



Fecha de presentación del Informe: Día Mes Año

Datos generales del Proyecto

Código del proyecto: C.I 2706			
Título del proyecto: Manejo de datos faltantes en el análisis de datos longitudinales usando un modelo lineal mixto: Comparación por medio de simulación de varias alternativas y aplicación a datos de salud			
Facultad o Instituto Académico: Facultad de Ingeniería			
Departamento o Escuela: Escuela de Estadística			
Grupo (s) de investigación: Grupo de investigación en estadística aplicada - INFERIR			
Investigadores ¹	Nombre	Tiempo asignado	Tiempo dedicado
Investigador Principal	Alvaro José Flórez ✓	300h/semestre	300h/semestre
Coinvestigadores	Mercedes Andrade ✓	30h/semestre	30h/semestre
Otros participantes	Karen Lizeth El Helou	10h/semana	10h/semana
	Marcela Lemus	10h/semana	10h/semana

1. Resumen ejecutivo:

La presencia de datos faltantes en un análisis de datos longitudinales es un problema bastante común, especialmente en estudios médicos donde se trabaja con sujetos humanos, que afecta la precisión de las estimaciones y además puede provocar sesgos, lo que puede llevar a conclusiones erróneamente acerca de un tratamiento o una terapia. Estas consecuencias negativas dependen, no solo de como se manejan los datos faltantes o de la proporción de los mismos, sino también de las razones por las cuales estos se presentan (llamados, mecanismos de datos faltantes). Esto hace que sea necesaria la utilización de alguna técnica que haga un buen trato a los datos faltantes, y además que sea robusta a los diferentes mecanismos de

¹ Todas las personas relacionadas en el informe y que participen en el proyecto deben haber suscrito el acta de propiedad intelectual de acuerdo con los formatos establecidos.



datos faltantes. Por esta razón este trabajo tiene como objetivo la comparación de diferentes técnicas para el manejo de datos faltantes en el análisis de datos longitudinales.

Para evaluar el comportamiento de estas técnicas se realizó un estudio de simulación donde se compararon las estimaciones de los efectos fijos y aleatorios de un modelo lineal de efectos mixtos usando cuatro métodos para el manejo de datos faltantes, bajo diferentes escenarios de datos faltantes (mecanismos), estos escenarios se basaron en el análisis de datos longitudinales obtenidos de recién nacidos con bajo peso entre febrero de 2002 y noviembre de 2010 que pertenecen al programa de seguimiento de alto riesgo y canguro del HUV.

En los resultados se pudo apreciar que los diferentes mecanismos de datos faltantes, así como la técnica que se utilice para su tratamiento, afectan las estimaciones que se hacen sobre el modelo, afectando de alguna manera las conclusiones del estudio. Pero las técnicas más simples como el análisis de datos completos y la imputación simple tuvieron el peor desempeño, proporcionando estimaciones de los parámetros sesgadas y con mayor varianza en la mayoría de las situaciones simuladas, mientras que las técnicas basadas en verosimilitud y la imputación múltiple mostraron mejores resultados en todos los casos. Solamente bajo el mecanismo de datos faltantes MNAR todas las técnicas proporcionaron estimaciones segadas.

Como recomendación se propone considerar una gama amplia de estructuras de datos y datos faltantes en los escenarios de simulación, para obtener unas conclusiones más generales. Además de la posible inclusión de valores faltantes en las covariables del modelo y variaciones de los modelos de imputación.

EXECUTIVE SUMMARY

The presence of missing data in a longitudinal data analysis is a very common problem, especially in medical trials when we work with human subjects. This problem affects the precision in the estimates and it may provide biased estimates, which can lead to misleading inferences about a treatment or therapy. These negative consequences depend on the reasons of the missing values (called missing data mechanisms), and not only in the percentage of missingness or how we treat them. This makes necessary the use of a technique that treats the missing data correctly and this technique have to be robust to different missing data mechanisms. For those reasons this project had the aim to compare different techniques to handle missing data in a longitudinal data analysis.

To evaluate the performance of these techniques, we performed a simulated study where we compared the estimates of the fixed and random effects of a linear mixed model using four different methods to handle missing data, under different scenarios of missing data (mechanisms). These scenarios were based on a longitudinal data analysis of low birth weight babies of the between February 2012 and November 2010 who belonged to the HUV high-risk and the kangaroo program.

