Trabajo de Fin de Máster: Tema: Estrategias para el tratamiento de datos faltantes ("missing data") en estudios con datos longitudinales

Memoria del Trabajo

Román Octavio Calafati
Máster Universitario en Bioinformática y Bioestadística
Área Bioestadística

Directora: Nuria Pérez Álvarez
Profesor Responsable de la Asignatura: Alexandre Sánchez Pla

6 de junio de 2017

FICHA DEL TRABAJO DE FIN DE MASTER

Título del trabajo:	*Estrategias para el tratamiento de datos faltantes ("missing data") en datos longitudinales*
Nombre del autor:	*Román Octavio Calafati*
Nombre del consultor/a:	*Nuria Pérez Álvarez*
Nombre del PRA:	*Alexandre Sánchez Pla*
Fecha de entrega (mm/aaaa):	05/2017
Titulación:	*Máster Universitario en Bioinformática y Bioestadística*
Área del Trabajo Final:	*Bioestadística*
Idioma del trabajo:	*Castellano*
Palabras clave:	*missing data, longitudinal studies*

Resumen del Trabajo:

En este trabajo, se analiza y estudia la problemática que pueden generar los datos faltantes ("missing data"): un problema recurrente en bases de datos de estudios longitudinales, sobre las cuales se realiza un análisis estadístico metódico. Asimismo, se evalúan y caracterizan los distintos mecanismos por los cuales se producen estos datos faltantes, los cuales desembocan en diferentes patrones de ausencia. Además, se determinan los métodos disponibles para analizar estos datos en presencia de estos datos ausentes y las bondades y limitaciones de estos métodos. Finalmente, se aborda un análisis de datos para ejemplificar la aplicación de los métodos estudiados.

Abstract:

In this paper, the missing data problem is studied and analysed: this is a returning problem in data bases and data sets from longitudinal studies, where methodical statistical analysis are to be performed. Furthermore, different mechanisms, arousing from this missing data problem, which result in different missing patterns, are evaluated and characterized. Also, several available methods in order to analyze these data, where absent data are present, are determined to define the goodness and limits of these methods. Finally, a data analysis is executed, in order to illustrate the application of these researched methods.

Índice

Lista de figuras

1. Introducción

1.1 Contexto y justificación del Trabajo

1.1.1 Contexto y descripción general

En este trabajo se estudia y analiza la problemática originada por los datos faltantes ("missing data") en bases de datos de estudios longitudinales, sobre las cuales se desea realizar un análisis estadístico metódico. El mismo, es un problema recurrente en este tipo de estudios y en muchos otros, ya que sobre los datos faltantes: "everybody has them, nobody wants them" ("todo el mundo los tiene, nadie los desea").[1]
Este asunto, relacionado a la forma en que se recolectan y obtienen los datos, puede desembocar en que estos missings tengan un efecto significativo en las conclusiones que se obtengan del análisis de los datos, ya que se reduzca la representatividad de la muestra obtenida y por lo tanto se distorsionen las inferencias sobre la población.

1.1.2 Justificación del Trabajo de Fin de Máster

Por lo tanto, en este trabajo, se analiza y estudia la problemática que pueden generar los datos faltantes. Asimismo, se evalúan y caracterizan los distintos patrones ("mecanismos"), por los cuales se producen los datos faltantes, y se determinan los métodos disponibles para analizar estos datos en presencia de datos faltantes y las bondades y limitaciones de estos métodos.

[1] Ref. [1], página 2.

1.2 Objetivos del Trabajo

1.2.1 Objetivos generales:

- se desea conocer la casuística involucrada en el tratamiento de datos faltantes en estudios longitudinales, entendiendo como llevar a cabo dicho tratamiento,

- y lograr la implementación de las herramientas y procedimientos que permiten realizar el mencionado tratamiento.

1.2.2 Objetivos específicos:

- evaluar en distintos datasets si los datos faltantes se corresponden con alguno de los siguientes patrones: MCAR (missing completely at random - faltantes completamente aleatorios), MAR (missing at random - faltantes aleatorios), MNAR (missing not at random - faltantes no aleatorios).

- determinar si la imputación de los datos faltantes de acuerdo a un patrón detectado en el punto anterior puede realizarse mediante: imputación parcial, eliminación parcial, análisis completo, interpolación.

1.3 Enfoque y metodología seguida para la consecución del TFM

La estrategia a seguir para llevar adelante este TFM, será estudiar e implementar las distintas estrategias factibles para tratar datos faltantes en estudios longitudinales.

Para ello, se trabajará con bibliografía que analiza y describe a fondo las distintas estrategias para la realización del tratamiento de datos faltantes (véase el apartado 5 en este mismo documento).

Además, para la implementación de estas estrategias se trabajará utilizando el lenguaje de programación R, orientado al tratamiento de datos con fines estadísticos y de análisis, utilizando para el alcance de este TFM los paquetes especializados en tratamiento de datos faltantes, tales como Amelia, BaBoon, Mi, Hmisc, Mice, en principio (*), aunque cabe mencionar que existen tantos paquetes R disponibles (más de 8000 en la actualidad), que a los paquetes mencionados podría agregarse o bien cambiarse algunos de ellos por otros con más funcionalidades.

Para la visualización de los datos faltantes en los datasets, un paquete apropiado podría ser VIM, una potente herramienta para graficación en R, que permite la visualización de los datos faltantes y la posterior visualización de la imputación de los mismos.

Para el registro y seguimiento de la planificación del TFM se trabajará con el software xPlan, disponible para plataformas Apple Macintosh y sistemas operativos macOS.

(*) Una descripción de los paquetes estadísticos que piensan utilizarse para este TFM, se explica en el anexo 1: "Resumen de paquetes estadísticos para tratar datos faltantes en R".

2. Datos faltantes: elementos teóricos para el tratamiento de los mismos.

2.1 Introducción a la teoría de datos faltantes

Los estudios sobre análisis estadístico de datos con valores faltantes han florecido desde principio de los años 70, estimulados por los avances en la tecnología informática que lograron facilitar los cálculos numéricos, previamente muy laboriosos.[2]

Muchos investigadores utilizan métodos ad hoc, tales como el de "análisis de caso completo", "análisis de caso disponible" (borrado por parejas), o el de "imputación de valor único". Aunque estos métodos se implementan facilmente, requieren de suposiciones ("assumptions") sobre los datos, que rara vez se mantienen.

Todos los investigadores se han encontrado con el problema de datos cuantitativos faltantes en algún momento de su trabajo. Los sujetos investigados, pueden haberse rehusado o olvidado de responder a una pregunta en una encuesta, tal vez se hayan perdido los archivos, o tal vez los datos no hayan sido grabados apropiadamente. Dado el precio asociado a la recolección de datos, no podemos permitirnos empezar otra vez o esperar hasta que se hayan desarrollado métodos a prueba de tontos para la recolección de información, lo cual es un objetivo inalcanzable. Por lo tanto, nos encontraremos que tenemos que decidir como analizar los datos cuando no tenemos información completa de todos los informantes. ¿Cuáles son entonces las opciones disponibles para analizar datos con información faltante? [3]

El caso más común - y el más fácil para aplicar - es la utilización de aquel en el cual se tienen en cuenta sólo aquellos casos con información completa. Los investigadores conscientemente o por defecto en un análisis estadístico eliminan los datos que no están completos en las variables de interés. Una alternativa a la utilización sólo de casos completos, es la cual en que los investigadores rellenan con valores factibles las observaciones incompletas, como por ejemplo utilizando la media para los casos observados sobre la variable en cuestión. Más recientemente, los investigadores han respaldado métodos que están basados en modelos de distribución para los datos (tales como el de máxima probabilidad y el de imputación múltiple).

Los métodos para analizar datos faltantes requieren suposiciones acerca de la naturaleza de los datos y también tener en cuenta las razones por las cuales las observaciones están incompletas. Muchas veces estas razones no son debidamente consideradas. Cuando los investigadores utilizan métodos sin considerar cuidadosamente las suposiciones requeridas para ese método, corren el riesgo de obtener resultados sesgados y desviados. El repaso de las etapas de recolección de datos, preparación de los mismos, el análisis, y la interpretación de los resultados dará luz en la cuestiones que los investigadores deben considerar al tomar una decisión acerca de como manejar los datos en su trabajo. Los métodos más utilizados para el tratamiento de datos faltantes son: casos completos, casos disponibles, imputación de valores únicos, y más recientemente,

[2] Ref. [3], página 13.
[3] Ref. [2], página 1.

los métodos basados en modelos, tales como el de máxima probabilidad para datos normales multivariantes, y el método de imputación múltiple. [4]

2.1.1 Razones para la casuística de datos faltantes

Durante la recolección de datos, el investigador tiene la oportunidad de observar las posibles razones para los datos faltantes, evidencia que le ayudará a tomar la decisión de cual es el método de tratamiento de datos faltantes más apropiado para el análisis.[5]

Para ilustrar la casuística de datos faltantes, se utilizarán los datos presentados en la tabla que se muestra en la figura siguiente:

Datos sobre selección de empleados		
IQ	Bienestar psicológico	Performance en el trabajo
78	13	-
84	9	-
84	10	-
85	10	-
87	-	-
91	3	-
92	12	-
94	3	-
94	13	-
96	-	-
99	6	7
105	12	10
105	14	11
106	10	15
108	-	10
112	10	10
113	14	12
115	14	14
118	12	16
134	11	12

Ilustración 1: Datos sobre selección de empleados

Esta tabla se ha diseñado para imitar un escenario en el cual se realiza una selección de candidatos a un puesto de trabajo. Durante la entrevista, los candidatos completan un test de IQ (IQ test: Intelligence Quotient test: test de cociente de inteligencia), y también un cuestionario de bienestar psicológico. A continuación, la compañía contrata a los candidatos que tuvieron un score (“puntuación”) que se encuentre en la mitad superior de la distribución de los datos, y un supervisor califica su performance (“rendimiento” o “desempeño”) en el trabajo a posteriori de un periodo de prueba de 6 meses.
Nótese que los scores de performance en el trabajo están sistematicamente faltantes en función de los scores de IQ (i.e., los individuos en la mitad inferior de la distribución no fueron contratados, con lo cual no tienen calificación de performance). Además, para

[4] Ref. [2], página 2.
[5] Ref. [2], página 3.

mejorar la simulación, tres scores de bienestar psicológico han sido eliminados (aleatoriamente), para simular la situación en la cual 3 cuestionarios de candidatos fueron extraviados.

2.1.2 Patrones de datos faltantes

Como punto de partida, es útil distinguir entre patrones de datos faltantes y mecanismos de datos faltantes. Estos términos tienen realmente distintos significados, pero los investigadores muchas veces los utilizan indistintamente. Un patrón de datos faltantes se refiere a la configuración de los datos observados y faltantes en un conjunto de datos (de aquí en adelante, para conjunto de datos se utilizará el término en inglés "data set", para utilizar el término que se utiliza más comunmente, incluso en estudios escritos en castellano).

En el data set, los mecanismos de datos faltantes, describen posibles relaciones entre variables medidas y la probabilidad de datos faltantes. Nótese que un patrón de datos faltantes sólo describe la localización de los "agujeros" en los datos, y no explica porqué los datos están faltantes.

Aunque los mecanismos de datos faltantes no ofrecen una explicación de la causa de datos faltantes, si representan relaciones matermáticas genéricas entre los datos observados y los ausentes (por ej., en el diseño de una encuesta, puede existir una relación sistemática entre el nivel de educación del sujeto encuestado y la propensión a producir datos faltantes). Los mecanismos de datos faltantes juegan un rol vital en la teoría de datos faltantes elaborada por Donald Rubin y Roderick Little. [6]

[6] Ref. [17], página 2.

2.1.3 Patrones típicos de datos faltantes

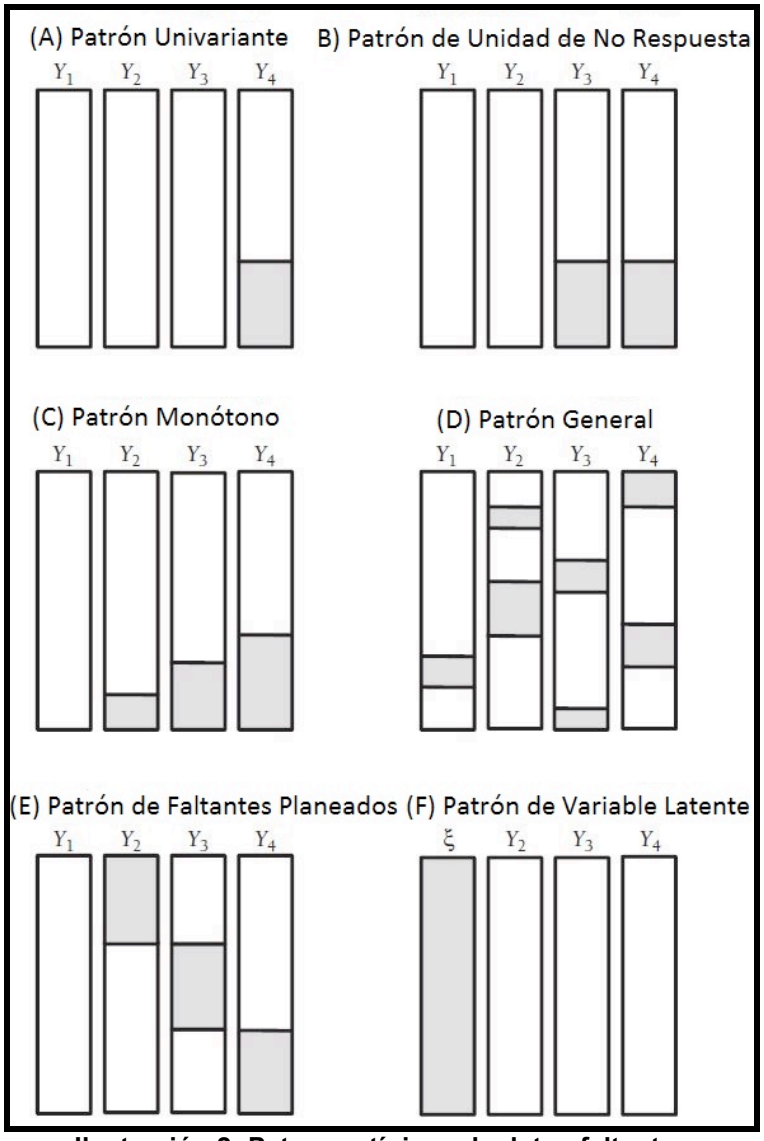

Ilustración 2: Patrones típicos de datos faltantes

La ilustración 7 muestra 6 tipos de patrones típicos de datos faltantes que pueden encontrarse en las investigaciones con datos faltantes, con áreas sombreadas representando la localización de los valores faltantes en el data set con 4 variables. El patrón univariante en el panel A tiene valores faltantes aislados pertenecientes a una única variable.

Un patrón univariante es relativamente raro en algunas disciplinas pero puede surgir en estudios experimentales. Por ejemplo, supongase que Y1 hasta Y3 son variables manipuladas (por ej, factores entre sujetos en un diseño ANOVA) e Y4 es una variable de salida incompleta. El patrón univariante de datos faltantes es uno de los problemas primarios en recibir atención en la literatura estadística, y una cantidad importante de artículos, ya considerados clásicos, están dedicados a este asunto.
El panel B muestra una configuración de valores faltantes conocida como patrón de unidad de "no respuesta" (non-response en el término original en inglés).[7]

[7] Ref. [17], página 3.

Este patrón ocurre a menudo en investigaciones que incluyen encuestas, donde Y1 e Y2 son características que están disponibles para cada miembro del marco de muestreo (por ej., datos de un censo), e Y3 e Y4 son encuestas que algunos sujetos se han negado a responder.

Un patrón monótono de datos faltantes en el panel C está típicamente asociado con un estudio longitudinal en el cual los participantes se retiran del estudio y nunca vuelven. La literatura algunas veces se refiere a esto como "deserción" (attrition en inglés). Por ejemplo, considérese un estudio clínico para un nuevo medicamento en el cual los participantes se retiran del mismo porque están teniendo reacciones adversas al principio activo. Visualmente, el patrón monótono se asemeja a una escalera, con lo que los casos con datos faltantes en un nivel en particular son siempre medidas faltantes subsecuentes.

Los patrones monótonos de datos faltantes han recibido mucha atención en la literatura de datos faltantes porque reducen muchísimo la complejidad matemática de la máxima probabilidad y de la imputación múltiple y al mismo tiempo puede eliminar la necesidad para los algoritmos iterativos de estimación (Véase Schafer, 1997, páginas 218–238).

Un patrón general de datos faltantes es tal vez la configuración más común de datos faltantes. Como se ve en el panel D, un patrón general tiene datos faltantes dispersos a través de la matriz de datos de un modo azaroso. El aparente patrón azaroso es engañoso porque los valores pueden ser faltantes de modo sistemático (por ej., podría haber una relación entre los valores de Y1 y la propensión a datos faltantes en Y2). Así y todo, es importante recordar que el patrón de datos faltantes describe la localización de los valores faltantes, y no la razón por la cual están faltando.[8]

El dataset en la tabla de la ilustración 6 es otro ejemplo de un patrón general de datos faltantes, e incluso se puede separar este patrón general en 4 patrones de datos faltantes: casos con solo los IQ scores (n = 2), casos con IQ y scores de "bienestar" (well-being) (n = 8), casos con IQ y scores de performance de trabajo (n = 1), y casos con datos completos en las tres variables (n = 9).
A posteriori en este TFM, se mostrarán una serie de diseños que producen datos faltantes de manera intencional. El patrón de datos faltantes planeado en el panel E, corresponde a un cuestionario de 3 formularios diseñado por Graham, Hofer, y MacKinnon en el año 1996. La idea básica detrás del diseño de tres formularios es distribuir cuestionarios a través de distintos formularios y administrar un subconjunto de estos formularios a cada sujeto. Por ejemplo, el diseño en el panel E distribuye los 4 cuestionarios en tres formularios, tal que cada formulario incluye Y1 pero no tiene Y2, Y3, o Y4. Los patrones de datos faltantes planeados son útiles para recolectar un número grande de ítems de cuestionario, mientras simultáneamente se reduce la carga para el sujeto que responde.[9]

Finalmente, el patrón de variable latente en el panel F es únivoco al análisis de variable latente, tales como los modelos estructurales de ecuaciones. Este patrón es interesante porque los valores de las variables latentes están faltando para la muestra entera. Por ejemplo, un modelo de análisis de factor confirmatorio utiliza un factor latente para explicar las asociaciones entre un conjunto de variables manifiestas de indicador (por ej., Y1 hasta Y3), pero los scores de factor en sí mismos, están faltando completamente. Aunque no es necesario ver los modelos de variable latente como problemas de datos faltantes, los investigadores han adaptado los algoritmos de datos faltantes para estimar

[8] Ref. [17], página 4.
[9] Ref. [17], página 5.

estos modelos, por ej., en modelos multinivel (Véase Raudenbush y Bryk, 2002, páginas 440–444).

Históricamente, los investigadores han desarrollado técnicas analíticas que tratan un patrón de datos faltantes particular. Por ejemplo, Little y Rubin (2002), dedican un capítulo entero a métodos antiguos que fueron desarrollados específicamente para estudios experimentales con un patrón de datos faltante univariante. Similarmente, los investigadores especializados en encuestas han desarrollado aproximaciones de "cubierta caliente" ("hot-deck") para tratar unidades de no respuesta (Véase Scheuren, 2005).
Desde un punto de vista práctico, distinguir entre patrones de datos faltantes no es más demasiado importante porque la estimación de máxima probabilidad y la imputación múltiple se adaptan bien a virtualmente cualquier patrón de datos faltantes.[10]

2.2 Una visión general conceptual de la teoría de datos faltantes

Rubin et al., en 1976, introdujeron un sistema de clasificación para los problemas de datos faltantes que se utiliza ampliamente en la literatura actual. Ese trabajo ha generado tres mecanismos de datos faltantes que describen la probabilidad que un dato faltante se relacione con los datos. Desafortunadamente, la terminología de Rubin (estándar en la actualidad) es un poco confusa, y hay investigadores que mal utilizan la misma. [11]

A continuación, se presenta una visión general conceptual en la teoría de datos faltantes, utilizando algunos ejemplos de investigación hipotética, para ilustrar mecanismos de datos faltantes, diseñados por Rubin.
Por más que se han propuesto otros esquemas de clasificación, los fundamentales son los 3 diseñados por Rubin: MAR, MCAR, MNAR.

[10] Ref. [17], página 5.
[11] Ref. [17], página 6.

2.2.1 Datos faltantes en forma aleatoria ("Missing at Random Data: MAR")

Los datos se consideran faltantes en forma aleatoria ("MAR"), cuando la probabilidad de datos faltantes en la variable Y está relacionada o otra variable (o variables) en el módelo de análisis, pero no a los valores de Y en si misma.
Dicho de otro modo, no existe relación entre la propensión a tener datos faltantes en Y y en los valores de Y, luego de haber dejado fuera otras variables. El término "datos faltantes en forma aleatoria" es un poco confuso, ya que implica que los datos están faltando de una manera azarosa, recordando a tirar una moneda.
De todos modos, MAR significa en realidad que existe una relación sistemática entre una o más variables y la probabilidad de tener datos faltantes. Para ilustrar, considérese el dataset en la ilustración 8.[12]

Calificaciones de performance en el trabajo con valores faltantes MCAR, MAR y MNAR				
	Calificaciones de performance en el trabajo			
IQ	Completo	MCAR	MAR	MNAR
78	9	-	-	9
84	13	13	-	13
84	10	-	-	10
85	8	8	-	-
87	7	7	-	-
91	7	7	7	-
92	9	9	9	9
94	9	9	9	9
94	11	11	11	11
96	7	-	7	-
99	7	7	7	-
105	10	10	10	10
105	11	11	11	11
106	15	15	15	15
108	10	10	10	10
112	10	-	10	10
113	12	12	12	12
115	14	14	14	14
118	16	16	16	16
134	12	-	12	12

Ilustración 3: Calificaciones de performance con MCAR, MAR y MNAR

Estos datos ejemplifican un escenario en donde se seleccionan candidatos a un empleo, y los candidatos completan un test de IQ, durante su entrevista de trabajo y posteriormente un supervisor evalúa su performance luego de un periodo de prueba de 6 meses.
Supóngase que la compañía utiliza los scores de IQ como medida de selección y no contrata candidatos que obtienen un score en el cuartil más bajo de la distribución. Puede observarse que las tasas de performance en el trabajo en la columna MAR de la tabla de la ilustración 8 están faltando para los candidatos con los scores de IQ más bajos.

[12] Ref. [17], página 6.

Consecuentemente, la probabilidad de una tasa de performance faltante es sólo una función de los scores de IQ y no está relacionada a la performance del individuo.

Hay muchas situaciones de la vida real en la cual una medida de selección tal como el IQ determina si los datos están faltando, pero es fácil generar ejemplos adicionales donde la propensión a los datos faltantes es menos determinística. Por ejemplo, supongase que un investigador en eduación está estudiando el éxito en la facilidad de lectura en alumnos jóvenes, y encuentra que los estudiantes de origen hispánico tienen una tasa alta de datos faltantes, en comparación con los alumnos de origen caucásico.
Como segundo ejemplo, supongáse que un psicólogo está estudiando la calidad de vida en un grupo de pacientes de cáncer y encuentra que los pacientes mayores y pacientes con un grado educativo más bajo, tienen una propensión alta a rehusarse al cuestionario de calidad de vida. Estos ejemplos califican como MAR, siempre y cuando no exista una relación residual entre la propensión a los datos faltantes y la variable incompleta de salida (por ej., después de haber dejado fuera la edad y la educación, la probabilidad de ausencia de datos no está relacionada a la calidad de vida).

