# Introducción.

# 2. Metodología.

2.1 Los datos longitudinales.

Los datos longitudinales están conformados por mediciones repetidas de una misma variable realizadas a la misma unidad. Estas mediciones surgen de observar unidades en diferentes ocasiones, es decir en diferentes momentos o condiciones experimentales.

Dado que las mediciones repetidas son obtenidas de la misma unidad, los datos longitudinales están agrupados. Las observaciones dentro de un mismo agrupamiento generalmente están correlacionadas positivamente.

El objetivo principal de estos estudios es estudiar los cambios en el tiempo y los factores que influencian el cambio.

Las ocasiones en las que se registran las mediciones repetidas no necesariamente serán iguales para todos los individuos, por lo tanto se pueden obtener tanto estudios balanceados (todos los individuos tienen el mismo número de mediciones durante un conjunto de ocasiones comunes) como desbalanceados (la secuencia de tiempos de observaciones no es igual para todos los individuos). Otra característica de estos datos es que en ocasiones se pueden obtener valores perdidos, obteniendo datos incompletos aunque se cuente con un estudio balanceado.

Con el fin de simplificar la notación, se asumirá que los tiempos de medición son los mismos para todas las unidades y que no hay datos faltantes.

Se obtiene una muestra de unidades cada una con n mediciones repetidas de la variable en estudio, observadas en los tiempos , siendo entonces el número total de observaciones . Se le llama a la realización sobre la unidad i en la ocasión j, .

Asociadas a cada unidad se observan las covariables , de las cuales existen dos tipos: variables en el tiempo (estocásticas), invariables en el tiempo (estacionarias).

Existen estudios empíricos que llevan a pensar que existen tres fuentes potenciales de variabilidad que influyen sobre la correlación entre medidas repetidas:

* Heterogeneidad entre las unidades: Refleja la propensión natural de las unidades a responder. Los individuos tienen diferentes reacciones frente a los mismos estímulos.
* Variación biológica intra-unidad: Se piensa que la secuencia de medidas repetidas de una unidad tiene un comportamiento determinado, que produce que las mediciones más cercanas sean más parecidas.
* Error de medición: Surge debido a los errores de medida, se puede disminuir usando instrumentos de medición más precisos.

Estas tres fuentes de variación se clasifican “variabilidad entre”, es decir, la variación entre las distintas unidades (heterogeneidad entre unidades) y “variabilidad intra”, es decir, la variación entre las mediciones de la misma unidad (variación biológica intra-unidad y error de medición)

Dado que, como se mencionó anteriormente, las mediciones están correlacionadas entre sí, si se utilizaran las técnicas habituales basadas en la independencia entre mediciones, los errores estándares nominales van a ser incorrectos, lo cual nos llevaría a inferencias incorrectas sobre los parámetros del modelo. En base a esto, surgen técnicas que consideran esa correlación modelando los datos considerando la modelación de dos estructuras: la parte media y la estructura de covariancia.

2.1 Análisis exploratorio

Antes de ajustar algún modelo, lo primero siempre es realizar un análisis exploratorio para estudiar cómo se comportan los datos. A continuación se presentan técnicas gráficas para cada estructura.

Evaluación de la parte media.

* *Perfil individual:* Consiste en un gráfico de dispersión en el cual se representan las respuestas vs las ocasiones. Cada respuesta tiene un punto y se unen con un segmento los puntos de la misma unidad. Sirven para detectar si hay mucha variabilidad entre y dentro de las unidades y si hay valores atípicos.
* *Perfiles promedio por grupo:* En general son más informativos. Para cada tiempo calculamos un promedio para cada grupo y luego se unen los puntos. Permiten ver la tendencia de las variables a través de las ocasiones. Se superponen en un mismo gráfico los perfiles promedio para cada grupo.

Estructura de covariancia.

* *Matriz de diagrama de dispersión:*

Para cada par de ocasiones se grafican los valores esperados de la respuesta y todos estos gráficos se acomodan dentro de una matriz. En general se utiliza cuando las ocasiones son las mismas para todas las unidades.

* *Gráfico de Draftman:*

Es similar al gráfico anterior pero utilizando las variables estandarizadas:

Donde es la variable respuesta del individuo i en la ocasión j, y

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.