In the results we observed that the different mechanisms of missing data and the technique used to treat them affect the estimated model, affecting the conclusion of the study. But the simplest techniques, such as the complete cases analysis or simple imputation, had the worst performance, providing biased and inefficient estimates of the parameters of the model in most of the simulated situations, while likelihood-based techniques and multiple imputation showed better performance in all cases. Only on MNAR mechanism all the techniques provided biased estimates.

As a recommendation we propose a wider range of data and missing data structures in the simulation scenarios with the purpose to obtain more general conclusions. We also recommend the inclusion of missing data in the covariates of the model and variations of the imputation model

2. Síntesis del proyecto:

Planteamiento del problema

La característica que define a un análisis de datos longitudinales es que tienen observaciones de un mismo individuo a lo largo de un periodo de tiempo determinado, con el objetivo principal de observar los cambios en las variables observadas a lo largo del periodo de estudio, aunque también se puede estar interesado de determinar si estos cambios están relacionados con algunas covariables observadas (Fitzmaurice *et al.* 2004). El método más utilizado para el manejo de este tipo de datos es el modelo de regresión de efectos mixtos, puesto que no tiene las limitantes que presentan algunas otras técnicas como el ANOVA para medidas repetidas o el MANOVA.

En cualquier análisis de datos, especialmente cuando son longitudinales, la presencia de datos faltantes (observaciones planeadas que no se pudieron obtener) es un inconveniente muy frecuente. El problema se incrementa cuando se trabaja con individuos humanos y en largos periodos de observación, como lo afirman Daniels & Hogan (2008) y Penny & Atkinson (2012). La presencia de datos faltantes puede, bajo ciertas circunstancias, afectar no solamente la precisión de las estimaciones, sino que también puede introducir sesgos, los cuales pueden llevar a conclusiones erróneas acerca de un tratamiento o una terapia, las cuales tendrá un impacto en la práctica de la medicina (Fielding *et al.* 2012). Por lo cual la utilización adecuada de alguna técnica para su manejo resulta aconsejable en estos casos.

Hay varias técnicas disponibles para el tratamiento de los valores faltantes en una base de datos longitudinales. Una de estas opciones es: desechar todos los individuos con algún dato faltante (Análisis de datos completos), otra sería utilizar técnicas para la imputación de los datos faltantes, que pueden ser simples o múltiples. La escogencia de estos métodos puede depender del tipo de variable que se va a imputar y de los motivos que llevaron a que un dato

fuera no observado (llamado mecanismo de datos faltantes) (Penny & Atkinson, 2012).

El objetivo de este proyecto es observar cómo la presencia de datos faltantes afecta a las estimaciones de los parámetros en un modelo lineal mixto para datos longitudinales, teniendo en cuenta los diferentes mecanismo de datos faltantes planeados en la literatura (principalmente por Little & Rubin (2002)), además de probar la efectividad de las técnicas para el manejo de datos faltantes más comúnmente utilizadas. Todo esto basándose en el estudio de casos simulados que nos permita abarcar diferentes escenarios de interés y concluir sobre la efectividad de dichas técnicas. También se pretende aplicar algunas de estas técnicas en un caso real de investigación en el campo de la salud que presente datos faltantes.

Objetivos

Comparar, por medio de simulaciones, diferentes técnicas para el manejo de datos faltantes en el análisis de datos longitudinales usando un modelo lineal mixto.

Objetivos específicos

- Evaluar el efecto de los diferentes mecanismos de datos faltantes en las propiedades del modelo lineal mixto (insesgamiento, eficiencia, consistencia)
- Evaluar el efecto de la aplicación de diferentes técnicas de imputación de datos sobre las estimaciones de los parámetros del modelo lineal mixto, teniendo en cuenta los diferentes mecanismos de datos faltantes.
- Determinar con cuáles de estas técnicas se obtienen mejores estimaciones del modelo dependiendo del mecanismo de datos faltantes y mostrar las ventajas y desventajas de cada una de ellas.
- Aplicar las técnicas que obtuvieron mejores resultados en un caso de estudio del campo de la salud, para observar su comportamiento en una situación real.

Metodología

Paso 1. Escenario de simulación para las bases de datos.

