Facultad de Ciencias Económicas y Estadística

Universidad Nacional de Rosario

Impacto de estrategias para el tratamiento de información faltante sobre la estimación de modelos de regresión de Cox

Lic. Luciana Carla Chiapella

Tesis presentada para obtener el grado de Doctora en Estadística

Directora:
Dra. María Eugenia Mamprin

Co-Directora:
Dra. Marta Quaglino

- 2019 -

Agradecimientos

Ante todo, quiero agradecer a Maru por la predisposición y la buena onda que recibí de su parte desde que le mandé aquel e-mail, hace ya varios años atrás. Por los almuerzos hablando de viajes, las recomendaciones sobre todos los aspectos de la vida (desde cómo pelear precios hasta cuestiones relacionadas con esta tesis) y por hacerme sentir siempre como una más de sus hijas. Por el tiempo y la confianza brindada en todo momento, gracias por tanto!

A Marta Q. por motivarme a seguir esta carrera de Doctorado y hacer que vuelva a estudiar en esta hermosa Facultad de Ciencias Económicas y Estadística. Gracias por la paciencia, el incentivo y el tiempo dedicado hasta el último minuto.

A la Dra. Adriana Torres y todas las integrantes del Área Farmacología de la FBIOyF por permitirme compartir el espacio de trabajo durante estos años.

Al Dr. Federico Daray y su equipo de trabajo por haber confiado en mi y haberme facilitado tan gentilmente su base de datos para realizar esta investigación.

A mis compañeros de doctorado Juli A.[1]*, Pau, Bel, Marquitos, Juli F., Nati, Euge y Gabi, porque no tengo dudas de que, sin ellos, esto hubiese sido imposible. Gracias por compartir el mate, los almuerzos, las reuniones para resolver las evaluaciones, los stickers de Whatsapp y los memes, las buenas noticias y las frustraciones. Pronto les tocará a ustedes estar en este lugar y tienen todo mi apoyo para lograrlo (si la más vieja del grupo pudo, ustedes también!).*

A Vani por estar hace más de 20 años bancándome con su amistad, más allá de las idas y venidas, los momentos dificiles, el calor o el frío extremo, gracias por tu apoyo constante. A Maru y Jesi por tratarme de loca por seguir estudiando pero aún así seguir permitiéndome compartir nuestros caminos. A mis cubanitas hermosas por sus locuras, su aguante y su buena onda, desde aquel viaje que nos juntó para siempre.

A mi vieja y mi hermano por estar siempre, incondicionalmente y por mirarme como si estuviera hablando en chino cuando cuento mi vida "académica" y mis tantas actividades. A mi abuela y mi tío, por sus locuras y su apoyo.

A todas las personas que me crucé durante estos años, que me motivaron a seguir en este camino, porque realmente es lo que me apasiona. Cada consejo, cada palabra de aliento, cada momento de ocio para despejarme de tanto trabajo, hicieron que esta tesis pueda ser finalizada y una etapa más concluya para dar lugar a lo que vendrá.

[1]Esta mención vale por 5312.

Resumen

En esta tesis se realiza una investigación sobre el impacto de distintos métodos para el tratamiento de datos faltantes sobre la estimación de los parámetros de modelos de regresión de Cox, cuando las pérdidas ocurren en variables explicativas mixtas.

Por medio de un estudio por simulación y otro por remuestreo a partir de datos de un caso real del área de Psiquiatría, se avanza en el estudio de propiedades distribucionales de los estimadores de los parámetros, obtenidos mediante datos completos y a partir de matrices imputadas siguiendo distintas estrategias. Se analizan comparativamente el error cuadrático medio y el sesgo de los estimadores tanto en relación a la esperanza como a la varianza, correspondientes a la distribución asintótica teórica de los estimadores máximo-verosímiles a partir de matrices completas. También se evalúa la forma de la distribución y la capacidad predictiva del modelo estimado. Finalmente, se aplican para el caso-estudio que motivó esta tesis, las recomendaciones surgidas de los resultados de los mencionados estudios.

Los resultados obtenidos permiten obtener resultados generales para diversos escenarios, es decir, para distintos tamaños de muestra y diferentes porcentajes y mecanismos de pérdida. Esto facilita la elección de estrategias adecuada para tratar información faltante en situaciones complejas.

Índice general

1. Introducción

Resulta frecuente en todo estudio de investigación, independientemente de su diseño metodológico, la pérdida o no disponibilidad de una proporción variable de los datos correspondientes a los sujetos o unidades seleccionadas [1]. Los motivos por los que se producen los datos faltantes son múltiples y se encuentran ligados, en general, a la naturaleza del problema y al campo de estudio. Por ejemplo, en el área de Medicina, cuando se realizan estudios retrospectivos basados en registros preexistentes como pueden ser las historias clínicas, es común que no conste alguno de los datos requeridos para un conjunto de individuos, especialmente cuando los datos recolectados rutinariamente son usados a posteriori con otros objetivos [2]. Cuando la información se obtiene mediante preguntas directas a los sujetos, el desconocimiento o el no recuerdo por parte del respondente de alguno de los aspectos de interés, constituye otro de los motivos que puede llevar a la pérdida de datos. Cuando se trata de cuestionarios extensos y/o autoadministrados, es posible que los individuos eludan brindar su respuesta [3, 4]. En el campo de la Psiquiatría estas situaciones son habituales, más aún debido a la indagación sobre aspectos sensibles y al estado de vulnerabilidad y fragilidad de los pacientes y sus familiares directos, quienes suelen ser consultados a fin de obtener información de interés sobre el sujeto padeciente de una determinada conducta o enfermedad [5, 6]. Esta fue una de las características surgidas en un estudio llevado a cabo por el Instituto de Farmacología de la Facultad de Medicina de la Universidad de Buenos Aires. El propósito del mismo fue evaluar los factores influyentes sobre el tiempo hasta el reintento de suicidio en pacientes con trastorno límite de la personalidad, que habían sido ingresados por intento de quitarse la vida [7]. El conjunto de mediciones recogidas, que presentaba datos faltantes, era de naturaleza mixta, algunas dicotómicas, otras nominales, otras continuas con asimetrías, con pérdidas que podían suponerse aleatorias y otras que podían intuirse relacionadas con la misma medición, planteando un escenario complejo para cuyo análisis no se encontraban

recomendaciones metodológicas publicadas.

La problemática del tratamiento de datos incompletos es tema de interés en investigación desde la década de 1970 a partir de los trabajos de Dempster *et al.* [8], Heckman [9] y Rubin [10], entre otros, los cuales permitieron iniciar una importante línea de estudio. Sin embargo, aún con los numerosos avances realizados, no están respondidos todos los interrogantes, ya que la influencia de los datos faltantes y su imputación puede no ser la misma en los diferentes análisis estadísticos y, en consecuencia, se requiere de investigaciones particulares para cada uno de ellos.

Los datos perdidos, según sus características y proporción, pueden afectar en forma importante tanto la precisión como la validez de las estimaciones [11]. Es por esto que el desarrollo de métodos para la *imputación de datos perdidos* y el estudio de su eficiencia constituyen un campo de investigación en sí mismo y requieren especial atención [12, 13]. Las pérdidas pueden corresponder a los valores que toman algunas de las variables en algunos sujetos o, incluso, a la totalidad de los datos para algunos de ellos. Little & Rubin [14] clasifican a los datos perdidos en tres clases según los *mecanismos de pérdida*, es decir, según el modo en que se generan: "perdidos completamente al azar" (*missing completely at random*, MCAR) cuando la probabilidad de pérdida de una observación para todos los individuos es la misma y no depende de los valores de otras variables, "perdidos al azar" (*missing at random*, MAR) cuando la probabilidad de pérdida de una observación de un individuo depende de la información observada para otra u otras variables y "perdidos no al azar" (*missing not at random*, MNAR) si la probabilidad de que la observación de un individuo esté perdida está relacionada con los mismos valores perdidos.

La propuesta más simple frente a la aparición de información incompleta es omitir del análisis a aquellos individuos que presentan datos faltantes. Este método se denomina a menudo "análisis de casos completos" (*complete-case analysis*, CCA). Cuando el tamaño muestral es grande, la proporción de datos faltantes es pequeña y el mecanismo de pérdida puede suponerse MCAR, esta técnica puede considerarse una buena opción [15]. Sin embargo, el empleo de CCA resulta en una disminución del tamaño muestral o poblacional y en la consecuente pérdida de precisión en las estimaciones [16]. Aún reconociendo las limitaciones de esta propuesta, ella constituye un método de referencia para contrastar el comportamiento de otros métodos superadores.