* *Gráfico PRISM (Partial Regression on Intervenors Scatterplot Matrix):*

Utilizando la variable estandarizada se crea una matriz de gráficos de dispersión. En la primer diagonal se encuentran gráficos de dispersión entre la variable respuesta en los tiempos y , . Luego, en la k-ésima diagonal, se obtienen gráficos de regresión parcial de las respuestas en los tiempos y , ajustadas por las respuestas en los tiempos intermedios. Estos gráficos permiten ver con mayor claridad ciertos tipos de estructuras seriales que se dan entre las medidas repetidas.

* *Correlograma:*

Representa las características que existen entre las respuestas de los individuos de cada grupo en tiempos que están separados una cantidad de periodos. Permite analizar cómo evoluciona la correlación a medida que aumenta el número de rezagos.

* *Semivariograma:*

Cuando los datos están desbalanceados el semivariograma permite distinguir las 3 fuentes de variabilidad. Después de haber estimado un modelo, el mismo permite confirmar si la estructura de correlación es adecuada. El semivariograma se define como una función:

y son los residuos estandarizados obtenidos después de ajustar un modelo de regresión considerando las observaciones independientes.

Se va a obtener un gráfico donde la variabilidad total va a estar dividida en 3 partes. Si la curva no empieza en cero significa que hay error de medición, si tiene pendiente quiere decir que hay un error debido a una causa biológica (correlación serial) y si la misma no llega a la variancia total significa que se debe explicar la variabilidad “entre”.

2.2 Modelo lineal general para datos longitudinales.

Si se piensa que existe una tendencia en el tiempo de las respuestas, y esta tendencia se puede expresar como una función, se puede escribir o representar las medidas repetidas de una unidad en un vector . Entonces, un modelo lineal para representar la evolución en el tiempo va a ser:

Donde:

: Respuesta obtenida de la i-ésima unidad en la ocasión .

: Matriz de diseño de la i-ésima unidad, de dimensión ()

: Vector de parámetros de dimensión ()

: Vector de errores aleatorios de la i-ésima unidad, de dimensión ()

representa todas las fuentes de variabilidad de los datos longitudinales.

: Vector de parámetros desconocidos de covariancia, de dimensión (

2.3 Modelación de la estructura de covariancia.

Al tenerse tantos parámetros de variancia (n) y covariancia (n(n-1)/2) para estimar, se proponen modelos específicos para la estructura de correlación. Se trata de elegir una estructura que no tenga tantos parámetros. Sin embargo, se debe tener cuidado de no seleccionar estructuras demasiado parcas con las que se pierda información.

La matriz de covariancia de cada unidad va a ser función de . El número de parámetros de este vector depende de la estructura de la matriz.

A continuación se mencionan algunas estructuras que se pueden utilizar, se llamara a la matriz de correlaciónes:

* *Arbitraria o no estructurada (datos balanceados):*

Considera variancias y covariancias distintas entre las mediciones repetidas. Siendo se expresa como:

La ausencia de restricciones hace que haya que estimar una cantidad muy grande de parámetros.

* *Simetría Compuesta (datos balanceados o no balanceados):*

La correlación entre pares de observaciones es la misma, sin importar la cantidad de rezagos entre ellas, para todo .

* *Toeplitz (datos equiespaciados):*

Se plantea que para cualquier par de respuestas que estén igualmente separadas en el tiempo la correlación es la misma, para todo

* *Autorregresiva de primer orden (datos equiespaciados):*

Es un caso especial de la estructura anterior, en la que para todo Esta estructura asume que la correlación entre medidas repetidas disminuye a medida que aumenta el número de rezagos entre ellas.

* *Markov (datos no equiespaciados):*

Es una generalización de la estructura autorregresiva para mediciones no equiespaciadas. donde para todo .

2.4 Modelos lineales mixtos.

En estos modelos, cada unidad tiene una trayectoria individual caracterizada por parámetros y un subconjunto de esos parámetros ahora se consideran aleatorios. La respuesta media es modelada como una combinación de características poblacionales que son comunes a todos los individuos (efectos fijos) y efectos específicos de la unidad que son únicos de ella (efectos aleatorios).

Se consideran las dos fuentes de variación (intra y entre) presentes en los datos longitudinales. Entonces, este modelo va a ser similar al modelo lineal general con respecto a la parte media del mismo, pero se va a diferenciar en cuanto a la estructura de covariancia.