El problema práctico con el mecanismo MAR es que no hay manera de confirmar que la probabilidad de los datos faltantes en Y es sólo una función de otras variables. Volviendo al ejemplo de la educación, supóngase que niños de origen hispánico con habilidades de lectura pobres tienen altas tasas de datos faltantes en los tests de éxito en la facilidad de lectura. Esta situación es inconsistente con el mecanismo MAR, ya que no existe relación entre el éxito en la facilidad de lectura y la ausencia de datos, incluso después de haber controlado el factor étnico. De todos modos, el investigador no tendría manera de verificar la presencia o ausencia de esta relación sin saber los valores ausentes en los scores de éxito. Consecuentemente, no hay modo de probar el mecanismo MAR o de verificar que los scores son MAR. Esto representa un problema práctico importante para los análisis de datos faltantes, ya que la estimación por máxima probabilidad y la imputación múltiple (dos técnicas usualmente recomendadas) asumen un mecanismo MAR.[13]

2.2.2 Datos faltantes en forma completamente aleatoria ("Missing Completely at Random Data: MCAR")

El mecanismo de datos faltantes en forma completamente aleatoria (MCAR) es lo que los investigadores toman como datos ausentes en forma puramente azarosa. La definición formal de MCAR requiere que la probabilidad de los datos faltantes en la variable Y no esté relacionada con otras variables y que no esté relacionada con los valores de Y en si misma. Dicho de otro modo, los puntos de datos observados son una muestra simple aleatoria de los scores que serían analizados en el casos que los datos estuvieran completos. Nótese que el mecanismo MCAR determina una condición más restrictiva que el mecanismo MAR, porque asume que la ausencia de datos no está relacionada en absoluto con los datos.[14]

Con respecto a la performance en el trabajo de la tabla de la ilustración 8, se ha creado una columna MCAR borrando scores basados en el valor de un número aleatorio. Los números aleatorios no estaban correlados con el IQ y con la performance en el trabajo, por lo tanto la ausencia de datos no está relacionada con los datos. Puede verse que los valores faltantes no están aislados a una localización particular en las distribuciones del

[13] Ref. [17], página 6.
[14] Ref. [17], página 7.

IQ y de la performance en el trabajo; con lo cual los 15 casos completos son relativamente representativos del grupo entero de candidatos.
Es fácil pensar en situaciones del mundo real donde las tasas de performance en el trabajo podrían estar ausentes de un modo azaroso.[15]

Por ejemplo, una empleada podría coger una baja por maternidad previamente a su evaluación de 6 meses, y el supervisor responsable de asignar la tasa de performance, podría ser ascendido a otra división dentro de la compañía, o bien un empleado/a podría renunciar porque su esposo/a ha aceptado un trabajo en otra lugar del país. Volviendo al ejemplo previo sobre educación juvenil, nótese que los jóvenes podrían tener scores de éxito de tipo MCAR, a causa de eventos personales inesperados (por ej., una enfermedad, un funeral, vacaciones familiares, cambio a otro centro educativo), dificultades de calendario (por ej., la clase estaba fuera realizando un trabajo de campo, cuando los investigadores visitaron el centro educativo), o bien caos administrativos (por ej., los investigadores inadvertidamente extraviaron los tests antes que los datos pudieran ser ingresados). Tipos similares de problemas podrían producir datos MCAR en la calidad de un estudio en ciencias de la vida.

En principio, es posible verificar que un grupo de scores sean MCAR. Por ejemplo, reconsidérense los datos en la tabla de la ilustración 8. La definición de MCAR requiere que los datos observados sean una muestra simple aleatoria del dataset hipotéticamente completo. Esto implica que los casos con tasas observadas de performance en el trabajo, no sean diferentes de los casos en los cuales faltan las evaluaciones de performance, en promedio. Para probar esta idea, se pueden separar los casos faltantes y los casos completos, y luego examinar las diferencia de media grupales en la variable IQ.
Si los patrones de datos faltantes son aleatoriamente equivalentes (i.e., los datos son MCAR), entonces las medias de IQ deberían ser las mismas, dentro del error natural de muestreo.

Para ilustrar este asunto, se han clasificado los scores en la columna MCAR tanto en observados como en faltantes, y comparado las medias de IQ para los dos grupos. Los casos completos tienen una media de IQ de 99.73, y los casos faltantes tienen una media de 100.80.

Esta pequeña diferencia en la media sugiere que los dos grupos son equivalentes aleatoriamente, y provee evidencia que los scores de performance en el trabajo son MCAR. Como contraste, se han utilizado las tasas de performance en la columna MAR para formar los grupos de datos faltantes. Los casos completos ahora tienen una media de IQ de 105.47, y los casos faltantes tienen una media de 83.60. Esta disparidad grande, sugiere que los dos grupos son sistemáticamente diferentes en la variable IQ, con lo cual hay evidencia en contra del mecanismo MCAR. Comparar los casos faltantes y los casos completos es una estrategia que es común en los tests MCAR.[16]

2.2.3 Datos faltantes en forma no aleatoria ("Missing Not at Random Data: MNAR")

Finalmente, los datos son faltantes en forma no aleatoria (MNAR), cuando la probabilidad de datos faltantes en una variable Y está relacionada a los valores de Y en si misma, incluso después de haber controlado otras variables.

[15] Ref. [17], página 7.
[16] Ref. [17], página 8.

Como ejemplo, reconsidérese los datos de performance en el trabajo de la tabla de la ilustración 8. Supóngase que la compañia contrató a los 20 candidatos y a continuación despidió a un número de individuos por baja performance previamente a las evaluaciones a los 6 meses. Puede observarse que las tasas de performance en el trabajo en la columna MNAR están faltando para los candidatos con las tasas de performance más bajas. Consecuentemente, la probabilidad de una tasa de performance faltante depende de la performance en el trabajo de la persona, luego de haber controlado la variable IQ.

Es relativamente simple generar ejemplos adicionales donde los datos MAN podrían ocurrir. Volviendo al ejemplo previo sobre educación, supóngase que los estudiantes con habilidad de lectura pobres tienen scores de tests faltantes porque experimentaron dificultades de comprensión de lectura durante el exámen. Similarmente, supóngase que un número de pacientes en el ensayo clínico del cáncer se ponen tan enfermos (por ej., su calidad de vida se vuelve muy pobre), que abandonan y no participan más del estudio. En los dos ejemplos, los datos son MNAR porque la probabilidad de un valor faltante depende de la variable que está faltando. Como en el mecanismo MAR, no hay manera de verificar que los scores son MNAR sin saber los valores de las variables faltantes.[17]

[17] Ref. [17], página 8.

3. Datasets que se abordan en este TFM. Descripción de la estructura de datos de los mismos. Aplicación de distintas estrategias de tratamiento de datos faltantes y métodos de análisis.

3.1 Datasets que se abordan en este TFM: repositorios y paquetes de los cuales se obtienen los datos

Los repositorios de los cuales se obtendrán los datasets (al momento de la realización de este TFM) son:

a) “KEEL-dataset repository”[18], el cual se encuentra disponible en la siguiente URL:
http://sci2s.ugr.es/keel/datasets.php

Los datasets con valores faltantes (“Experimental studies with data sets with missing values”) que se pueden descargar de este servidor se encuentran en la siguiente URL:

http://sci2s.ugr.es/keel/study.php?cod=28

b) "UC Irvine Machine Learning Repository", disponible en la URL:
https://archive.ics.uci.edu/ml/

c) "Rdatasets Repository", disponible en la URL:
https://vincentarelbundock.github.io/Rdatasets/datasets.html

Además, algunos datasets a utilizarse en este TFM serán recogidos de los paquetes de R "DATASETS", "MICE" y "MASS", ya que estos datasets que se encuentran incluidos como ejemplos en los mencionados paquetes, son de fácil acceso, ya que se cargan en memoria por defecto al cargar el paquete en el entorno de R.

De los repositorios y paquetes mencionados, los datasets con valores faltantes a utilizarse en este TFM son:

- Horse-colic
- Air Quality
- Boys
- Walking
- Nhanes
- Pattern4

Si fuera necesario los paquetes de R pueden descargarse de las siguientes URLs:

MICE: Multivariate Imputation by Chained Equations:
https://cran.r-project.org/web/packages/mice/index.html

[18] Ref. [18].

DPLYR: A Grammar of Data Manipulation
https://cran.r-project.org/web/packages/dplyr/index.html

VIM: Visualization and Imputation of Missing Values
https://cran.r-project.org/web/packages/VIM/index.html

LATTICE: Trellis Graphics for R
https://cran.r-project.org/web/packages/lattice/index.html

MASS: Support Functions and Datasets for Venables and Ripley's MASS
https://cran.r-project.org/web/packages/MASS/index.html

Cabe comentar que el paquete "DATASETS" se instala en R por defecto, por lo tanto se puede acceder al mismo ejecutando simplemente:

data(datasets)

3.1.1 Estructura de datos básica de los datasets a utilizarse

A continuación, una tabla con una descripción de la estructura de datos (básica) contenida en los datasets mencionados previamente[19]:

Dataset	Abreviatura	Registros	Variables	% VF	% Registros con VF
Horse-colic	HOC	300	29	21.82	98.1
Boys	BOYS	748	9	24.08	70.16
Walking	WALK	434	17	13.62	67.41
Air Quality	AIR	153	6	6.62	26.33
Nhanes	NHAN	25	4	21.6	48.0
Pattern4	PATT	8	3	33.3	75.0

Ilustración 4: Estructura de datos básica de los datasets a utilizarse

Con fines de practicidad, los datasets a utilizarse se mencionarán de ahora en adelante por la abreviatura mencionada en la tabla previa.
Además, en la tabla anterior, la columna "Registros" se refiere a la cantidad de registros (cantidad de filas) del dataset. Luego, la columna "Variables" se refiere a la cantidad de variables; % VF se refiere al porcentaje de valores faltantes en el dataset; y por último, la columna % Registros con VF, significa el porcentaje de registros con valores faltantes en el dataset.

[19] Ref. [19], página 6.

3.2 Estrategias para el tratamiento de datos faltantes

Para esbozar y explicar las distintas estrategias para el tratamiento de datos faltantes, se expondrán las instrucciones pertinentes en lenguaje R, utilizando para ello los datasets mencionados en el apartado anterior.

3.2.1 Visión práctica del problema[20]

Supongamos que se desea calcular la media de los números 4, 6 y 7. Utilizando lenguaje R sería:

```
> y <- c(4, 6, 7)
> mean(y)
[1] 5.666667
```

Ahora supongamos que el último valor se encuentra faltante. En R, se identifica a los valores faltantes con el símbolo NA (del inglés "Not Available": "No disponible").

```
> y <- c(4, 6, NA)
> mean(y)
[1] NA
```

R no devuelve ningún resultado (devuelve "NA"). Es posible agregar un argumento extra: na.rm = TRUE, en la llamada a la función. Lo cual significa que eliminará cualquier dato faltante, y calculará en este caso la media con los valores que se encuentren disponibles:

```
> y <- c(4, 6, NA)
> mean(y, na.rm=TRUE)
[1] 5
```

El resultado obtenido es 5, ya que 4+6=10, y luego 10/2=5. Con lo cual se puede observar que efectúa el cociente por la cantidad de valores disponibles, en este caso 2 valores.

¿Qué sucede cuando se desea realizar por ejemplo una regresión lineal, y en la variable predictora se tienen valores faltantes?

Utilizando el dataset HOC (véase apartado 3.1.1):

```
> fit<-lm(pulse ~ type_of_lesion_1, data=HOC)
Error in lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :  NA/NaN/Inf in 'y'
```

El código anterior intenta calcular una regresión lineal para predecir el pulso (latidos del corazón por minuto) de acuerdo al tipo de lesión que tenga el caballo.

No se obtiene un resultado, ya que, existen valores faltantes en la variable "pulse".

[20] Ref. [15], página 3.

3.2.2 Listwise Deletion[21]

Se pueden omitir completamente los registros u observaciones que tengan valores faltantes si se especifica como argumento na.action = na.omit para la llamada a lm():

```
> fit <- lm(pulse ~ type_of_lesion_1, data=HOC, na.action = na.omit)
> coef(fit)
   (Intercept)      type_of_lesion_1
  7.037461e+01     4.271436e-04
```

El objeto fit guarda los resultados del análisis de regresión, y la función coef muestra los valores para los coeficientes obtenidos. Como no sería práctico poner el argumento na.action = na.omit cada vez que se desee ejecutar una regresión lineal, vale aclarar que se pueden cambiar las opciones por defecto de R, mediante el comando:

```
> options(na.action = na.omit)
```

Al ejecutar el mencionado comando, no se volverán a obtener mensajes de error, ya que por defecto ahora quedará establecido que se omitirán las observaciones que tengan valores faltantes.

El procedimiento anterior (eliminar completamente registros cuando contienen datos faltantes) se conoce como "listwise deletion" o "complete case analysis", y es uno de los procedimientos más utilizados.

De todos modos, este procedimiento puede traer algunos problemas en la interpretación. Es importante tener en cuenta cuantas observaciones han sido eliminadas, lo cual puede hacerse del siguiente modo:

```
> eliminados <- na.action(fit)
> naprint(eliminados)
[1] "24 observations deleted due to missingness"
```

Con lo cual queda claro que fueron eliminados 24 registros al realizar el ajuste del modelo de regresión lineal.

Si se agregan más variables predictoras al modelo de regresión lineal, y en las nuevas variables también hay valores faltantes, se eliminarán incluso más registros del modelo de regresión lineal:

```
> fit2 <- lm(pulse ~ type_of_lesion_1 + abdominal_distension, data=HOC)
> naprint(na.action(fit2))
[1] "71 observations deleted due to missingness"
```

Con lo cual, cabe observar que al cambiar el modelo, se ha alterado la muestra con la cual opera el modelo.

Los coeficientes obtenidos ahora son distintos:

[21] Ref. [15], página 8.

> coef(fit2)

(Intercept)	type_of_lesion_1	abdominal_distension
4.690759e+01	2.045349e-04	1.094311e+01

Hay algunos problemas estadísticos y metodológicos asociados a este procedimiento. Por ejemplo, algunas cuestiones pueden entrar en juego:

a) ¿Son comparables los coeficientes de regresión de los dos modelos?
b) ¿Las diferencias en los coeficientes se atribuyen al cambio de modelo o al cambio de la muestra con la que se opera?
c) ¿Los coeficientes estimados se pueden generalizar a la población?
d) ¿Se tiene la cantidad de registros suficientes para detectar efectos?

Sin embargo, como se ha comentado, el procedimiento estándar es eliminar los registros que contienen datos faltantes. Es lo que se hace comunmente en los estudios científicos.

Lo cual determina las siguientes casuísticas:

1) La presencia de datos faltantes la mayoría de las veces no se menciona en los papers de los estudios científicos.
2) Procedimientos como el mencionado previamente (listwise deletion) se utilizan y tampoco se menciona que han sido utilizados.
3) Diferentes tablas con distintos resultados obtenidos provienen de muestras distintas.
4) Algunos métodos de tratamiento de datos faltantes, tales como direct likelihood (probabilidad directa), full information maximum likelihood (probabilidad directa de información completa) o el tratamiento mediante imputación múltiple, son escasamente utilizados por la comunidad científica.

3.2.3 Pairwise Deletion[22]

El método de "pairwise deletion", también llamado "available case analisys", intenta mejorar de algún modo los problemas que se producen utilizando el método listwise deletion.

Este método calcula las medias y las covarianzas de todos los datos observados. Por lo tanto, la media de la variable V1 se basa en todos los casos que se observan en esa variable, luego la media de la variable V2 sigue el mismo procedimiento, y así para todas las variables.

Para la correlación y la covarianza, se toman todos los datos en los cuales V1 y V2 no tienen scores faltantes. Así, la matriz de sumario de estadísticos, se carga en un programa para el análisis de regresión, análisis de factores o otros procedimientos de modelado.

Se puede calcular la media, las covarianzas y las correlaciones del dataset HOC utilizando pairwise deletion del siguiente modo:

```
colMeans(HOC, na.rm = TRUE)
cor(HOC, use = "pair")
cov(HOC, use = "pair")
```

Nota: para los resultados obtenidos por estos comandos, véase el Anexo 2: "Anexo de prácticas en R para este TFM".

El método listwise deletion es simple, utiliza toda la información disponible y produce estimaciones consistentes de la media, las correlaciones y las covarianzas bajo el supuesto que los datos faltantes se han producido mediante el mecanismo MCAR, en donde se considera que los datos ausentes se han producido en forma puramante azarosa (Véase apartado 2.2.2).

De todos modos, cabe comentar que las estimaciones podrían resultar sesgadas, si los datos faltantes no se han producido mediante el mecanismo MCAR.

La idea subyacente detrás del método listwise deletion, es utilizar toda la información disponible. Aunque la idea es buena, en un principio, un análisis de la matriz generada requerirá algunas técnicas de optimización avanzada y algunas fórmulas especiales para calcular los errores estándar. Con ello, se pierde la pretendida simplicidad de este método.

[22] Ref. [15], página 9.

3.2.4 Imputación de la media[23]

La imputación de la media es un método rápido en el cual se reemplazan los valores faltantes por la media de la variable en la cual se encuentren estos valores. Supóngase que se desea imputar la media en las variables "pulse" y "abdominal_distension" del dataset HOC.

Para ello, puede utilizarse el paquete MICE (Multivariate Imputation by Chained Equations), el cual habrá que instalar y cargar en el entorno de R mediante las siguientes instrucciones:

```
> install.packages("mice", dependencies = TRUE)
> library("mice")
```

Antes de imputar la media, puede visualizarse el patrón de datos faltantes del dataset mediante la instrucción "md. pattern", también disponible en el paquete MICE:

```
> md.pattern(HOC)
```

Habida cuenta que el dataset HOC tiene 300 observaciones y 29 variables, a efectos explicativos, utilizaré el paquete DPLYR para seleccionar sólo algunas observaciones y algunas variables, y luego poder visualizar mejor el funcionamiento del comando md.pattern:

```
> library("dplyr")
> subsample <- slice(HOC,1:10) %>% select(pulse, abdominal_distension, pain ,
nasogastric_reflux)
```

Previamente, mediante el paquete DPLYR, he realizado una selección de las 10 primeras observaciones y 4 variables que contienen valores faltantes. Esa selección la he guardado en un dataframe llamado "subsample". Para ver el contenido del dataframe "subsample" puede ejecutarse:

```
> View(subsample)
```

	pulse	abdominal_distension	pain	nasogastric_reflux
1	66	4	5	NA
2	88	2	3	NA
3	40	1	3	NA
4	164	4	2	2
5	104	NA	NA	NA
6	NA	2	2	1
7	48	3	3	1
8	60	2	NA	1
9	80	4	4	1
10	90	1	5	1

Ilustración 5: Dataframe "subsample"

[23] Ref. [15], página 10.

Ahora, con una muestra más pequeña de los datos, ejecuto el comando md.pattern, guardando su resultado en un dataframe llamado "pattern".

> pattern <- md.pattern(subsample)

Al visualizar el contenido del dataframe pattern, se obtiene:

	row.names	pulse	abdominal_distension	pain	nasogastric_reflux	V5
1	4	1	1	1	1	0
2	1	0	1	1	1	1
3	1	1	1	0	1	1
4	3	1	1	1	0	1
5	1	1	0	0	0	3
6		1	1	2	4	8

Ilustración 6: Dataframe "pattern" (patrón de datos faltantes en dataset "subsample")

La matriz de la figura anterior, presenta un resumen de los valores observados y los valores faltantes en las observaciones y en las variables.

Un valor 1 en la matriz presentada, significa que un valor ha sido observado, mientras que un valor 0 significa que ha habido un valor faltante.

El resumen determina que en el dataframe subsample:

1) En 4 filas sólo han habido valores presentes en las 4 variables.
2) En 1 fila ha habido un valor faltante para la variable pulse.
3) En 1 fila ha habido un valor faltante para la variable pain.
4) En 3 filas han habido 3 valores faltantes para la variable nasogastric_reflux.
5) En 1 fila ha habido un valor faltante en las variables abdominal_distension, pain y nasogastric_reflux.
6) En la última fila se presenta un resumen de las varialbles, lo que viene a indicar que en pulse ha habido un valor faltante, en abdominal_distension ha habido un valor faltante, en pain han habido 2 valores faltantes y en nasogastric_reflux han habido 4 valores faltantes.
7) En la última columna se presenta un resumen de las observaciones, con la misma información que la comentada en los puntos anteriores 1, 2, 3, 4 y 5.
8) Por último, puede observarse que el valor 8 de la última fila y última columna, indica que en el dataframe subsample han habido en total 8 valores faltantes.

Otro modo interesante de visualizar el patrón de datos faltantes es utilizar el paquete VIM ("Visualization and Imputation of Missing Values") (Véase el anexo 2 para el código fuente que realiza esta graficación):

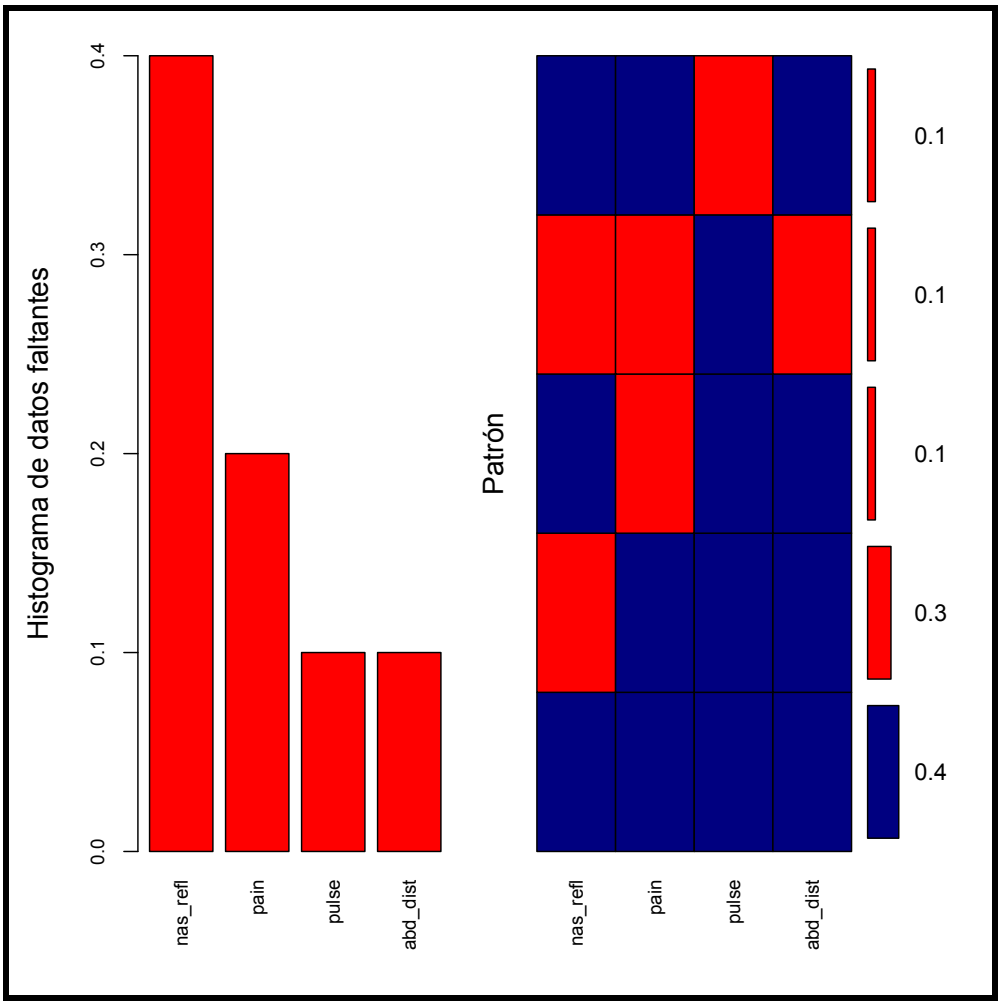

Ilustración 7: Gráfico con patrón de datos faltantes en dataset "subsample"

Para este TFM, se utilizará el estándar de colores de Abayomi (Abayomi et al., 2008). El color azul indica los valores realmente observados de los datos, el color rojo muestra los valores "imputados" (i.e., a imputar) de los datos (independientemente del método de imputación que se utilice), mientras que el color negro indica los valores imputados efectivamente en el conjunto de datos ("datos completados").

Así, un gráfico como el anterior permite descubrir rápidamente cuales variables en el dataset tienen mayor cantidad de datos faltantes (barras rojas del histograma en la figura de la izquierda, indicando el porcentaje de valores faltantes en cada una de las variables).

Además, en la figura de la derecha, se pueden advertir claramente las proporciones de valores faltantes en relación a las proporciones de datos observados (en color azul) en el conjunto de datos.

Por ejemplo:
a) el 40 % de los datos en las 4 variables no tienen valores faltantes, sólo tienen valores observados.

b) En la variable nas_refl existen un 30 % de valores faltantes (en las otras 3 variables el 30 % de valores se corresponden a valores observados).
c) Luego se advierte que en la variable pain hay un 20 % de valores faltantes, y 10 % de valores faltantes para las variables nas_refl, pulse y abd_dist.

