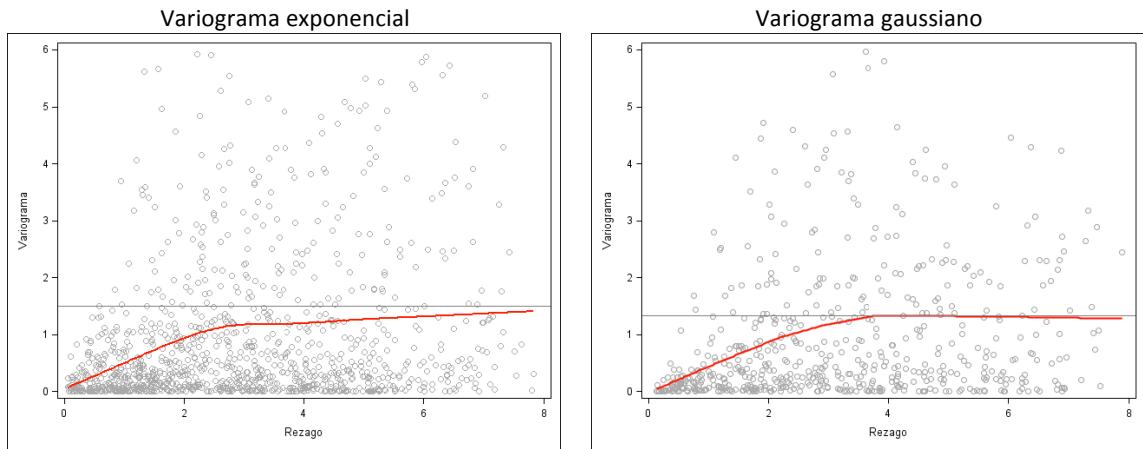
considerando que los errores ε_{ij} están no correlacionados, y tienen variancia constante y $E(\varepsilon_{ij}) = 0$, es decir, ignorando cualquier dependencia entre las medidas repetidas.

Luego se calculan los residuos de esta regresión. Con estos residuos, los cuales no tiene ninguna estructura de covariancia fijada, se construye un variograma muestral.

Mediante las simulaciones se desea comprobar si el variograma muestral refleja el verdadero modelo de covariancia que tienen los datos.

Algunos de los variogramas muestrales se presentan a continuación:

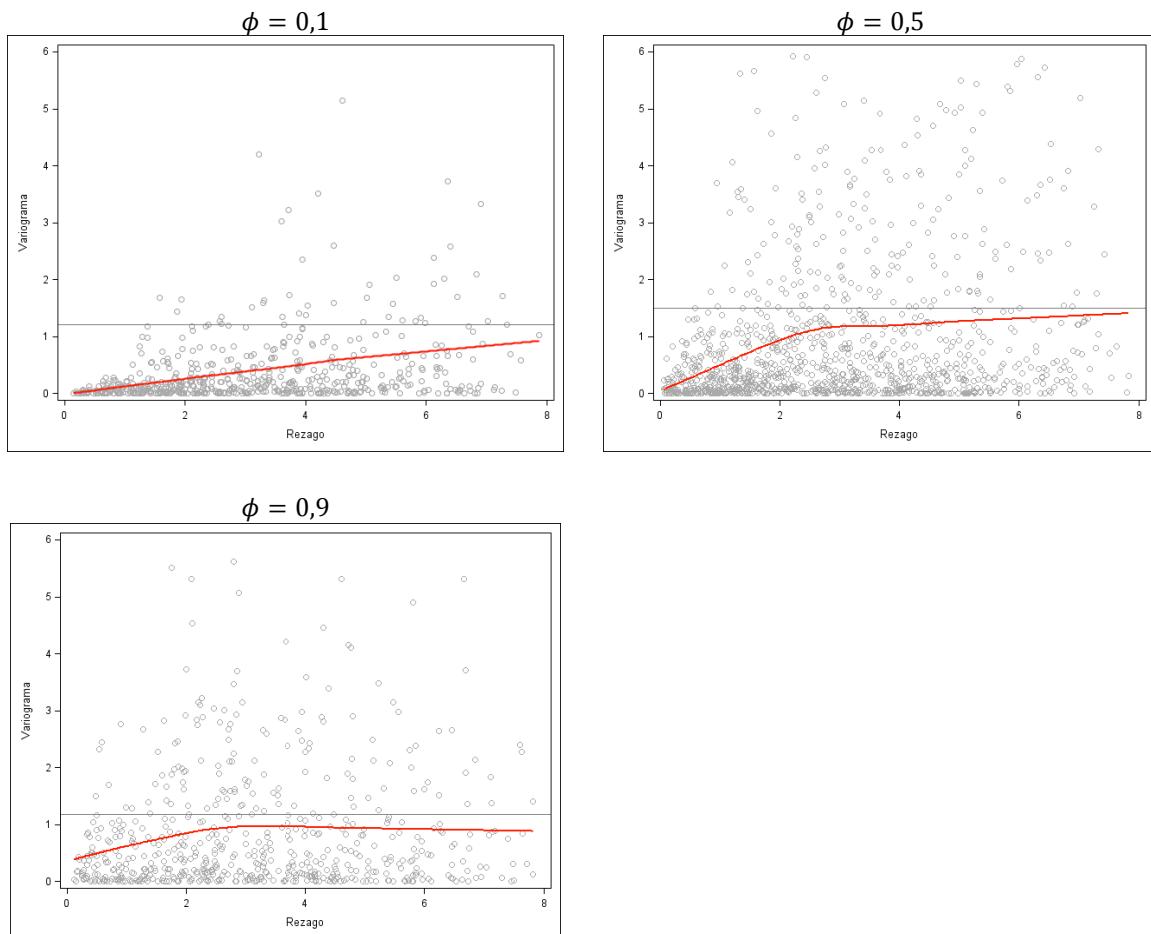
Gráfico 4.1: Variogramas muestrales de los conjuntos de datos generados con correlación serial y con función de correlación exponencial y gaussiana. Datos no balanceados.
($\phi = 0,5$)



Los variogramas presentados en el Gráfico 4.1 permiten identificar las características de las dos estructuras de correlación utilizadas en la generación:

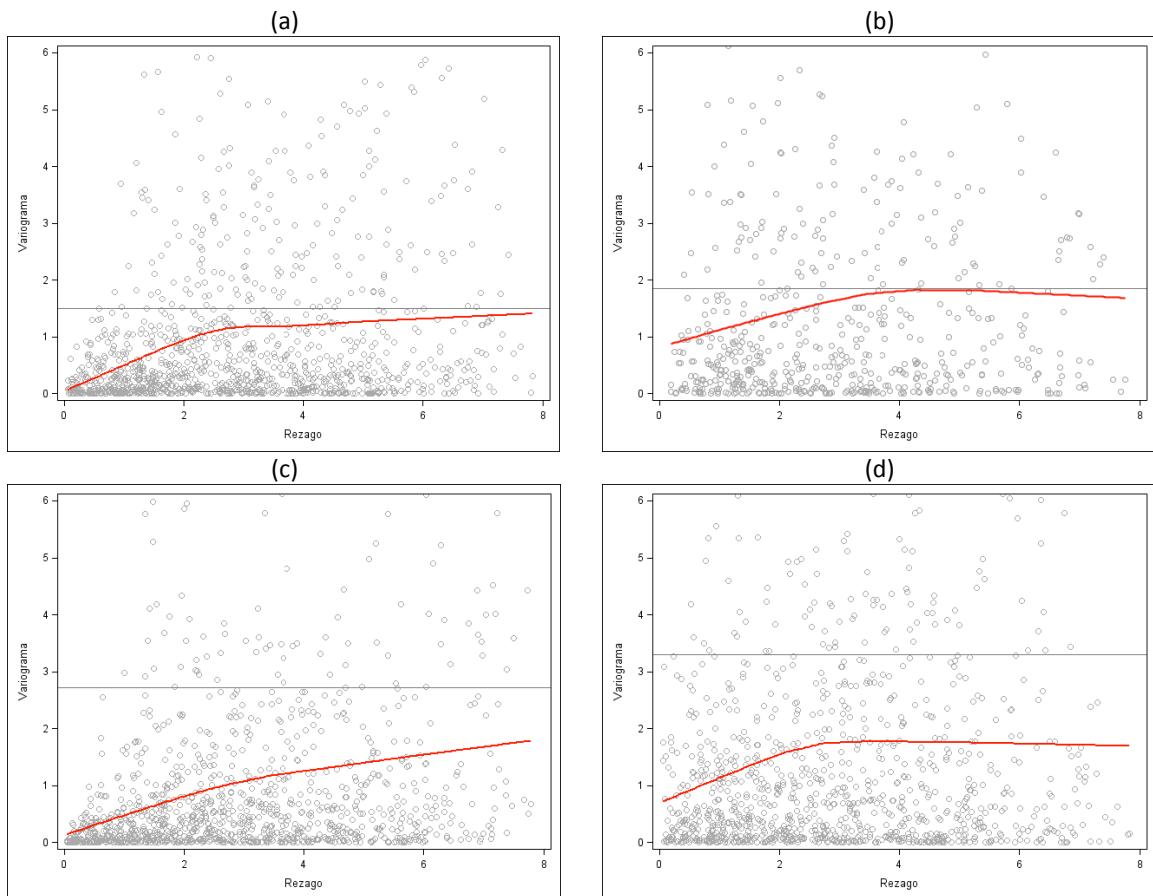
- El variograma muestral correspondiente al modelo exponencial aumenta suavemente hasta llegar a la asíntota.
- Para el modelo gaussiano, el variograma muestra un aumento a medida que los rezagos crecen hasta alcanzar la asíntota. Este aumento es más rápido que el observado para el modelo exponencial.
- No se observan las diferencias entre los dos modelos en la cercanías de $u_{ijj'} = 0$ observadas en los Gráficos 3.4 y 3.6.