Los métodos desarrollados a fin de tratar los datos faltantes pueden ser agrupados en tres grandes categorías: *métodos de eliminación*, en los cuales las observaciones o las variables con datos faltantes son excluídas del análisis, tal como en el caso de CCA; *métodos basados en modelos* o *métodos condicionales*, donde se utilizan procedimientos de máxima verosimilitud adaptados para estimar los parámetros de un modelo definido para datos completos y *métodos de imputación*, en los que los datos faltantes son reemplazados por valores posibles, determinados a partir de los datos disponibles para la variable en cuestión [17]. Sin embargo, solo algunos de los métodos de imputación desarrollados permiten asignar valores simultáneamente a datos faltantes de variables cualitativas y cuantitativas, restringiendo de esta forma las técnicas de imputación disponibles cuando se trabaja con matrices de información con variables mixtas. En particular, los modelos de regresión de Cox suelen admitir como variables explicativas del tiempo hasta el evento, una combinación de distintos tipos de variables.

Dentro de los métodos de imputación, la primera propuesta fue la *sustitución por la media*, que consiste en asignar a los datos faltantes la media de los valores observados en la variable correspondiente. Es un método simple que no depende de los datos observados para otras variables pero solo puede usarse para variables cuantitativas y puede verse afectado por la falta de simetría en la distribución de la variable o por la presencia de valores extremos [16]. En estos casos, los valores faltantes en una variable pueden reemplazarse por su mediana o, si la variable es cualitativa, pueden ser reemplazados por su modo. En la presente tesis, esta técnica se denomina "imputación por valores representativos" (*representative values imputation*, RVI). Aunque proporciona un método rápido y simple para imputar valores perdidos, subestima la varianza de la variable con pérdidas, compromete la relación entre las variables y sesga las estadísticas de resumen [2]. Además, implica la suposición de que los datos perdidos son MCAR.

Otro método de imputación es "k-vecinos cercanos" (*k-nearest neighbours*, KNN), propuesto por Troyanskaya *et al.* en 2001 [18]. KNN es un método eficiente para imputar los datos faltantes, que consiste en asignar a cada dato perdido en un individuo un valor obtenido a partir de la información disponible de los k individuos más cercanos o parecidos a este (*donantes* o *vecinos*). El parecido se establece en función de todas las variables que tengan mediciones completas. Dada la capacidad de sustituir los datos faltantes con valores plausibles que se suponen cercanos al valor real, KNN ayuda a conservar la estructura

original de los datos y evita distorsionar la distribución de la variable imputada [16]. Para su aplicación, es necesario elegir correctamente el número k de vecinos, la medida de distancia que se utilizará para encontrarlos y el valor que se empleará para completar los datos faltantes (media, mediana, modo, otros). Este método ofrece la ventaja de permitir la imputación simultánea de variables cuantitativas y cualitativas utilizando una distancia como la de Gower para encontrar los donantes [19, 20].

Una propuesta más reciente es *missForest* (MF) [21]. Este método hace uso de modelos de bosques aleatorios [22], que son altamente flexibles y versátiles, para lograr la imputación de valores faltantes. MF crea un modelo de bosque aleatorio para cada variable con pérdidas empleando el resto de las variables presentes en el conjunto de datos y lo utiliza para imputar los valores faltantes. Esto se hace en forma cíclica para todas las variables con pérdidas y el proceso se repite iterativamente hasta que se alcanza un criterio de detención. Entre las ventajas de utilizar el modelo de bosque aleatorio se mencionan que puede considerar simultáneamente variables cuantitativas y cualitativas, no requiere supuestos sobre la estructura de los datos y proporciona una estimación del error con validación interna.

Los métodos de imputación simple reemplazan cada valor faltante por un valor adecuado para la variable en cuestión, pero algunas estrategias consisten en hallar m valores plausibles para imputar cada dato faltante. Estos métodos, denominados de "imputación múltiple" (*multiple imputation*, MI), generan m conjuntos de datos completos, cada uno de los cuales es considerado para realizar estimaciones de parámetros de interés utilizando CCA. Las estimaciones obtenidas en base a cada conjunto de datos son combinadas al final de proceso para arribar a estimaciones que tienen en cuenta la incertidumbre producida por las observaciones faltantes [23]. La "imputación múltiple por ecuaciones encadenadas" (*multiple imputation by chained equations*, MICE) es un enfoque para generar imputaciones basada en un conjunto de modelos de regresión, uno para cada variable con valores faltantes [24]. Una característica importante de MICE es su capacidad para tratar diferentes tipos de variables, ya que cada una de ellas se imputa utilizando su propio modelo. Así, cada variable con datos faltantes es considerada como una variable respuesta en un modelo adecuado, que considera como explicativas a las variables restantes. Si bien este método es ampliamente utilizado y recomendado en diversas publicaciones [25, 26, 27], su aplicación práctica implica un adecuado estudio del cumplimiento de los supuestos

involucrados en cada uno de los modelos de regresión elegidos [28, 29].

La selección del mecanismo de imputación a utilizar no puede aislarse del contexto en el que se está trabajando, ya que deben considerarse los motivos por los cuáles existen dichos datos faltantes (mecanismos de pérdida), las características de la o las variables sobre la que se realizarán las imputaciones, las técnicas estadísticas con la que se analizarán los datos y otros aspectos vinculados con cada problema en particular. Para generar inferencia estadística válida, la imputación debe considerarse parte del proceso de investigación con el propósito de arribar a conclusiones sustentadas en evidencia empírica sólida [10].

Si bien son numerosos los trabajos que presentan técnicas para imputar datos faltantes, son pocos los que comparan el efecto que producen los distintos métodos sobre la estimación de los parámetros de un modelo de regresión de Cox y, además, suelen limitarse a considerar variables explicativas de un solo tipo, ya sean cualitativas o cuantitativas [17, 30, 31, 32, 33]. Más aún, la mayoría de los artículos publicados se restringen a la comparación de los resultados obtenidos bajo el uso de CCA y uno o dos métodos de imputación bajo ejemplos particulares, limitando la evaluación de los resultados a un escenario predeterminado sin permitir una generalización [31, 34, 35]. Otros trabajos dentro de esta línea, se centran en comparar solamente la capacidad predictiva de los modelos estimados luego de la imputación por medio de diversas técnicas [36, 37]. Son escasos los trabajos que evalúan comparativamente situaciones que involucran en forma simultánea variables mixtas, distintos mecanismos y porcentajes de pérdida y diferentes tamaños de muestra, aplicando una variedad de métodos de imputación que permitan arribar a recomendaciones más generales. Además, en estos casos, el enfoque se restringe especialmente a la comparación entre el valor imputado y el valor real, sin profundizar en las propiedades de los estimadores involucrados.

En esta tesis se propone realizar estudios por simulación que permitan evaluar el impacto del empleo de distintos métodos de imputación sobre las propiedades distribucionales de los estimadores de los coeficientes de un modelo de regresión de Cox. Los estudios planteados consideran distintas situaciones a partir de un modelo especificado suponiendo variables explicativas mixtas provenientes de escenarios amplios. También se realizan estudios por remuestreo a partir de los datos de una investigación de particular interés sobre el tiempo hasta el reintento de suicidio en pacientes con trastorno límite

de la personalidad [7], motivadora de esta tesis. En todo el trabajo se enfocan y evaluan también aspectos de interés en las aplicaciones, como la facilidad de la implementación de los métodos de imputación, la flexibilidad de los mismos en cuanto a los supuestos que requieren respecto a la estructura de los datos y los tiempos de ejecución computacional requeridos por cada uno ellos.

El contenido de esta tesis se encuentra organizado en seis capítulos, un apartado de Bibliografía y los anexos. Los dos primeros capítulos corresponden a esta Introducción y al enunciado de los Objetivos de la investigación.

El Capítulo 3 está dedicado a una síntesis de los temas involucrados en la investigación, conceptos y definiciones vinculados al análisis estadístico de datos de supervivencia, incluyendo el caso particular de los modelos de riesgos proporcionales de Cox y de la estimación de sus parámetros, conceptos y definiciones sobre los datos faltantes y su tratamiento, y una revisión de los antecedentes referidos al uso de modelos de Cox frente a datos con información faltante.

En el Capítulo 4, se presenta el diseño y los resultados de un estudio por simulación que considera un modelo de regresión de Cox que es usado como modelo poblacional, adoptando distintas distribuciones para las variables explicativas que abarcan un espectro amplio de variables mixtas, a partir del cual se obtienen muestras de distintos tamaños y se generan pérdidas con distintos mecanismos y porcentajes. En estos escenarios se aplica una variedad de cinco métodos para el tratamiento de datos faltantes (CCA, RVI, KNN, MF y MICE) y se compara su eficiencia a partir de un conjunto de medidas especialmente definidas que enfocan la calidad de las estimaciones, la capacidad predictiva del modelo y la precisión de las imputaciones.