Los escenarios de simulación propuestos se basaron en el análisis datos longitudinales obtenidos de historias clínicas de recién nacidos con bajo peso entre febrero de 2002 y noviembre de 2010 que pertenecen al programa de seguimiento de alto riesgo y canguro del HUV. Esto datos fueron estudiados por Bermudez *et al.* (2013). En este estudio se observó el

peso de los recién nacidos en cinco momentos del tiempo durante 3 años (recién nacidos, a los 12, 18, 24 y 36 meses).

La simulación de las bases de datos se hizo a partir de los resultados del análisis de estos datos, donde las estimaciones del modelo ajustado servirán de parámetros del modelo con el cual se generan las bases de datos.

Luego de generar las diferentes bases de datos se procedió a la simulación de los datos faltantes. Para esta simulación se tuvo en cuenta las definiciones de los patrones y mecanismos de datos faltantes de Rubin (1976), Verbeke & Molenberghs (2000) y Daniels & Hogan (2008). Cada escenario se hizo bajo cada mecanismo de datos faltantes: Valores faltantes completamente aleatorios (MCAR), aleatorios (MAR) y no aleatorios (MNAR) y los patrones monótonos y no monótonos.

Paso 2. Simulación de datos y comparación de técnicas.

Teniendo en cuenta el mecanismo de datos faltantes (4 casos) y el patrón (2 casos), se tienen 8 escenarios de simulación de datos faltantes. También se realizan simulaciones para diferentes porcentajes de datos faltantes (10 %, 30 % y 50 % de individuos con datos faltantes).

Para cada escenario, se simulan 1000 bases de datos con 154 individuos, 106 con control prenatal y 48 sin control prenatal (igual que en la base de datos original) y se estiman los modelos usando las técnicas de manejo de datos faltantes propuestas (análisis de casos completos (CC), Last Observation Carried Forward (LOCF), Imputación Múltiple (MI) (Modelación conjunta-JM y Especificación completamente condicionada - FCS) y Casos disponibles usando el algoritmo EM (AC).

Luego se observan las estimaciones para cada parámetro del modelo, tanto los efectos fijos, como los aleatorios. A partir de estos resultados se hicieron las comparaciones de las estimaciones de cada parámetro por los diferentes métodos por medio del sesgo relativo y el error estándar de las estimaciones de cada parámetro.

Resultados

En primera instancia se presentan los resultados del análisis de los datos de neonatos con bajo peso al nacer, usando diferentes técnicas para el manejo de datos faltantes. Esto con la finalidad de mostrar que los resultados con cada método pueden ser diferentes, en algunos casos pueden llevar a conclusiones contrarias. Posteriormente se presentan los resultados de las simulaciones bajo los diferentes escenarios de simulación propuestos.

Resultados del análisis de datos

La base de datos presenta gran cantidad de datos faltantes, de los 154 neonatos observados en el estudio hay 119 sujetos con al menos una observación faltante. Las mediciones en el tiempo 0 están observadas para casi todos los individuos, pero a medida que pasa el tiempo van aumentando las observaciones faltantes, al mes 36 solo se tiene el 30.7% de las mediciones observadas. De acuerdo con el comportamiento gráfico de los datos y a los resultados encontrados por Bermudez *et al.* (2013), se propuso el siguiente modelo lineal mixto:

$$y_{ij} = \beta_0 + \beta_1 t_j + \beta_2 t_j^2 + \beta_3 CP_i + b_{0i} + b_{1i} t_j + b_{2i} t_j^2 + \varepsilon_i$$

donde: y_{ij} es el peso del neonato i en el tiempo t_j , $i = 1, \dots, 154$. t_j es el tiempo de medición, $t_j = 0, 12, 18, 24, 36$. $CP_i = 1$ si la madre del neonato i recibió asistencia prenatal, $CP_i = 0$ en caso contrario. $\beta_0, \beta_1, \beta_2, \beta_3$ son los efectos fijos y b_{0i}, b_{1i}, b_{2i} son los efectos aleatorios.

Como resultado de la aplicación sobre los datos de peso de los neonatos, se pudo observar que las técnicas para el manejo de datos faltantes utilizadas proporcionan un ajuste diferente del modelo propuesto. Donde se observa que con las técnicas CC y LOCF se ajustan curvas de crecimiento con pendiente más baja que con los otros métodos, esto puede llevar a conclusiones erradas sobre el comportamiento medio del peso de los neonatos (posiblemente subestimando la verdadera curva), sugiriendo que el programa del hospital no ayuda a que los bebés ganen peso durante los primeros 36 meses.