Realizando las sumas correspondientes queda claro que la figura de la derecha del gráfico anterior expone los siguientes porcentajes:

Variables	Valores observados	Valores faltantes
nas_refl	**60 %**	**40 %**
pain	**80 %**	**20 %**
pulse	**90 %**	**10 %**
abd_dist	**90 %**	**10 %**

Todavía incluso, en la figura de la derecha, puede obervarse si puede llegar a existir algún patrón de co-ocurrencia entre variables para los datos faltantes, cuando coinciden los cuadros rojos en una misma fila.

Si sucede que los valores faltantes de dos o más variables tienden a aparecer en forma conjunta entonces puede sospecharse que existe alguna causa subyacente que provoque esta ausencia simultánea, con lo cual podemos encontrarnos frente a un mecanismo MAR y no MCAR (véanse los apartados 2.2.1 hasta 2.2.3).

Volviendo a la imputación de la media, para imputar la media mediante el paquete MICE, en el dataframe HOC, pueden ejecutarse las siguientes instrucciones:

> HOC_mean_imp <- mice(HOC, method="mean", m=1, maxit=1)

iter imp variable
1 1 surgery rectal_temp pulse
iter imp variable
1 1 surgery rectal_temp pulse resp_rate extremities_temp peripheral_pulse mucous_membranes capillary_refill_time pain peristalsis abdominal_distension nasogastric_tube nasogastric_reflux nasogastric_reflux_PH rectal_examination_feces abdomen capillary_refill_time pain peristalsis abdominal_distension nasogastric_tube nasogastric_reflux nasogastric_reflux_PH rectal_examination_feces abdomen packed_cell_volume total_protein abdominocentesis_appearance packed_cell_volume total_protein abdominocentesis_appearance abdomcentesis_total_protein outcome outcome

El argumento method="mean" es el que especifica que se realice la imputación mediante la media, mientras que los argumentos m=1 y maxit=1, indican que se imputará a un dataset, y que lo hará en una iteración, respectivamente.

Para una mejor comprensión del procedimiento, procederé a imputar la media en un dataframe más pequeño, de 200 observaciones, y con sólo 2 variables: pulso y resp_rate (tasa de respiración), al que llamaré subsample2:

> subsample2 <- slice(HOC,1:200) %>% select(pulse, resp_rate)

Antes de proceder a la imputación de la media en subsample2, visualizo mediante VIM el patrón de datos faltantes para este dataframe:

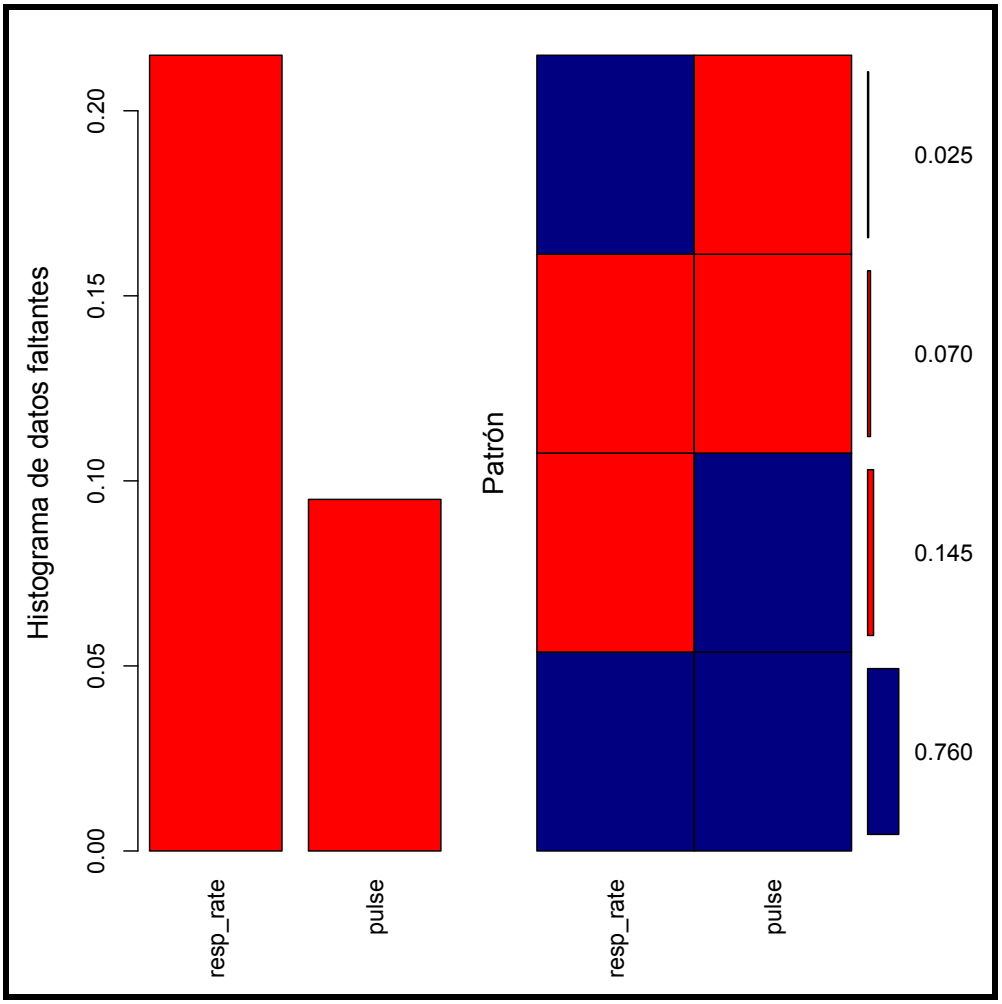

Ilustración 8: Gráfico con patrón de datos faltantes en dataset "subsample2"

Obsérvese como la variable resp_rate tiene el doble de datos faltantes que la variable pulse (figura izquierda del gráfico anterior). Por otro lado no se observa co-ocurrencia entre estas dos variables, como se puede apreciar en la figura de la derecha.

Ahora procedo a imputar la media al dataset subsample2:

```
> imp <- mice(subsample2, method="mean", m=1, maxit=1)

 iter imp variable
  1   1  pulse  resp_rate
```

Para graficar la imputación realizada, puede operarse con el paquete LATTICE, el cual permite observar, mediante la parametrización adecuada, la distribución resultante de la imputación (Véase el anexo 2 para el código fuente que realiza la graficación).

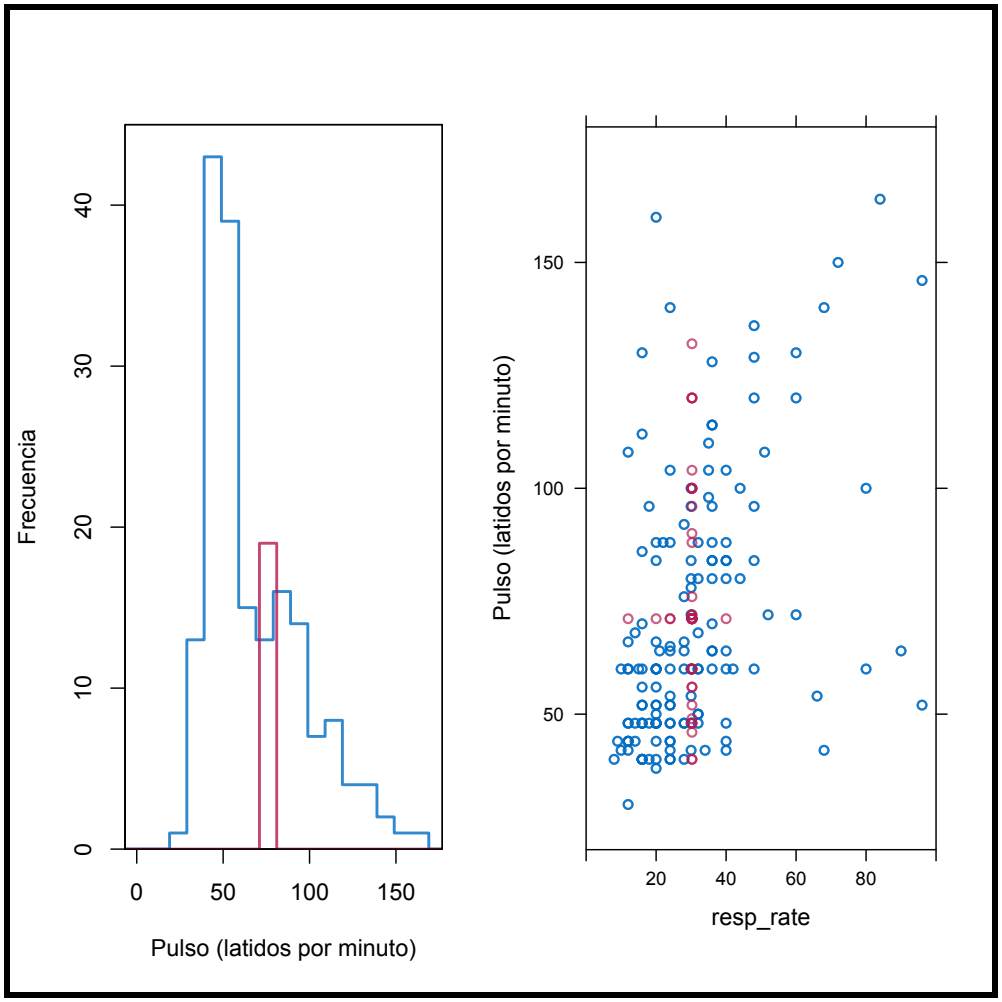

Ilustración 9: Imputación de la media en el dataset "subsample2"

La imputación de la media puede distorsionar en cierto sentido la distribución resultante.

En la ilustración anterior, puede observarse como el color azul corresponde a los valores observados realmente, mientras que el color rojo se corresponde con los valores imputados.

La figura de la izquierda muestra entonces la media imputada dentro del histograma de los datos reales, mientras que la figura de la derecha muestra la nube de puntos con los puntos rojos determinando la imputación de la media. Obsérvese como se genera una "cruz" de puntos rojos.

Como puede apreciarse, la imputación de la media es una forma rápida y simple para resolver el problema de los datos faltantes.

Así y todo, no tiene en cuenta la varianza, corrompe en cierto modo las relaciones entre variables, y sesga la estimación de la media cuando los datos faltantes no se han generado mediante el mecanismo MCAR.

Concluyendo, la imputación de la media debería sólo utilizarse como remedio rápido cuando sólo unos pocos valores están faltando, con lo cual debería evitarse en general este procedimiento.

3.2.5 Imputación mediante regresión[24]

La imputación mediante regresión, incorpora el conocimiento que se tiene de las otras variables con la idea de producir imputaciones más inteligentes.

El primer paso consiste en construir un modelo de los datos observados. Luego las predicciones para los casos incompletos se calculan de acuerdo al modelo ajustado, y se utilizan para reemplazar los datos faltantes.

Supóngase que se desea modelar el pulso mediante la función de regresión lineal de resp_rate, utilizando para ello el dataframe subsample2:

```
> fit <- lm(pulse ~ resp_rate, data=subsample2)
> pred <- predict(fit, newdata=ic(subsample2))
```

Obsérvense las gráficas resultantes de la imputación realizada:

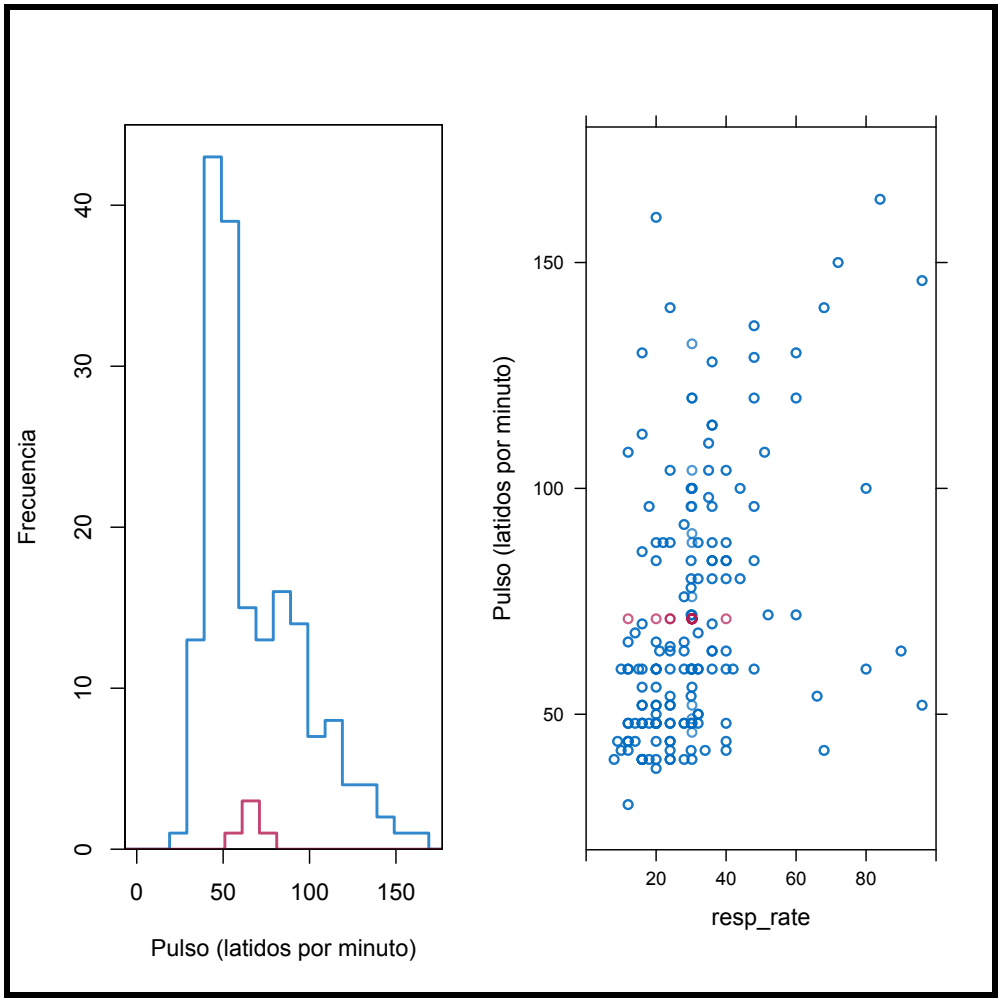

Ilustración 10: Imputación mediante regresión en el dataset "subsample2"

En la ilustración anterior, puede observarse como los valores imputados corresponden a los valores más probables de acuerdo al modelo de regresión generado.

[24] Ref. [15], página 11.

De todos modos, puede apreciarse que los valores imputados varían menos que los valores observados. Sería bastante improbable que los valores no observados de la variable pulso tuvieran esa distribución.
Además, los valores predichos e imputados, podrían tener un efecto en la correlación. Los puntos rojos tienen una correlación de 1, ya que se encuentran sobre una línea. Está claro que si los puntos rojos se combinaran con los azules, entonces crecería la correlación.

Por tanto, la imputación mediante regresión genera estimaciones sesgadas de las medias bajo el mecanismo MCAR, del mismo modo que la imputación de la media, y también sesga los coeficientes de regresión. Además, la variabilidad de los datos imputados se estima a la baja sistemáticamente. El grado de estimación a la baja depende de la varianza explicada y de la proporción de datos faltantes.

La única manera en la cual los valores predichos e imputados puedan dar imputaciones realistas es si las predicciones estuvieran cerca de la perfección. Si así fuera, el método lo que haría es reconstruir las partes faltantes dentro de los datos disponibles. De todos modos, es improbable encontrar este tipo de casuísticas, en la mayoría de los casos.

3.2.6 Imputación mediante regresión estocástica[25]

La imputación mediante regresión estocástica, es un tipo de imputación de regresión más sofisticada, ya que agrega ruido a las predicciones. Esto permitirá que caiga la correlación.

Si se procede a imputar pulse mediante regresión estocástica, utilizando el dataframe subsample2 y el paquete MICE, puede realizarse así:

```
> imp <- mice(subsample2[,1:2],method="norm.nob",m=1,maxit=1,seed=1)

 iter imp variable
  1   1  pulse  resp_rate
```

Obsérvense las gráficas resultantes de la imputación realizada:

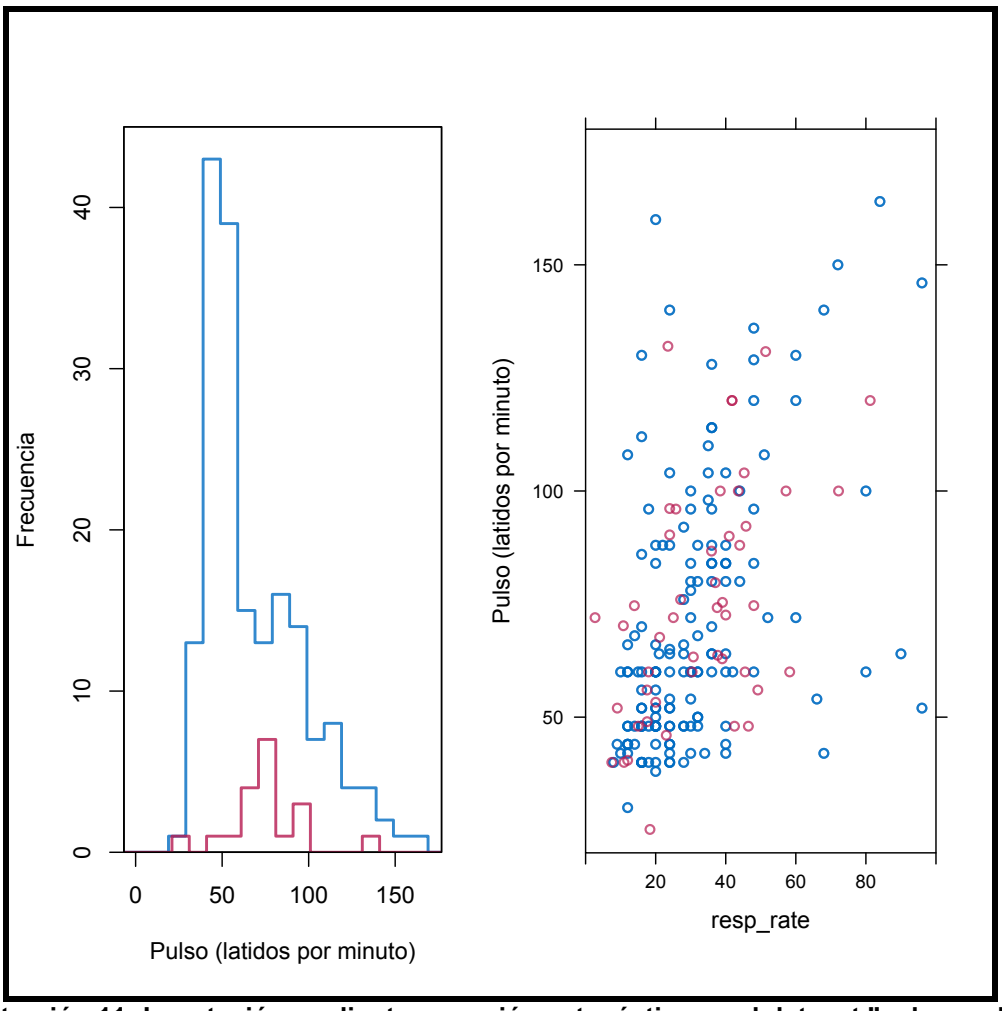

Ilustración 11: Imputación mediante regresión estocástica en el dataset "subsample2"

En la ilustración anterior, puede observarse como los valores imputados corresponden a los valores más probables de acuerdo al modelo de regresión estocástica, pero con más variabilidad, en comparación con el modelo de regresión.

[25] Ref. [15], página 13.

El argumento method="norm.nob", determina un modelo de regresión estocástica no Bayesiano. Este método primero estima el intercepto, la pendiente y la varianza residual bajo el modelo lineal, y a continuación genera los valores de imputación de acuerdo a esa especificación.

El agregado de ruido, agregó más variabilidad, como se comentó, y abrió la distribución de los valores imputados, lo cual era deseable.

La regresión estocástica es un importante paso adelante. Ya que preserva los coeficientes de correlación, pero también la correlación entre variables.

La regresión estocástica no resuelve todos los problemas, pero la idea principal que es determinar las imputaciones a partir de los residuos es muy potente, y es la base para otros métodos de imputación más avanzados.

3.2.7 Imputación mediante LOCF y BOFC[26]

Los métodos de imputación mediante LOCF (last observation carried forward) y BOFC (baseline observation carried forward) requieren datos longitudinales, i.e., datos que se desarrollan a lo largo de un periodo de tiempo.

La idea es utilizar la última observación anterior al valor faltante como reemplazo para los valores faltantes en cada una de las variables.

Utilizando por ejemplo la imputación mediante LOCF, y utilizando sólo 100 observaciones del dataframe subsample2, para la variable pulse se obtiene el siguiente gráfico (para el código fuente véase el Anexo 2 de este TFM):

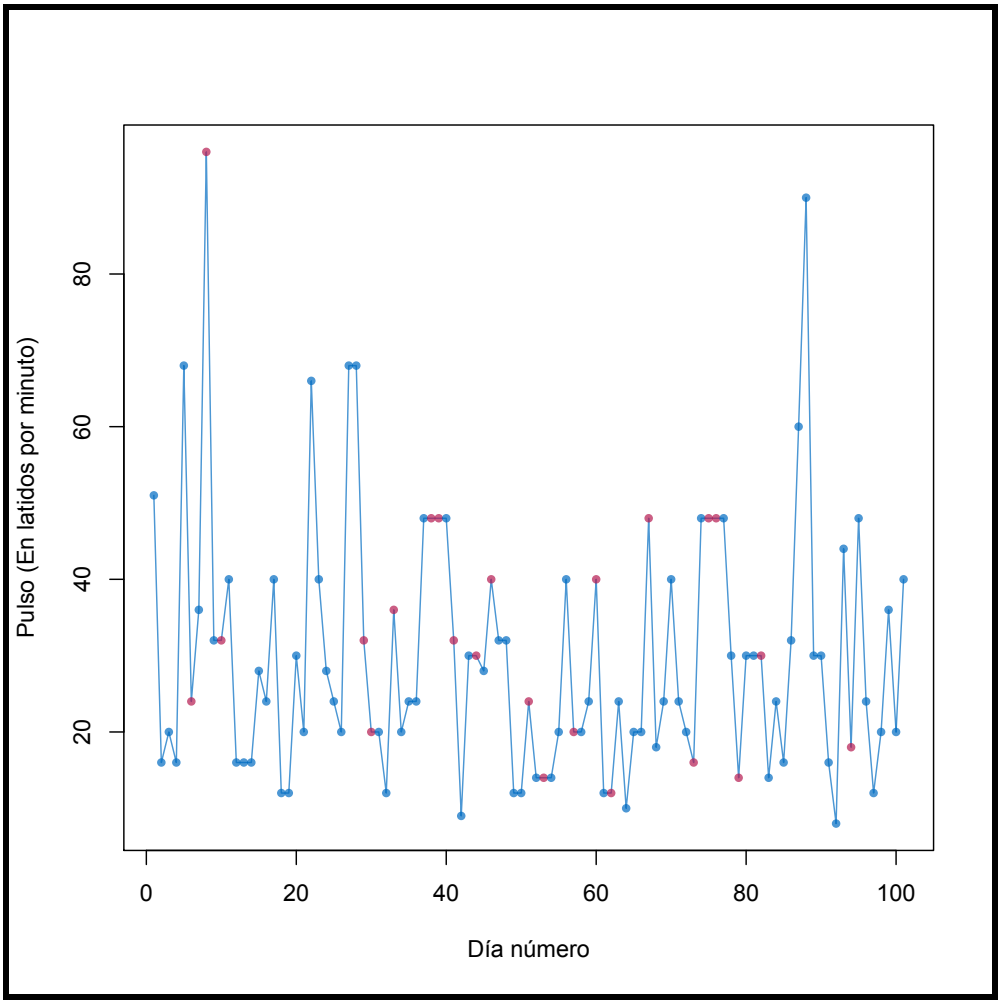

Ilustración 12: Imputación mediante LOCF en el dataset "subsample2" (variable "pulse")

Los puntos rojos son las imputaciones llevadas a cabo mediante LOCF. Este método se utiliza mucho en ensayos clínicos, y es el método preferido de imputación, considerando que es conservador y poco sesgado como otros métodos.

De todos modos, algunos autores consideran que el sesgo puede producirse en ambas direcciones ("hacia atrás y hacia adelante"), y se ha observado que el método de

[26] Ref. [15], página 14.

imputación LOCF puede producir estimaciones sesgadas incluso bajo el mecanismo MCAR.