Gráfico 4.2: Variogramas muestrales de los conjuntos de datos generados con correlación serial y con función de correlación exponencial con parámetros $\phi = 0,1; 0,5$ y $0,9$. Datos no balanceados.



El Gráfico 4.2 muestra el efecto que tiene el parámetro ϕ sobre el variograma. Cuando ϕ aumenta, el variograma crece más bruscamente alcanzando rápidamente la asymptota. Lo mismo se observa para la función gaussiana (Gráficos A.1 y A.2, Anexo). Esta característica hace que la identificación de la función de correlación sea más difícil a medida que ϕ aumenta.

Gráfico 4.3: Variogramas muestrales de los conjuntos de datos generados con distintas componentes de variancia y con función de correlación exponencial con parámetros $\phi = 0,5$. Datos no balanceados.



(a): Variograma con correlación serial

(b): Variograma con correlación serial y errores de medición

(c): Variograma con correlación serial y ordenada al origen aleatoria

(d): Variograma con correlación serial, errores de medición y ordenada al origen aleatoria

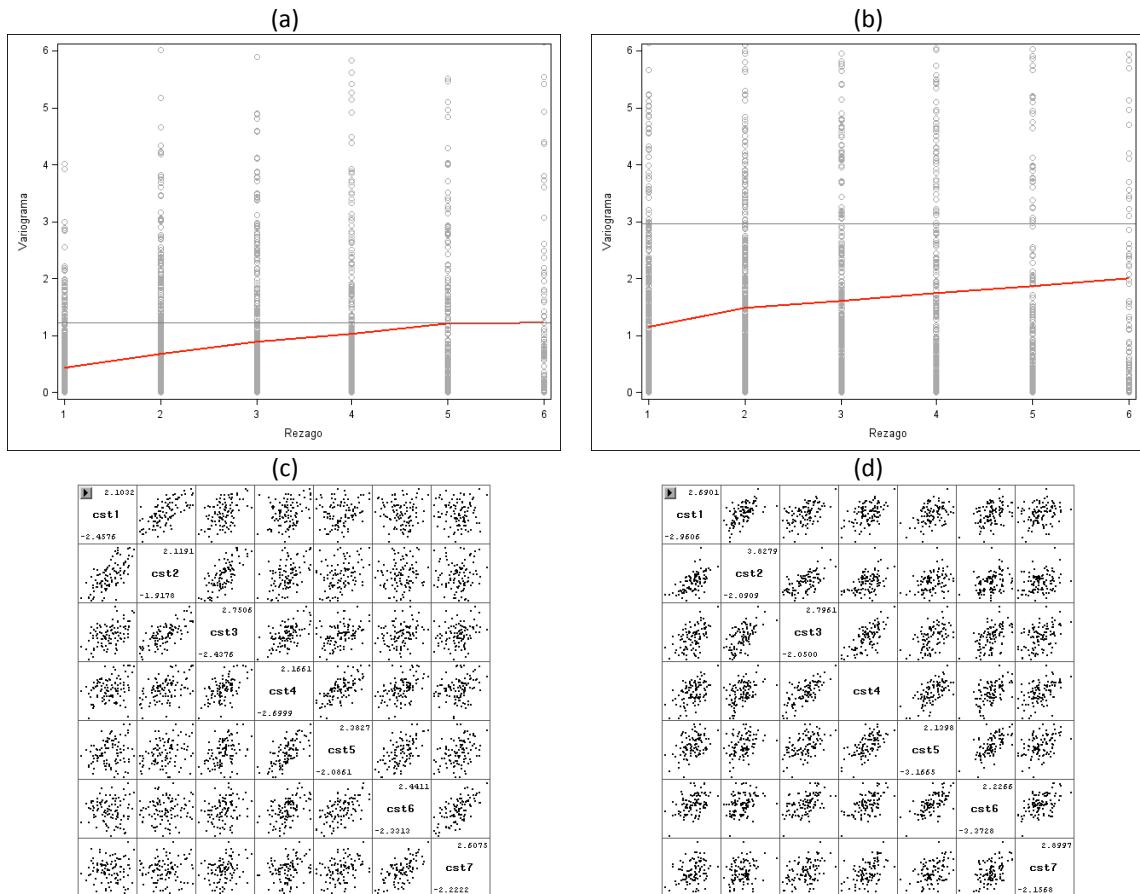
Los variogramas presentados en el Gráfico 4.3 permiten identificar claramente las componentes de variancia incluídas en cada uno de los modelos utilizados para generar los datos. Sin embargo, si bien determinar la presencia de correlación serial es sencillo, identificar la función de correlación se vuelve más difícil a medida que más fuentes de variación intervienen en la variabilidad de los datos. Además si el valor del parámetro de la función de correlación es grande, reconocer el modelo resulta más complicado. Lo mismo puede observarse para la función gaussiana (Gráficos A.3 a A.17, Anexo).

Como enfatizan Verbeke y Molenberghs (2000) una precisa caracterización de la correlación serial es difícil en presencia de efectos aleatorios, sugiriendo incluirla cuando el variograma indique que está presente, aunque no se pueda identificar claramente cuál es la función de correlación, para obtener inferencia correctas sobre los coeficientes de regresión y estimaciones eficientes de los parámetros.

Con el fin de comparar los gráficos de Draftman con los variogramas se presentan dos conjuntos generados de datos balanceados, el primero según un modelo de variancia que sólo contempla correlación serial, y el segundo considerando que las tres fuentes de variación intervienen en la variabilidad de los datos.

A través del Gráfico 4.4, se observa que en ambos variogramas se puede determinar con facilidad las componentes de variancia que intervinieron en la generación de los datos. Sin embargo, la identificación de la función de correlación utilizada es más difícil que en el caso de datos no balanceados.

Gráfico 4.4: Variogramas muestrales y gráficos de Draftman de los conjuntos de datos generados considerando dos modelos para la covariancia y función de correlación exponencial con parámetros $\phi = 0,5$. Datos balanceados.



Al igual que para el caso no balanceado, a medida que aumenta el valor de ϕ y mientras más fuentes de variación intervienen en la generación de los datos, más difícil es identificar la función de correlación, en algunos casos dificultando, incluso, determinar la presencia de correlación serial (Gráficos A.18 a A.61, Anexo).

Por otro lado, los gráficos de Draftman sólo permiten identificar la presencia de correlación en los datos. Éstos no permiten determinar cuáles son las fuentes de variación

consideradas en la generación, razón por las cual el variograma muestral es una herramienta valiosa para explorar la estructura de covariancias de los datos.

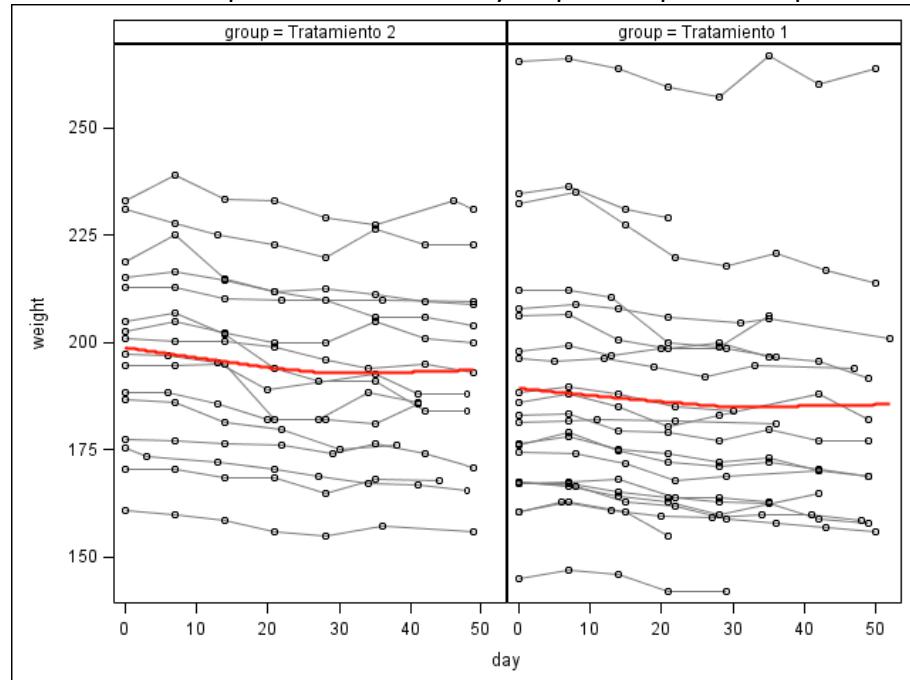
4.2 Caso 2: Análisis de datos referidos al estudio de pérdida de peso en mujeres

Los datos que se utilizan, 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 este un conjunto de datos no balanceados.

En el gráfico de perfiles individuales y promedio por grupo (Gráfico 4.5) 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 contrario, 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.