El modelo lineal mixto para la unidad i se puede expresar en forma matricial como:

Donde:

: Vector de la variable respuesta de la i-ésima unidad, de dimensión ()

: Matriz de diseño de la i-ésima unidad, que caracteriza la parte sistematica de la respuesta, de dimensión ()

: Vector de parámetros de dimensión ()

: Matriz de diseño de la i-ésima unidad, que caracteriza la parte aleatoria de la respuesta, de dimensión ()

: Vector de efectos aleatorios de la i-ésima unidad, de dimensión ()

: Vector de errores aleatorios de la i-ésima unidad, de dimensión ()

y son independientes

Las matrices y contienen las variancias y covariancias de los elementos de los vectores y respectivamente. A partir de este modelo se obtiene:

(media condicional o específica de la i-ésima unidad)

(media marginal)

(variancia condicional)

(variancia marginal)

Generalmente, la matriz adopta una estructura de covariancia arbitraria, mientras que la matriz adopta cualquiera de las vistas anteriormente.

2.5 Estimación de los parámetros del modelo.

Bajo el supuesto de que y se distribuyen normalmente se pueden usar métodos de estimación basados en la teoría de máxima verosimilitud, cuya idea es asignar a los parámetros el valor más probable en base a los datos que fueron observados. Se usarán para estimar los parámetros de la parte media y los de las estructuras de covariancia los métodos de máxima verosimilitud (ML) y máxima verosimilitud restringida (REML) respectivamente.

* *Método de máxima verosimilitud (ML):*

Bajo los supuestos de que y las son independientes entre sí, se obtiene la siguiente función de log-verosimilitud:

Siendo función del vector que contiene los parámetros de covariancia.

La ecuación anterior se debe derivar con respecto a y y luego debe igualarse a cero, de esta manera se obtienen sus estimadores. Cuando es deconocido (lo que generalmente sucede) se obtiene una ecuación no línea, por lo que no se puede obtener una expresión explicita de , para encontrar su solución se recurren a algoritmos numéricos. El estimador del vector resulta:

El estimador resulta insesgado de . Cuando es conocido se conoce la distribución exacta del estimador. Sin embargo, cuando es desconocido, no se puede calcular de manera exacta la matriz de covariancias de . Si el número de unidades es grande se puede demostrar que asintóticamente:

donde

* *Método de máxima verosimilitud restringida (REML):*

El inconveniente que posee el método de MV es que los parámetros de covariancia resultan sesgados. Es decir, a pesar que la estimación de resulta insesgada, no pasa lo mismo con . Si el tamaño de muestra es chico, los parámetros que representan las variancias van a ser demasiado pequeños, dando así una visión muy optimista de la variabilidad de las mediciones, es decir, se subestiman los parámetros de covariancia. El sesgo se debe a que en la estimación MV no se tiene en cuenta que es estimado a partir de los datos.

El método REML separa la parte de los datos usada para estimar de aquella usada para estimar los elementos de , la función log-verosimilitud restringida que se propone es:

Maximizando esta función con respecto a y se obtiene:

En donde es el estimador REML de .

2.6 Pruebas de hipótesis.

Generalmente, la inferencia en estos modelos se centra en los parámetros de la parte media. Es decir, en combinaciones lineales de los parámetros de . La hipótesis lineal general para los test se plantea construyendo dichas combinaciones a través de , siendo una matriz de dimensión r x p.

El estimador de resulta y asintóticamente su distribución muestras es aproximadamente:

Para probar las hipótesis se proponen dos métodos basados en la función de verosimilitud: El test de Wald y el del cociente de verosimilitud.

* *Test de Wald:*

Se plantean las hipótesis:

La estadística de prueba resulta:

Donde se distribuye aproximadamente como .

Este test provee inferencias válidas cuando el N es grande, ya que utiliza la aproximación asintótica a la distribución Normal. Si el N es chico se propone reemplazar la distribución de la chi-cuadrado por una F de Snedecor. El problema con el uso de esta estadística es que no se conocen los grados de libertad del denominador y deben ser calculados con los datos. Algunos métodos para estirmar los grados de libertad son: Contaiment, Within-Between, Satterthwaite y Kenward y Roger, entre otros (Fitzmaurice et al., 2004, pág 98).

* *Test de cociente de verosimilitud:*

Se basa en la teoría asintótica y va a suplir las dificultades del test de wald. El test se obtiene comparando las verosimilitudes de dos modelos, uno de los cuales incorpora la restricción (modelo reducido) y el otro no tiene resitricciones (modelo completo). Estos dos modelos están anidados, ya que el modelo reducido es un caso particular del modelo completo.