3.2.8 Imputación múltiple[27]

La imputación múltiple crea m > 1 datasets. Cada uno de estos datasets es luego analizado. Los m resultados se agrupan en una estimación puntual final más el error estándar de acuerdo a reglas simples de agrupación ("Reglas de Rubin").

La ilustración a continuación muestra los 3 pasos principales en la imputación múltiple: imputación, análisis y agrupación.

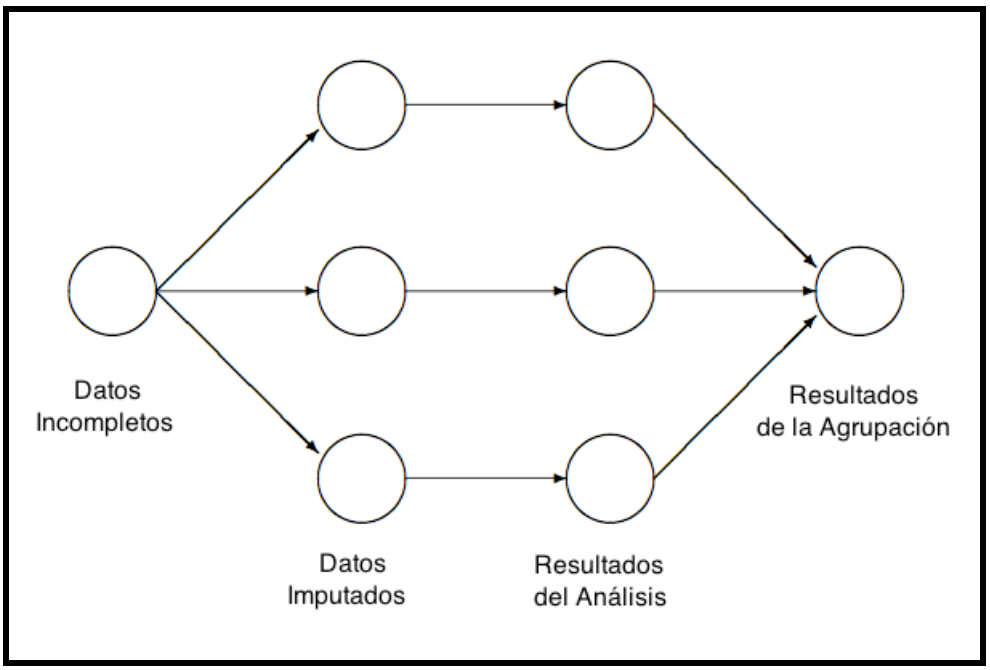

Ilustración 13: Pasos principales en una imputación múltiple

El análisis inicia con los datos observados e incompletos. La imputación múltiple crea varias versiones completas de los datos, reemplazando los valores faltantes con valores factibles.

Estos valores factibles se obtienen de una distribución modelada específicamente para cada valor faltante. En la ilustración anterior, los datasets imputados son 3, i.e., m = 3. En la práctica, se utilizan una mayor cantidad de datasets m.

Siguiendo con el ejemplo, en los 3 datasets imputados, aunque provienen del mismo dataset original, los valores imputados en cada uno de ellos, son distintos.

El paso siguiente es estimar los parámetros de interés para cada uno de los datasets imputados. Esto se realiza típicamente mediante la aplicación de métodos analíticos que se hubieran utilizado en el caso que los datos estuvieran completos.

El último paso es agrupar las estimaciones para el parámetro m en una única estimación, y luego estimar su varianza. La varianza combina la varianza de muestreo convencional

[27] Ref. [15], página 16.

(varianza dentro de imputaciones) con la varianza causada por los valores faltantes (varianza entre imputaciones).

Bajo las condiciones apropiadas, las estimaciones agrupadas no están sesgadas, y tienen las propiedades estadísticas correctas.
Para ejemplificar la imputación múltiple, generaré un nuevo dataset subsample3, con 5 variables y 100 observaciones, y luego efectuaré una imputación múltiple para los valores faltantes del mismo. Antes de realizar la imputación múltiple observaré el patrón de datos faltantes del dataset generado subsample3.

Genero subsample3:
> subsample3 <- slice (HOC,1:100) %>% select (pulse, resp_rate, pain, packed_cell_volume, total_protein)

Investigo el patrón de datos faltantes en subsample3:

> pattern3 <- md.pattern(subsample3)
> View (pattern3)

	row.names	packed_cell_volume	pulse	total_protein	pain	resp_rate	V6
1	58	1	1	1	1	1	0
2	2	1	0	1	1	1	1
3	12	1	1	1	1	0	1
4	11	1	1	1	0	1	1
5	1	0	1	1	1	1	1
6	1	1	1	0	1	1	1
7	3	1	0	1	1	0	2
8	1	1	1	1	0	0	2
9	2	1	1	0	1	0	2
10	1	0	1	0	1	1	2
11	2	1	0	1	0	0	3
12	3	0	1	0	0	1	3
13	1	0	0	0	1	0	4
14	1	0	1	0	0	0	4
15	1	0	0	0	0	0	5
16		8	9	10	19	23	69

Ilustración 14: Dataframe "pattern3" (patrón de datos faltantes dataset "subsample3")

Puedo observar que en el subsample3 hay 69 valores faltantes.

Además, visualizando el patrón mediante el paquete VIM, se obtiene el siguiente gráfico:

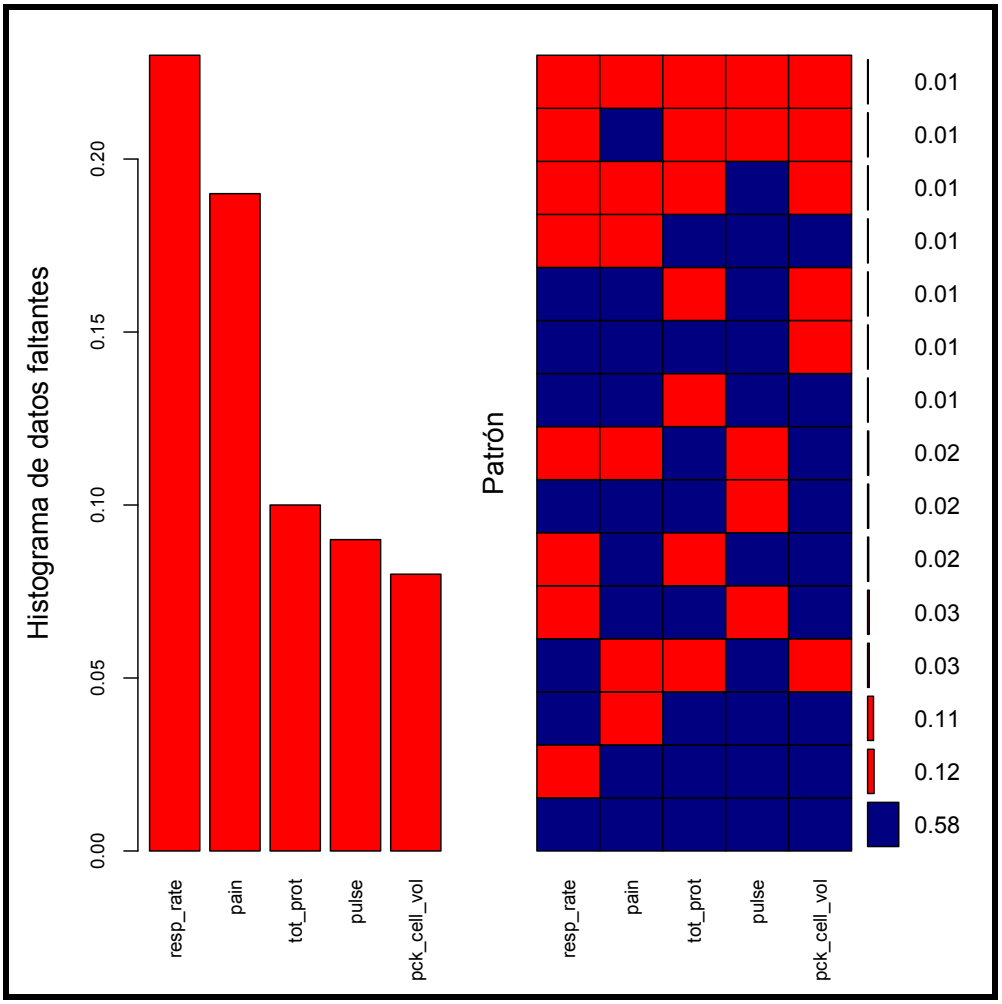

Ilustración 15: Gráfico con patrón de datos faltantes en dataset "subsample3"

Puede observarse como existen dos variables (resp_rate y pain) que tienen prácticamente el doble de valores faltantes, en comparación con las otras 3 variables. Pareciera existir una co-ocurrencia en los valores faltantes de las 5 variables, pero en un porcentaje tan bajo (0.01), que resulta despreciable.

Aunque no parece haber co-ocurrencia que deba ser tenida en cuenta, se pueden realizar un gráfico mediante el cual analizar los valores faltantes que se producen en una variable en relación a los valores faltantes de otra variable. Suponiendo que existiera una co-ocurrencia entre las variables resp_rate y pain, el siguiente gráfico muestra la relación de los valores faltantes entre estas dos variables:

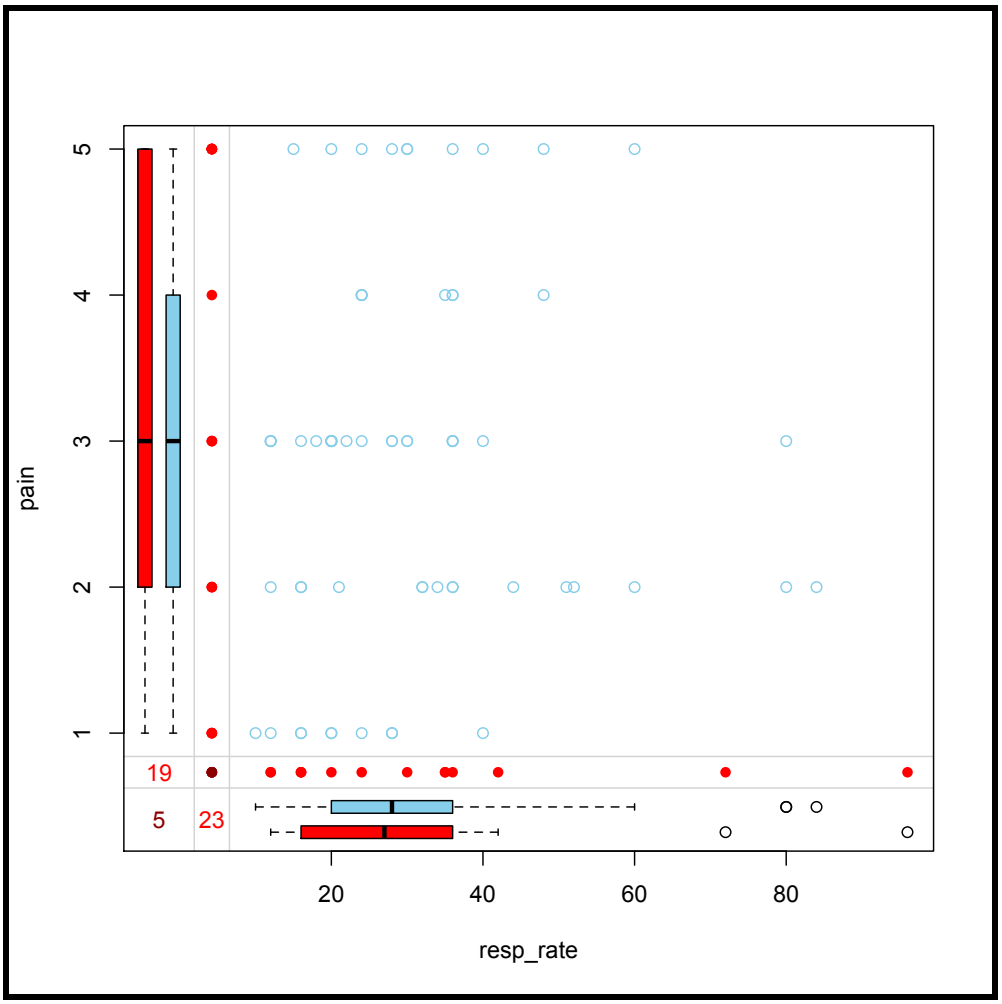

Ilustración 16: Relación entre valores faltantes de las variables resp_rate y pain del dataframe "subsample3"

Como se mencionó, no existe co-ocurrencia de datos faltantes entre estas dos variables, sin embargo vale recalcar la interesante forma de visualización del gráfico anterior para representar una situación semejante. Como en todos los gráficos anteriores, los puntos rojos son los valores faltantes, mientras que los azules son los valores observados realmente. Pueden observarse los box-plots que indican la distribución de los datos faltantes y los datos observados para las dos variables, así como la cantidad de datos faltantes para una o otra variable (19 y 23). El valor 5 indica que hay una co-ocurrencia de 5 valores faltantes entre las dos variables. Se ven incluso puntos outliers en los boxplots, en este caso para la variable resp_rate.

Ahora realizo la imputación múltiple, utilizando el paquete MICE:

```
> imp <- mice(subsample3, seed=1, print=FALSE)
> fit <- with(imp, lm(pulse ~ resp_rate + pain + packed_cell_volume + total_protein))
> tab <- round(summary(pool(fit)),3)
> tab[,c(1:3,5)]
```

	est	se	t	Pr(>|t|)
(Intercept)	-17.543	15.719	-1.116	0.276
resp_rate	0.855	0.148	5.760	0.000
pain	1.881	2.010	0.936	0.353
packed_cell_volume	1.210	0.284	4.268	0.000
total_protein	0.153	0.104	1.463	0.153

Gráficos resultantes de la imputación realizada:

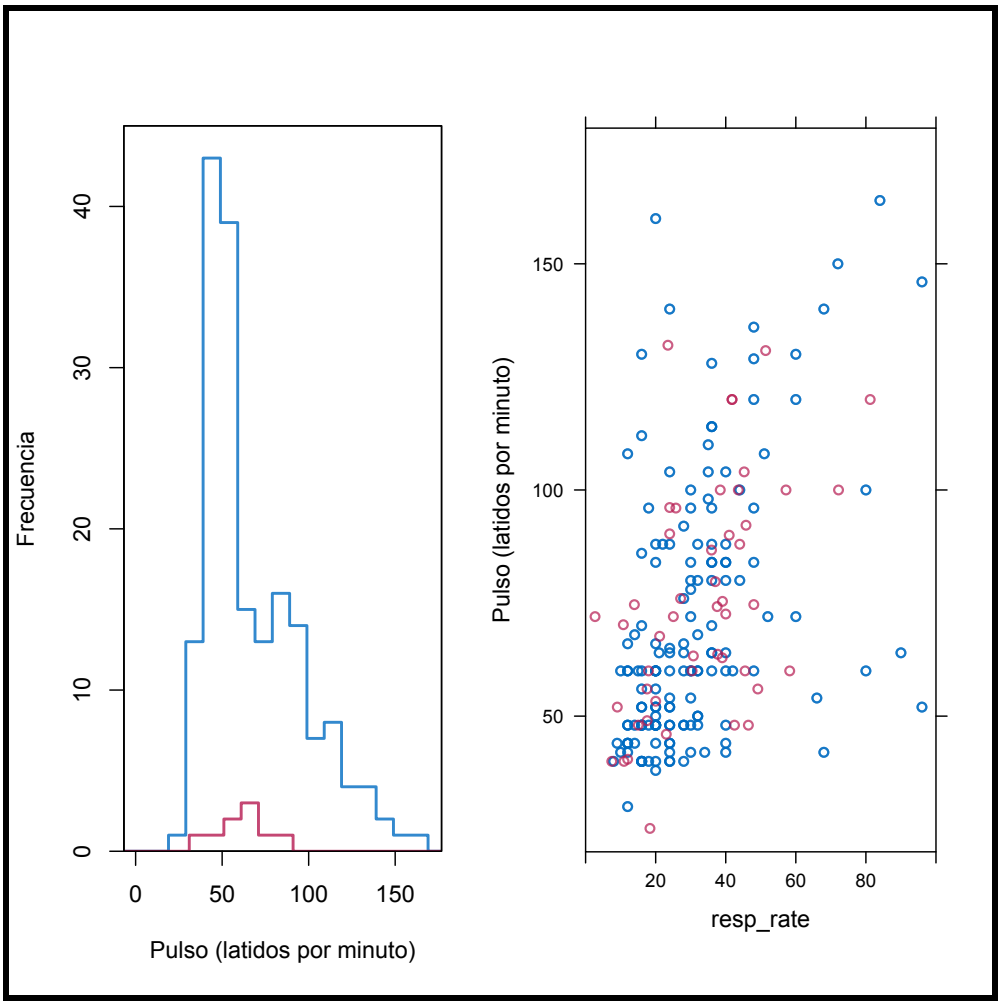

Ilustración 17: Modelo de Imputación múltiple en el dataset "subsample3"

En la ilustración anterior, pueden verse los datos observados y los datos imputados mediante la imputación múltiple. Obsérvese como las distribuciones de los puntos rojos y azules son bastante similares, en la figura derecha.

Para verificar el modelo de imputación empleado en cada una de las variables puedo ejecutar:

```
> imp$meth
     pulse  resp_rate       pain pck_cell_vol   tot_prot
     "pmm"      "pmm"      "pmm"        "pmm"      "pmm"
```

El programa me informa que todas las variables fueron imputadas mediante el método pmm (Predictive Mean Matching, la imputación por defecto que realiza el paquete MICE).

4. Imputación múltiple sobre los datasets abordados.

De todos las metodologías explicadas en los apartados previos, hay que subrayar que la estrategia más completa para resolver la casuística de datos faltantes en los datasets que conlleven este aspecto, es el método de imputación múltiple.

La imputación múltiple se basa en la reintroducción del error o variabilidad original de los datos en los valores imputados. Por lo tanto, se estudiará la imputación de los datasets mencionados en el apartado 3.1, mediante este tipo de imputación.

El código fuente en lenguaje R, de los procedimientos para las imputaciones y sus visualizaciones, se encuentran detallados en el Anexo 2.

Habida cuenta que las demostraciones sobre distintos métodos de imputación se ha realizado sobre muestras del dataset HOC, procederé ahora a realizar una visualización y una imputación de este dataset (con todos sus datos: 300 observaciones con 29 variables).

4.1 Dataset "HOC"

4.1.1 Visualización de patrón de datos faltantes del dataset HOC

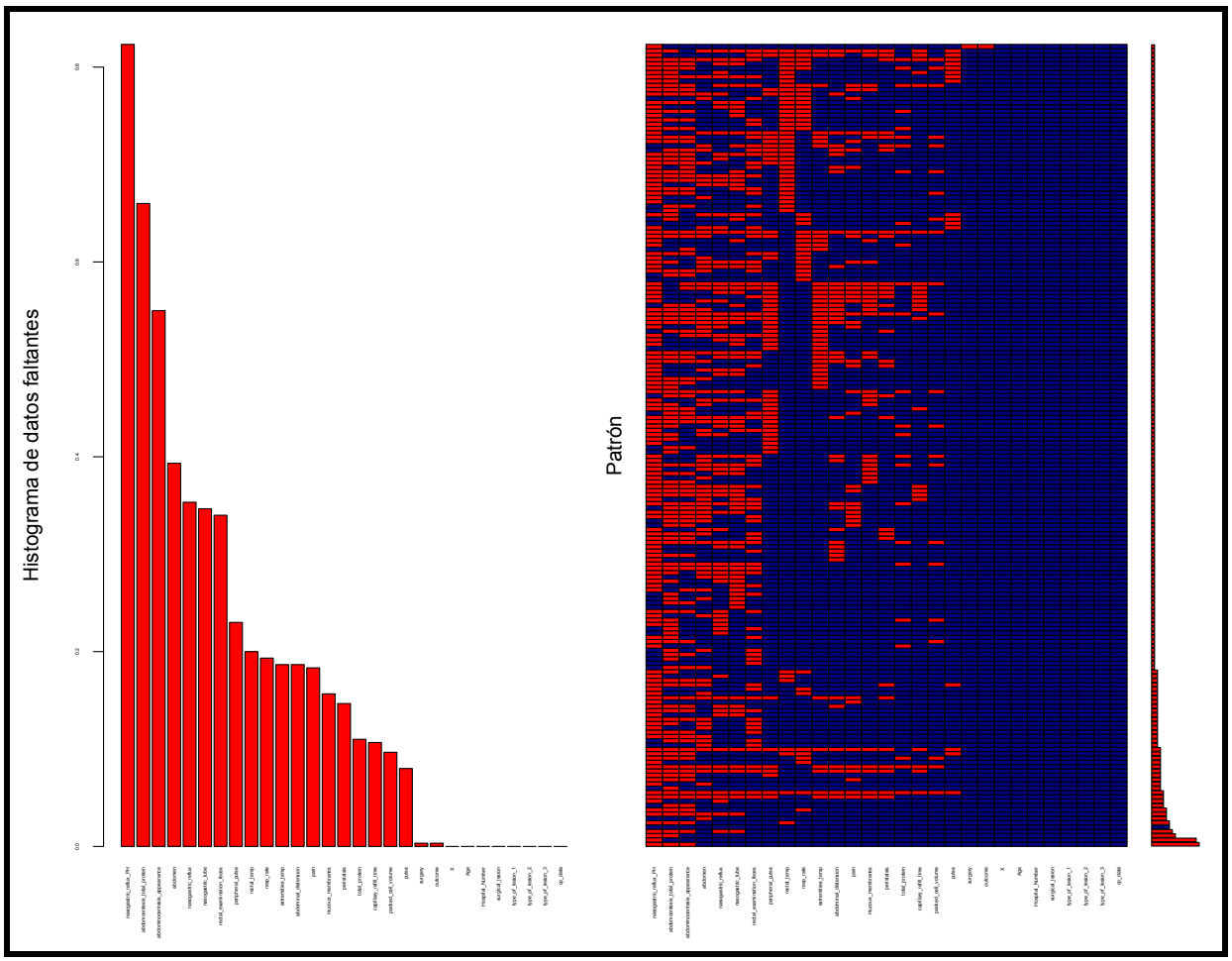

Ilustración 18: Gráfico con patrón de datos faltantes en dataset "HOC"

Variables ordenadas por cantidad de datos faltantes	
Variable	Cantidad
nasogastric_reflux_PH	0.823333333
abdomcentesis_total_protein	0.66
abdominocentesis_appearance	0.55
abdomen	0.393333333
nasogastric_reflux	0.353333333
nasogastric_tube	0.346666667
rectal_examination_feces	0.34
peripheral_pulse	0.23
rectal_temp	0.2
resp_rate	0.193333333
extremities_temp	0.186666667
abdominal_distension	0.186666667
pain	0.183333333
mucous_membranes	0.156666667
peristalsis	0.146666667
total_protein	0.11
capillary_refill_time	0.106666667
packed_cell_volume	0.096666667
pulse	0.08
surgery	0.003333333
outcome	0.003333333
Age	0
Hospital_Number	0
surgical_lesion	0
type_of_lesion_1	0
type_of_lesion_2	0
type_of_lesion_3	0
cp_data	0

Obsérvese que aunque el dataset tenga 29 variables, los datos faltantes se concentran en algunas variables, en particular en las variables:

nasogastric_reflux_PH: 82.33 % de datos faltantes
abdomcentesis_total_protein: 66 % de datos faltantes
abdominocentesis_appearance: 55 % de datos faltantes

Las variables ***abdomen, nasogastric_reflux, nasogastric_tube y rectal_examination_feces*** tienen entre 40 % y 30 % de datos faltantes, mientras que todas las demás tienen menos de 20 % de datos faltantes.

Incluso existen 7 variables que no contienen datos faltantes.

Aunque no es una imputación múltiple, a efectos demostrativos, procederé a realizar una imputación de la media sobre el dataset HOC, para poder exponer como la imputación

conlleva 2 pasos: la generación del "modelo de imputación" (mal llamada imputación a secas) y la imputación efectiva del dataset con los valores de imputación (en este TFM mencionada como "imputación efectiva").

Cabe mencionar que en la mayoría de los textos a veces no se presentan con claridad los dos prodedimientos, mencionando a los mismos simplemente como "imputación".