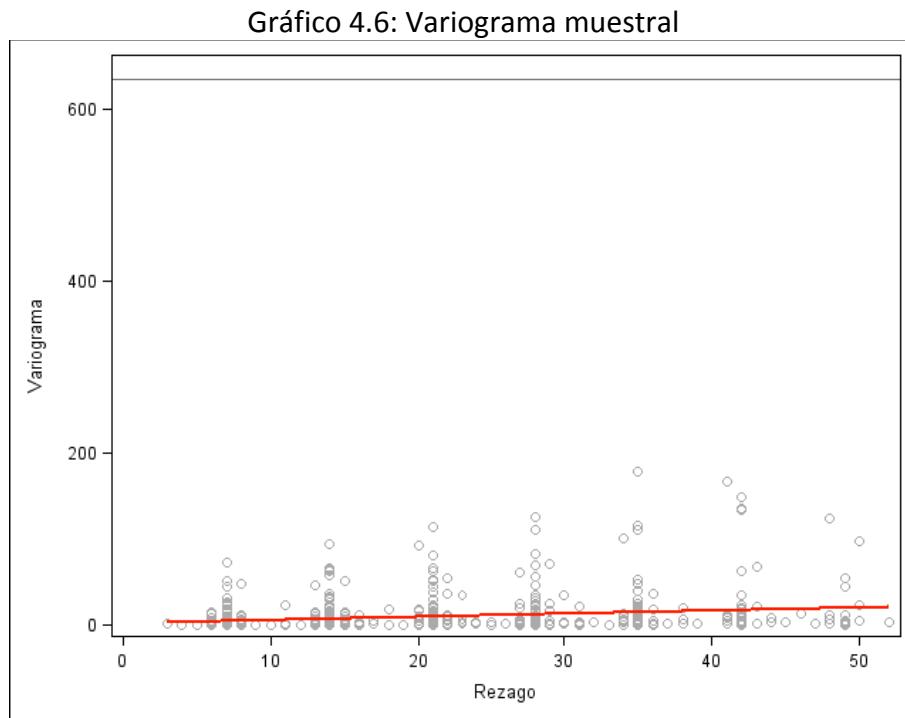
Gráfico 4.5: Gráfico de perfiles individuales y de perfiles promedio por tratamiento



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 4.6 del variograma muestral se observa que resulta apropiado agregar al modelo una ordenada al origen aleatoria 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 4.5). La curva ajustada presenta una pendiente distinta de cero, sugiriendo que es necesario incluir una componente de

correlación serial en el modelo. Se puede apreciar mejor la forma de la misma en el Gráfico A.62 (Anexo), que se focaliza en la curva ajustada.



La identificación de la función de correlación en base al Gráfico 4.6 resulta dificultosa.

Visualmente se asemeja a una línea recta que aumenta lentamente pero en base a los resultados obtenidos con los datos simulados en la sección anterior, el patrón observado se correspondería con una función exponencial, con parámetro cercano a cero.

Por otro lado, en el Gráfico 4.6, 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,

$$Y_{ij} = (\beta_0 + \beta_{01}g_i + b_{0i}) + (\beta_1 + \beta_{11}g_i)t_{ij} + e_{ij} \quad i = 1, \dots, 38 \quad j = 1, \dots, n_i, \quad (4.1)$$

siendo,

- Y_{ij} el peso de la i -ésima mujer en la j -ésima ocasión de medición.
- t_{ij} representa la fecha en que se realiza la j -ésima medición de la participante i .
- g_i una variables indicadora que vale uno para las mujeres asignadas al tratamiento uno y cero para las mujeres asignadas al tratamiento dos.

Los supuestos del modelos son:

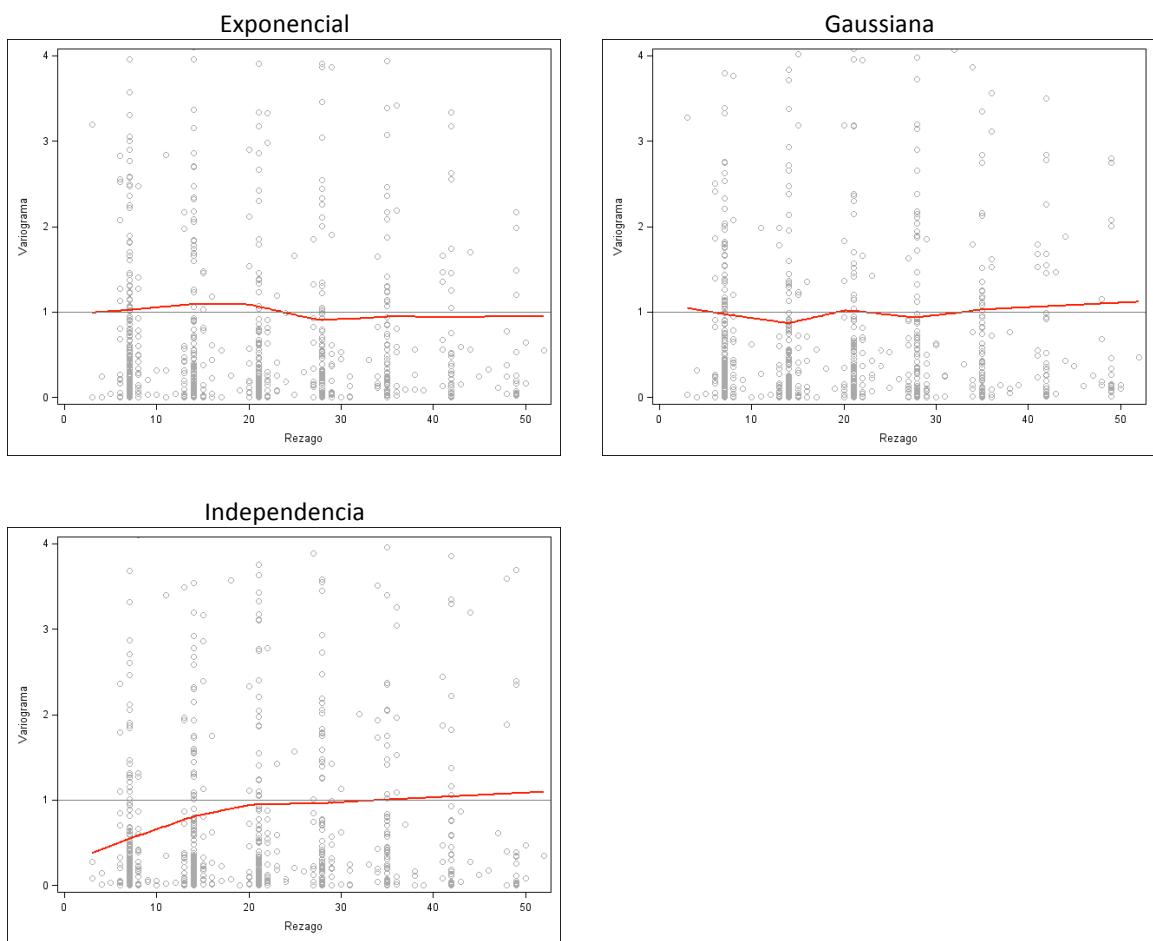
- $b_{0i} \sim N(0, v^2)$
- $\mathbf{e}_i \sim N(0, \tau^2 \mathbf{H}_i)$
- b_{0i} y $\mathbf{e}_i = (e_{i1}, \dots, e_{in_i})'$ son independientes.

Como en el Gráfico 4.6 no resultó claro qué estructura de correlación serial sería apropiada para representar los datos se consideraron dos funciones de correlación serial (exponencial y gaussiana) y una de independencia para representar la variabilidad biológica intra individuo.

Una vez ajustados los modelos se realizan los variogramas usando los residuos de Cholesky marginales. El Gráfico 4.7 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 transformados correspondientes al modelo que supone errores independientes (Gráfico

4.7) 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.

Gráfico 4.7: Variogramas de los residuos transformados para los distintos modelos de correlación



La tabla 4.1 presenta los criterios de información para comparar las tres estructuras de covariancias consideradas. 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.

Tabla 4.1: 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

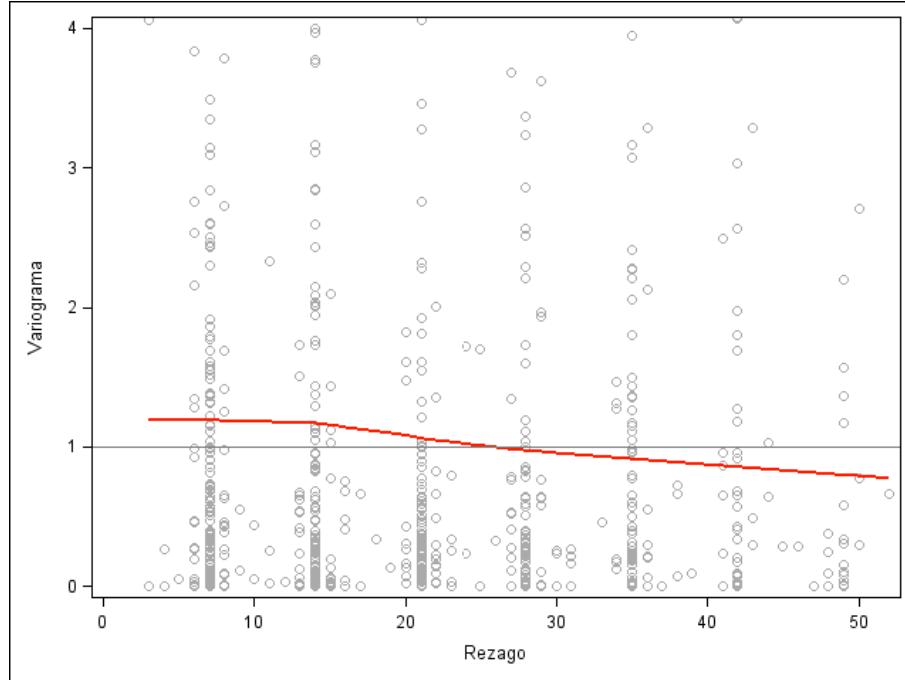
Tanto las diferencias entre los criterios de información como entre los variogramas de los residuos transformados correspondientes a las dos funciones de correlación serial son muy pequeñas, con lo cual la elección de la misma no influye en gran medida sobre las estimaciones e inferencias que se realicen.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

Para determinar si es necesario incluir un efecto aleatorio para la ordenada al origen se plantea un modelo que no incluya el mismo, y se comparan los valores de AIC y BIC para probar la bondad del ajuste de ambos modelos.