Al igual que anteriormente, se plantean las hipótesis:

La función de log-verosimilitud maximizada del modelo completo es:

Y la del modelo reducido es:

Donde es el vector de parámetros resultante de haber impuesto .

Si el N es grande, la distribución muestral aproximada de la estadística es:

(

Donde es el nº de parámetros del modelo completo menos el nº de parámetros del modelo reducido.

Si lo que se quiere comparar mediante este test es la estructura media de dos modelos se recomienda estimar los parámetros mediante el método ML. En cambio, si lo que se desea comparar son patrones de covariancia anidados entre dos modelos se recomienda estimar los parámetros con el método REML.

Cuando se desea hacer inferencia sobre las componentes de variancia (por ejemplo, postular que una variancia es nula), la diferencia entre los dos modelos va a ser de k+1 parametros (1 de var y k de cov). Si sucede esto, la distribución de la estadística ya no será una chi cuadrado común, sino que será una mezcla entre dos chi cuadrados, una de k grados de libertad y otra de k+1 grados de libertad. Si bien hay tablas para ésta estadística, se recomienda en general utilizar en vez de un nivel de significación del 5%, uno del 10%, para no tener problemas de elegir un modelo demasiado simple.

2.7 Elección entre modelos de covariancia.

Cuando los modelos no están anidados, como sucede generalmente cuando se plantean modelos con distintas estructuras de covariancia, no se puede usar el método del cociente de verosimilitud para compararlos.

Como en la matriz de covariancia siempre intervienen los residuos y en ellos aparecen los parámetros de la parte media del modelo, para asegurarse que la parte media esté bien especificada se elige un modelo maximal (modelo con todos los parámetros que queremos incorporar). Dado dicho modelo, se plantean distintos modelos que se van a diferenciar únicamente en la estructura de la matriz de covariancia.

Los enfoques que se proponen a continuación se basan en la comparación de versiones penalizadas de las log-verosimilitudes de los modelos. Como es conocido, a medida que se incorporan parámetros a los modelos, mayor va a ser la verosimilitud. Para comparar modelos con distinto número de parámetros se penalizan a los mismos, surgiendo varios criterios. A continuación se destacan dos de estos: el criterio de Akaike y de Schwarz.

Criterio de Akaike (AIC) =

Criterio bayesiano de Schwarz (BIC) =

Donde:

: log-verosimilitud maximizada del modelo.

: número de parámetros del modelo.

El BIC penaliza más la verosimilitud dando modelos más sencillos. El criterio para seleccionar un modelo es aquel que minimice los valores de AIC o BIC.

2.8 Predicción de los efectos aleatorios.

En muchas aplicaciones la inferencia está centrada solamente en los efectos fijos. Estos parámetros se interpretan como los cambios de la respuesta media en el tiempo, pero muchas veces se desea conocer además los perfiles individuales, por lo que necesitamos conocer los valores de . Como es una cantidad aleatoria se habla de predicción de .

Llamando al predictor de los efectos aleatorios, se debe minimizar . La función que hace que esa esperanza sea lo más chica posible se llama esperanza condicional de dado y se simboliza:

Esto se conoce como mejor predicción lineal insesgada (BLUP) y el BLUP empírico (EBLUP) o estimador empírico de Bayes resulta:

A través de esto se pueden conocer los perfiles individuales, siendo:

, operando algebraicamente se llega a la expresión:

Resulta que el perfil individual estimado es un promedio ponderado entre el perfil de la respuesta media poblacional () y el perfil de la respuesta observada en la unidad i ().

2.9 Examen de residuos.

Como en todo análisis de datos en donde se utiliza un modelo estadístico, es necesario realizar un estudio del cumplimiento de los supuestos del modelo utilizando los gráficos de los residuos.

El modelo lineal mixto para la unidad i es:

Con y independientes y

En los modelos lineales mixtos se va a distinguir entre tres tipos de residuos:

* *Residuos marginales:*

Se definen como la diferencia entre el vector de respuestas y su media marginal estimada.

* *Residuos condicionales:*

Se definen como la diferencia entre el vector de respuestas y su media condicional estimada.

* *Residuos de los efectos aleatorios (EBLUP):*

Estos residuos así calculados reciben el nombre de residuos ordinarios, van a estar correlacionados y sus variancias no van a ser necesariamente contrastes. Para evitar este inconveniente es necesario realizar una estandarización de los mismos:

* Si se divide el residuo por se obtienen los residuos estudentizados.
* Si se divide el residuo por se obtienen los residuos de Pearson.