Realizando entonces la Imputación de la media en el dataset HOC:

imp <- mice(HOC, method = "mean", seed=1, print=FALSE)

Se genera el modelo de imputación en el objeto "imp", el cual puede ser llamado ejecutando simplemente "imp":

```
> imp
Multiply imputed data set
Call:
mice(data = subsample3, printFlag = FALSE, seed = 1)
Number of multiple imputations:  5
Missing cells per column:
     pulse    resp_rate         pain pck_cell_vol     tot_prot
         9           23           19            8           10
Imputation methods:
     pulse    resp_rate         pain pck_cell_vol     tot_prot
     "pmm"        "pmm"        "pmm"        "pmm"        "pmm"
VisitSequence:
     pulse    resp_rate         pain pck_cell_vol     tot_prot
         1            2            3            4            5
PredictorMatrix:
             pulse         resp_rate    pain   pck_cell_vol              tot_prot
pulse        0             1            1          1                   1
resp_rate    1             0            1          1                   1
pain         1             1            0          1                   1
pck_cell_vol       1             1            1          0                   1
tot_prot     1             1            1          1                   0
Random generator seed value:  1
```

Con lo cual, para proceder a la imputación efectiva del dataset, debe ejecutarse:

> HOC_completed = complete(imp)

Y realizando un gráfico del estado del dataset con los valores imputados se obtiene:

4.1.2 Visualización de patrón de datos en dataset HOC después de la imputación efectiva

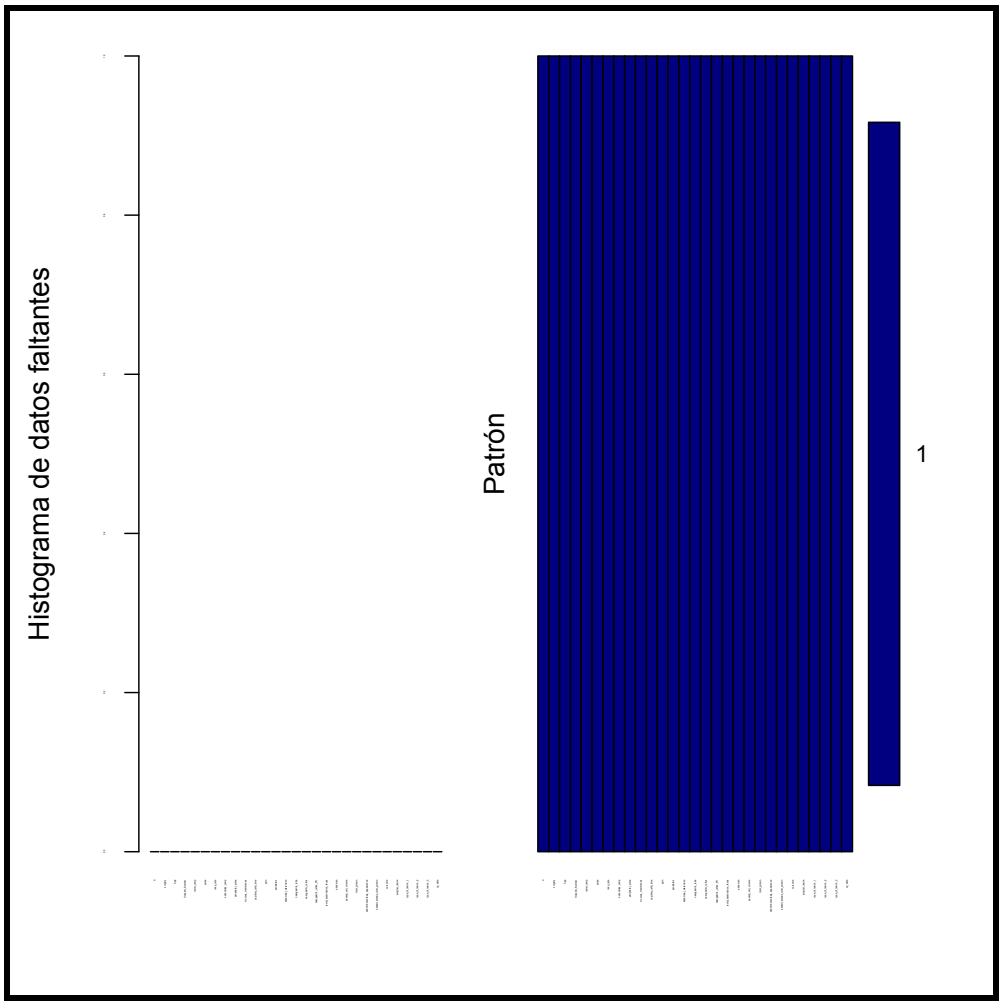

Ilustración 19: Gráfico posterior a la imputación efectiva del dataset "HOC"

Ya no quedan valores faltantes en el dataset HOC, los valores ausentes han sido reemplazados por la media de su variable correspondiente.

Para verificar el modelo de imputación empleado en las 6 primeras variables puedo ejecutar:

```
> head(imp$meth)
```

X	surgery	Age	Hospital_Number	rectal_temp	pulse
"mean"	"mean"	"mean"	"mean"	"mean"	"mean"

El programa me informa que efectivamante las variables fueron imputadas mediante el método mean (media).

4.2 Dataset "subsample3"

4.2.1 Visualización de patrón de datos faltantes del dataset subsample3

Para una imputación múltiple utilizando el paquete MICE con su algoritmo por defecto PMM (Predictive mean matching: un algoritmo desarrollado por Rubin a finales de los 80, y uno de los más populares para realizar imputación), utilizaré un dataset más pequeño (subsample3) en donde se pueda apreciar mejor el mecanismo de imputación:

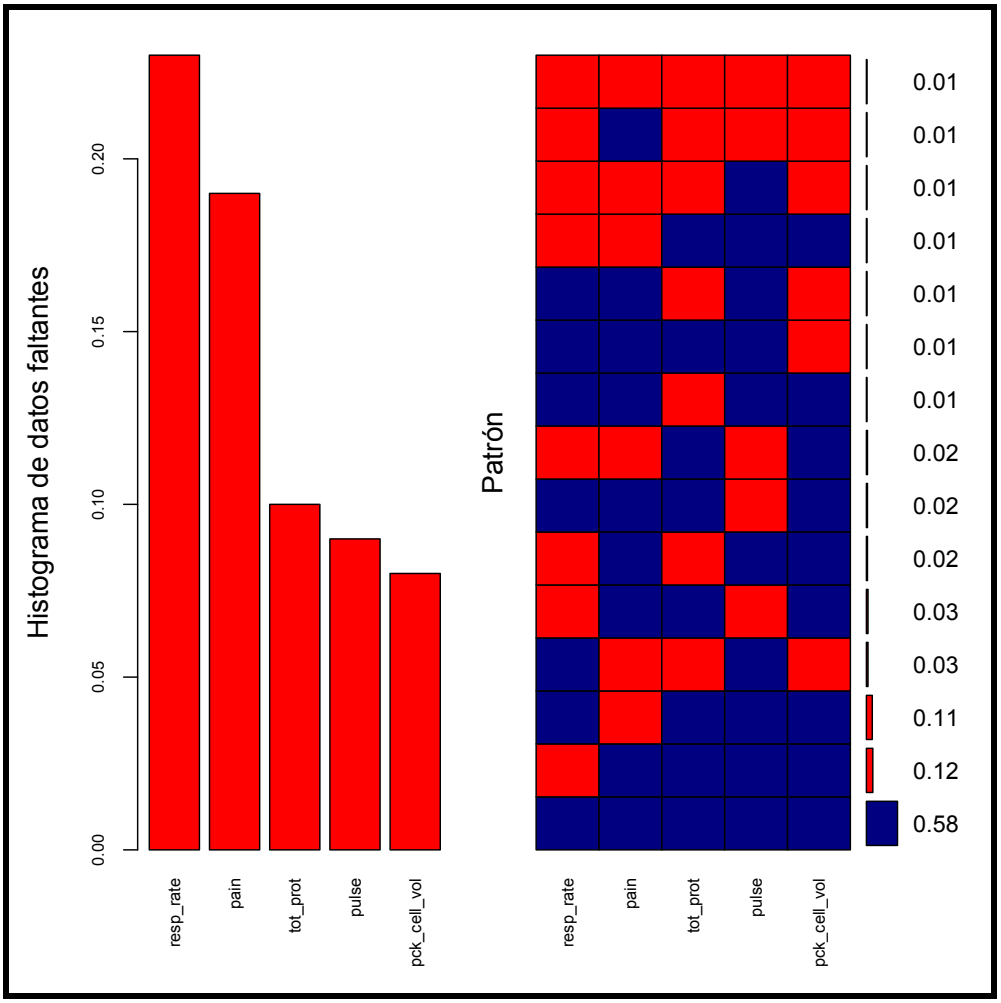

Ilustración 20: Gráfico con patrón de datos faltantes en dataset "subsample3"

Genero el modelo de imputación:

imp <- mice(subsample3, method = "pmm", seed=1, print=FALSE)

Imputación efectiva del dataset subsample3

subsample3_completed = complete(imp)

4.2.2 Visualización de patrón de datos en dataset subsample3 después de la imputación efectiva

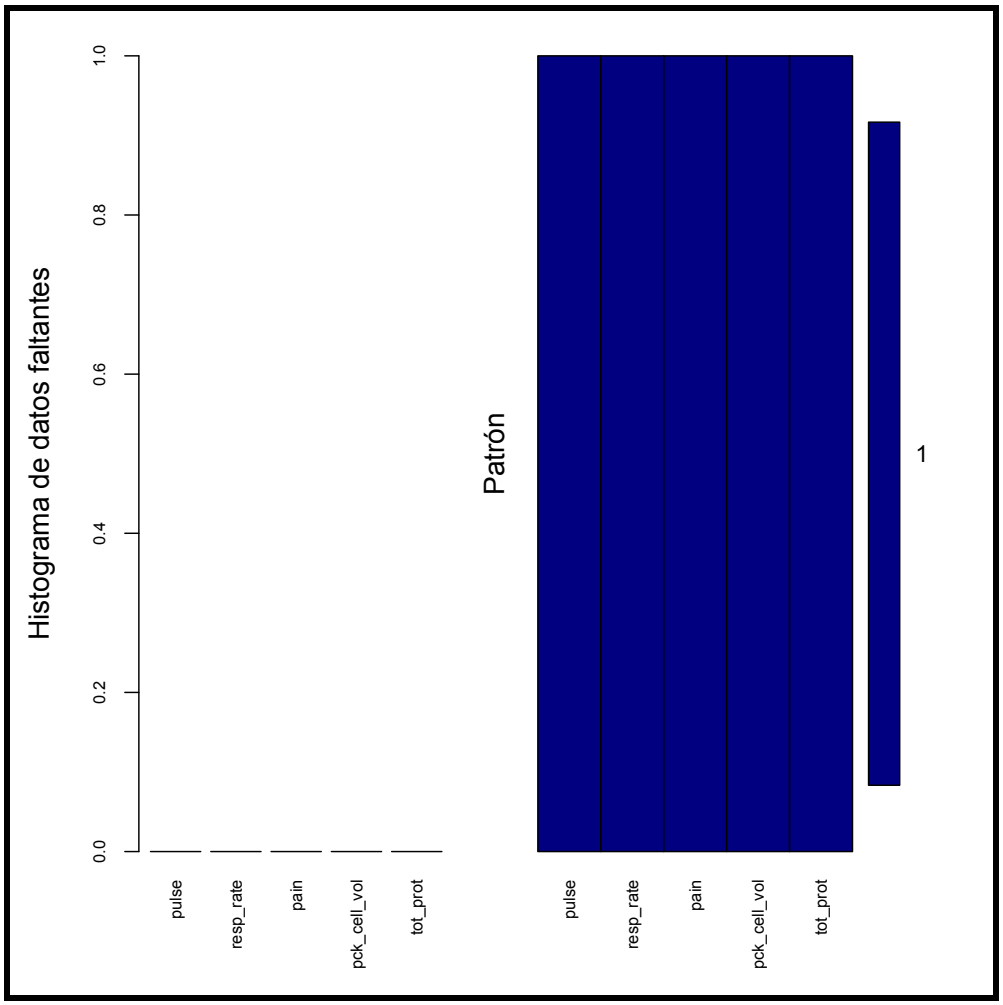

Ilustración 21: Gráfico posterior a la imputación efectiva del dataset "subsample3"

También aqui para verificar el modelo de imputación empleado en las variables puedo ejecutar:

```
> imp$meth
pulse      resp_rate    pain         pck_cell_vol    tot_prot
"pmm"      "pmm"        "pmm"        "pmm"           "pmm"
```

El programa me informa que efectivamante las variables fueron imputadas mediante el método pmm.

4.3 Dataset "AIR"

4.3.1 Visualización de patrón de datos faltantes del dataset AIR

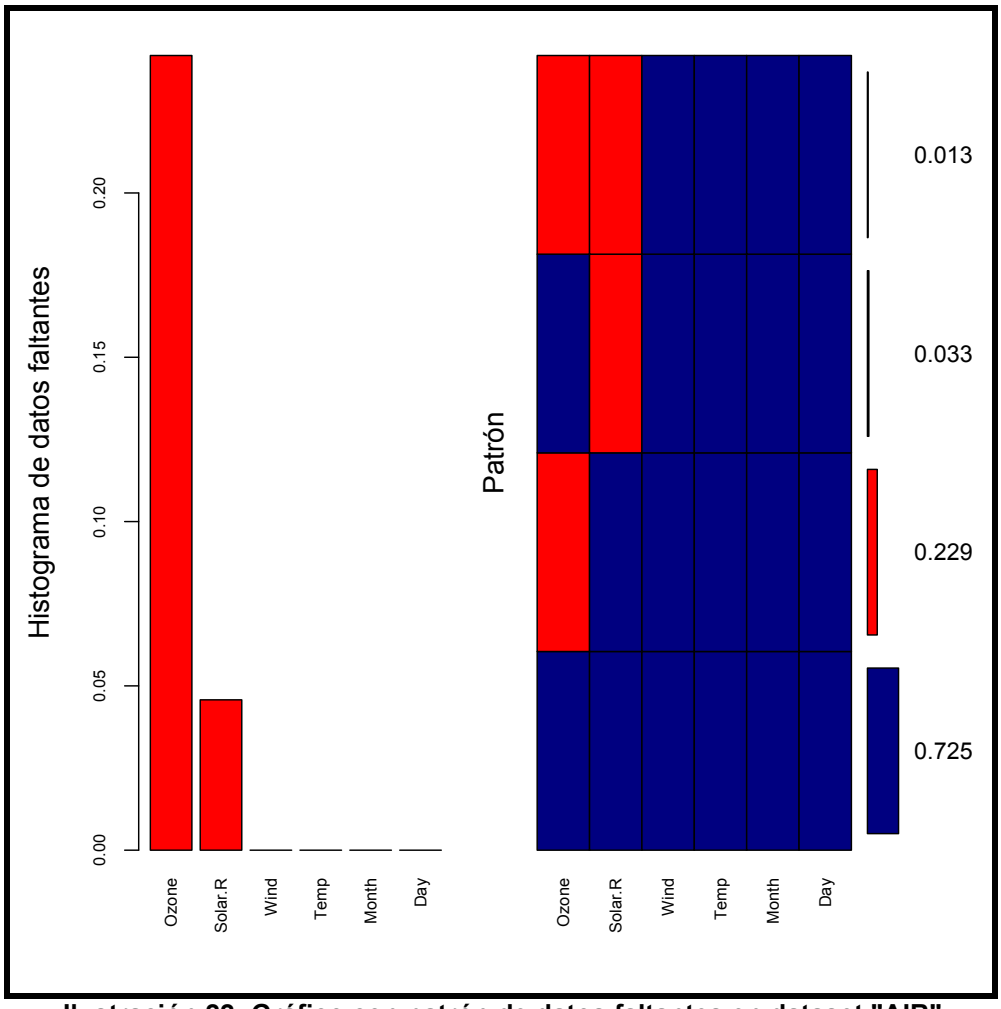

Ilustración 22: Gráfico con patrón de datos faltantes en dataset "AIR"

Variables ordenadas por cantidad de datos faltantes	
Variable	Cantidad
Ozone	0.24183007
Solar.R	0.04575163
Wind	0
Temp	0
Month	0
Day	0

Imputación múltiple del dataframe AIR:

imp <- mice(airquality, seed=1, print=FALSE)

Gráfico:

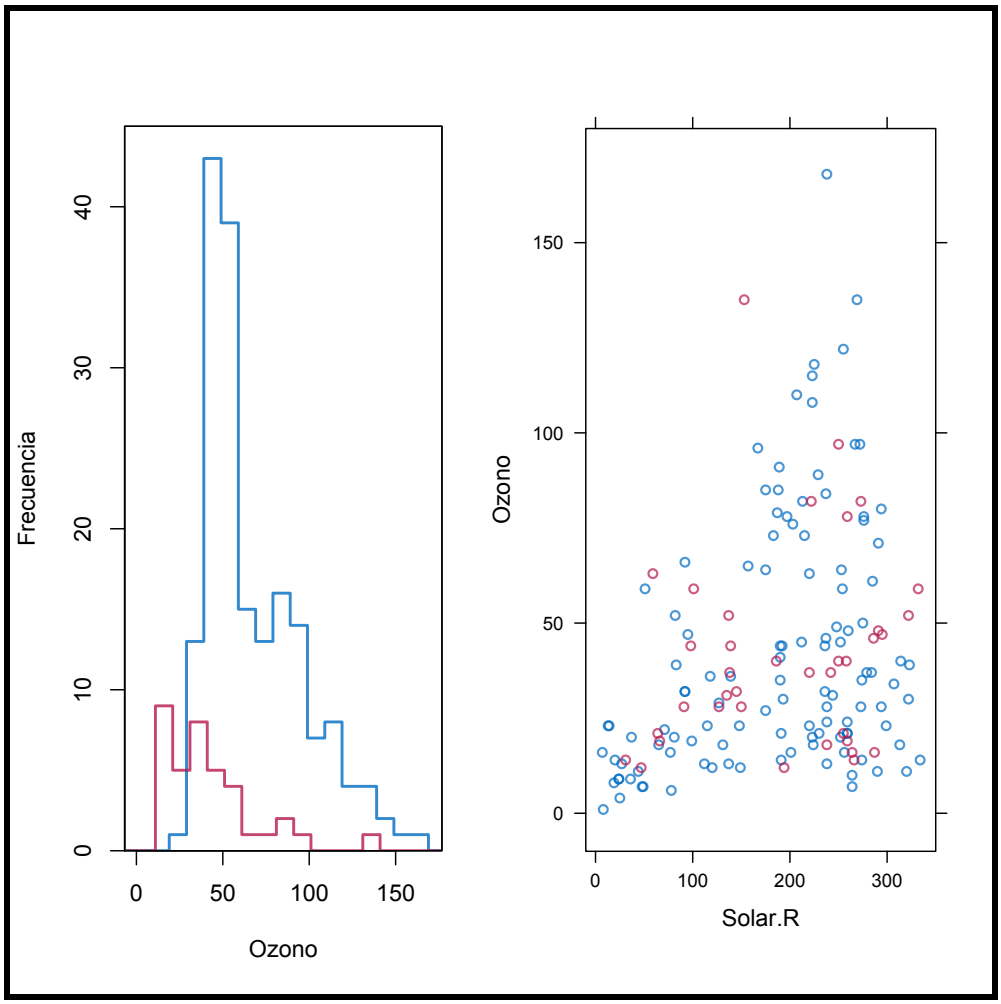

Ilustración 23: Modelo de Imputación múltiple en el dataset "AIR"

Puede observarse que en la ilustración anterior (figura de la derecha), en donde se muestran las imputaciones (en rojo) que son candidatas a ser imputadas, que se mantiene muy bien la variabilidad de los datos, en este caso, de la nube de puntos formada por la visión bidimensional de la variable Ozono contra la variable Solar.R. Recuérdese que los puntos azules son los valores reales, mientras que los valores en rojo son los valores a ser imputados de acuerdo al modelo de imputación generado.

Procedo ahoar a realizar una imputación efectiva del dataset AIR:

airquality_completed = complete(imp)

4.3.2 Visualización de patrón de datos en dataset AIR después de la imputación efectiva:

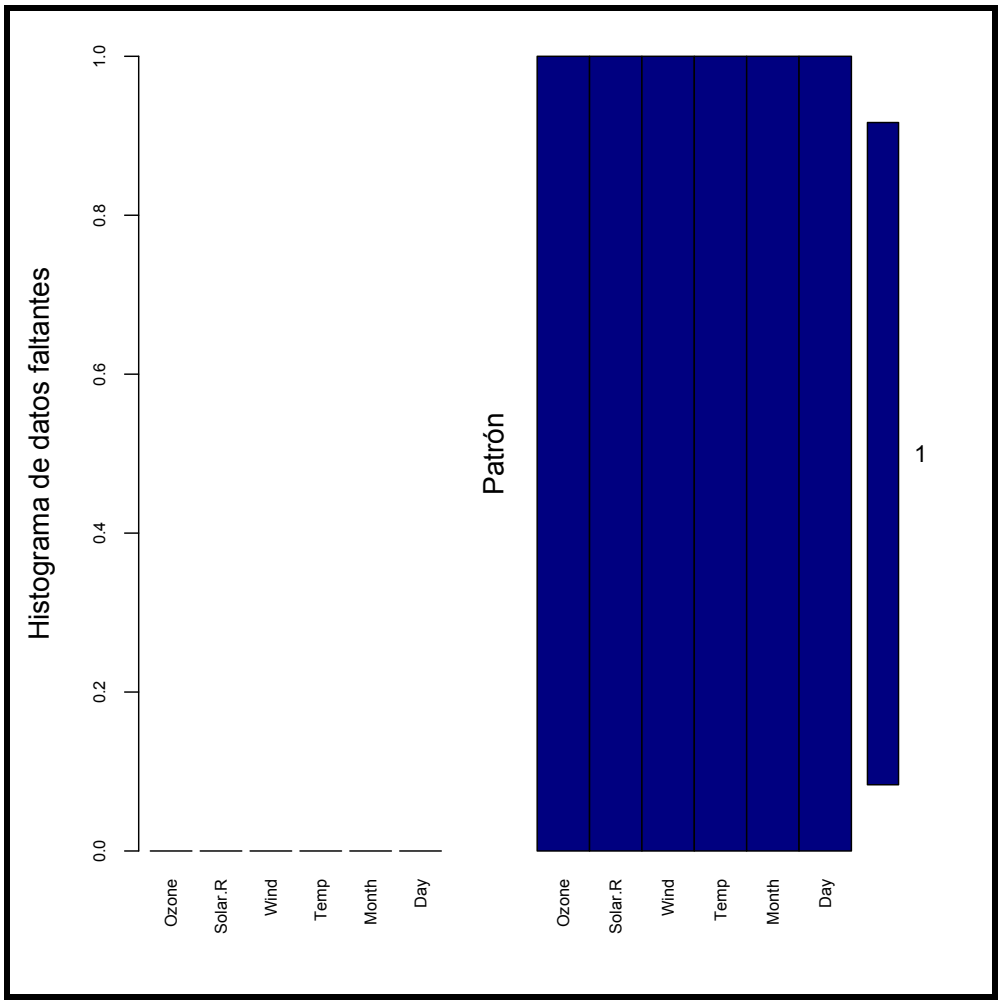

Ilustración 24: Gráfico posterior a la imputación efectiva del dataset "AIR"

Si verifico el modelo de imputación empleado en cada una de las variables obtengo:

```
> imp$meth
Ozone   Solar.R    Wind    Temp   Month     Day
"pmm"   "pmm"      ""      ""     ""        ""
```

El programa me informa que las dos primeras variables fueron imputadas mediante el método pmm. El resto de las variables no fueron imputadas, ya que no contenían valores faltantes (Véase la ilustración 22).

Una manera interesante de observar la eficacia del modelo de imputación es comparando, mediante gráficos de densidad, las variables antes de ser imputadas, y luego de ser imputadas. Los gráficos que se muestran a continuación fueron generados utilizando el paquete ggplot2 (Para el código fuente en lenguaje R que realiza estos gráficos, véase el Anexo de Prácticas).