Resultados de las simulaciones hechas

En las simulaciones bajo el mecanismo MCAR se pudo apreciar que solamente la técnica LOCF proporciona estimaciones sesgadas (sesgos negativos), especialmente del parámetro β_1 y β_4 , asociados a la pendiente de la curva estimada y al efecto del control prenatal, y estos tienden a aumentar a medida que la proporción de datos faltantes es mayor. Las otras tres técnicas presentan estimaciones insesgadas en todos los casos, pero cabe resaltar que los errores estándar mayores se encuentran para las estimaciones por CC, Esto debido a que al descartar todos los individuos con al menos un faltante disminuye drásticamente el tamaño de muestra. Los otros tres métodos (AC, MI-JM y MI-FCS) proporcionan estimaciones insesgadas y con errores estándar similares.

Bajo el mecanismo MAR, se pudo apreciar que tan solo las técnicas AC y MI-JM producen estimaciones insesgadas para todos los coeficientes en todos los casos simulados. La técnica MI-FCS proporciona sesgos negativos pequeños para los parámetros β_2 y β_3 , los cuales aumentan al tener mayor porcentaje de datos faltantes. Las técnicas CC y LOCF proporcionan estimaciones sesgadas en todos los casos, proporcionando curvas con menor pendiente y efecto de la variable control prenatal, pues en ambos casos los sesgos son negativos. Al igual que en los escenarios MCAR, el método CC proporciona las estimaciones con mayor error

estándar, mientras que las técnicas AC y MI-JM presentaron las estimaciones con menor variabilidad, seguidos por MI-FCS.

Tal como era de esperarse, todas las técnicas proporcionan estimaciones sesgadas de los parámetros asociados a los efectos fijos bajo el mecanismo MNAR. Al igual que los casos anteriores, las técnicas AC y MI-JM proporcionan las mejores estimaciones, en este caso con pequeños sesgos negativos para los parámetros β_1 y β_3 (máximo del 8% para una proporción del 50% de datos faltantes). La técnica MI-FCS proporciona sesgos un poco más grandes, especialmente para el parámetro β_3 . Tanto CC y LOCF, al igual que en los escenarios anteriores, siguen presentando los resultados más pobres, pues presentan los mayores sesgos relativos y errores estándar.

Los resultados de los escenarios MCAR, MAR y MNAR bajo patrón de datos faltantes intermitentes no difieren en gran medida de los encontrados en los escenarios de patrón de datos faltantes monótonos, bajo las mismas circunstancias.

Principales conclusiones y recomendaciones

Aunque las técnicas de análisis de datos completos (CC) e imputación simple (como el caso de LOCF) son atractivas por su simplicidad y disponibilidad en los software estadísticos, estas presentan grandes desventajas frente a técnicas más robustas ante los datos faltantes, como imputación múltiple (Modelación conjunta - JM y especificación completamente condicionada - FCS) o análisis de casos disponibles usando estimaciones de máxima verosimilitud por medio del algoritmo EM (AC), como se pudo apreciar en los resultados de esta investigación. Estos resultados son acordes a múltiples investigaciones que también recomiendan evitar el uso de estas técnicas tradicionales (Olsen *et al.* 2012, Fielding *et al.* 2012, Penny & Atkinson 2012).

Por medio de los resultados de las simulaciones de los diferentes escenarios se puede observar que el método CC proporcionó estimaciones insesgadas de los efectos fijos bajo el mecanismo MCAR, sin embargo, estas estimaciones tuvieron el error estándar más alto en todos los casos estudiados; esto era de esperarse puesto que no hace uso de toda la información disponible, ya que elimina observaciones válidas de los individuos que presentan al menos un dato faltante. La imputación simple LOCF proporciona estimaciones de los coeficientes sesgadas incluso bajo el mecanismo MCAR. Ambos métodos presentan estimaciones sesgadas bajo los mecanismos MAR y MNAR, subestimando el estimador asociado a la pendiente y al control prenatal. Esto sesgo produce que el modelo subestime la curva del peso de los neonatos, mostrando pesos medios por debajo de los valores reales, y poca importancia al efecto del control prenatal en el futuro peso de un neonato. Por medio de estos resultados pueden llevar a conclusiones erróneas sobre el rendimiento real de los programas de bajo peso del hospital.