Se define como residuo puro al residuo que depende de solo los componentes fijos y del error que predice, y residuo confundido si además de depender de las componentes fijas y del error que predicen están en función de otros residuos.

A continuación se presenta una tabla resumen sobre algunos diagnósticos que se pueden realizar utilizando los gráficos de residuos.

|  |  |  |
| --- | --- | --- |
| Diagnóstico | Tipo de residuo | Gráfico |
| Linealidad | Marginal | vs Covariables |
| Presencia de observaciones atípicas | Condicional | estudentizado vs nº de la observación |
| Homocedasticidad del error condicional | Condicional | estudentizado vs valores ajustados |
| Normalidad del error condicional | Condicional | Gráfico probabilístico normal de estudentizado con mínima confusión |
| Presencia de unidades atípicas | EBLUP | vs nº de la unidad |
| Normalidad de los efectos aleatorios | EBLUP | Gráfico probabilístico normal de |

Lo de minima confusión ver porque es un quilombo de explicar? Esto hacer referencia bibliografica

2.10 Covariables que varían en el tiempo.

Las covariables que varían en el tiempo (CVT) pueden ser estocásticas o no. Las CVT no estocásticas son covariables que varían sistemáticamente a través del tiempo pero son fijas por diseño del estudio o bien su valor puede predecirse. En cambio, las CVT estocásticas son covariables que varían aleatoriamente a través del tiempo, es decir, los valores en cualquier ocasión no pueden ser estimados ya que son gobernados por un mecanismo aleatorio.

Cuando una CVT es estocástica, surgen nuevos problemas con respecto a la interpretación y estimación de los parámetros de regresión en modelos para datos longitudinales.

Para covariables estacionarias (no dependientes del tiempo):

Cuando una CVT es estocástica puede que no necesariamente se mantenga. Por ejemplo, la suposición será violada cuando el valor actual de , dado , predice el valor posterior de . En ese caso

y se dice que confunde la relación entre y

En general, cuando esto no se cumple, los valores precedentes y / o posteriores de la CVT estocástica confunden la relación entre y ; esto puede llevar a estimaciones sesgadas de .

2.10.1 Exogeneidad

Se dice que una CVT es exógena cuando los valores actuales y anteriores de la respuesta en la ocasión (, ..., ), dados los valores actuales y precedentes de la covariable (, ... ., ), no predice el valor posterior de . Más formalmente, una CVT es exógena cuando

, ... ., , ..., , ... ., (1)

de lo contrario, se dice que la covariable es endógena.

Teniendo en cuenta que cuando una covariable es exógena,

,

En principio, es posible examinar la suposición de que una CVT es exógena al considerar modelos de regresión para la dependencia de en , ..., (o alguna función (es) conocida (s) de , ..., ) y , ..., (o alguna función conocida de , ..., ). La ausencia de cualquier relación entre y , ..., , dado el perfil de covariable anterior, , ..., , proporciona soporte para la validez de la suposición de que el proceso de covariable es exógeno.

A los parámetros de regresión se les puede dar una interpretación causal sólo cuando se puede asumir además que las covariables variables en el tiempo son exógenas con respecto a la variable de respuesta.

Si es distinto a el estimador de beta puede ser inconsistente del verdadero valor

2.10.2 Tipos de CVT

Trabajos más recientes han definido una nueva categorización de las CVT para facilitar las interpretaciones y los métodos de estimación adecuados para el modelo. Se pueden definir cuatro tipos de CVT relacionados con el grado de no exogeneidad con respecto a la respuesta.