El modelo con ordenada al origen aleatoria presenta un valor de AIC de 1471,8 y de BIC 1476,7, resultando mejor que el modelo sin ordenada al origen (AIC=1484,9 y BIC=1488,1). Lo cual está en correspondencia con lo observado en el Gráfico 4.6. Otra forma de verificar esto es a través del variograma de los residuos transformados obtenidos de un modelo sin ordenada al origen aleatoria.

En el Gráfico 4.8 se observa que el variograma no fluctúa aleatoriamente alrededor de la línea centrada en uno, sugiriendo que este modelo no está correctamente especificado.

Gráfico 4.8: Variograma de los residuos transformados para un modelo sin ordenada al origen aleatoria



Una vez seleccionada la estructura de covariancias y debido a que es de interés 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. Para probar esta hipótesis se utiliza el test de razón de verosimilitud con un nivel de significación (α) de 0,05.

La hipótesis a probar es si el peso medio es el mismo para ambos tratamientos a través de todo el periodo en estudio.

$$H_0) \beta_{01} = \beta_{11} = 0$$

$$H_1) Al \ menos \ uno \neq 0$$

El valor de la estadística resulta,

$$G^2 = -2\log\hat{L}_{reducido} + 2\log\hat{L}_{completo} = 1466,8 - 1465,2 = 1,6.$$

Dado que $G^2 < \chi^2_{2;0,05} \cong 6$, no se rechaza la hipótesis nula, y 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. El modelo estimado resulta,

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

La variancia estimada del efecto aleatorio es $\hat{D} = \hat{\tau}^2 = 627,44$ y las estimaciones de los parámetros de la matriz de covariancia intra individuo,

$$Var(\mathbf{e}_i) = \hat{\tau}^2 \hat{H}_i,$$

siendo los elementos de $\hat{H}_i = e^{(-\hat{\phi} u_{ijj'})}$,

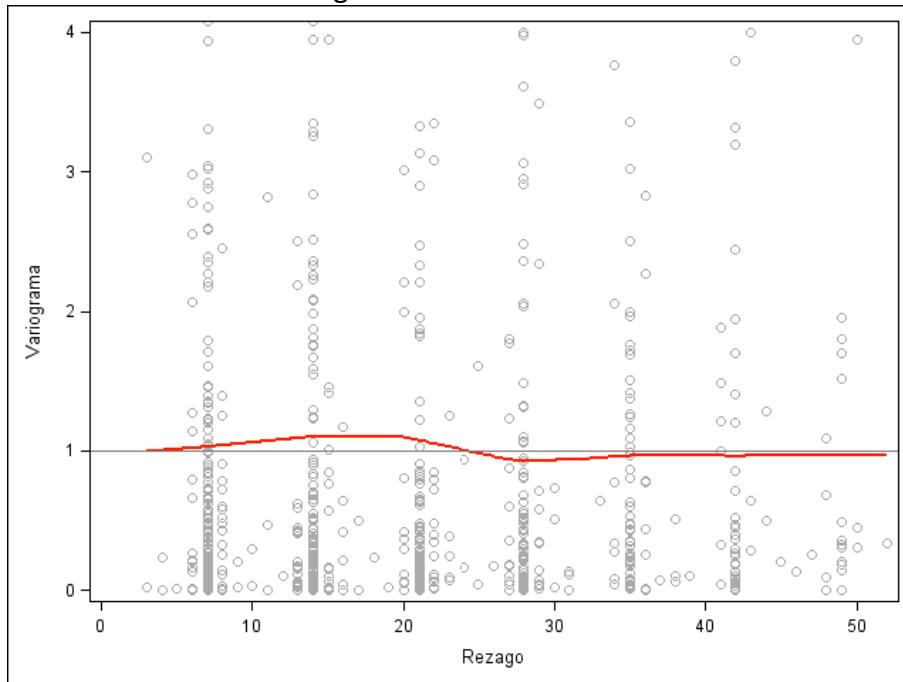
donde $\hat{\tau}^2 = 11,2497$ y $\hat{\phi} = \frac{1}{15,2049} = 0,0658$.

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 4.9 se puede observar que la curva ajustada fluctúa alrededor del uno, lo cual indica que el modelo para la covariancia está correctamente especificado.

Gráfico 4.9: Variograma de los residuos transformados



El Gráfico 4.10 de los residuos marginales 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.

En base al Gráfico 4.11 de los residuos condicionales estudentizados versus el número de observación se puede evaluar la presencia de observaciones atípicas. En el mismo se observan algunas observaciones más alejadas del resto.

Gráfico 4.10: Gráfico de los residuos marginales estudentizados versus los días

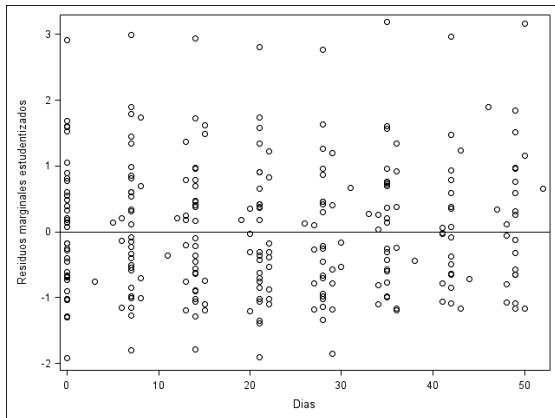
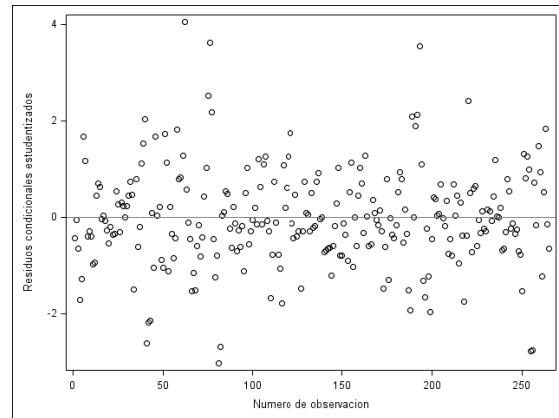
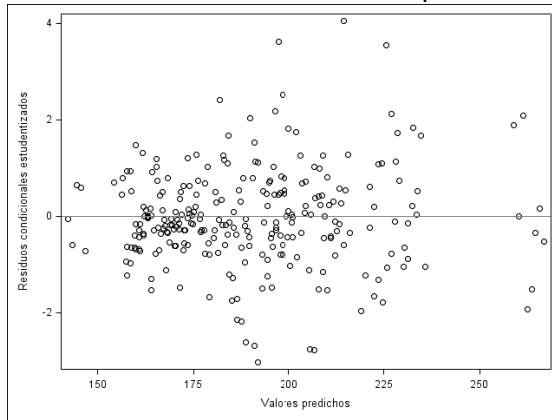


Gráfico 4.11: Gráfico de los residuos condicionales estudentizados vs. el número de observación



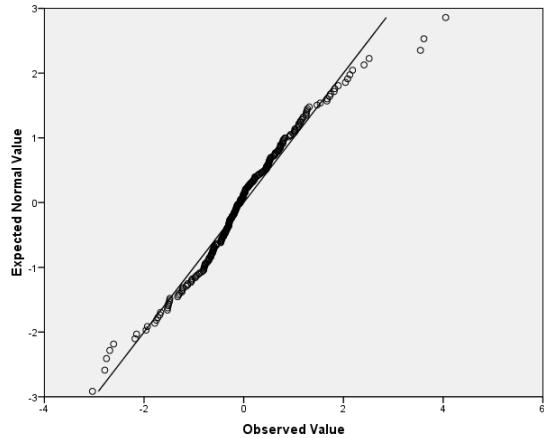
En el Gráfico 4.12 de los residuos condicionales estudentizados versus los valores predichos se observa que los residuos fluctúan aleatoriamente alrededor de cero con un rango de variación constante, indicando que no se evidencia heterocedasticidad de variancias.