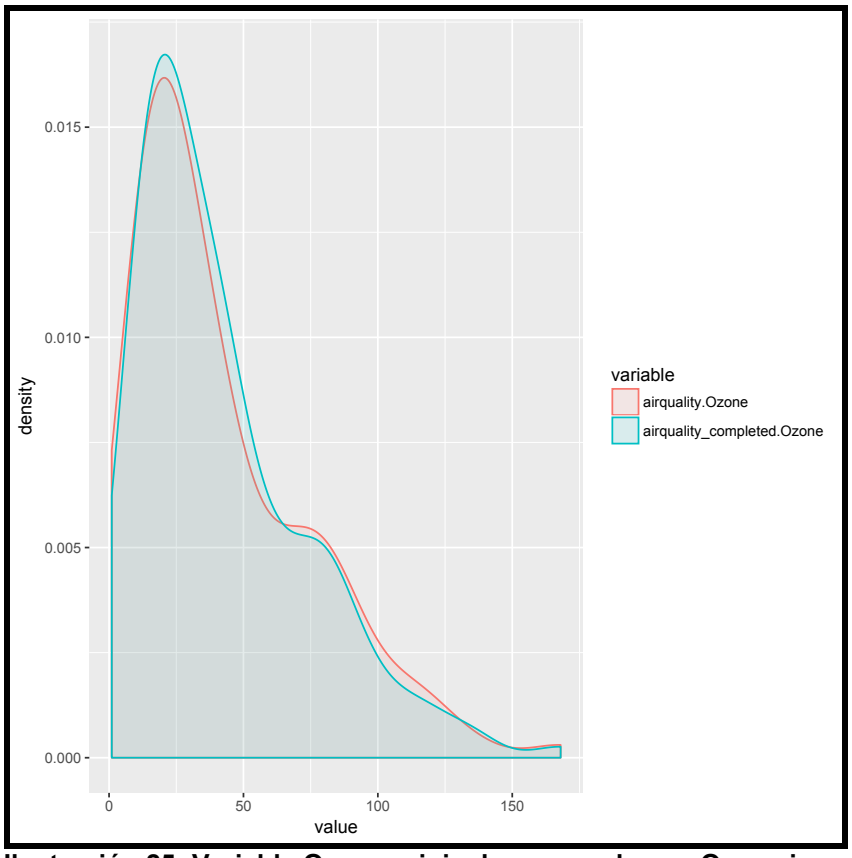

Ilustración 25: Variable Ozone original comparada con Ozone imputada

En la ilustración anterior, puede observarse que la imputación de la variable Ozone es muy eficaz, ya que la forma de la distribución es prácticamente igual a la variable original.

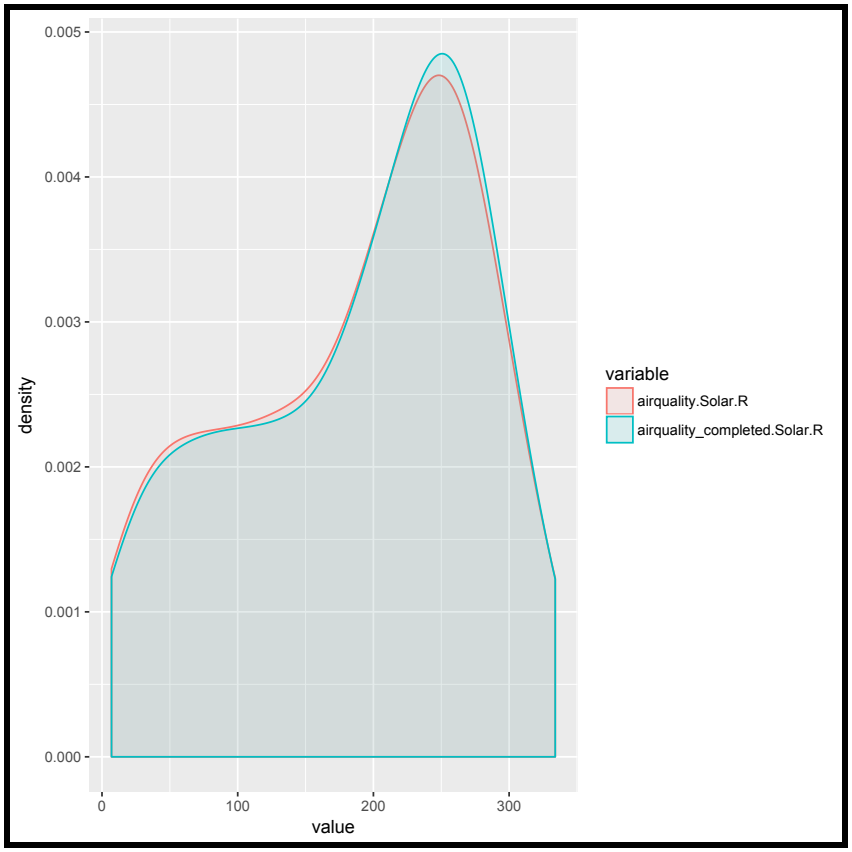

Ilustración 26: Variable Solar.R original comparada con Solar.R imputada

En la ilustración anterior, puede observarse que la imputación de la variable Solar.R es muy eficiente, ya que la densidad de la distribución imputada es prácticamente igual a la variable original.

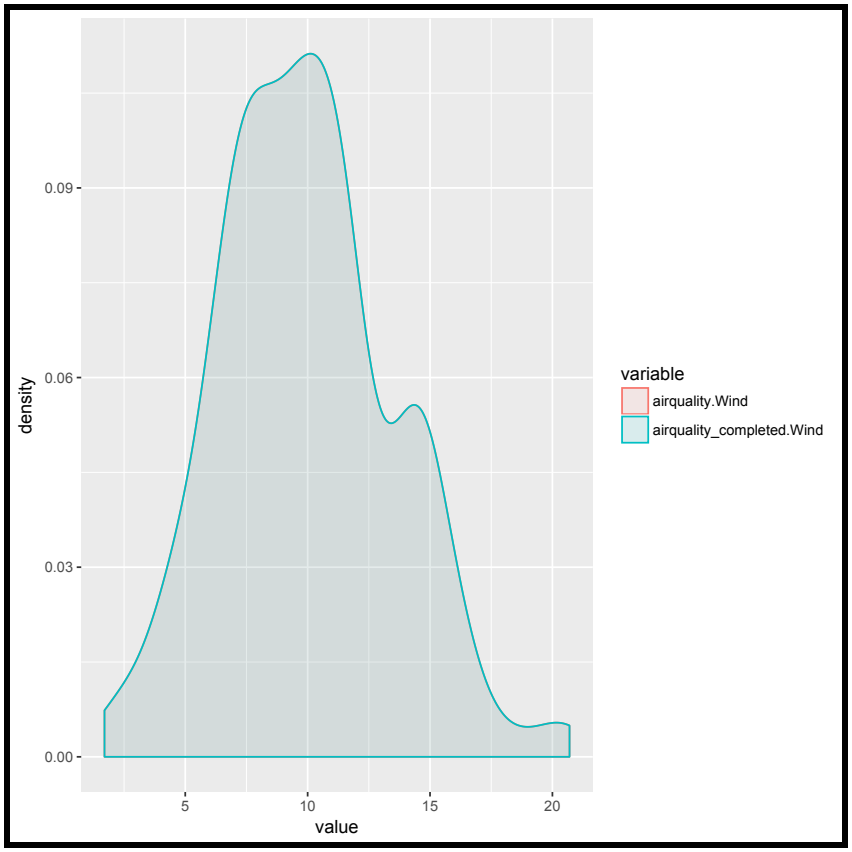

Ilustración 27: Variable Wind original comparada con Wind imputada

En la ilustración anterior, puede observarse que la variable Wind queda exactamente igual, en el dataset imputado, esto es evidente, ya que no se ha producido imputación alguna, habida cuenta que la variable Wind no contenía valores faltantes (Véase también la ilustración 22, donde se observa que la variable Wind no contiene valores faltantes).

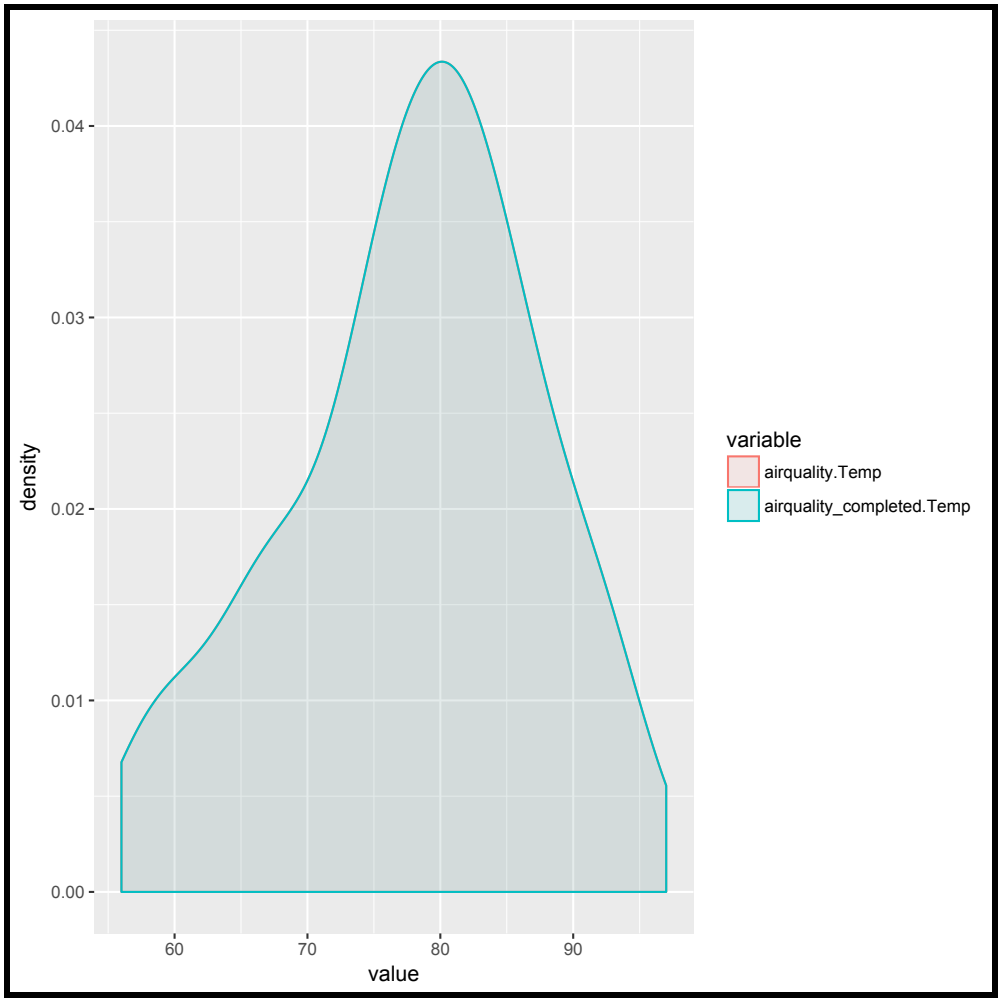

Ilustración 28: Variable Temp original comparada con Temp imputada

Del mismo modo que en la variable comentada anteriormente, puede observarse en esta ilustración que la variable Temp queda exactamente igual, en el dataset imputado, ya que no se ha producido imputación alguna, habida cuenta que la variable Temp no contenía datos faltantes.

4.4 Dataset "NHAN"

4.4.1 Visualización de patrón de datos faltantes del dataset NHAN

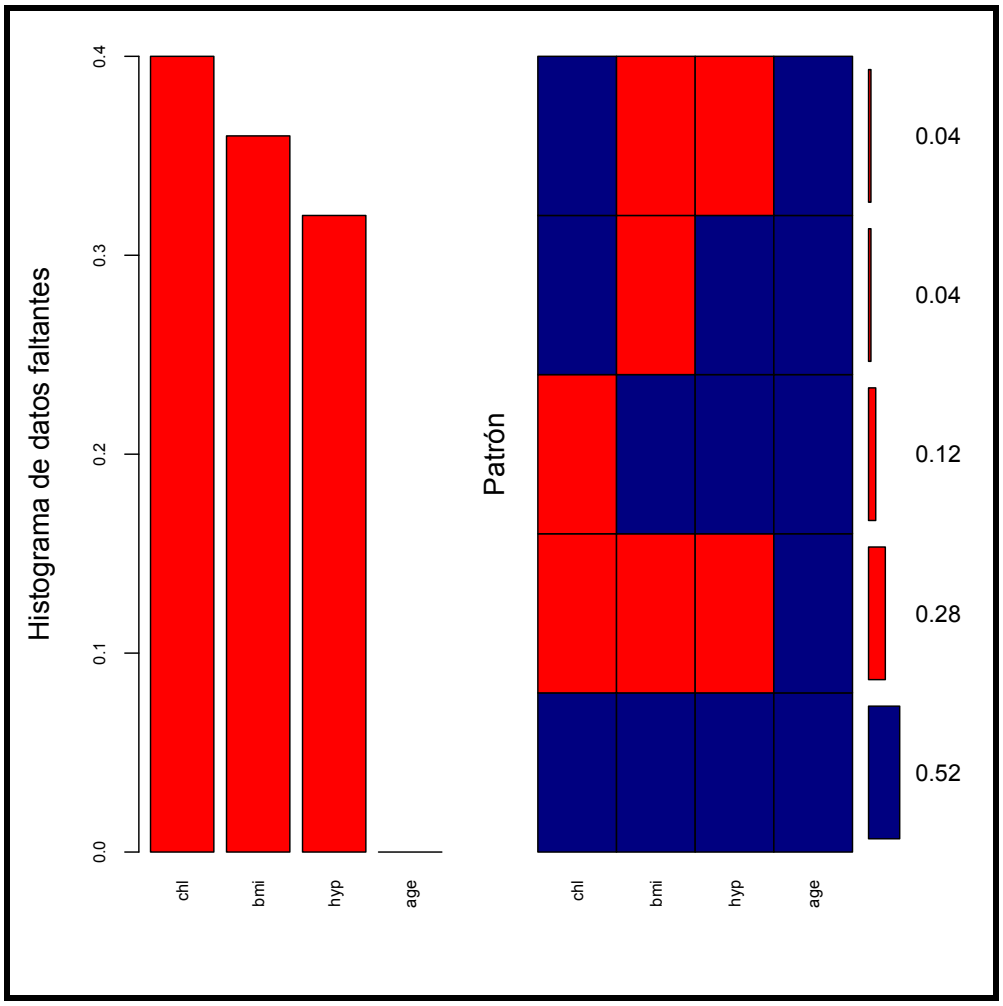

Ilustración 29: Gráfico con patrón de datos faltantes en dataset "NHAN"

Variables ordenadas por cantidad de datos faltantes	
Variable	Cantidad
chl	0.40
bmi	0.36
hyp	0.32
age	0

Imputación múltiple del dataframe NHAN:

```
> imp <- mice(nhanes, seed=5, print=FALSE)
> fit <- with(imp, lm(bmi ~ age + hyp + chl))
> tab <- round(summary(pool(fit)),3)
> tab[,c(1:3,5)]
```

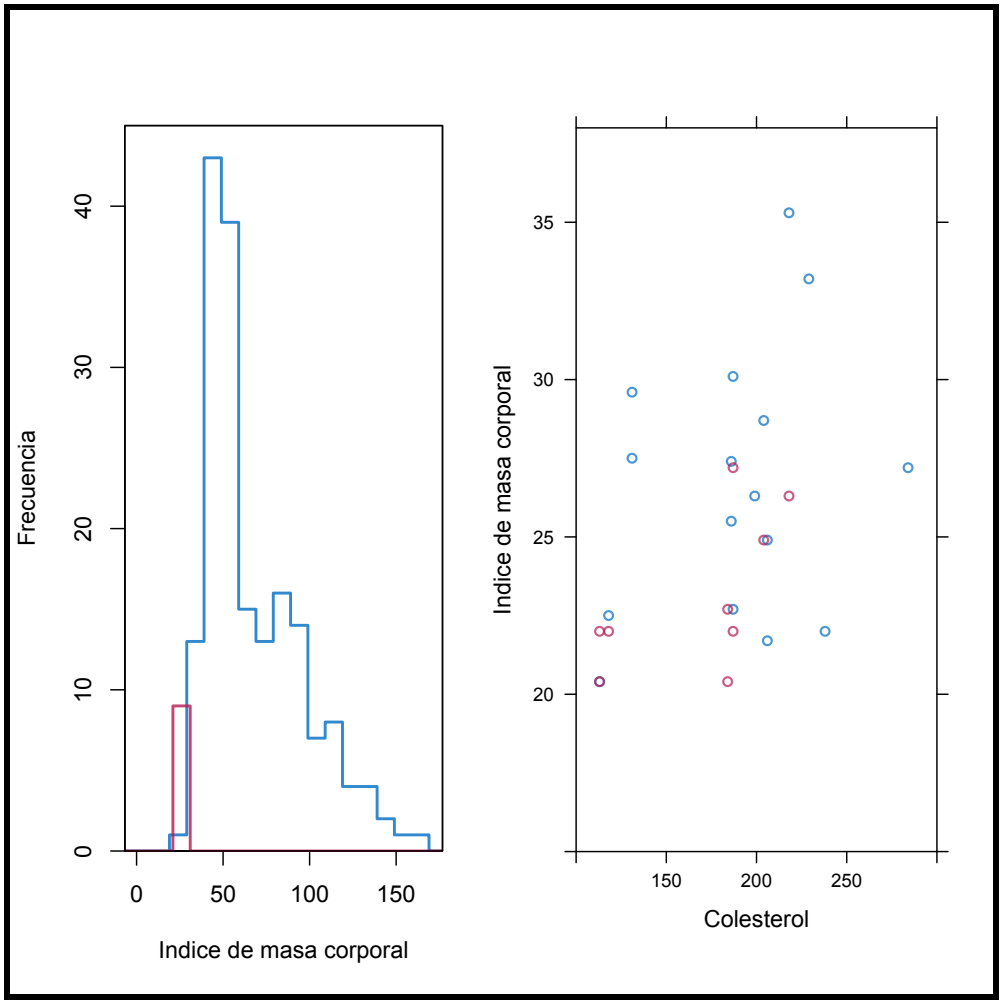

Ilustración 30: Modelo de Imputación múltiple en el dataset "NHAN"

Imputación efectiva del dataset NHAN:

> nhanes_completed = complete(imp)

Si verifico el modelo de imputación empleado en cada una de las variables obtengo:

> imp$meth

age	bmi	hyp	chl
""	"pmm"	"pmm"	"pmm"

El programa me informa que la primera variable "age" no fue imputada, ya que no contenía valores faltantes (Véase la ilustración 29). Las otras 3 variables fueron imputadas mediante el método pmm.

4.4.2 Visualización de patrón de datos en dataset NHAN después de la imputación efectiva

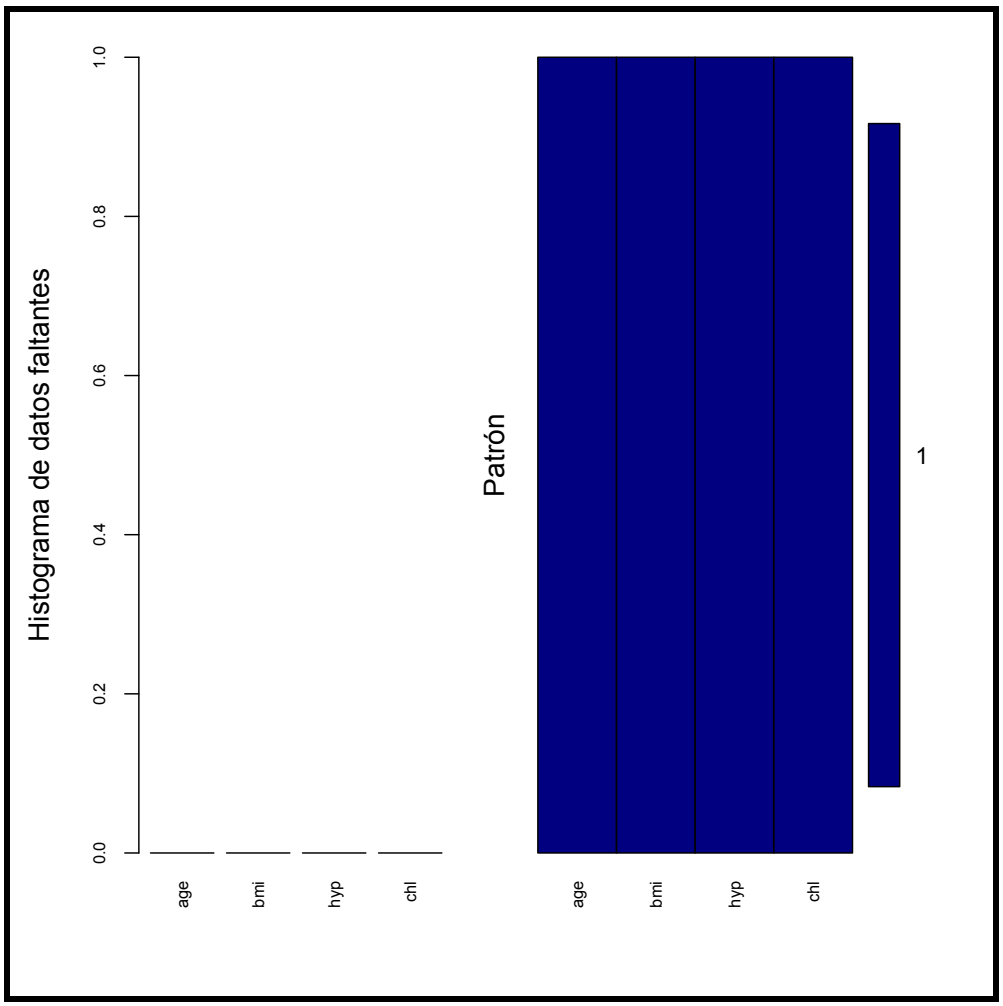

Ilustración 31: Gráfico posterior a la imputación efectiva del dataset "NHAN"

Si observo las primeras 6 filas de los dataframes (el original y el imputado), puedo observar como los NAs han sido reemplazados por valores de acuerdo al modelo PMM:

```
> head(nhanes)
  age  bmi hyp chl
1   1   NA  NA  NA
2   2 22.7   1 187
3   1   NA   1 187
4   3   NA  NA  NA
5   1 20.4   1 113
6   3   NA  NA 184

> head(nhanes_completed) # Imputado con PMM
  age  bmi hyp chl
1   1 22.0   1 113
2   2 22.7   1 187
```

3 1 27.2 1 187
4 3 24.9 2 204
5 1 20.4 1 113
6 3 20.4 2 184

El mejor modo de ver la eficiencia del modelo de imputación es comparar mediante gráficos de densidad, las variables antes de ser imputadas, y luego de ser imputadas. A continuación los gráficos de densidad de las variables bmi, hyp y chl (la variable age no tuvo imputación, ya que no contenía NAs, con lo cual no tiene sentido mostar el gráfico de densidad, ya que el mismo no muestra cambios entre la variable imputada y la original).

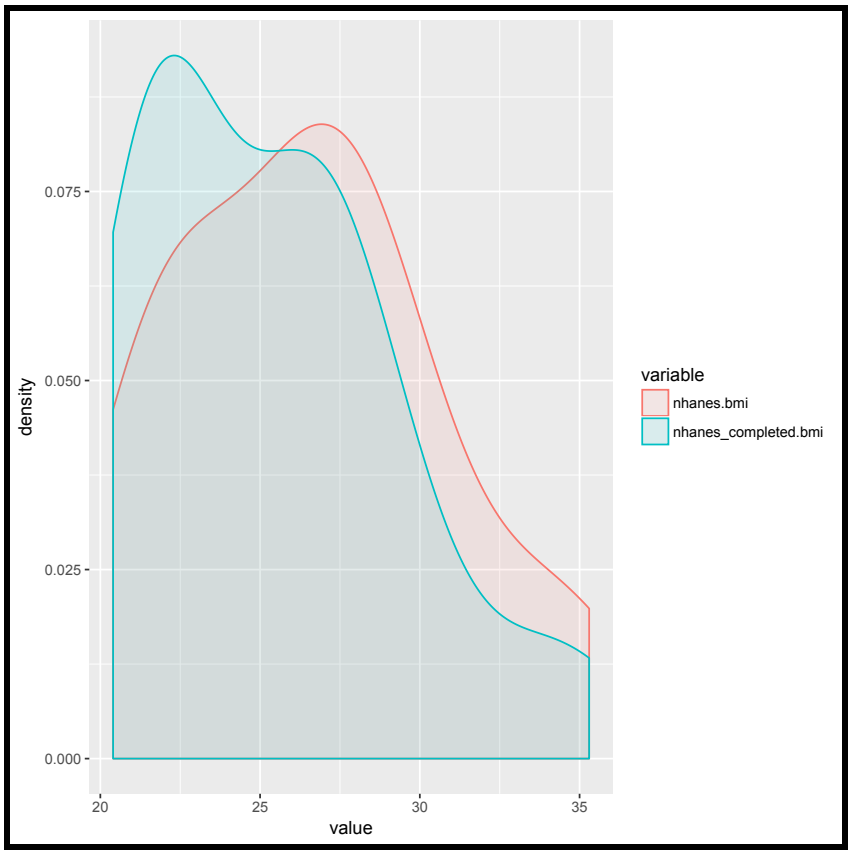

Ilustración 32: Variable bmi original comparada con bmi imputada

En la ilustración anterior, puede observarse que la imputación de la variable bmi no es demasiado eficaz, esto es así porque el algoritmo pmm no puede reproducir con fiabilidad la distribución de la variable original, ya que no cuenta con suficientes observaciones para calcular en forma eficaz la media predictiva y al mismo tiempo mantener la variabilidad (el dataset tiene sólo 25 observaciones, y la variable bmi tiene 9 datos faltantes en esas 25 observaciones).