En el Capítulo 5 se realiza un estudio por remuestreo en forma similar al planteado en el Capítulo 4, a partir de un caso real que considera específicamente los resultados del estudio sobre pacientes en riesgo de suicidio realizado en el Instituto de Farmacología. El remuestreo se realiza a partir del modelo estimado con la información completa del estudio. Finalmente, se presentan los resultados del análisis de la información del caso real considerado, siguiendo los lineamientos que surgen de las recomendaciones derivadas de los capítulos anteriores.

En el Capítulo 6 se discuten los resultados obtenidos, en relación con los hallazgos reportados por otros autores. Además, se establecen las conclusiones y resultados originales

más importantes de esta investigación, a partir de los cuales se realizan recomendaciones de uso de los distintos métodos de imputación considerados. Finalmente se plantean posibles líneas futuras de investigación en la temática abordada.

2. Objetivos

2.1. Objetivo General

Esta tesis plantea como objetivo evaluar el impacto de diferentes métodos para el tratamiento de información faltante sobre la estimación de los modelos de regresión de Cox en estudios de tiempos de sobrevida cuando intervienen variables explicativas mixtas considerando una amplia variedad de escenarios que combinan distintos tamaños muestrales, mecanismos de pérdida, incidencia de las pérdidas respecto a los tamaños de muestra y distribución de las pérdidas.

2.2. Objetivos Específicos

1. Profundizar en el estudio sobre metodología que permita el tratamiento de información faltante en contextos de investigación que requieran realizar análisis de supervivencia y exista pérdida de información debida a distintas causas.

2. Diseñar estudios comparativos que permitan evaluar la eficacia de métodos para el tratamiento de información faltante y derivar recomendaciones de uso según el contexto de variables explicativas disponibles, porcentaje y mecanismo de pérdida y tamaño muestral.

3. Definir medidas de comparación adecuadas que consideren distintos aspectos de la eficiencia de los métodos de imputación.

4. Comparar los métodos de tratamiento de información faltante CCA, RVI, KNN, MF y MICE según las medidas definidas en 3. frente a diferentes escenarios.

5. Analizar la robustez de los métodos de imputación frente a cambios en aspectos particulares como número de donantes considerados al utilizar KNN y la consideración del *tiempo* y el *estado* como variables informativas para completar los valores faltantes.

3. Marco conceptual

Este capítulo tiene como objetivo focalizar al lector en los aspectos teóricos fundamentales sobre los que se apoya esta tesis. Se compone de cuatro secciones, de las cuales la primera repasa los conceptos básicos del análisis de datos de supervivencia y de los modelos de regresión de Cox, incluyendo la estimación de sus parámetros y las propiedades distribucionales de sus estimadores. En la segunda sección, se mencionan conceptos relacionados con datos perdidos y se realiza una clasificación detallada de los métodos de imputación para completar la matriz de información. En la tercer sección se desarrollan en profundidad métodos que permiten asignar valores en forma simultánea a variables mixtas, enfatizando en los que se utilizarán en este trabajo de investigación. Por último, se reseñan las publicaciones científicas revisadas que abordan la problemática de datos faltantes en el contexto del análisis de regresión de Cox, a fin de poner en foco los resultados que anteceden a los obtenidos en la presente tesis.

3.1. Análisis de supervivencia

Se denomina *análisis de supervivencia* al conjunto de técnicas utilizadas para analizar datos donde la variable de interés es el tiempo t transcurrido desde un momento inicial bien definido (t_0) hasta la ocurrencia de un evento [38]. Si bien el concepto de supervivencia, en ciencias de la salud, se relaciona con la finalización de la vida de un sujeto, puede tratarse de eventos diferentes, por ejemplo: la curación, aparición de la enfermedad, falla de un producto, recepción de un servicio, etc. En los últimos casos, los tiempos de sobrevida suelen indicarse como *tiempo hasta el fallo* o *tiempo de espera*.

Generalmente, los *tiempos hasta la ocurrencia de un evento* o *tiempos de supervivencia* observados para un conjunto de unidades tienen una distribución con asimetría positiva. Como consecuencia, no es razonable asumir que provienen de una distribución

Normal, por lo que las técnicas habituales para el análisis de variables continuas pueden no resultar adecuadas. Una alternativa de análisis podría ser transformar los valores a fin de obtener un comportamiento aproximadamente Normal, sin embargo existe otra característica a tener en cuenta: en los estudios de supervivencia, los datos pueden estar *censurados*.

El tiempo de supervivencia de un individuo se denomina *censurado por derecha* cuando el evento no ha sido observado en ese individuo al momento de la finalización del estudio o bien hasta que se produce una pérdida de seguimiento del individuo. En estos casos, solo se conoce que el sujeto no ha presentado el evento hasta el momento de la censura. Otro tipo de censura es la *censura por izquierda*, que ocurre cuando se conoce que el evento ha ocurrido antes de cierto tiempo de observación pero no se conoce con exactitud cuándo ha sucedido, es decir, solo se sabe que el tiempo hasta el evento es menor al tiempo máximo de observación. Finalmente, existe la *censura por intervalo*. En ese caso, se sabe que el evento ha ocurrido dentro de un intervalo temporal, pero no se sabe con exactitud cuándo ha sucedido [39]. Los datos censurados contribuyen con información valiosa y ellos no deben ser omitidos en el análisis.

Un supuesto importante que se realiza durante el análisis de supervivencia es que el tiempo de supervivencia t es independiente de cualquier mecanismo que causa que dicho tiempo sea censurado en el momento c, siendo $c < t$. Esto implica que si se considera un grupo de individuos, los cuales tienen los mismos valores de las variables relevantes para explicar el tiempo de supervivencia, un individuo cuyo tiempo de supervivencia es censurado en el tiempo c debe ser representativo de todos los otros individuos del grupo que no han presentado el evento hasta ese momento. Un sujeto cuyo tiempo de supervivencia es censurado será representativo de los sujetos en riesgo de presentar el evento en el tiempo de censura si el proceso de censura es aleatorio.

Similarmente, cuando los datos de supervivencia se analizan en un punto determinado del calendario o luego de un intervalo de tiempo fijo a partir del momento de ingreso al estudio, el pronóstico de los individuos que no han presentado el evento en dicho punto puede ser considerado independiente de la censura, siempre que el tiempo del análisis sea especificado antes de examinar los datos. Sin embargo, este supuesto no puede ser realizado en casos en los que el motivo de la censura está relacionado con las variables en estudio, por ejemplo, cuando un determinado tratamiento médico provoca deterioro físico

en un individuo. Este tipo de censura se denomina *censura informativa* y debe ser tenida en cuenta cuando se analizan los datos, ya que los métodos habitualmente utilizados en el análisis de supervivencia consideran que la censura es *no informativa* [38]. En esta tesis, la censura se considera siempre *no informativa*.

3.1.1. Funciones de supervivencia y riesgo

Definición 3.1. *Sea T una variable aleatoria que representa el tiempo de supervivencia y que puede asumir valores reales no negativos, $t \geq 0$. Sea $f(t)$ la función de densidad de probabilidad de T y $F(t) = P(T \leq t) = \int_0^t f(u)\, du$ su función de distribución acumulada. Se denomina* ***función de supervivencia*** *para la variable aleatoria T a $S(t) = P(T > t) = 1 - F(t)$.*

La *función de supervivencia* representa la probabilidad de que un individuo sobreviva, desde el tiempo de origen, hasta más allá de cierto tiempo t. Dado que S es una probabilidad, está acotada entre 0 y 1, y como T no puede tomar valores negativos, luego $S(0) = 1$. Además, S es una función monótona decreciente. Considerando estas restricciones, S puede asumir una amplia variedad de formas.

Definición 3.2. *Sea T una variable aleatoria que representa el tiempo de supervivencia y que puede asumir valores reales no negativos, $t \geq 0$. Si $\Delta t > 0$, se denomina* ***función de riesgo*** *a: $h(t) = \lim_{\Delta t \to 0} \left[\frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \right]$.*

La *función de riesgo* o *tasa de falla instantánea* es comúnmente utilizada para expresar el riesgo instantáneo de muerte o de presentación del evento de interés de un individuo en un momento dado, t, condicionada a que ese individuo ha sobrevivido hasta ese momento. Como el tiempo es una variable continua, la probabilidad de que el evento ocurra exactamente en el tiempo t es necesariamente cero, por lo cual se considera la probabilidad de que dicho evento ocurra en un intervalo muy pequeño de tiempo, comprendido entre t y $t + \Delta t$. Si un individuo presentó el evento antes del tiempo t, entonces ya no está en riesgo de presentarlo luego de ese tiempo, por lo tanto interesa considerar la probabilidad de supervivencia condicionada a que el individuo no haya presentado el evento hasta ese momento. Esta probabilidad condicional es expresada como una probabilidad por unidad de tiempo al dividirla por el intervalo de tiempo, Δt. Luego, la función de riesgo, $h(t)$,

está dada por el límite de dicho cociente cuando Δt tiende a cero.