Gráfico 4.12: Gráfico de los residuos condicionales estudentizados vs. los valores predichos



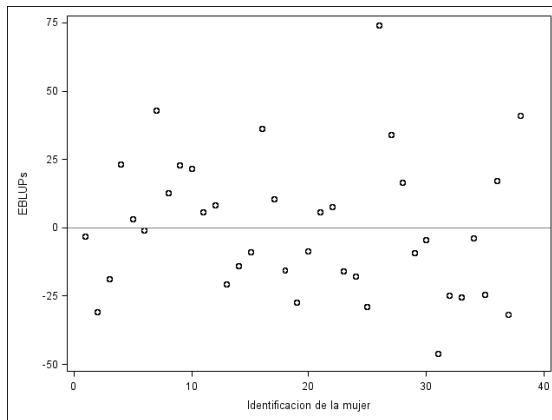
El supuesto de normalidad de los errores intra individuos se evalúa a través de un gráfico probabilístico normal de los residuos condicionales estudentizados. En el Gráfico 4.13 se observa que los residuos presentan una distribución aproximadamente normal con un leve alejamiento de la normalidad de las colas.

Gráfico 4.13: Gráfico probabilísticos normal de los residuos condicionales estudentizados



Se evalúa la existencia de individuos atípicos utilizando el gráfico de los EBLUPs versus el número de unidad. De acuerdo al Gráfico 4.14 se puede observar que la participante número 26 es un posible individuo atípico, dado que se aleja considerablemente del resto.

Gráfico 4.14: Gráfico de los EBLUPs vs. el número de unidad



5. Conclusiones

El uso de los modelos lineales mixtos para modelar los datos longitudinales es atractivo debido a su flexibilidad para representar las múltiples fuentes de variación y correlación y para manejar datos incompletos y no balanceados.

Un paso 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 esta tesina se ilustra el uso del variograma en la etapa exploratoria, para la identificación de las componentes estocásticas que deben incluirse 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 para evaluar el ajuste del modelo de covariancia seleccionado.

A partir de los datos simulados, los variogramas muestrales permitieron:

- La identificación de la presencia de efectos aleatorios y errores de medición, que fue clara en todos los casos considerados.
- La identificación del modelo de correlación serial no resultó sencilla. A medida que el parámetro de las funciones de correlación aumenta, las diferencias que caracterizan a las distintas funciones no se detectan con facilidad.
- La identificación de la función de correlación se vuelve más dificultosa en los casos en que se tienen varias fuentes de variación. Sin embargo, reconocer la existencia de correlación serial en los datos es más importante que la caracterización precisa de la misma. (Diggle et al., 1994)

- Para los diseños no balanceados el comportamiento del variograma resultó más concluyente que para el caso balanceado.

En el análisis del conjunto de datos referido a la pérdida de peso en mujeres, el variograma muestral permitió identificar claramente qué componentes estocásticas debían incluirse 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 permitió evaluar de forma sencilla y concluyente el modelo de covariancia seleccionado.

Se identificó un individuo atípico, la paciente 26, que presenta pesos superiores al resto durante todo el período en estudio. Se debe considerar seriamente realizar el análisis removiendo tal individuo del conjunto de datos para identificar si afecta las conclusiones obtenidas.

6. Bibliografía

- 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: Akadémiai Kiadó, p 267-281.
- Dawson, K. S.; Gennings, C.; Carter, W. H. (1997). *Two graphical techniques useful in detective correlation structure in repeated measures data*. 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.

7. Anexo

Gráfico A.1: Variograma muestral del conjunto de datos generados con correlación serial y función de correlación gaussiana con $\phi = 0,1$

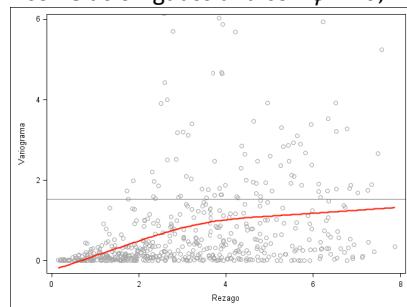


Gráfico A.2: Variograma muestral del conjunto de datos generados con correlación serial y función de correlación gaussiana con $\phi = 0,9$

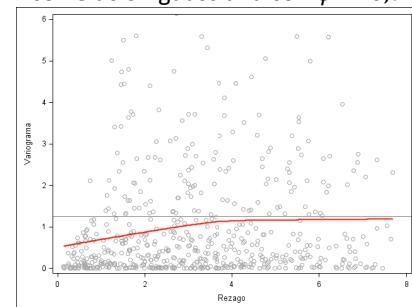


Gráfico A.3: Variograma muestral del conjunto de datos generados con correlación serial y errores de medición y función de correlación exponencial con $\phi = 0,1$

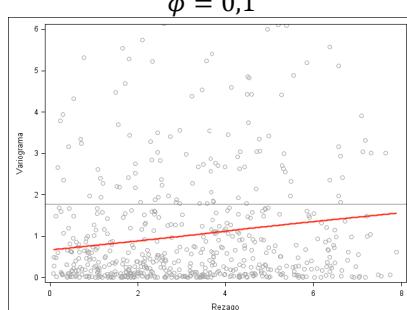


Gráfico A.4: Variograma muestral del conjunto de datos generados con correlación serial y errores de medición y función de correlación exponencial con $\phi = 0,9$

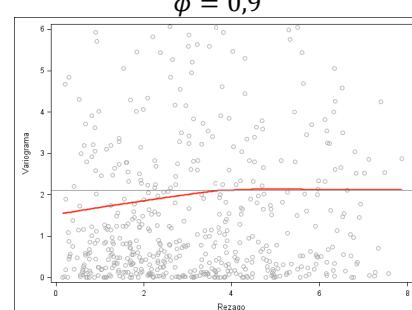


Gráfico A.5: Variograma muestral del conjunto de datos generados con correlación serial y errores de medición y función de correlación gaussiana con $\phi = 0,1$

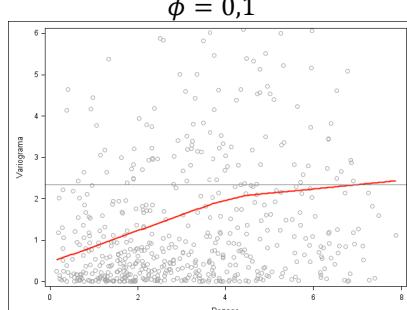


Gráfico A.6: Variograma muestral del conjunto de datos generados con correlación serial y errores de medición y función de correlación gaussiana con $\phi = 0,5$

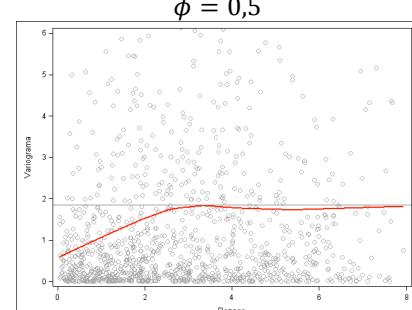


Gráfico A.7: Variograma muestral del conjunto de datos generados con correlación serial y errores de medición y función de correlación gaussiana con $\phi = 0,9$

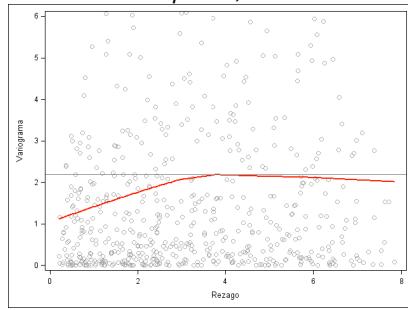


Gráfico A.8: Variograma muestral del conjunto de datos generados con correlación serial, errores de medición y ordenada al origen aleatoria y función de correlación exponencial con $\phi = 0,1$

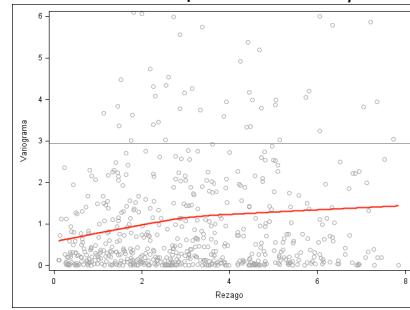


Gráfico A.9: Variograma muestral del conjunto de datos generados con correlación serial, errores de medición y ordenada al origen aleatoria y función de correlación exponencial con $\phi = 0,9$

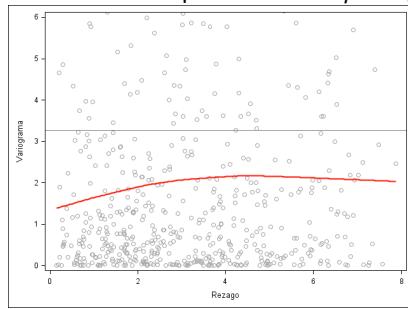


Gráfico A.10: Variograma muestral del conjunto de datos generados con correlación serial, errores de medición y ordenada al origen aleatoria y función de correlación gaussiana con $\phi = 0,1$

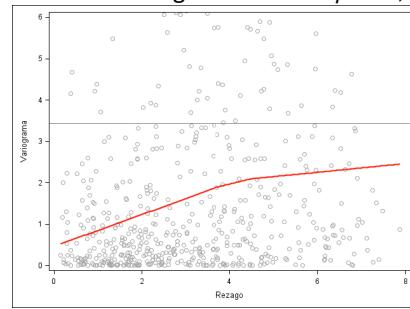


Gráfico A.11: Variograma muestral del conjunto de datos generados con correlación serial, errores de medición y ordenada al origen aleatoria y función de correlación gaussiana con $\phi = 0,5$