A diferencia de los métodos tradicionales, las técnicas AC y MI-JM mostraron mejores



resultados, proporcionando estimaciones insesgadas de los efectos fijos en los escenarios con mecanismo MAR, aunque estimaciones con sesgos pequeños bajo los escenarios MNAR. Como se pudo apreciar en los resultados, estos sesgos fueron mayores a medida que la proporción de datos faltantes es mayor. La técnica de imputación múltiple usando el modelo de imputación FCS, presentó unos resultados un poco inferiores a las técnicas mencionadas anteriormente, al presentar algunos sesgos pequeños bajo el mecanismo MAR.

Los escenarios de simulación planteados tuvieron en cuenta diferentes patrones de datos faltantes bajo una misma estructura de datos, por lo cual para ampliar las conclusiones sería conveniente plantear escenarios de simulación que consideren una gama amplia de estructuras de datos, diferentes estructuras de media y correlación. Aunque la presencia de datos faltantes es más frecuente en la variable respuesta, o en las variables tomadas de forma longitudinal, también pueden presentarse situaciones donde hayan datos faltantes en algunas de las covariables del estudio, por lo cual sería de gran interés observar que técnicas de imputación pueden utilizarse en este caso y observar el efecto que tienen sobre los resultados del modelo. Teniendo en cuenta las ventajas que tiene la imputación múltiple de utilizar variables auxiliares y variaciones al modelo de imputación, también se pueden plantear escenarios donde se incluyan este tipo de variables.

Aunque los métodos AC y MI parecen mostrar mejores resultados, no se puede estar seguro que estos resultados sean válidos para el análisis longitudinal del peso de los neonatos, pues no se sabe si el mecanismo de datos faltantes es al menos MAR. Los datos observados no proporcionan ninguna información para determinar un posible mecanismo, además no se tiene ninguna explicación sobre las razones que llevaron a que algunos neonatos no se siguieran observando. Por esta razón, se recomienda recolectar información adicional que pueda explicar posibles razones que lleven a datos faltantes. Schafer & Graham (2002) recomiendan adicionar en las encuestas preguntas que permitan determinar la posibilidad de que un paciente no asista al próximo tiempo de medición y así anticiparse a las posibles causas de retiro. El uso de esta información puede convertir con conjunto de datos MNAR en uno MAR.



3. Productos:

Tabla No. 1. Cantidad y tipo de productos pactados en el *Acta de Trabajo y Compromiso* y productos finalmente presentados

TIPO DE PRODUCTOS	No. de PRODUCTOS PACTADOS		No. de PRODUCTOS PRESENTADOS	
Productos de nuevos conocimientos				
Artículo completo publicado en revistas A1 o A2			1 (en evaluación)	
Artículo completo publicados en revistas B	1			
Artículo completo publicados en revistas C				
Libros de autor que publiquen resultados de investigación				
Capítulos en libros que publican resultados de investigación				
Productos o procesos tecnológicos patentados o registrados				
• Prototipos y patentes				
• Software				
Productos o procesos tecnológicos usualmente no patentables o protegidos por secreto industrial				
Normas basadas en resultados de investigación				
Formación de recursos humanos	No. de estudiantes vinculados	No. de tesis	No. De estudiantes Vinculados	No. De tesis
Estudiantes de pregrado	2	1	2	1
Semillero de Investigación				
Estudiantes de maestría				



TIPO DE PRODUCTOS	No. de PRODUCTOS PACTADOS		No. de PRODUCTOS PRESENTADOS	
Estudiantes de doctorado				
Productos de divulgación				
Publicaciones en revistas no indexadas				
Ponencias presentadas en eventos (congresos, seminarios, coloquios, foros)	No. de ponencias nacionales	No. de ponencias internacionales	No. de ponencias nacionales	No. de ponencias internacionales
Propuesta de investigación	1		2	1
Propuestas presentadas en convocatorias externas para búsqueda de financiación.	1		1	

Tabla No. 2. Detalle de productos.