A partir de la Definición (3.2), se tiene que $h(t)\Delta t$ es aproximadamente la probabilidad de que un individuo presente el evento en el intervalo $[t, \Delta t)$, condicionada a que el individuo haya sobrevivido al tiempo t. Por esta razón, la función de riesgo es habitualmente interpretada como el riesgo de presentar el evento en el tiempo t.

En base a la definición de probabilidad condicional y distribución acumulada, se tiene que:

$$P(t \leq T < t + \Delta t \mid T \geq t) = \frac{P(t \leq T < t + \Delta t)}{P(T \geq t)} = \frac{F(t + \Delta t) - F(t)}{S(t)} \,. \tag{3.1}$$

Luego, reemplazando (3.1) en la Definición (3.2), se tiene:

$$h(t) = \lim_{\Delta t \to 0} \left[\frac{F(t + \Delta t) - F(t)}{\Delta t} \right] \frac{1}{S(t)} \,. \tag{3.2}$$

Siendo $F'(t) = f(t) = -S'(t)$, entonces:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}[\log S(t)]. \tag{3.3}$$

Derivando ambos miembros en (3.3) se tiene que:

$$S(t) = \exp\left[-\int_0^t h(u)\, du \right]. \tag{3.4}$$

Definición 3.3. *Sea $S(t)$ la función de supervivencia y $h(t)$ la función de riesgo para una variable aleatoria T. Se define la* ***función de riesgo acumulada*** *o* ***tasa de falla acumulada*** *como: $H(t) = \int_0^t h(u)\, du = -\log S(t)$.*

3.1.2. Modelo de regresión de Cox

El *modelo de regresión de Cox* permite estimar la probabilidad de los individuos de presentar el evento en estudio en un cierto tiempo t, en función de un conjunto de *variables explicativas*. Este modelo fue propuesto por Cox en 1972 [40] y, debido a su popularidad, es conocido como *modelo de regresión de Cox* o, simplemente, *regresión de Cox*. Se trata de

una regresión donde el tiempo está siempre presente y la variable dependiente es siempre una función de riesgo o una función de supervivencia.

El modelo de regresión de Cox se basa en el *supuesto de riesgos proporcionales*, pero no requiere ningún supuesto sobre la forma que asume la función de riesgo o de supervivencia por lo que, en general, se menciona como un modelo *semiparamétrico*.

Se dice que, en la presencia de dos grupos poblacionales I y II, los riesgos en ambas poblaciones, $h_1(t)$ y $h_2(t)$ respectivamente, son proporcionales si cumplen con $h_1(t) = \psi h_2(t)$, donde ψ es una constante que no depende de t. Por lo tanto: $\exp\left[-\int_0^t h_1(u)\,du\right] = \exp\left[-\int_0^t \psi h_2(u)\,du\right]$. Si $S_1(t)$ y $S_2(t)$ son las funciones de supervivencia para los individuos del grupo I y II, respectivamente, de acuerdo a la ecuación (3.4) se llega a $S_1(t) = S_2(t)^{\psi}$. Dado que la función de supervivencia se encuentra acotada entre 0 y 1, luego $S_1(t)$ será mayor o menor que $S_2(t)$ de acuerdo a si ψ es menor o mayor a 1, para cualquier tiempo t. Esto significa que si dos funciones de riesgo son proporcionales, sus respectivas funciones de supervivencia no se cruzan en ningún punto. Esto es una condición necesaria pero no suficiente para el *supuesto de riesgos proporcionales*.

Supóngase que el riesgo de presentar el evento de interés en un momento particular t ($h(t)$) depende de los valores observados para r variables explicativas, $X_1, X_2, \ldots, X_r$. El conjunto de los valores observados en el i-ésimo individuo para las variables explicativas se representa mediante el vector $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ir})$. En este trabajo se asume que los valores de estas variables han sido obtenidos en el tiempo de origen del seguimiento de los individuos.

Definición 3.4. *Sea $h_0(t)$ la función de riesgo para un individuo para el cual los valores observados en las r variables explicativas son nulas ($\boldsymbol{x}_j = \mathbf{0}$). La función $h_0(t)$ se denomina* ***función de riesgo basal****.*

La función de riesgo para el i-ésimo individuo puede ser escrita como:

$$h_i(t) = \psi(\boldsymbol{x}_i) h_0(t), \tag{3.5}$$

donde $\psi(\boldsymbol{x}_i)$ es una función de los valores del vector de variables explicativas correspondientes al i-ésimo individuo y puede interpretarse como el riesgo en el tiempo t, relativo al riesgo de un individuo que tiene $\boldsymbol{x}_j = \mathbf{0}$. Por lo tanto, $\psi(\boldsymbol{x}_i)$ no puede tomar valor negati-

vos, por lo que puede expresarse como $\exp(\eta_i)$, donde η_i es una combinación lineal de las r variables explicativas de $\boldsymbol{x}_i$, es decir, $\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_r x_{ir}$. En notación matricial, se expresa $\eta_i = \boldsymbol{\beta}' \boldsymbol{x}_i$, donde $\boldsymbol{\beta}$ es el vector de coeficientes correspondientes a las variables explicativas $X_1, X_2, \ldots, X_r$ del modelo. La cantidad η_i se denomina *componente lineal* del modelo, aunque también suele denominarse *score de riesgo* para el i-ésimo individuo. Por lo tanto, la función de riesgo presentada en la Definición (3.4) puede reescribirse como:

$$h_i(t) = \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_r x_{ir}) h_0(t). \tag{3.6}$$

El modelo de riesgos proporcionales puede se visto como un modelo lineal para el logaritmo de la razón de riesgos ya que (3.6) puede expresarse como:

$$\log \left[\frac{h_i(t)}{h_0(t)} \right] = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_r x_{ir}. \tag{3.7}$$

Además, la razón de riesgos entre el i-ésimo y el i'-ésimo individuo resulta constante a lo largo del tiempo:

$$\theta_{i,i'} = \frac{h_i(t)}{h_{i'}(t)} = \frac{h_0(t) \exp(\eta_i)}{h_0(t) \exp(\eta_{i'})} = \frac{\exp(\eta_i)}{\exp(\eta_{i'})} = \exp[\boldsymbol{\beta}'(\boldsymbol{x}_i - \boldsymbol{x}_{i'})]. \tag{3.8}$$

3.1.3. Estimación de parámetros del modelo de regresión de Cox

Estimar el modelo de riesgos proporcionales, o modelo de regresión de Cox, dado en la ecuación (3.6), por medio de la observación de un conjunto de datos de supervivencia, implica estimar los coeficientes correspondientes a las variables explicativas $X_1, X_2, \ldots, X_r$, es decir, estimar $\beta_1, \beta_2, \ldots, \beta_r$ para la componente lineal del modelo. La función de riesgo basal, $h_0(t)$, también puede ser estimada. Ambas componentes del modelo pueden ser estimadas separadamente. Primero se estima el vector $\boldsymbol{\beta}$ y esa estimación es utilizada para encontrar una estimación para la función de riesgo basal. Por lo tanto, si se desea realizar inferencia acerca del efecto de las variables explicativas $X_1, X_2, \ldots, X_r$ sobre el riesgo relativo, $h_i(t)/h_0(t)$, no es necesario estimar $h_0(t)$.

Los coeficientes de $\boldsymbol{\beta}$ correspondientes al modelo de regresión de Cox pueden ser estimados mediante el *método de máxima verosimilitud*. Supóngase que se cuenta con información correspondiente a n individuos, de los cuales a han presentado el evento de

interés en distintos momentos durante el estudio (sin empates en dichos tiempos) y $n-a$ son censurados por derecha. Los a tiempos de ocurrencia del evento son ordenados de modo que $t_{(1)} < t_{(2)} < \ldots < t_{(a)}$, de modo que $t_{(j)}$ es el j-ésimo tiempo ordenado de ocurrencia del evento. El conjunto de individuos en riesgo al tiempo $t_{(j)}$ es denotado como $R(t_{(j)})$, es decir, $R(t_{(j)})$ es el conjunto de individuos que no han presentado el evento ni han sido censurados en el momento inmediatamente anterior a $t_{(j)}$ y es denominado *conjunto de riesgo*.