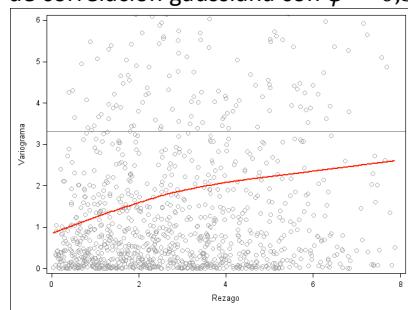


Gráfico A.12: Variograma muestral del conjunto de datos generados con correlación serial, errores de medición y ordenada al origen aleatoria y función de correlación gaussiana con $\phi = 0,9$

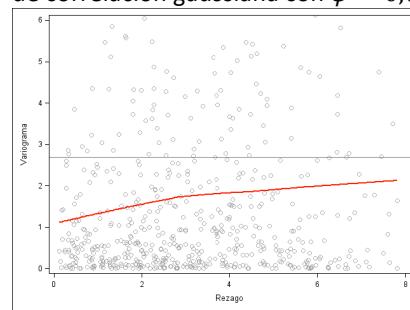


Gráfico A.13: Variograma muestral del conjunto de datos generados con correlación serial y ordenada al origen aleatoria y función de correlación exponencial con $\phi = 0,1$

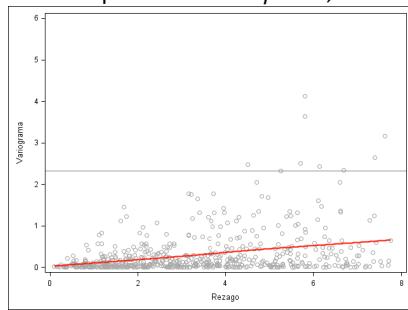


Gráfico A.14: Variograma muestral del conjunto de datos generados con correlación serial y ordenada al origen aleatoria y función de correlación exponencial con $\phi = 0,9$

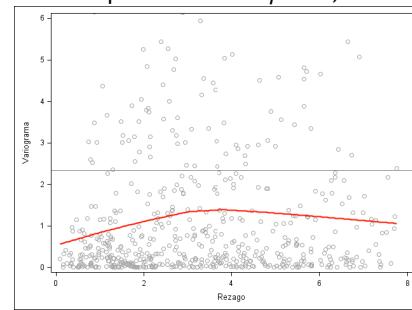


Gráfico A.15: Variograma muestral del conjunto de datos generados con correlación serial y ordenada al origen aleatoria y función de correlación gaussiana con $\phi = 0,1$

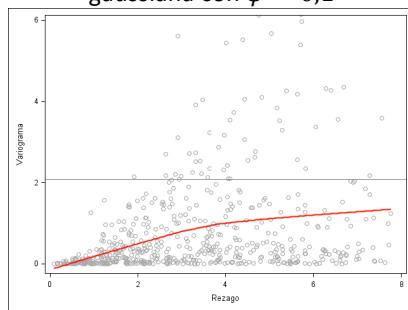


Gráfico A.16: Variograma muestral del conjunto de datos generados con correlación serial y ordenada al origen aleatoria y función de correlación gaussiana con $\phi = 0,5$

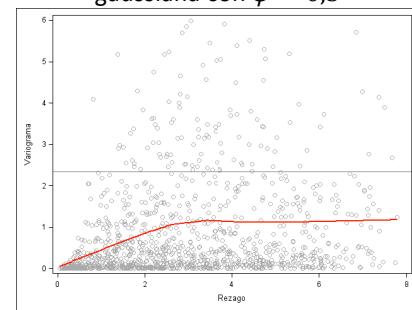


Gráfico A.17: Variograma muestral del conjunto de datos generados con correlación serial y ordenada al origen aleatoria y función de correlación gaussiana con $\phi = 0,9$

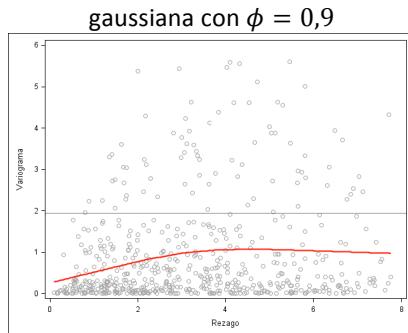


Gráfico A.18: Variograma muestral del conjunto de datos generados con correlación serial y función de correlación exponencial con $\phi = 0,1$. Datos balanceados

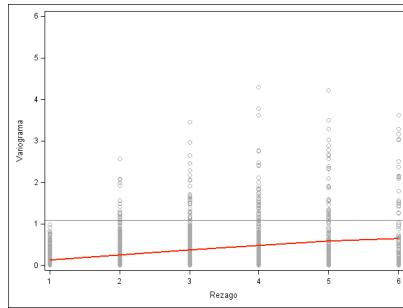


Gráfico A.20: Variograma muestral del conjunto de datos generados con correlación serial y función de correlación exponencial con $\phi = 0,9$. Datos balanceados

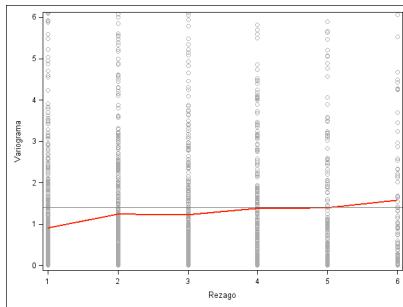


Gráfico A.22: Variograma muestral del conjunto de datos generados con correlación serial y función de correlación gaussiana con $\phi = 0,1$. Datos balanceados

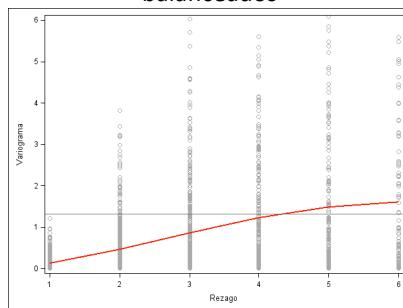


Gráfico A.19: Gráfico de Draftman del conjunto de datos generados con correlación serial y función de correlación exponencial con $\phi = 0,1$

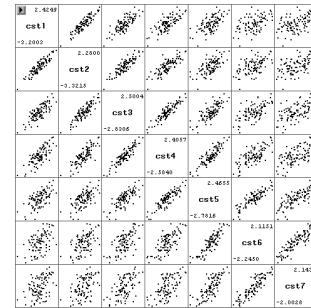


Gráfico A.21: Gráfico de Draftman del conjunto de datos generados con correlación serial y función de correlación exponencial con $\phi = 0,9$

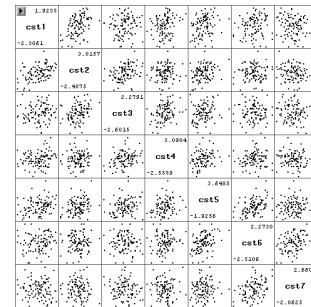


Gráfico A.23: Gráfico de Draftman del conjunto de datos generados con correlación serial y función de correlación gaussiana con $\phi = 0,1$

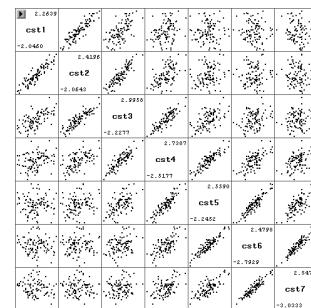


Gráfico A.24: Variograma muestral del conjunto de datos generados con correlación serial y función de correlación gaussiana con $\phi = 0,5$. Datos balanceados

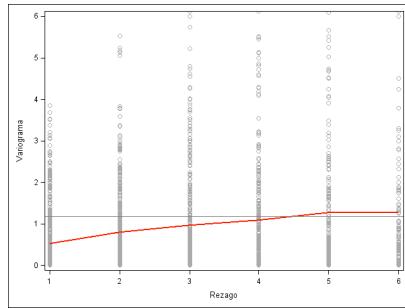


Gráfico A.25: Gráfico de Draftman del conjunto de datos generados con correlación serial y función de correlación gaussiana con $\phi = 0,5$

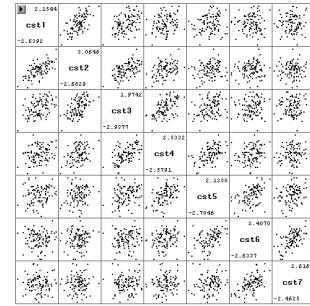


Gráfico A.26: Variograma muestral del conjunto de datos generados con correlación serial y función de correlación gaussiana con $\phi = 0,9$. Datos balanceados

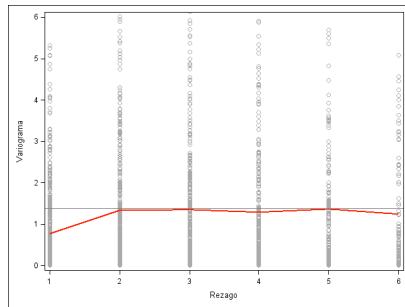


Gráfico A.27: Gráfico de Draftman del conjunto de datos generados con correlación serial y función de correlación gaussiana con $\phi = 0,9$

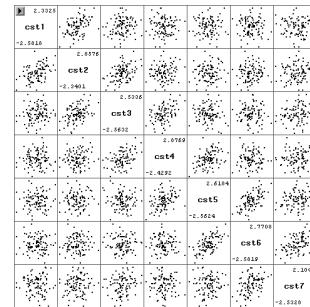


Gráfico A.28: Variograma muestral del conjunto de datos generados con correlación serial y errores de medición y función de correlación exponencial con $\phi = 0,1$. Datos balanceados