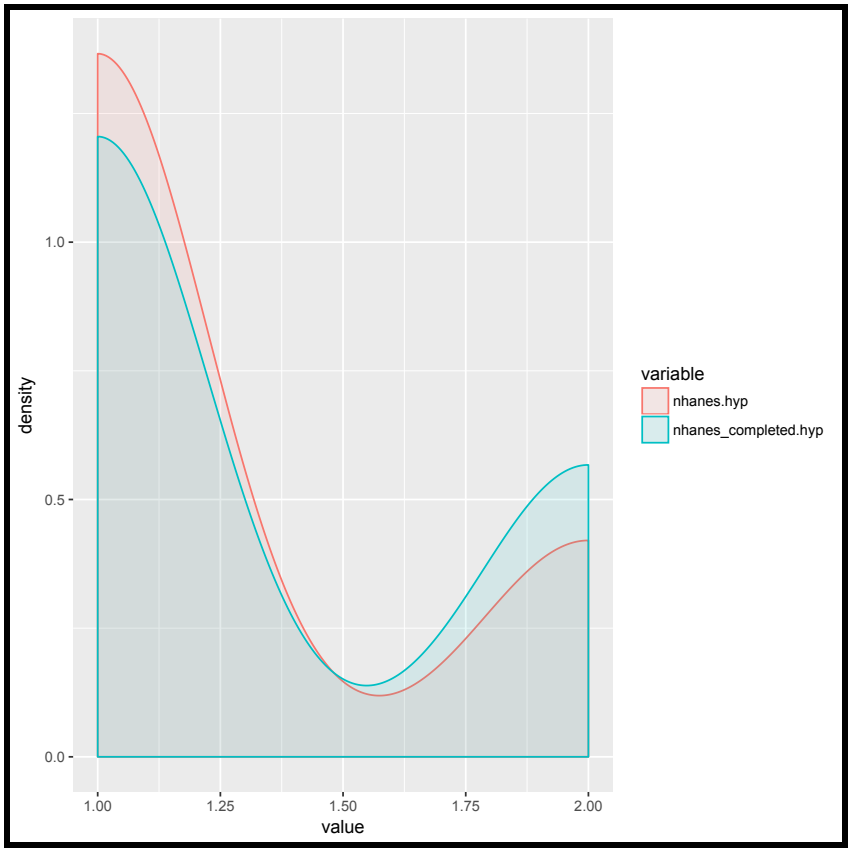

Ilustración 33: Variable hyp original comparada con hyp imputada

En la ilustración anterior, puede observarse que la imputación de la variable bmi es medianamente eficaz, no logra que la forma de la distribución de la variable imputada sea exactamente igual a la distribución de la variable original. Sin embargo, se acerca bastante, y hay que decir que mantiene bastante bien la variabilidad original, habida cuenta que el dataset tiene sólo 25 observaciones, y la variable hyp tiene 8 datos faltantes en esas 25 observaciones.

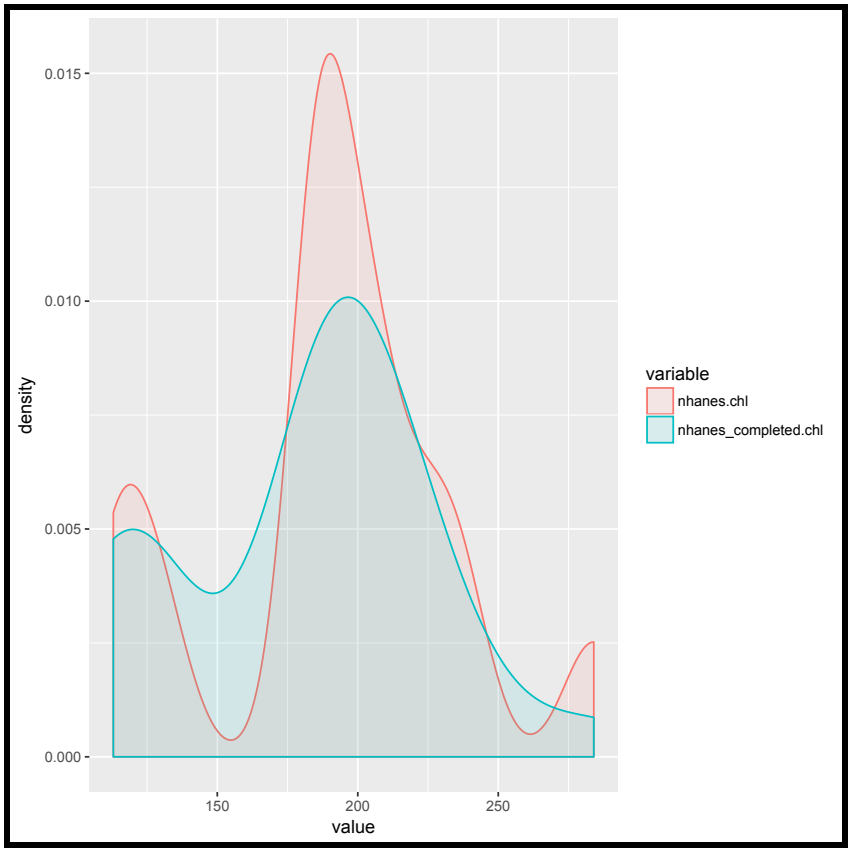

Ilustración 34: Variable chl original comparada con chl imputada

En la ilustración anterior, puede verse que la imputación de la variable chl no logra ser eficaz, esto es porque esa variable contiene 10 datos faltantes de 25 observaciones, y el algoritmo pmm no cuenta con suficientes datos reales para imputar en forma eficaz la media predictiva en los valores faltantes, y al mismo tiempo mantener la variabilidad.

4.5 Dataset "BOYS"

4.5.1 Visualización de patrón de datos faltantes del dataset BOYS

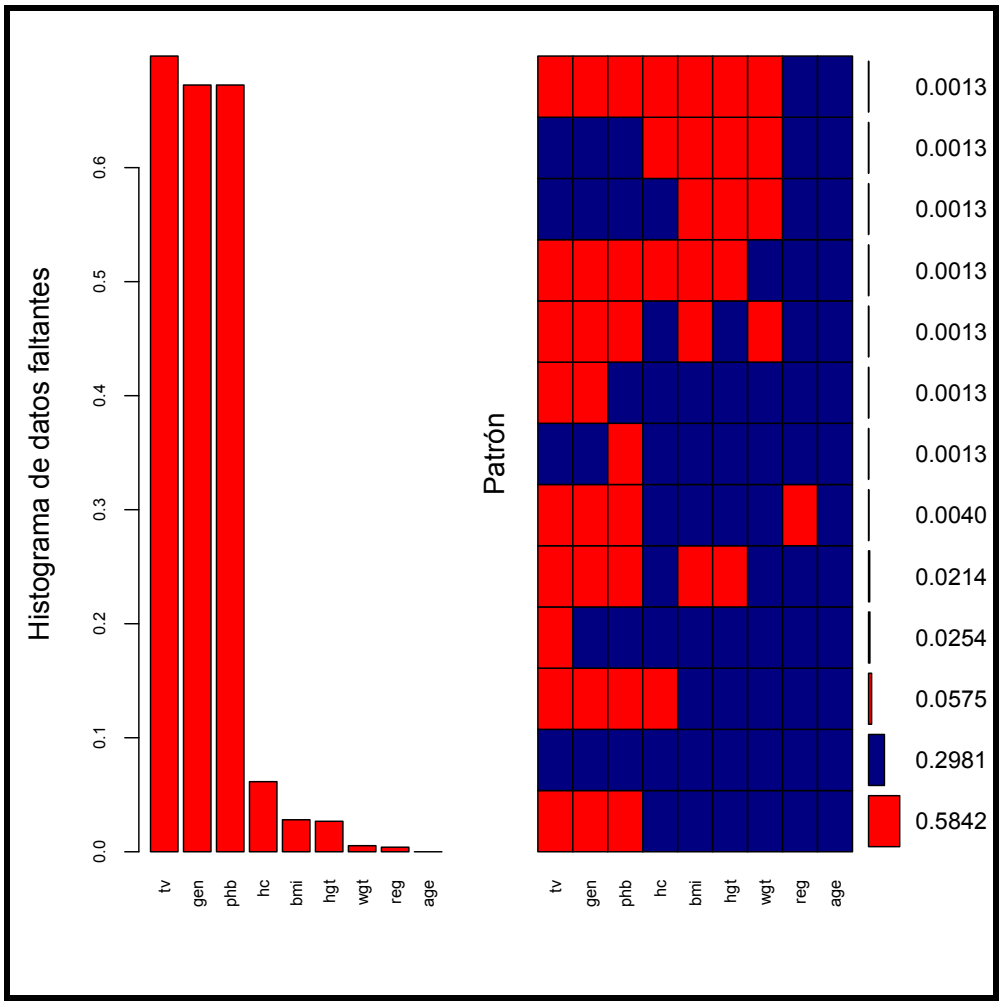

Ilustración 35: Gráfico con patrón de datos faltantes en dataset "BOYS"

Variables ordenadas por cantidad de datos faltantes	
Variable	Cantidad
tv	0.697860963
gen	0.672459893
phb	0.672459893
hc	0.061497326
bmi	0.028074866
hgt	0.026737968
wgt	0.005347594
reg	0.004010695
age	0.000000000

Imputación múltiple del dataframe BOYS:

> imp <- mice(boys, seed=5, print=FALSE)

Imputación efectiva del dataset BOYS:

> boys_completed = complete(imp)

4.5.2 Visualización de patrón de datos en dataset BOYS después de la imputación efectiva

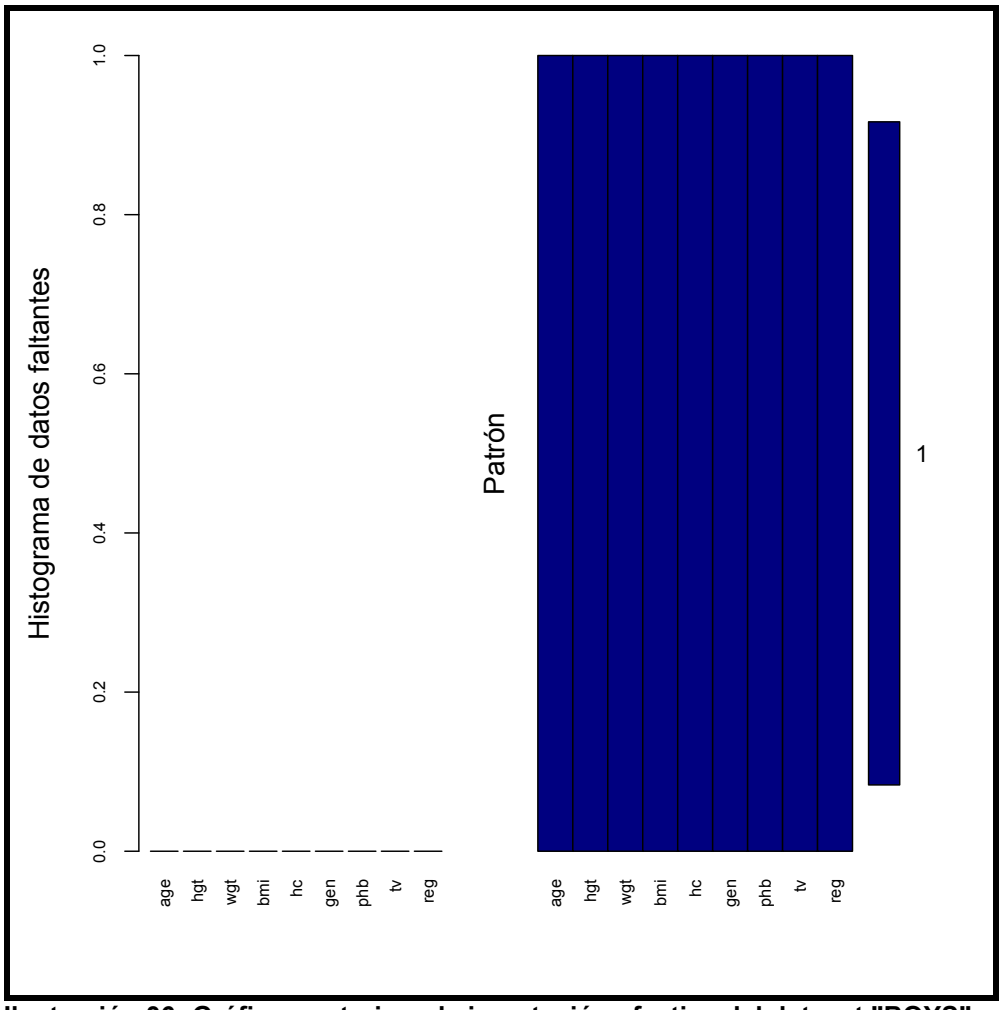

Ilustración 36: Gráfico posterior a la imputación efectiva del dataset "BOYS"

Si verifico el modelo de imputación empleado en cada una de las variables del dataset "BOYS" obtengo:

> imp$meth

age	hgt	wgt	bmi	hc	gen	phb	tv	reg
""	"pmm"	"pmm"	"pmm"	"pmm"	"polr"	"polr"	"pmm"	"polyreg"

El programa me informa que la primera variable "age" no fue imputada, ya que no contenía valores faltantes (Véase la ilustración 35). Luego las variables numéricas fueron imputadas mediante el método pmm (variables hgt, wgt, bmi, hc y tv). Las variables gen y phb, que son variables factores ordinales fueron imputadas mediante el método polr (polytomous regression - ordered), y por último la variable reg (variable factor) fue imputada mediante el método polyreg (polytomous regression - unordered).

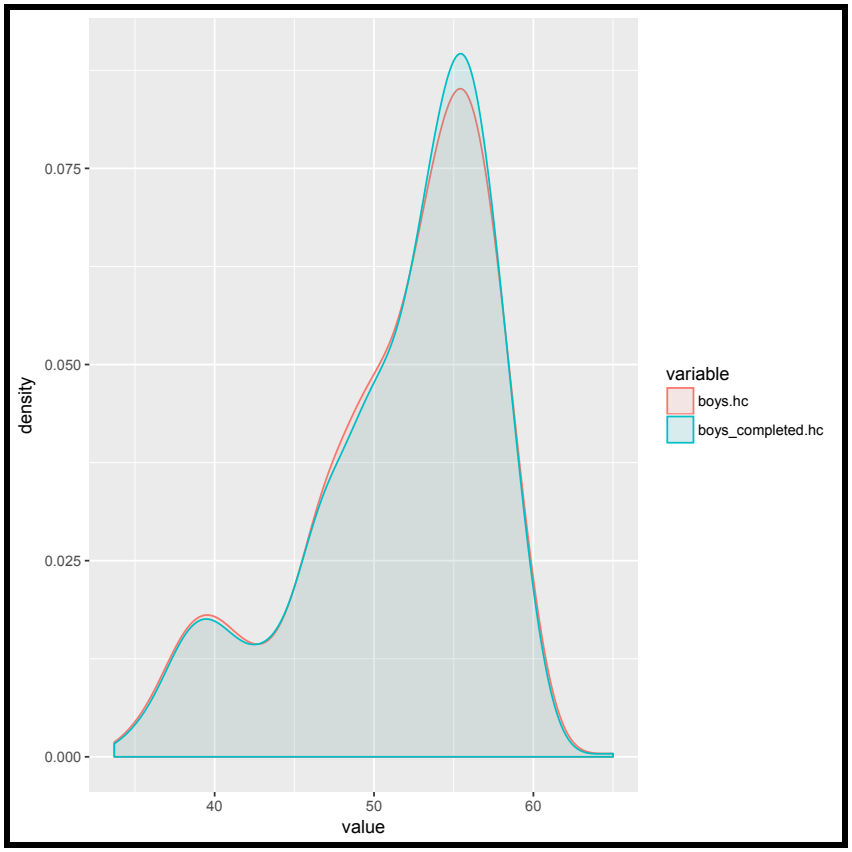

Ilustración 37: Variable hc original comparada con hc imputada

En la ilustración anterior, puede observarse que la imputación de la variable hc es muy eficiente, ya que la densidad de la distribución imputada es prácticamente igual a la variable original. Esto es así porque de 748 observaciones, en la variable hc, sólo 46 contienen datos faltantes (6.1 % de los valores).

4.6 Dataset "WALK"

4.6.1 Visualización de patrón de datos faltantes del dataset WALK

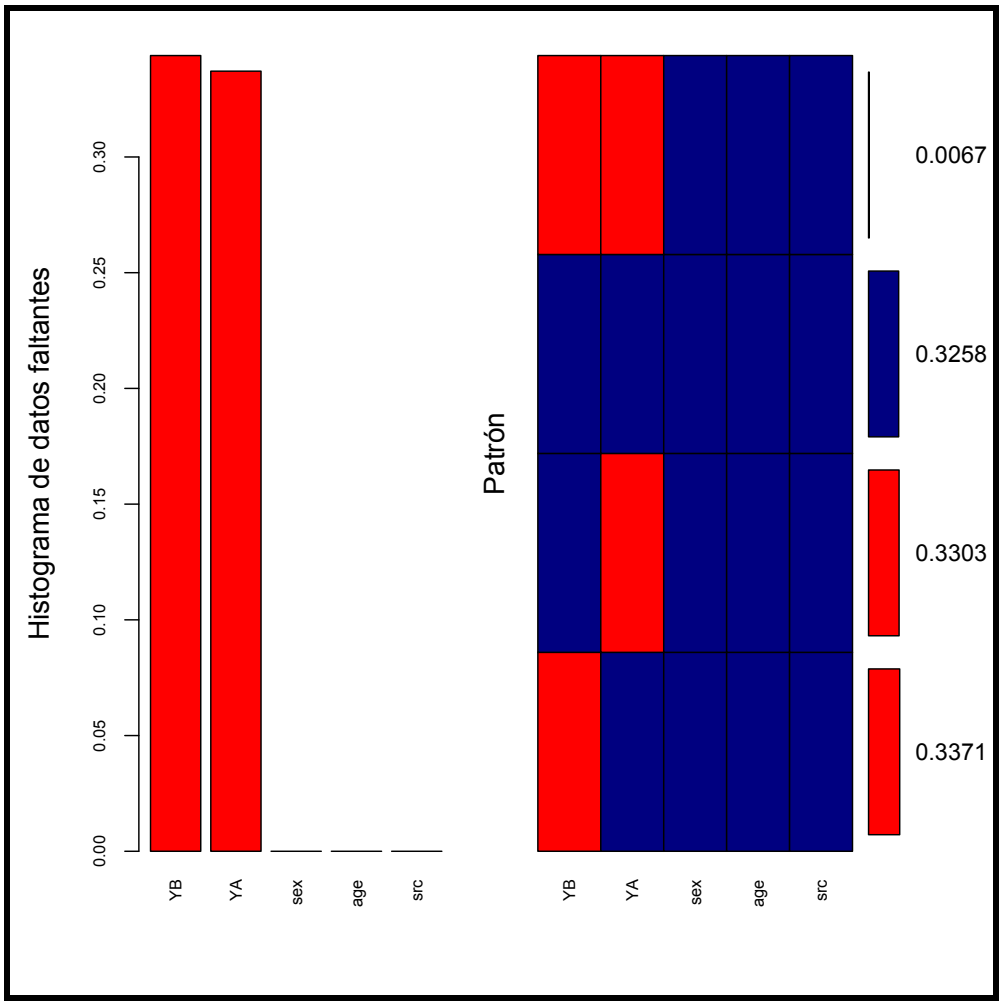

Ilustración 38: Gráfico con patrón de datos faltantes en dataset "WALK"

Variables ordenadas por cantidad de datos faltantes	
Variable	Cantidad
YB	0.3438202
YA	0.3370787
sex	0
age	0
src	0

Imputación múltiple del dataframe WALK:

> imp <- mice(walking, seed=5, print=FALSE)

Imputación efectiva del dataset WALK:

> walking_completed = complete(imp)

4.6.2 Visualización de patrón de datos en dataset WALK después de la imputación efectiva

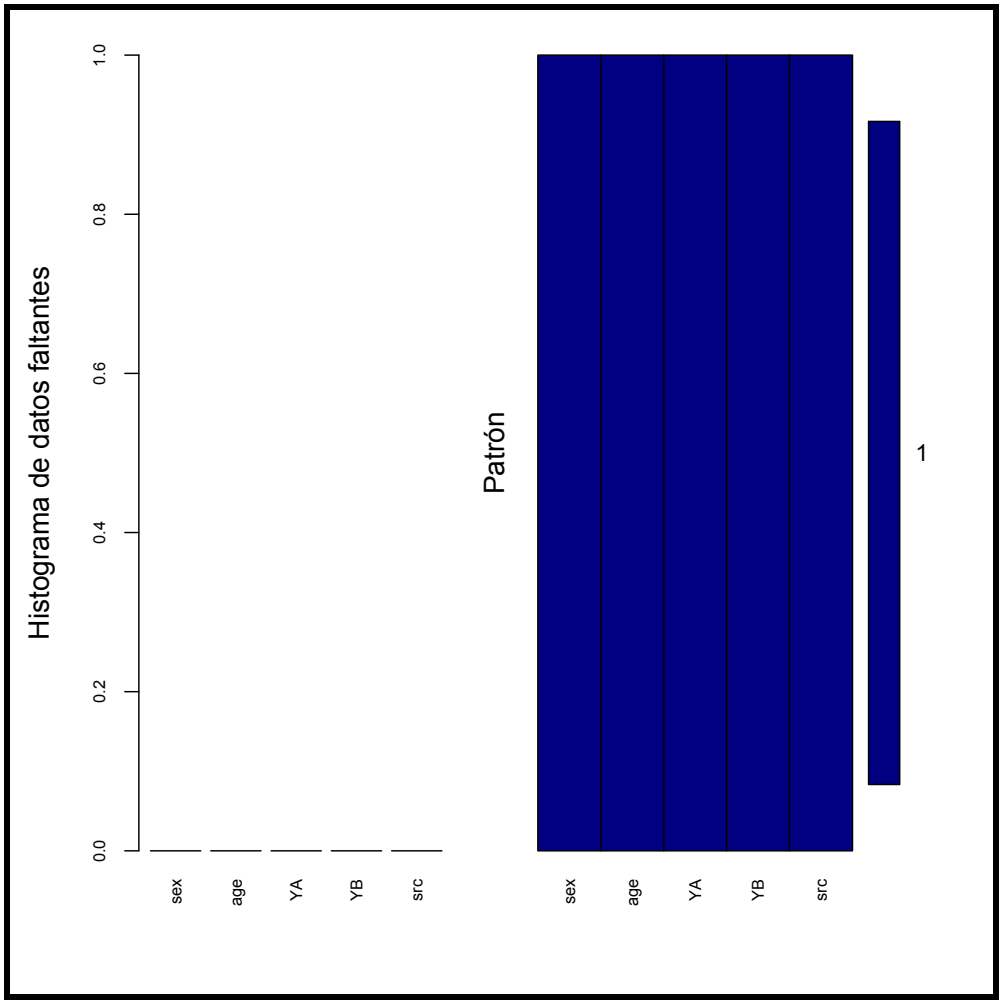

Ilustración 39: Gráfico posterior a la imputación efectiva del dataset "WALK"

Si verifico el modelo de imputación empleado en cada una de las variables del dataset "WALK" obtengo:

```
> imp$meth
  sex       age       YA        YB        src
   ""        ""      "polr"    "polr"     ""
```

El lenguaje R informa que las variables sex, age y src no fueron imputadas, ya que no contenían valores faltantes (Véase la ilustración 38). Las variables YA y YB, que son variables factores ordinales fueron imputadas mediante el método polr (polytomous regression - ordered.