Cox [40] demostró que la función de verosimilitud para el modelo de riesgos proporcionales dado en la Definición (3.4) es:

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{a} \frac{\exp(\boldsymbol{\beta}'\boldsymbol{x}_{(j)})}{\sum_{l \in R(t_{(j)})} \exp(\boldsymbol{\beta}'\boldsymbol{x}_{(l)})} \ , \tag{3.9}$$

donde $\boldsymbol{x}_{(j)}$ es el vector de valores de las variables explicativas correspondiente a un individuo que presentó el evento en el j-ésimo tiempo ordenado, $t_{(j)}$. La sumatoria en el denominador de la función de verosimilitud es la suma de los valores de $\exp(\boldsymbol{\beta}'\boldsymbol{x})$ correspondientes a todos los individuos en riesgo de presentar el evento al momento $t_{(j)}$.

Supóngase que los datos consisten en observaciones de n tiempos de supervivencia, indicados como $t_1, t_2, \ldots, t_n$ y δ_i es una variable indicadora que toma el valor cero si el i-ésimo tiempo de supervivencia t_i, con $i = 1, 2, \ldots, n$ se encuentra censurado por derecha y el valor uno en otro caso. Siendo $R(t_i)$ el conjunto de individuos en riesgo en el tiempo t_i, la función de verosimilitud en (3.9) puede expresarse como:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)}{\sum_{l \in R(t_i)} \exp(\boldsymbol{\beta}'\boldsymbol{x}_l)} \right]^{\delta_i} . \tag{3.10}$$

Luego, la función de log-verosimilitud está dada por:

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[\boldsymbol{\beta}'\boldsymbol{x}_i - \log \sum_{l \in R(t_i)} \exp(\boldsymbol{\beta}'\boldsymbol{x}_l) \right]. \tag{3.11}$$

Los estimadores máximo-verosímiles de los parámetros de $\boldsymbol{\beta}$ del modelo de riesgos proporcionales pueden ser obtenidos maximizando esta función de log-verosimilitud usando métodos numéricos. En general, el método elegido es el de *Newton-Raphson*.

El modelo de riesgos proporcionales asume que la función de riesgo es continua, por lo que no sería posible observar *empates* en los tiempos de supervivencia. Sin embargo, de acuerdo a la escala de medida y la unidad de tiempo utilizada, en la práctica es habitual encontrar individuos que presentan el mismo tiempo hasta la ocurrencia de un evento. También es posible observar coincidencias en los tiempos de censura o, incluso, encontrar la ocurrencia de eventos y censuras en un mismo tiempo. En este último caso, se asume que las censuras ocurren después de la ocurrencia de los eventos.

Con la finalidad de poder considerar las observaciones *empatadas*, la función de verosimilitud (3.9) ha sido modificada por Kalbfleisch & Prentice [41]. Sin embargo, la función de verosimilitud definida por estos autores resulta compleja y dificil de maximizar computacionalmente, por lo que suelen utilizarse aproximaciones numéricas para la función de verosimilitud [38].

Sea $\boldsymbol{s}_j$ el vector de sumas de cada una de las r variables explicativas para los individuos que presentaron el evento en el j-ésimo tiempo, $t_{(j)}$, con $j = 1, 2, \ldots, a$. Si en el tiempo $t_{(j)}$ ocurrieron d_j eventos, el h-ésimo elemento de $\boldsymbol{s}_j$ es $s_{hj} = \sum_{k=1}^{d_j} x_{hjk}$, donde x_{hjk} es el valor de la h-ésima variable explicativa ($h = 1, 2, \ldots, r$) para el k-ésimo ($k = 1, 2, \ldots, d_j$) de los d_j individuos que presentaron el evento en el tiempo $t_{(j)}$. La aproximación más simple de la función de verosimilitud es debida a Breslow [42], quien propuso como verosimilitud aproximada:

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{a} \frac{\exp(\boldsymbol{\beta}' \boldsymbol{s}_j)}{\left[\sum_{l \in R(t_{(j)})} \exp(\boldsymbol{\beta}' \boldsymbol{x}_l)\right]^{d_j}} \,. \tag{3.12}$$

En esta aproximación, los d_j eventos del tiempo $t_{(j)}$ son considerados diferentes y que ocurren subsecuentemente. Las probabilidades de todas las secuencias posibles de eventos son sumadas para dar lugar a la función de verosimilitud (3.12). La misma es sencilla de calcular y resulta una aproximación adecuada cuando el número de observaciones empatadas no es muy alto. Por estas razones, este método es el procedimiento utilizado por defecto en los distintos programas estadísticos para la estimación de parámetros de un modelo de Cox. Otras aproximaciones alternativas han sido propuestas por Cox y por Efron [40, 43]. Cuando no hay observaciones empatadas, esto es, cuando $d_j = 1$ para cada tiempo de ocurrencia de eventos, estas aproximaciones coinciden con la función de verosimilitud (3.9).

Una vez estimado el vector $\boldsymbol{\beta}$, la función de riesgo para el i-ésimo individuo resulta $\widehat{h}_i(t) = \widehat{h}_0(t)\exp(\widehat{\boldsymbol{\beta}}'\boldsymbol{x}_i)$. Supóngase que el vector de valores de las variables explicativas del i-ésimo individuo, $\boldsymbol{x}_i$, difiere del vector del i'-ésimo individuo solo en el elemento correspondiente a la k-ésima variable, es decir, $x_{ij} = x_{i'j}$ para $j \neq k$ y $x_{ij} \neq x_{i'j}$ para $j = k$. Entonces, la razón de riesgo estimada entre estos individuos resulta:

$$\widehat{\theta}_{i,i'} = \frac{\widehat{h}_i(t)}{\widehat{h}_{i'}(t)} = \frac{\widehat{h}_0(t)\exp(\widehat{\boldsymbol{\beta}}'\boldsymbol{x}_i)}{\widehat{h}_0(t)\exp(\widehat{\boldsymbol{\beta}}'\boldsymbol{x}_{i'})} = \exp\left[\widehat{\boldsymbol{\beta}}_k(x_{ik} - x_{i'k})\right]. \tag{3.13}$$

En la situación anterior, supóngase además que X_k es una variable cuantitativa, y que los individuos i e i' difieren en b unidades para dicha variable, es decir, $x_{ik} - x_{i'k} = b$. Entonces $\widehat{\theta}_k = \exp(b\widehat{\beta}_k)$ puede ser interpretado como el cambio en el riesgo de presentar el evento de interés cuando la variable X_k se incrementa en b unidades y las demás variables se mantienen constantes.

Si se incluye en el modelo una variable explicativa cualitativa con c categorías, $c \geq 2$, se puede pensar que existen c grupos de individuos, cada uno compuesto por quienes presentan la misma categoría para esa variable. La función de riesgo para un individuo del j-ésimo grupo, $j = 1, 2, \ldots, c$, está dada por $h_j(t) = \exp(\gamma_j)h_0(t)$, donde γ_j es el efecto de la j-ésima categoría de la variable. Generalmente, para evitar una sobreparametrización del modelo, se considera $\gamma_1 = 0$, por lo que la función de riesgo basal es la función para individuos del primer grupo y que tienen todas las demás variables explicativas con valor cero. La razón de riesgos para un individuo del j-ésimo grupo, $j \geq 2$, relativo a un individuo del primer grupo, es entonces $\exp(\gamma_j)$. Para incluir una variable cualitativa en el modelo, se definen $c-1$ variables indicadoras, $X_2, X_3, \ldots, X_c$, de modo que los coeficientes que las acompañan corresponden a $\gamma_2, \gamma_3, \ldots, \gamma_c$. Por lo tanto, la estimación de γ_j, es decir, $\widehat{\gamma}_j$, representa el cambio en el riesgo instantáneo del evento al comparar a un individuo del j-ésimo grupo en relación a uno del primer grupo, cuando ambos coinciden en los valores del resto de las variables explicativas.

La distribución asintótica de los estimadores máximo-verosímiles de los parámetros de un modelo de regresión de Cox fue estudiada por Cox, Tsiatis, Næs y Bailey [44, 45, 46, 47]. El vector de parámetros estimados $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \widehat{\beta}_2, \cdots, \widehat{\beta}_r)$, bajo condiciones muy generales, tiene una distribución asintótica Normal con esperanza $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_r)$ y matriz de varianzas y covarianzas $\boldsymbol{I}^{-1}(\widehat{\boldsymbol{\beta}})$, donde $\boldsymbol{I}(\widehat{\boldsymbol{\beta}})$ se denomina *matriz de información*

observada y corresponde a una matriz de derivadas parciales segundas del logaritmo de la verosimilitud. Sin embargo, no es completamente eficiente, es decir no alcanza la cota de *Cramer-Rao*.