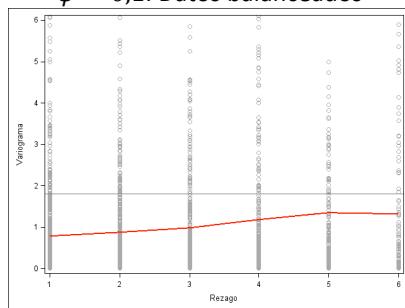


Gráfico A.29: Gráfico de Draftman del conjunto de datos generados con correlación serial y errores de medición y función de correlación exponencial con $\phi = 0,1$

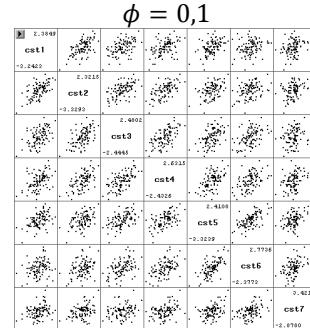


Gráfico A.30: Variograma muestral del conjunto de datos generados con correlación serial y errores de medición y función de correlación exponencial con $\phi = 0,5$. Datos balanceados

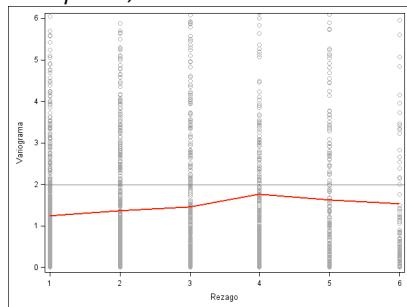


Gráfico A.31: Gráfico de Draftman del conjunto de datos generados con correlación serial y errores de medición y función de correlación exponencial con $\phi = 0,5$

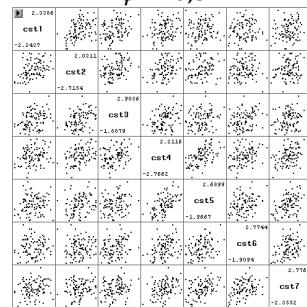


Gráfico A.32: Variograma muestral del conjunto de datos generados con correlación serial y errores de medición y función de correlación exponencial con $\phi = 0,9$. Datos balanceados

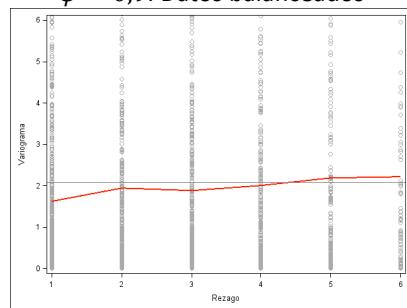


Gráfico A.33: Gráfico de Draftman del conjunto de datos generados con correlación serial y errores de medición y función de correlación exponencial con $\phi = 0,9$

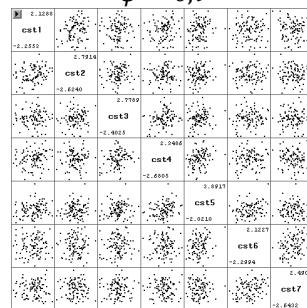


Gráfico A.34: Variograma muestral del conjunto de datos generados con correlación serial y errores de medición y función de correlación gaussiana con $\phi = 0,1$. Datos balanceados

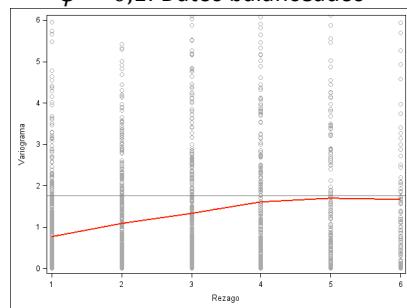


Gráfico A.35: Gráfico de Draftman del conjunto de datos generados con correlación serial y errores de medición y función de correlación gaussiana con $\phi = 0,1$

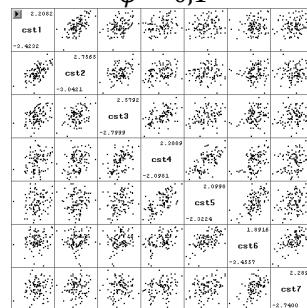


Gráfico A.36: Variograma muestral del conjunto de datos generados con correlación serial y errores de medición y función de correlación gaussiana con $\phi = 0,5$. Datos balanceados

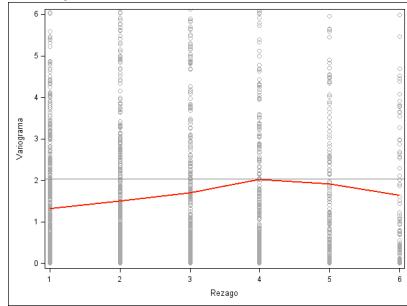


Gráfico A.38: Variograma muestral del conjunto de datos generados con correlación serial y errores de medición y función de correlación gaussiana con $\phi = 0,9$. Datos balanceados

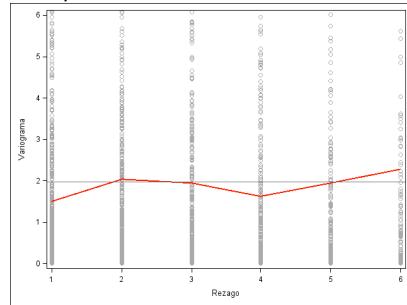


Gráfico A.40: Variograma muestral del conjunto de datos generados con correlación serial, errores de medición y ordenada al origen aleatoria y función de correlación exponencial con $\phi = 0,1$. Datos balanceados

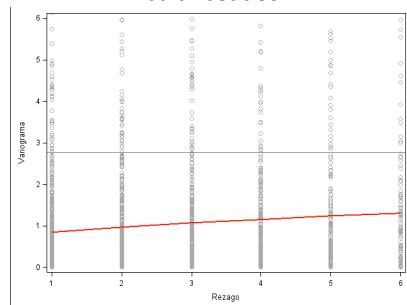


Gráfico A.37: Gráfico de Draftman del conjunto de datos generados con correlación serial y errores de medición y función de correlación gaussiana con $\phi = 0,5$

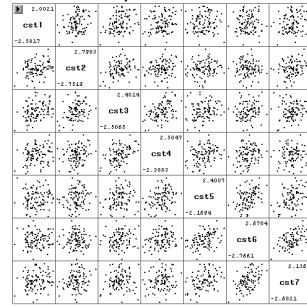


Gráfico A.39: Gráfico de Draftman del conjunto de datos generados con correlación serial y errores de medición y función de correlación gaussiana con $\phi = 0,9$

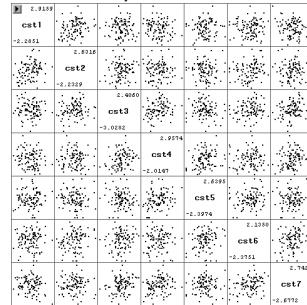


Gráfico A.41: Gráfico de Draftman del conjunto de datos generados con correlación serial, errores de medición y ordenada al origen aleatoria y función de correlación exponencial con $\phi = 0,1$

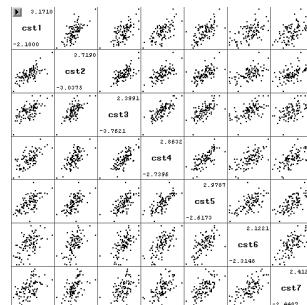


Gráfico A.42: Variograma muestral del conjunto de datos generados con correlación serial, errores de medición y ordenada al origen aleatoria y función de correlación exponencial con $\phi = 0,9$. Datos balanceados

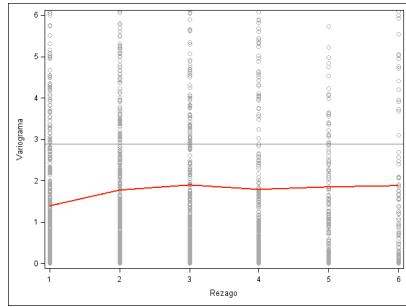


Gráfico A.43: Gráfico de Draftman del conjunto de datos generados con correlación serial, errores de medición y ordenada al origen aleatoria y función de correlación exponencial con $\phi = 0,9$

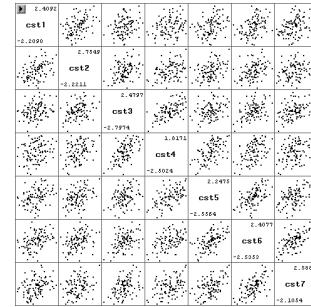


Gráfico A.44: Variograma muestral del conjunto de datos generados con correlación serial, errores de medición y ordenada al origen aleatoria y función de correlación gaussiana con $\phi = 0,1$. Datos balanceados

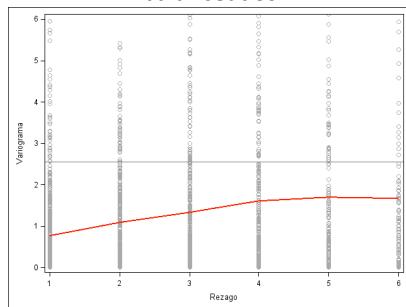


Gráfico A.45: Gráfico de Draftman del conjunto de datos generados con correlación serial, errores de medición y ordenada al origen aleatoria y función de correlación gaussiana con $\phi = 0,1$