4.7 Dataset "PATTERN4"

4.7.1 Visualización de patrón de datos faltantes del dataset PATTERN4

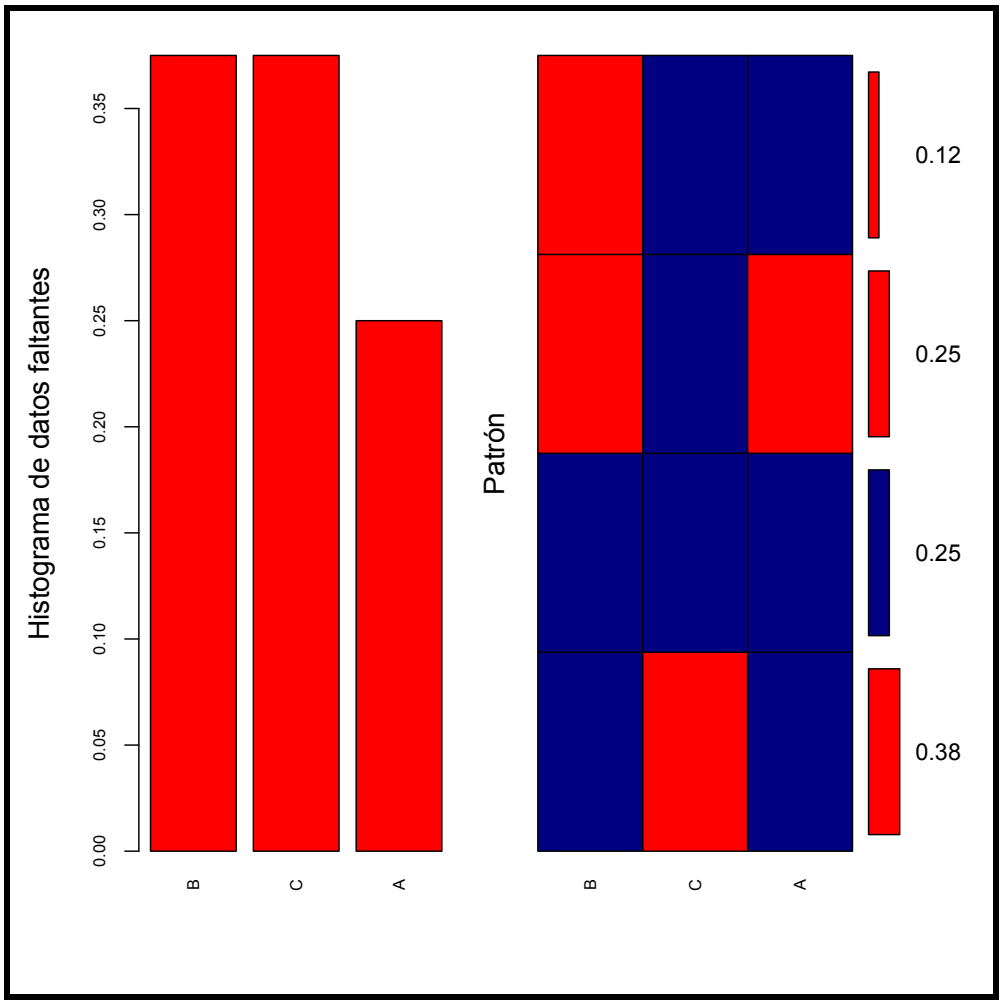

Ilustración 40: Gráfico con patrón de datos faltantes en dataset "PATTERN4"

Variables ordenadas por cantidad de datos faltantes	
Variable	Cantidad
B	0.375
C	0.375
A	0.250
hgt	0.026737968
wgt	0.005347594
reg	0.004010695
age	0.000000000

Imputación múltiple del dataframe PATT:

> imp <- mice(pattern4, seed=4, print=FALSE)

Imputación efectiva del dataset PATT:

> pattern4_completed = complete(imp)

4.7.2 Visualización de patrón de datos en dataset PATT después de la imputación efectiva

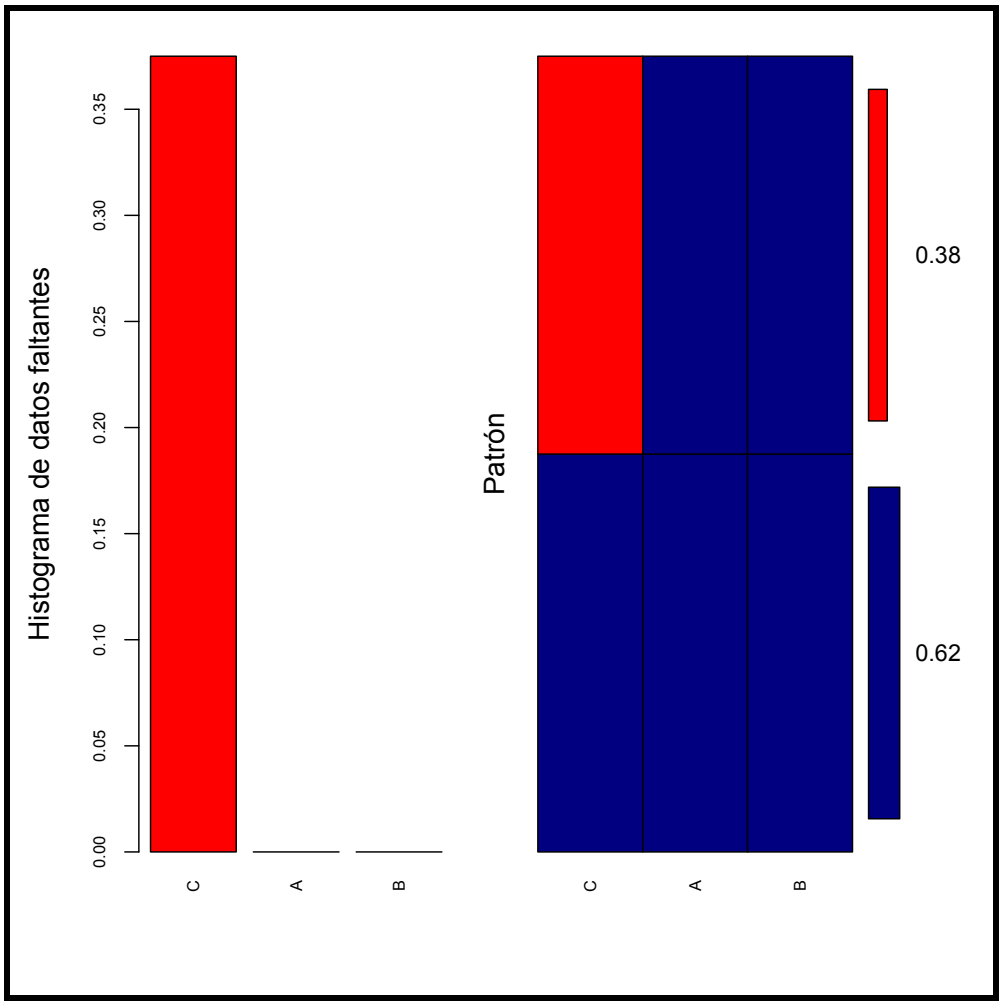

Ilustración 41: Gráfico posterior a la imputación efectiva del dataset "PATT"

En el gráfico anterior, puede observarse que no realizó la imputación efectiva de la variable "C", esto ha sucedido así porque el paquete MICE no ha encontrado la suficiente cantidad de registros con valores reales en esa variable, y por lo tanto no ha podido efectuar una PMM confiable, con lo cual no ha imputado nada en esa variable.

Si verifico el modelo de imputación empleado en cada una de las variables del dataset "PATT" obtengo:

> imp$meth
A B C

"pmm" "pmm" "pmm"

Efectivamente puede observarse que ha utilizado el método PMM en las tres variables, y en el caso de la variable C ese método no imputa, ya que el software interpreta que los valores imputados no serían robustos, habida cuenta la poca cantidad de valores reales en proporción a los valores faltantes, necesarios para predecir imputaciones confiables.

5. Discusión y debate sobre los resultados obtenidos.

Para establecer una discusión apropiada sobre los resultados obtenidos, se hace necesario comprender mínimamente como funciona el algoritmo PMM[28]:

El algoritmo PMM es una manera atractiva de realizar imputaciones múltiples para datos faltantes, especialmente para imputar variables cuantitativas que no están normalmente distribuidas.

En comparación con métodos estándar basados en regresión lineal y bajo suposición de distribución normal, el método PMM produce valores de imputación que son mucho más parecidos a los valores reales. Si la variable original es asimétrica ("skewed"), los valores imputados también serán asimétricos. Si la variable original se encuentra entre 0 y 100, los valores de imputación también se encontrarán entre 0 y 100. Y si los valores reales son discretos (como por ej. el número de hijos), los valores a imputar también serán discretos.

El método PMM ha sido definido hace bastante tiempo (Rubin 1986, Little 1988), pero sólo recientemente se encuentra disponible y además es práctico de utilizar. Originalmente, podía ser utilizado cuando una única variable contenía datos faltantes, o bien cuando el patrón de datos faltantes era monótono. En la actualidad, el método PMM está incluido en varios paquetes de software que implementan una aproximación a la imputación múltiple conocida como MICE (multiple imputation by chained equations: imputación múltiple por ecuaciones en cadena), regresión generalizada secuencial, o bien FCS (fully conditional specification: especificación completamente condicional).

Se encuentra disponible en varios paquetes estadísticos: SAS, Stata, SPSS, y R, los cuales permiten utilizar el método PMM para practicamente cualquier patrón de datos faltantes.

A continuación, una breve descripción de como funciona el método PMM.

Supóngase que hay una variable X que tiene algunos casos con datos faltantes, y un conjunto de variables z (sin valores faltantes) que se utilizan para imputar x. Se realiza lo siguiente:

1) Para los casos con valores faltantes, se estima una regresión lineal de x sobre z, produciendo un conjunto de coeficientes b.

2) Luego se crea una "distribución predictiva posterior" (aleatoria) de b, produciendo un nuevo conjunto de coeficientes b*. Típicamente esto sería una representación aleatoria de una distribución normal multivariante con media b y la matriz estimada de covarianzas de b (con una representación adicional aleatoria para la varianza residual). Este paso es necesario para producir suficiente variabilidad en los valores a imputar, y esto es lo común en todos los métodos "apropiados" para la imputación múltiple.

[28] Ref. [20].

3) Utilizando b*, generar valores predictivos para x para todos los casos, ambos con valores faltantes en x y con datos presentes.

4) Para cada caso con x faltantes, identificar un conjunto de casos con x observados cuyo valores predichos son cercanos al valor predicho para el caso con datos faltantes.

5) Para aquellos casos cercanos, elegir aleatoriamente uno y asignar su valor observado para sustituir el valor faltante.

6) Repetir los pasos 2 a 5 para cada dataset completado.

A diferencia de varios métodos de imputación, el propósito de la regresión lineal no es realmente generar valores de imputación. En realidad, sirve para construir una métrica para equiparar casos con valores faltantes, con casos similares con valores presentes.

Como se comenta en la referencia anterior, el método PMM es muy eficiente manteniendo la variabilidad original de los datos, y por ello, produce valores imputados mucho más parecidos a los reales, en comparación con otros métodos de imputación, simples o múltiples.

En los datasets analizados en este TFM, se ha podido apreciar la potencia de este algoritmo, lo cual ha quedado reflejado en los gráficos de densidad, comparando variables originales con variables imputadas.

Es evidente que los datasets a imputar deben tener una cantidad "coherente" de datos presentes, en proporción a los datos ausentes, y al mismo tiempo tener un número razonable de observaciones que permitan que el algoritmo despliegue su capacidad predictiva.

La cantidad "coherente" de observaciones con datos presentes, en relación a la cantidad de observaciones con datos ausentes, ha sido estudiada y analizada por diversos autores. Por ejemplo, en el artículo de Marshall et al. (2010), se declara que el método PMM de la librería MICE "produce las estimaciones menos sesgadas y las mejores mediciones en cuanto a performance, con lo cual es la aproximación preferida para realizar imputaciones múltiples, si menos del 50 % de las observaciones tienen datos faltantes y los datos faltantes no son MNAR".[29]

Y respecto de la cantidad razonable de observaciones (tamaño de la muestra), el método obtiene mejores resultados con muestras grandes, y no tan buenos resultados con muestras pequeñas.[30] Por ejemplo, con 1000 observaciones, con datos faltantes contenidos entre 100 y 500 observaciones, PMM obtuvo los mejores resultados. Cuando la cantidad de observaciones con datos faltantes fue menor a 50 (5%), los resultados fueron similares y aceptables entre distintos métodos de imputación. [31]

[29] Ref. [21], página 1.
[30] Ref. [15], página 74.
[31] Ref. [21], página 1.

6. Conclusiones

Como se ha comentado previamente en esta memoria, actualmente existen muchos estudios e investigaciones en los cuales se realizan análisis estadísticos, y estos estudios pueden realizarse fundamentalmente gracias a los avances del poder de cálculo de los ordenadores disponibles actualmente.

Se ha demostrado a lo largo del texto previo, que en estas investigaciones, por distintos motivos, es bastante común encontrar datos faltantes en los conjuntos de datos, resultados de las investigaciones mencionadas.

La mayoría de los investigadores no trata esta casuística de manera metódica.

Lamentablemente, la mayoría de ellos utilizan métodos básicos y simples, como son los métodos de borrado completo de casos o bien de imputación simple (moda, media, mediana).

Este tipo de soluciones simplistas, muchas veces perjudica el análisis posterior de los datos, ya que los datos pueden haberse sesgado, y las distribuciones resultantes pueden haber cambiado su variabilidad.

Se hace necesario en entornos de investigación, un mayor rigor en el tratamiento de datos, utilizando en forma apropiada métodos que mantengan la variabilidad original de los datos, así como también es necesario que los investigadores sean formados en cuestiones de fondo, como por ejemplo, que interpreten que si van a realizar imputaciones simples, deben primero comprobar si los datos cumplen con las suposiciones básicas, tales como: normalidad de distribución de los datos, homogeneidad de la varianza, independencia de las observaciones, y que los datos estén medidos como mínimo al nivel de intervalo (i.e., los intervalos conserven proporcionalidad).

Los investigadores, cuentan actualmente con herramientas de software, tales como el lenguaje R con sus 8000 paquetes, que facilitan mucho la utilización de opciones y métodos más sofisticados para resolver la casuística de datos faltantes de manera más coherente, razonable y estadísticamente correcta.

Como se ha demostrado en este TFM, el método de imputación múltiple, de la herramienta MICE, con su algoritmo de media predictiva, es un prodecimiento robusto para la implementación de una solución plausible para reemplazar los datos faltantes, en conjuntos de datos de estudios longitudinales.

7. Glosario

IQ: Quotient test: cociente de inteligencia (test de)

MI: multiple imputation: imputación múltiple (método de)

ML: maximum-likelihood: máxima probabilidad (método de)

MAR: Missing At Random: datos faltantes por causas aleatorias

NMAR: Not Missing At Random: datos faltantes por causas no aleatorias

KEEL: Knowledge Extraction based on Evolutionary Learning: Extracción de conocimiento basado en Aprendizaje Evolutivo (repositorio de datasets)

AMV: artificial missing values: valores faltantes introducidos de manera artificial

NA: Not availabe: No disponible. Denominación que se da a los valores faltantes, ya que evidentemente no son valores de los que se pueda disponer.

LD / CCA: Listwise deletion o complete case analysis: Eliminación inteligente por lista o Análisis de caso completo: se refiere a cuando se eliminan completamente los casos (registros u observaciones), cuando contienen valores faltantes en alguna de sus variables.

DL: Direct likelihood: Probabilidad directa: método de imputación de valores faltantes.

FIML: Full information maximum likelihood: Probabilidad directa de información completa: otro método de imputación de valores faltantes.

PD / ACA: Pairwise deletion, también llamado available case analisys: Eliminación inteligente por pares o Análisis de caso disponible: se refiere a cuando se eliminan completamente los casos (registros u observaciones), cuando contienen valores faltantes en alguna de sus variables.

MICE: Multivariate Imputation by Chained Equations: Imputación multivariante mediante ecuaciones en cadena.

VIM: Visualization and Imputation of Missing Values: Visualización e imputación de valores faltantes.

LOCF: Last observation carried forward: Última observación llevada hacia adelante: se refiere a que se utiliza para las imputaciones la última observación.

BOFC: Baseline observation carried forward: Observación de línea base llevada hacia adelante: se refiere que se utiliza para las imputaciones la observación de línea base.

8. Bibliografía

[1] Presentación: Handling Missing Data for Indicators - Susanne Rässler.
URL: https://www.uni-trier.de/fileadmin/fb4/projekte/SurveyStatisticsNet/KEI_UT_Raessler.pdf - Web accedida el 2/04/2017.

[2] Artículo: A Review of Methods for Missing Data - Therese D. Pigott - 2001 - DOI: 1380-3611/01/0704-353

[3] Libro: Statistical Analysis with Missing Data - 2da. Edición - Roderick J Little, Donald B. Rubin - 2002 - ISBN: 0-471-18386-5

[4] Libro: Clinical Trials with Missing Data - A Guide for Practitioners - Michael O'Kelly, Bohdana Ratitch - 2014 - ISBN: 978-1-118-46070-2

[5] Capítulo de libro: Data Analysis Using Regression and Multilevel-Hierarchical Models - 2006 - Capítulo 25 - Missing-data imputation (with R)

[6] Libro: Handbook of Missing Data Methodology - Geert Molenberghs et al. - 2015 - ISBN: 978-1-4398-5462-4

[7] Libro: Handling Missing Data in Ranked Set Sampling - Carlos N. Bouza-Herrera - 2013 - ISBN: 978-3-642-39898-8

[8] Libro: ISS-2012 Proceedings Volume On Longitudinal Data Analysis Subject to Measurement Errors, Missing Values, and/or Outliers - Brajendra C. Sutradhar et al. - 2013 - ISBN 978-1-4614-6870-7

[9] Libro: Missing Data - Analysis and Design - John W. Graham - 2012 - ISBN: 978-1-4614- 4017-8

[10] Libro: Missing Data and Small-Area Estimation - Modern Analytical Equipment for the Survey Statistician - Nicholas T. Longford - 2005 - ISBN: 978-185233-760-5

[11] Libro: Missing Data in Clinical Studies - Geert Molenberghs et al. - 2007 - ISBN: 978-0- 470-84981-1

[12] Libro: Semiparametric Theory and Missing Data - Anastasios A. Tsiatis - 2006 - ISBN: 978-0387-32448-7
[13] Artículo: An introduction to modern missing data analyses - Amanda N. Baraldi, Craig K. Enders - 2009 - DOI: 10.1016/j.jsp.2009.10.001

[14] Artículo: Much Ado About Nothing - A Comparison of Missing Data Methods - Nicholas J. Horton, Ken P. Kleinman - 2007 - DOI: 10.1198/000313007X172556

[15] Libro: Flexible Imputation of Missing Data - Stef van Buuren - 2012- ISBN: 978-1-4398-6825-6

[16] Libro: Missing Data Analysis in Practice - Trivellore Raghunathan - 2015 - ISBN: 978-1-4822-1193-1

[17] Libro: Applied Missing Data Analysis - Craig K. Enders - 2010- ISBN: 978-1-60623-639-0

[18] Repositorio de datos KEEL: J. Alcalá-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. García, L. Sánchez, F. Herrera. KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. Journal of Multiple-Valued Logic and Soft Computing 17:2-3 (2011) 255-287.
URL: http://sci2s.ugr.es/keel/datasets.php - Web accedida el 5/04/2017.

[19] Artículo: A study on the use of imputation methods for experimentation with Radial Basis Function Network classifiers handling missing attribute values: The good synergy between RBFNs and EventCovering method - Julián Luengoa, Salvador García, Francisco Herrera - 2010 - DOI: 10.1016/j.neunet.2009.11.014.

[20] Artículo: Predictive Mean Matching (Equiparación de media predictiva) - Paul Allison. URL: https://statisticalhorizons.com/predictive-mean-matching - Web sobre temas estadísticos - URL accedida el 14/05/2017.

[21] Artículo: Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model - A resampling study - Marshall et al. - 2010 - DOI: 10.1186/1471-2288-10-112.

9. Anexos

Anexo 1: Resumen de paquetes estadísticos ("packages") en R para la operación/imputación de datos faltantes

Para el lenguaje R se encuentran disponibles una gran cantidad de paquetes estadísticos, de los cuales algunos de ellos vienen en la instalación por defecto, mientras que otros pueden descargarse e instalarse de acuerdo a la necesidad del usuario. Al mes de enero del año 2016, se encontraban disponibles para instalar en R, 7801 paquetes adicionales. Los paquetes en general se pueden descargar del repositorio CRAN (Comprehensive R Archive Network), aunque algunos también pueden descargarse desde otros repositorios, tales como el repositorio de paquetes para Bioconductor y GitHub.

Los paquetes especializados en el tratamiento de datos faltantes son los siguientes:

Amelia: Implementa el algoritmo "Amelia II", el cual asume que los datos observados y los datos faltantes del dataset son normales multivariantes. Las imputaciones se realizan mediante el algoritmo EMB (expectation-maximization with bootstraping).

BaBoon: Provee dos variantes para la media predictiva bayesiana mediante bootstrap, permitiendo imputar múltiples datos faltantes. Este paquete se desarrolló originalmente para datos de encuestas, los algoritmos de imputación en el mismo se describen como robustos con respecto a la especificación del modelo de imputación.

Mi: Realiza un aproximación bayesiana para la imputación de los datos faltantes. El algoritmo de imputación ejecuta múltiples cadenas de Markov-Monte Carlo, para obtener iterativamente valores imputados a partir de distribuciones condicionales de datos observados e imputados. Además, contiene funciones para la visualización de patrones de datos faltantes en un dataset, así como para la evaluación de la convergencia de las cadenas de Markov-Monte Carlo.

Hmisc: Contiene varias funciones muy útiles para la imputación de valores faltantes, tales como agreImpute(), impute() y transcan().

Mice: Mice es una abreviatura que significa: multivariate imputation of chained equations (imputación multivariante de ecuaciones en cadena). Este paquete es probablemente el mejor de los disponibles, ya que implementa el método FCS (Fully Conditional Specification) de imputación múltiple.

Vim: Vim es una abreviatura que se refiere a: visualization and imputation of missing values: Visualización e imputación de valores faltantes. Es un paquete que permite, como su nombre indica, visualizar patrones de datos faltantes en los datasets, y luego proceder a la imputación de los mismos, mediante distintas técnicas.

Para este TFM, el paquete seleccionado para realizar las demostraciones de patrones de datos faltantes es el paquete VIM, y el paquete elegido para la realización de las imputaciones es el paquete MICE.

Además, para el lenguaje R existen otros paquetes altamente especializados para otras funcionalidades, tales como la generación de representaciones gráficas (ggplot, ggplot2,

igraph), la importación/exportación de datos (foreign), la generación de reportes (Knitr, Sweave), la administración y manejo de dataframes como si fueran bases de datos (Dplyr). Este último paquete se utiliza en este TFM para la generación de muestras (con menor cantidad de registros) de los datasets que se abordan.

Anexo 2: Anexo de prácticas en R para este TFM

Se adjunta a este trabajo el fichero "TFM_Anexo_de_Prácticas.rmd" (generado en lenguaje R y con formato RMarkdown), donde se efectúan las prácticas necesarias para la resolución de los ejercicios desarrollados a lo largo del texto.

Asimismo, se adjuntan los ficheros "TFM_Anexo_de_Prácticas.doc" y "TFM_Anexo_de_Prácticas.pdf", para una mayor comodidad de lectura y visualización de todas las sentencias de programación efectuadas.

La mayoría de los datasets que se utilizan en este TFM se encuentran disponibles en internet, de acuerdo a lo comentado y detallado en el apartado 3.1.

Sin embargo, también para comodidad del usuario que desee reproducir el código fuente de este TFM, se adjuntan los datasets utilizados en formato .rds (formato nativo de R):

a) HOC.rds (Dataset "Horse-Colic")
b) AIR.rds (Dataset "Air Quality")
c) BOYS.rds (Dataset "Boys")
d) WALK.rds (Dataset "Walking")
e) NHAN.rds (Dataset "Nhanes")
f) PATT.rds (Dataset "Pattern4")

Anexo 3: Anexo de planificación para este TFM

Se adjunta asimismo a este trabajo el fichero "TFM_Anexo_de_Planificación.doc", donde se detalla la planificación temporal calendarizada, con tareas e hitos, que se ha llevado a cabo para el desarrollo de este TFM.

En el anexo mencionado, se encuentran desglosadas las tareas con sus subtareas, así como los recursos materiales que fueron utilizados para la consecución del trabajo.

Por último, también se ha puntualizado en el anexo de planificación, un análisis de posibles incidencias y riesgos, con un plan de contingencia y mitigación, para las posibles eventualidades que pudieran repercutir negativamente.