Definición 3.5. *Se denomina matriz de información de un vector de parámetros β a la matriz $r \times r$ de los opuestos de las derivadas segundas de la log-verosimilitud (3.9), tal que el elemento (j,k) de $\boldsymbol{I}(\boldsymbol{\beta})$ es $-\frac{\partial \log L(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k}$. Si el vector de parámetros es reemplazado por sus estimadores, la matriz se denomina* ***matriz de información observada****.*

De acuerdo al procedimiento de *Newton-Raphson* los estimadores del vector de parámetros $\boldsymbol{\beta}$ en la $(s+1)$-ésima iteración del proceso son:

$$\widehat{\boldsymbol{\beta}}_{s+1} = \widehat{\boldsymbol{\beta}}_s + \boldsymbol{I}^{-1}(\widehat{\boldsymbol{\beta}}_s)\boldsymbol{u}(\widehat{\boldsymbol{\beta}}_s), \tag{3.14}$$

para $s = 0, 1, \cdots$, donde $\boldsymbol{u}(\boldsymbol{\beta}_s)$ es el vector de derivadas primeras de la log-verosimilitud (3.9) con respecto a $\boldsymbol{\beta}$ y $\boldsymbol{I}(\boldsymbol{\beta}_s)$ corresponde a la Definición (3.5), ambos evaluados en $\widehat{\boldsymbol{\beta}}_s$. El procedimiento finaliza cuando el cambio en la log-verosimilitud o en los valores de los parámetros estimados es suficientemente pequeño [38]. Cuando el proceso iterativo converge, la matriz de varianzas y covarianzas de los parámetros estimados se aproxima por la inversa de la matriz de información observada y, por lo tanto, las raíces cuadradas de los elementos diagonales de dicha matriz son los desvíos estándares estimados para $\widehat{\beta}_1, \widehat{\beta}_2, \cdots, \widehat{\beta}_r$.

3.2. Datos faltantes o perdidos

3.2.1. Mecanismos de pérdida

Cuando en una matriz de datos no se registra el valor de una variable para un individuo, se dice que el mismo es un *dato faltante* o *perdido*. Los datos perdidos pueden aparecer por múltiples causas, como ser fallas o errores en instrumentos de medición, olvido de un dato al momento del registro de la información, decisión de un individuo de no responder algún aspecto específico de una encuesta, falta de memoria sobre alguna información de interés, etcétera.

Resulta relevante preguntarse si el hecho de que se encuentren perdidos los valores de una variable, para algunos individuos en particular, está relacionado con el mismo valor del dato perdido o bien con los valores de otras de las variables consideradas. Esta pregunta dió lugar a la primera definición formal de diferentes *mecanismos de pérdida* realizada por Rubin y, posteriormente, extendida por Little & Rubin, debido al rol crucial que tiene su identificación cuando se realizan análisis de datos con valores faltantes [10, 48].

Sea $\boldsymbol{X} = \{x_{ij}\}$ una matriz de dimensión $n \times r$, con la fila i-ésima $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ir})$ donde x_{ij} es el valor de la variable X_j para el individuo i y $\boldsymbol{M} = \{m_{ij}\}$ una matriz indicadora de igual dimensión donde $m_{ij} = 1$ si x_{ij} es un valor faltante y $m_{ij} = 0$ si x_{ij} es un valor observado. El mecanismo de datos faltantes se caracteriza por la distribución condicional de $\boldsymbol{M}$ dada $\boldsymbol{X}$, es decir $f(\boldsymbol{M}|\boldsymbol{X}, \phi)$ donde ϕ denota un conjunto de parámetros desconocidos.

Definición 3.6. *Se dice que los datos son* ***perdidos completamente al azar*** *y se indica como* MCAR *por sus siglas en inglés (missing completely at random) si la función de distribución condicional $f(\boldsymbol{M}|\boldsymbol{X}, \phi)$ coincide con la función de densidad marginal para todo $(\boldsymbol{X}, \phi)$, es decir, $f(\boldsymbol{M}|\boldsymbol{X}, \phi) = f(\boldsymbol{M}|\phi)$.*

Bajo MCAR, la pérdida de datos no depende de los valores del conjunto de variables, ya sean observados o perdidos. MCAR ocurre cuando un individuo con información faltante representa un subconjunto aleatorio del conjunto total de individuos. Dado que MCAR implica un supuesto muy fuerte sobre el mecanismo de pérdida, pocas veces se verifica.

Sea $\boldsymbol{X}$, subdividida en dos conjuntos: $\boldsymbol{Z}^{obs}$ el conjunto de los valores observados de $\boldsymbol{X}$ y $\boldsymbol{Z}^{miss}$ el conjunto de valores faltantes.

Definición 3.7. *Se dice que los datos son* ***perdidos al azar*** *y se indica como* MAR *por sus siglas en inglés (missing at random) si las pérdidas de los valores dependen solo de las componentes de $\boldsymbol{Z}^{obs}$, y no de las componentes de $\boldsymbol{Z}^{miss}$, es decir: $f(\boldsymbol{M}|\boldsymbol{X}, \phi) = f(\boldsymbol{M}|\boldsymbol{Z}^{obs}, \phi)$ para todo $\boldsymbol{Z}^{miss}, \phi$.*

Bajo el mecanismo MAR, las pérdidas de datos que ocurren en una variable solo dependen de los valores observados en otra u otras variables.

Definición 3.8. *Se dice que los datos son* ***perdidos no al azar*** *y se indica como* MNAR *por sus siglas en inglés (missing not at random) si las pérdidas de los valores dependen de las componentes de* $\boldsymbol{Z}^{obs}$*, y/o de las componentes de* $\boldsymbol{Z}^{miss}$*, es decir:* $f(\boldsymbol{M}|\boldsymbol{X},\phi) = f(\boldsymbol{M}|\boldsymbol{Z}^{obs},\boldsymbol{Z}^{miss},\phi) \neq f(\boldsymbol{M}|\boldsymbol{Z}^{obs},\phi)$ *para todo* $\boldsymbol{Z}^{miss}, \boldsymbol{Z}^{obs}, \phi$*.*

Por lo tanto, los datos son *perdidos no al azar* cuando la pérdida depende de los valores observados y/o perdidos de la variable correspondiente, por lo que la probabilidad de que un dato sea faltante depende del valor real del mismo. En este caso, es fundamental no ignorar estas pérdidas.

3.2.2. Tratamiento de datos perdidos

En general, los distintos métodos de análisis estadístico se basan en información completa, sin datos perdidos. Por lo tanto, en la etapa previa a dicho análisis, es necesario establecer cómo se tratarán los valores faltantes cuando estos existen, siendo esta etapa tan importante como la de depuración de los datos, la detección de valores extremos o erróneos y la corrección de incosistencias. De manera general, los métodos de tratamiento de datos faltantes se pueden dividir en tres categorías [48]:

1. *Métodos de eliminación:* Habitualmente se opta por elegir para el análisis estadístico aquellas unidades con observaciones completas en todas las variables, eliminando toda unidad donde exista al menos un dato faltante. La mayoría de los programas estadísticos utilizan este mecanismo por defecto, que suele denominarse *análisis de casos completos* (*case-complete analysis*, CCA, también conocido como *listwise deletion*). Otra opción es realizar un *análisis de casos disponibles*, método conocido como *pairwise deletion*, que consiste en considerar solamente aquellas variables sin datos perdidos en cada paso del análisis estadístico. Una condición necesaria para que estos métodos sean válidos es que el mecanismo de pérdida pueda suponerse MCAR ya que, de otro modo, se estarían eliminando casos con características particulares, lo que derivaría en un sesgo relevante al momento del análisis estadístico. Se debe considerar el tamaño de la muestra y la proporción de individuos con datos faltantes, a fin de definir si la cantidad de individuos retenidos es suficiente para alcanzar cierta potencia estadística, además de tenerse en cuenta el esfuerzo demandado para la recolección de datos que efectivamente se han conseguido.

2. *Métodos basados en verosimilitudes:* Cuando el objetivo es estimar los parámetros de un modelo estadístico, es posible realizar modificaciones sobre la verosimilitud de la muestra de modo que la misma se pueda maximizar utilizando solamente la información de los casos completos. Los métodos más conocidos son el *enfoque de grupo múltiple* (*multiple-group approach*) [49], el *algoritmo EM* [8] y el método de *máxima verosimilitud con información completa* (*Full information maximum likelihood*, FIML) [50].