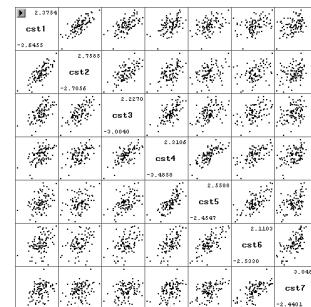


Gráfico A.46: Variograma muestral del conjunto de datos generados con correlación serial, errores de medición y ordenada al origen aleatoria y función de correlación gaussiana con $\phi = 0,5$. Datos balanceados

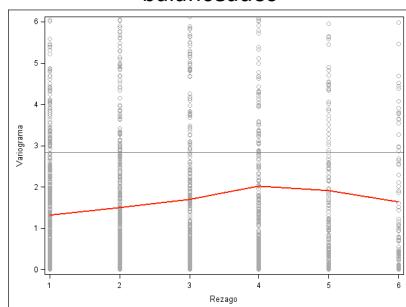


Gráfico A.47: Gráfico de Draftman del conjunto de datos generados con correlación serial, errores de medición y ordenada al origen aleatoria y función de correlación gaussiana con $\phi = 0,5$

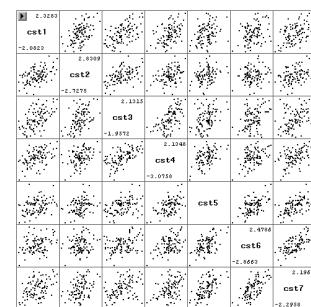


Gráfico A.48: Variograma muestral del conjunto de datos generados con correlación serial, errores de medición y ordenada al origen aleatoria y función de correlación gaussiana con $\phi = 0,9$. Datos balanceados

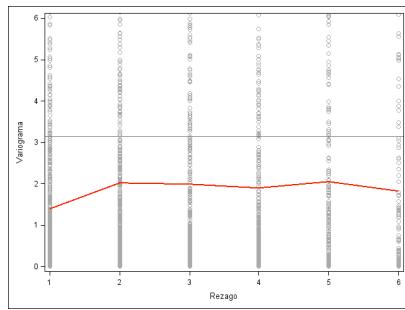


Gráfico A.49: Gráfico de Draftman del conjunto de datos generados con correlación serial, errores de medición y ordenada al origen aleatoria y función de correlación gaussiana con $\phi = 0,9$

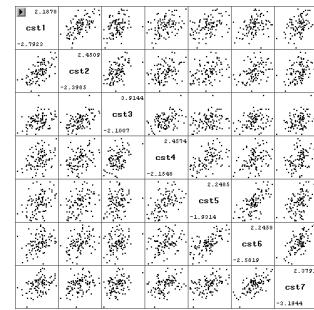


Gráfico A.50: Variograma muestral del conjunto de datos generados con correlación serial y ordenada al origen aleatoria y función de correlación exponencial con $\phi = 0,1$. Datos balanceados

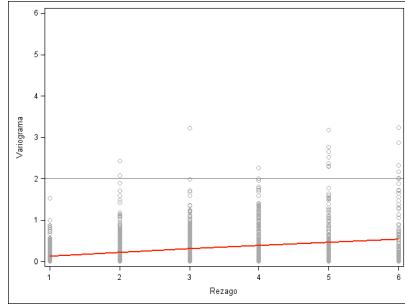


Gráfico A.51: Gráfico de Draftman del conjunto de datos generados con correlación serial y ordenada al origen aleatoria y función de correlación exponencial con $\phi = 0,1$

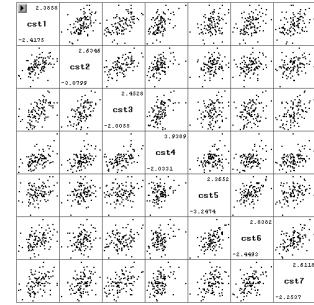


Gráfico A.52: Variograma muestral del conjunto de datos generados con correlación serial y ordenada al origen aleatoria y función de correlación exponencial con $\phi = 0,5$. Datos balanceados

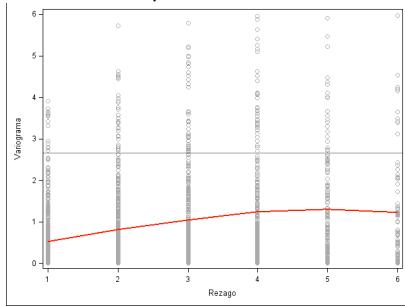


Gráfico A.53: Gráfico de Draftman del conjunto de datos generados con correlación serial y ordenada al origen aleatoria y función de correlación exponencial con $\phi = 0,5$

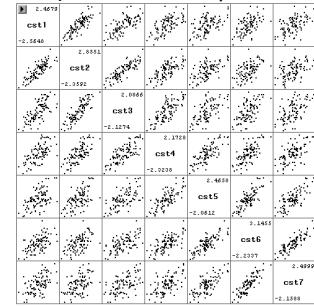


Gráfico A.54: Variograma muestral del conjunto de datos generados con correlación serial y ordenada al origen aleatoria y función de correlación exponencial con $\phi = 0,9$. Datos balanceados

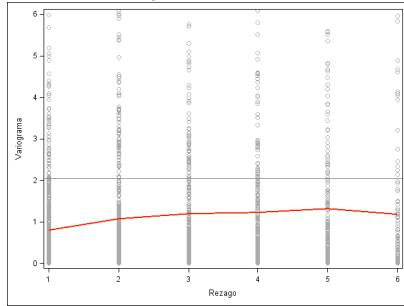


Gráfico A.56: Variograma muestral del conjunto de datos generados con correlación serial y ordenada al origen aleatoria y función de correlación gaussiana con $\phi = 0,1$. Datos balanceados

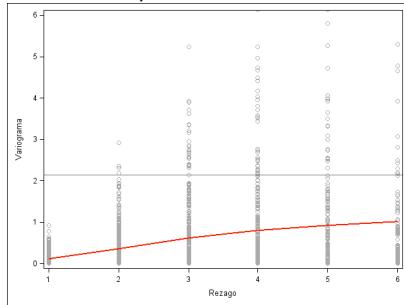


Gráfico A.58: Variograma muestral del conjunto de datos generados con correlación serial y ordenada al origen aleatoria y función de correlación gaussiana con $\phi = 0,5$. Datos balanceados

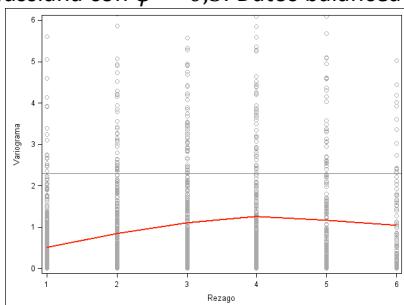


Gráfico A.55: Gráfico de Draftman del conjunto de datos generados con correlación serial y ordenada al origen aleatoria y función de correlación exponencial con $\phi = 0,9$

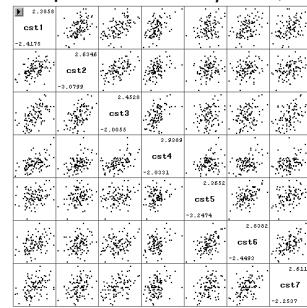


Gráfico A.57: Gráfico de Draftman del conjunto de datos generados con correlación serial y ordenada al origen aleatoria y función de correlación gaussiana con $\phi = 0,1$

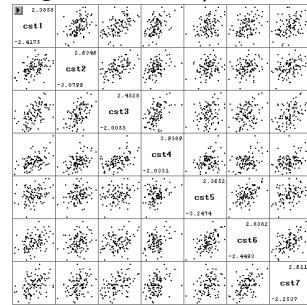


Gráfico A.59: Gráfico de Draftman del conjunto de datos generados con correlación serial y ordenada al origen aleatoria y función de correlación gaussiana con $\phi = 0,5$

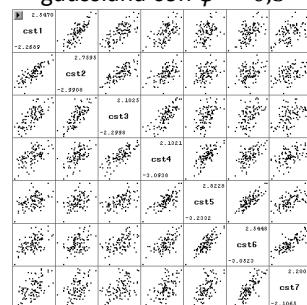


Gráfico A.60: Variograma muestral del conjunto de datos generados con correlación serial y ordenada al origen aleatoria y función de correlación gaussiana con $\phi = 0,9$. Datos balanceados

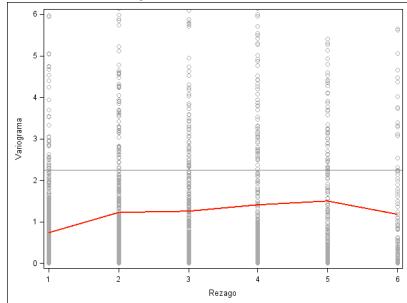
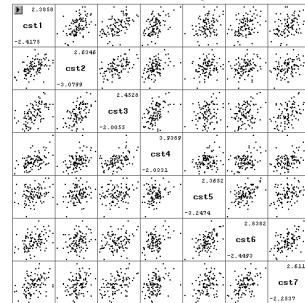


Gráfico A.61: Gráfico de Draftman del conjunto de datos generados con correlación serial y ordenada al origen aleatoria y función de correlación gaussiana con $\phi = 0,9$



A.62: Variograma muestral acercado