* CVT Tipo 1

# Aplicación.

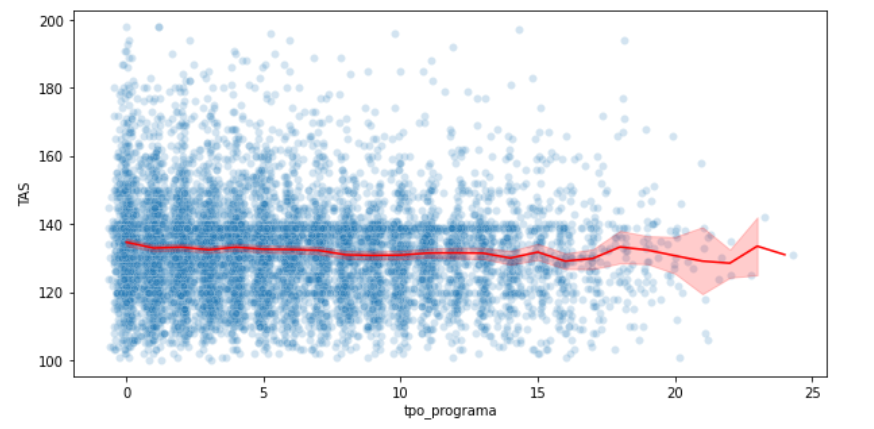
Un programa de atención y control de pacientes hipertensos iniciado en el año 2014 en Rosario realiza un seguimiento exhaustivo de aproximadamente 1000 pacientes. Este programa contempla: efectores no médicos supervisados, tratamiento farmacológico genérico para la hipertensión y utilización de un algoritmo terapéutico sistematizado. En cada visita se registran tanto características de los pacientes, del tratamiento y de los valores de la tensión arterial. En particular, se desea evaluar si la adherencia al tratamiento farmacológico influye en los valores de la tensión arterial sistólica a lo largo del seguimiento.

Como la variable “adherencia al tratamiento farmacológico” es una CVT estocástica se evaluarán diferentes enfoques para incluirla en un modelo longitudinal que pueda explicar el cambio en la tensión arterial sistólica media a lo largo del tiempo.

* 1. Descripción de la población en estudio.

En primera instancia se explora la estructura de la parte media. Para ello, a continuación se presentan gráficos que permiten observar el comportamiento de la TAS a través del tiempo de tratamiento.

Gráfico 3.1.1: Gráfico de perfiles individuales y perfiles promedio.



En base al gráfico se puede observar que la TAS en promedio disminuye levemente en los primeros meses del estudio, mientras que luego se mantiene más constante. También se observa gran variabilidad en el tiempo 0, lo que puede llevar a pensar en una ordenada al origen aleatoria. Ver qué hacer con los últimos meses

En cuanto a la exploración de la estructura de covariancia, se realizó un semivariograma en base a un modelo completo:

Gráfico 3.1.3: Semivariograma muestral.

HACER EL VARIOGRAMA

En base a este gráfico se puede pensar que no existe correlación serial en el tiempo y que probablemente haya que incluir efectos aleatorios.

Dada la gran cantidad de individuos que hay en el estudio se dificulta ver la pendiente a través del gráfico de perfiles individuales, si ésta debe ser aleatoria se analizará luego mediante test de hipótesis.

* 1. Análisis de la covariable dependiente del tiempo.

Como se mencionó anteriormente, se deberá analizar si la variable de adherencia al tratamiento es exógena, ya que ésta varía en el tiempo.

Se ajustó para cada tiempo un modelo logit teniendo en cuenta los 3 momentos anteriores:

Para los todos los tiempos las covariables dieron no significativas con un nivel de significación del 5% excepto sólo por uno, por lo que se puede decir que la variable de adherencia al tratamiento es exógena (ver anexo).

* 1. Análisis inferencial.

Debido a que el modelo de covariancia depende del modelo postulado para la parte media, para evitar alguna falta de ajuste de ésta se utiliza un modelo maximal:

Para este tipo de modelo se pueden incorporar efectos aleatorios para la ordenada al origen y para la pendiente. Se ajustaron modelos mixtos con estructura de covariancia independiente para la parte media y se obtuvo la siguiente tabla.

Tabla 3.3.2: Medidas resumen sobre los distintos modelos mixtos.

|  |  |
| --- | --- |
| **Efecto aleatorio** | **-2 Res Log Likelihood** |
| Ordenada al origen | 65331.90 |
| Pendiente | 66121.97 |
| Ordenada al origen y Pendiente | 65229.96 |

Luego, se realizaron test de hipótesis para probar la inclusión de los efectos aleatorios.

Siendo la matriz de covariancia de la parte aleatoria, se postula:

Se calculó la estadística para cada modelo y se obtuvo:

Comparando estos valores con una chi-cuadrado 50/50 de 1 y 2 grados de libertad (valor crítico de 6.84) se llega a que en ambos casos se rechaza la hipótesis nula.

En base a esto se plantearon efectos aleatorios para ambas partes.

Para la parte media se llevó a cabo un proceso de selección de variables, partiendo de un modelo completo y removiendo covariables no significativas en base a diferentes hipótesis. El modelo obtenido entonces es el siguiente:

+