3. *Métodos de imputación:* Para evitar la eliminación de información y sus consecuencias, es posible asignar a los datos faltantes valores obtenidos bajo algún criterio adecuado, según el mecanismo de pérdida, las características de las variables, la proporción de datos faltantes y el uso que se dará a la matriz completa conseguida. Esta acción se denomina habitualmente *imputación de datos faltantes* y múltiples técnicas para realizarla han sido presentadas en la bibliografía, fundamentalmente a medida que se produjeron avances computacionales. En general, pueden agruparse según las siguientes categorías [51]:

    a) *Sustitución de casos:* Generalmente utilizado en la realización de encuestas. Cada unidad con datos faltantes es reemplazada por una nueva unidad no incluída inicialmente y para la cual se dispongan de datos completos.

    b) *Imputación por valores representativos:* Consiste en asignar a cada dato faltante un valor relativo a los datos observados, como puede ser el promedio, la mediana, el mínimo, el máximo o, incluso, un valor aleatorio dentro del rango para cada variable [52].

    c) *Hot deck y cold deck:* Se reemplazan los datos faltantes por valores calculados a partir de uno o más casos completos del mismo conjunto de datos. Existen diferentes formas de asignar un valor de reemplazo, por ejemplo, eligiendo un donante al azar o calculando la media de valores correspondientes a los donantes. La imputación *cold deck* es similar a la imputación *hot deck*, pero la fuente de datos debe ser distinta de la fuente de datos actual. Históricamente, el término *hot deck* viene del uso de tarjetas perforadas de computadora para el almacenamiento de datos, y se refiere al conjunto de tarjetas donantes disponibles para un dato faltante. Ese conjunto de tarjetas estaba "caliente" ya que se estaba procesando en ese momento, a diferencia de la "plataforma fría"

que se refiere al uso de datos preprocesados como donantes, es decir, datos de una recopilación de datos anterior o un conjunto de datos diferente, de donde deriva la expresión *cold deck* [53].

d) *Modelos predictivos:* Se crea un modelo predictivo para estimar valores que sustituirán los datos faltantes. La variable con datos faltantes se usa como variable respuesta o predicha, y las variables restantes se usan como entrada para el modelo predictivo. Un argumento importante a favor de este enfoque es que, con frecuencia, las variables presentan correlaciones entre ellas que podrían usarse para crear un modelo predictivo para clasificación o regresión, según el tipo de variable con datos faltantes. Algunas de estas relaciones entre las variables se pueden mantener si fueron capturadas por el modelo predictivo. Un inconveniente importante de este enfoque es que los valores estimados del modelo suelen tener un mejor comportamiento que los valores reales, es decir, dado que los valores faltantes se predicen a partir de un conjunto de variables, es probable que los valores predichos sean más consistentes con este conjunto de variables que el valor verdadero (no conocido) correspondiente al dato faltante [54]. En caso de no existir relaciones entre una o más variables en el conjunto de datos y la variable con datos faltantes, entonces el modelo no será preciso para estimar los datos faltantes [29, 55].

Los métodos de imputación pueden ser clasificados en *simples* o *múltiples*. En el primer caso, reemplazan cada valor faltante por un valor adecuado para la variable en cuestión. El segundo caso consiste en hallar m valores plausibles para imputar cada dato faltante, generando así m conjuntos de datos completos, cada uno de los cuales es considerado para realizar estimaciones de parámetros de interés utilizando CCA. Las estimaciones obtenidas en base a cada conjunto de datos son combinadas al final de proceso para arribar a estimaciones que tienen en cuenta la incertidumbre producida por las observaciones faltantes [23].

En general, los métodos de imputación simple no tienen en cuenta la incertidumbre de los datos faltantes y, como resultado, los errores estándar de las estimaciones resultan subestimados, sobrestimando así la precisión de los resultados. Potencialmente, esto puede aumentar la probabilidad de error de Tipo I al probar hipótesis de independencia entre variables [56, 57]. La imputación múltiple es una técnica ampliamente reconocida y es

recomendada como el método estándar para tratar los datos faltantes en muchas áreas de investigación, especialmente en investigación clínica y epidemiología [27, 58, 59]. Además, se ha vuelto más popular con la creciente disponibilidad en los programas computacionales para análisis de datos [25, 26]. Sin embargo, para realizar imputación múltiple la elección del modelo y la relación existente entre las variables a considerar, resultan un paso fundamental en el proceso de imputación. Más aún, requieren de supuestos fuertes que muchas veces no son tenidos en cuenta, lo que ha llevado a diversos autores a realizar críticas a este método. Fay [60] ha sido pionero en este aspecto, cuestionando la aplicación de imputaciones múltiples en diversos trabajos, especialmente en relación al tratamiento de datos correspondientes a censos y encuestas. Muchos de los trabajos publicados que realizan imputaciones múltiples no informan si ha sido considerada la verificación de los supuestos, limitando así la *reproducibilidad* y validación de las inferencias realizadas [26].

Por otro lado, es importante tener en cuenta que algunos de los métodos de imputación disponibles restringen su capacidad de asignar valores a los datos faltantes a un solo tipo de variables, ya sean cuantitativas [61, 62, 63] o cualitativas [64, 65, 66, 67]. Dado que la presente tesis centra su atención en la imputación de datos faltantes correspondientes a variables explicativas de un modelo de regresión de Cox y que las mismas suelen corresponder a distintos tipos, se enfocan en la sección 3.3, métodos de imputación válidos para tratar simultáneamente ambos casos. Se indica en cada método descripto, considerado en los estudios comparativos, el anexo donde se agregan los pseudo-códigos desarrollados en esta tesis. Los mismos no se encuentran en forma completa u ordinada en las publicaciones que los proponen.

3.3. Métodos de imputación de datos perdidos para variables mixtas

3.3.1. Imputación por valores representativos

Una de las primeras ideas sugeridas en la literatura estadística, presumiblemente debida a Wilks [68] corresponde a variables cuantitativas y consiste en la una *sustitución por la media*. Se hace evidente que, al imputar todos los datos faltantes mediante el valor promedio, se produce una subestimación de la desviación estándar de la variable. Además, debido a la sensibilidad de la media respecto a la presencia de valores extremos,

la misma no es una medida representativa adecuada para variables cuya distribución de probabilidad es asimétrica, por lo que la sugerencia inmediata es utilizar el valor de la mediana de los datos observados para imputar los faltantes [16]. En ambos casos, la distribución de probabilidad de la variable resulta distorsionada por la adición del mismo valor tantas veces como datos faltantes tenga la variable. Para las variables cualitativas, la opción más simple, que se asemeja a esta estrategia, es asignar a los datos faltantes el valor modal entre los datos observados lo cual, claramente, afectará la distribución de frecuencias de estas variables.

En esta tesis, esta técnica de reemplazo se denomina *imputación por valores representativos* (representative values imputation, RVI). Si bien este es un enfoque simple y computacionalmente rápido, puede llevar a un rendimiento deficiente de los modelos resultantes. Además, no tiene en cuenta la relación entre las variables y solo resulta pertinente su uso si el mecanismo de pérdida puede suponerse como MCAR. En caso contrario, se está en riesgo de ignorar que los valores perdidos presentan características propias o ligadas a otras variables que los diferencian de los valores observados [2].

3.3.2. *k*-vecinos cercanos

La técnica denominada *k-vecinos cercanos* (*k-nearest neighbours*, KNN) es un algoritmo de imputación eficiente que asigna a cada dato perdido un valor obtenido en base a los datos observados para las k unidades más *cercanas* o *similares*, en relación a los valores observados para otras variables medidas sobre las mismas unidades. A estas unidades más cercanas se las suele llamar *vecinos* o *donantes* [18].

El concepto de *similitud* pone en juego la elección de una medida adecuada que permita elegir correctamente los k vecinos más cercanos, esto es, una métrica que permita cuantificar la similaridad o distancia entre la unidad a imputar y sus posibles donantes. Una mayor medida de similaridad implica mayor semejanza entre las unidades mientras que, cuanto mayor es la distancia, menor es la semejanza entre ellas.

Para conjuntos de datos cuantitativos, la *Distancia de Minkowski* es un método general utilizado para calcular la distancia entre dos puntos multivariados. Sean $\boldsymbol{x_i} = (x_{i1}, \cdots, x_{ir})$ y $\boldsymbol{x_j} = (x_{j1}, \cdots, x_{jr})$ los vectores de datos observados en las r variables para los individuos i y j respectivamente. La *distancia de Minkowski* de orden q entre

estos individuos se define como:

$$\delta_{m_q}(i,j) = \left(\sum_{k=1}^{r} |x_{ik} - x_{jk}|^q \right)^{1/q}, \ q > 0. \tag{3.15}$$

En particular, la Distancia de Minkowski de orden 1 (*Distancia de Manhattan*) y la de orden 2 (*Distancia Euclidiana*) son las dos medidas de distancia más utilizadas para datos cuantitativos. Otras opciones son la *Distancia de Camberra* que se define como:

$$\delta_C(i,j) = \sum_{k=1}^{r} \frac{|x_{ik} - x_{jk}|}{|x_{ik}| + |x_{jk}|}, \tag{3.16}$$

y la *Distancia de Mahalanobis* que se expresa como:

$$\delta_M(i,j) = (\boldsymbol{x_i} - \boldsymbol{x_j})' \boldsymbol{S}^{-1} (\boldsymbol{x_i} - \boldsymbol{x_j}), \tag{3.17}$$

donde $\boldsymbol{S}$ es la matriz de covarianzas del conjunto de r variables.