Para cada uno de los productos obtenidos y relacionados en la tabla anterior, indique la información solicitada para cada uno, anexando copia de las respectivas constancias. Como anexo a esta guía encontrará el instructivo para instructivo para la revisión de informes finales y productos

Tipo de producto:	Artículo <i>OK (A2)</i>
Nombre General:	Revista Ingeniería y Ciencias de la universidad EAFIT. (A2)
Nombre Particular:	Comparación por medio de simulación de algunas alternativas para el manejo de datos faltantes en el análisis de datos longitudinales usando un modelo lineal de efectos mixtos
Ciudad y fechas:	El artículo se encuentra en evaluación
Participantes:	Karen Lizeth El Helou, Marcela Lemus y Alvaro José Flórez
Sitio de información:	
Formas organizativas:	Universidad del Valle



Tipo de producto:	Ponencia Internacional <i>OK</i>
Nombre General:	28th International Workshop of Statistical Modeling
Nombre Particular:	Handling missing data in longitudinal studies: an application to healthcare data
Ciudad y fechas:	Palermo, Italia. 8 al 12 de julio de 2013
Participantes:	Alvaro José Flórez, Mercedes Andrade, Javier Torres, Nora Donaldson, Nairn Wilson
Sitio de información:	
Formas organizativas:	Universidad del Valle, King's College London.

Tipo de producto:	Ponencia Nacional <i>OK</i>
Nombre General:	XXIII Simposio de Estadística
Nombre Particular:	Comparación por medio de simulación de varias alternativas para el manejo de datos faltantes en el análisis de datos longitudinales usando un modelo lineal mixto
Ciudad y fechas:	Bogotá, 25 al 27 de julio de 2013
Participantes:	Karen Lizeth El Helou, Marcela Lemus, Alvaro José Flórez
Sitio de información:	
Formas organizativas:	Universidad del Valle

Tipo de producto:	Ponencia <i>OK</i>
Nombre General:	VI Escuela de Verano – CEAES



Nombre Particular:	Comparación por medio de simulación de varias alternativas para el manejo de datos faltantes en el análisis de datos longitudinales
Ciudad y fechas:	Cali, 7 al 9 de noviembre de 2013.
Participantes:	Alvaro José Flórez, Mercedes Andrade, Javier Torres, Nora Donaldson, Nairn Wilson
Sitio de información:	
Formas organizativas:	Universidad del Valle, King's College London.

Tipo de producto:	Tesis
Nombre General:	Trabajo de grado del programa de estadística
Nombre Particular:	Manejo de datos faltantes en el análisis de datos longitudinales usando un modelo lineal mixto: Comparación por medio de simulación de varias alternativas
Ciudad y fechas:	Cali,
Participantes:	Marcela Lemus, Karen Lizeth El Helou. Director: Alvaro José Flórez
Sitio de información:	Biblioteca Central.
Formas organizativas:	Universidad del Valle

4. Impactos actual o potencial:

A partir de los resultados de este proyecto, presentados como ponencias en tres eventos académicos diferentes y un artículo académico, se pudo mostrar a la comunidad científica en general las dificultades que se pueden presentar cuando hay presencia de datos faltantes dentro del estudio de datos longitudinales y además se alerta sobre los sesgos que se pueden cometer en las estimaciones, y por ende, en las conclusiones al usar algunas técnicas para su manejo (sobre todo las comúnmente utilizadas). Esto con el fin de que los investigadores hagan un buen uso de las técnicas estadísticas cuando estén en un estudio con estas características y puedan llegar a conclusiones que tengan validez estadística.



Se espera además que este proyecto sirva como referencia para un estudio sobre el tema de los datos faltantes no solamente en el caso de datos longitudinales usando un modelo mixto, sino también en diferentes tipos de estudios y técnicas estadísticas, puesto que la presencia de datos faltantes puede ocurrir en muchas otras situaciones.

Por medio de este proyecto de investigación se apoyó el desarrollo de un trabajo de grado de dos estudiantes del pregrado de estadística, que tuvo una conclusión satisfactoria. Esto contribuyó a la formación académica de estas estudiantes.

La presente versión del informe contiene las observaciones de los evaluadores: No

Firma del investigador principal

VoBo. Vicedecano de Investigaciones

Por favor presente su informe impreso y en formato digital en hoja tamaño carta, letra arial 11, con espacios de 1 1/2