En el caso de vectores aleatorios con componentes binarias o dicotómicas son de utilidad los coeficientes de *similaridad de emparejamiento simple* [69] y *Jaccard* [70]. Sea a el número de variables que toman el valor 1 en ambos individuos, x_i y x_j, y sea d el número de variables que toman el valor 0 en ambos individuos, el *coeficiente de similaridad de emparejamiento simple* se define como:

$$s_{ij} = \frac{a+d}{r}, \tag{3.18}$$

y el *coeficiente de Jaccard* como:

$$s_{ij} = \frac{a}{r-d}. \tag{3.19}$$

Para matrices con datos cualitativos, el grado de asociación entre cualquier par de unidades puede medirse utilizando las medidas de similitud propuestas para variables binarias si las variables cualitativas multiestados son reemplazadas por pseudo-variables binarias que toman el valor 0 si la característica está ausente y el valor 1 si está presente. Sin embargo, esta propuesta metodológica tiene el inconveniente de ser artificial ya que otorga mayor peso a las variables que posean más estados. Si, por el contrario, las categorías para cada variable son codificadas en forma numérica, el grado de asociación entre cualquier par de unidades puede medirse a través de la *expansión del emparejamien-*

to simple. Otras medidas de distancia o similitud para variables cualitativas pueden ser encontradas en el trabajo de Boriah *et al.* [71].

Para matrices de datos compuestas por variables mixtas, existen algunos enfoques que provienen del campo del aprendizaje automático [72, 73]. Éstos son bastante complejos y no están diseñados para variables ordinales. Por lo tanto, la medida de similaridad más frecuentemente utilizada es el *coeficiente general de Gower* [19], que permite incorporar simultáneamente variables cuantitativas, binarias, cualitativas nominales y cualitativas ordinales. Si se supone que, entre las r variables de cada vector de observaciones, hay r_1 variables cuantitativas, r_2 variables binarias y r_3 variables cualitativas nominales, el *coeficiente general de similitud de Gower* es:

$$s_{ij} = \frac{\sum_{h=1}^{r_1}(1 - |x_{ih} - x_{jh}|/rg_h) + a + \alpha}{r_1 + r_2 + r_3}, \tag{3.20}$$

donde x_{ih} es el valor para la h-ésima variable cuantitativa en el i-ésimo individuo, rg_h es el rango de la h-ésima variable cuantitativa, a es el número total de coincidencias entre la i-ésima y la j-ésima unidad, en 0 o en 1, para las r_2 variables binarias y α es el número de coincidencias entre la i-ésima y la j-ésima unidad para las r_3 variables cualitativas nominales. Para las variables cualitativas ordinales, cada categoría se codifica numéricamente siguiente el orden correspondiente y luego se las considera tal como si se tratase de variables cuantitativas. Las categorías se tratan como si fueran equidistantes, lo cual se puede modificar al momento de la asignación de valores numéricos a las categorías. También es posible ponderar cada variable de acuerdo al peso relativo que se quiera asignar a la misma en el cálculo de las distancias entre individuos.

Para aplicar el método KNN no basta con establecer la métrica para definir los parecidos, sino que también se debe fijar el valor k de donantes a considerar. En relación a este aspecto, el trabajo de Beretta & Santaniello sugiere que $k = 3$ es una elección razonable [74]. Cartwright *et al.*, por otro lado, sugieren valores pequeños para k, típicamente 1 o 2, aunque señalan que $k = 1$ es sensible a los valores atípicos y, en consecuencia, recomiendan utilizar $k = 2$ [75]. Batista & Monard informan mejores resultados al considerar $k = 10$ para grandes conjuntos de datos [51], mientras que el trabajo de Troyanskaya *et al.* sugiere que el método es bastante insensible a la elección de k [18]. Sin embargo, es lógico pensar que, a medida que aumenta el número de donantes a utilizar, la distancia media entre ellos y la unidad a imputar resulta mayor, lo que implica que los valores de

reemplazo podrían ser menos precisos.

Finalmente, debe determinarse el método de agregación a utilizar para el grupo de vecinos. Para imputar un dato correspondiente a una variable cuantitativa, generalmente se opta por asignar la media aritmética o la mediana de dicha variable calculada sobre los k donantes seleccionados para la unidad a imputar. Además, es posible optar por ponderar cada donante mediante un peso relativo en base a la distancia con la unidad a imputar. Cuando el dato faltante corresponde a una variable cualitativa o binaria, el valor elegido para la imputación es el modo observado sobre los k vecinos más cercanos [51].

El método KNN tiene la ventaja de permitir considerar variables mixtas mediante la adaptación de la medida de distancia o similitud utilizada. Así mismo, es un método flexible que no requiere de la estimación de modelos predictivos y, por ende, no requiere de supuestos respecto a las variables. La posibilidad de elegir el método de agregación permite evitar cierto sesgo de imputación debido a la presencia de valores extremos o a la asimetría de alguna variable. Sin embargo, tiene la desventaja de ser computacionalmente costoso ya que, para cada unidad con datos faltantes, debe calcular la distancia respecto a todas las unidades con datos completos [74].

En el Anexo I, se presenta el pseudo-código correspondiente al mecanismo de imputación KNN (Algoritmo 1).

3.3.3. missForest

El algoritmo *missForest* [21], desarrollado para trabajar en forma simultánea con datos mixtos, se basa en la metodología *Random Forest* [22], considerando como datos de entrenamiento la sub-matriz de datos completos de una matriz $\boldsymbol{X}$ de tamaño $n \times r$, con la finalidad de predecir los datos faltantes.

Sea X_j una variable cuyas observaciones se encuentran en el vector columna $\boldsymbol{x}_j$ de matriz $\boldsymbol{X}$ y cuyas observaciones faltantes corresponden al conjunto de subíndices $\boldsymbol{i}_j^{miss} \subseteq \{1, 2, \ldots, n\}$. Luego se definen los siguientes elementos:

- $\boldsymbol{x}_j^{obs}$: sub-vector de $\boldsymbol{x}_j$ con valores observados (o completos) para la variable X_j;
- $\boldsymbol{x}_j^{miss}$: sub-vector de $\boldsymbol{x}_j$ con valores faltantes (o incompletos) para la variable X_j;
- $\boldsymbol{X}_{-j}^{obs}$: matriz de variables X_l, con $l \neq j$, con observaciones (filas) correspondientes al

conjunto de subíndices $\boldsymbol{i}_j^{obs} = \{1, 2, \ldots, n\} \backslash \boldsymbol{i}_j^{miss}$. Dado que este conjunto depende de los valores observados de X_j, puede que $\boldsymbol{X}_{-j}^{obs}$ contenga observaciones faltantes.

- $\boldsymbol{X}_{-j}^{miss}$: matriz de variables X_l, con $l \neq j$, con observaciones correspondientes al conjunto de subíndices $\boldsymbol{i}_j^{miss}$. Dado que este conjunto depende de los valores faltantes de X_j, puede que $\boldsymbol{X}_{-j}^{miss}$ no contenga observaciones faltantes.

Para comenzar, se completan los valores faltantes de la matriz $\boldsymbol{X}$ mediante imputación por algún método simple. Luego, se ordenan las variables X_j, $j = 1, \ldots, r$, de acuerdo con la cantidad de valores faltantes, empezando con la que contenga menor cantidad de estos. Para cada variable X_j, los datos faltantes son imputados mediante Random Forest considerando como respuesta $\boldsymbol{x}_j^{obs}$ y como predictores $\boldsymbol{X}_{-j}^{obs}$, una matriz completa en base a los datos observados y a los imputados en el paso anterior. Luego, se predicen los valores $\boldsymbol{x}_j^{miss}$ aplicando el modelo Random Forest entrenado a $\boldsymbol{X}_{-j}^{miss}$ (matriz formada por datos observados y datos imputados en el paso anterior), obteniendo $\boldsymbol{x}_j^{*(k)}$, donde k indica el paso del proceso iterativo. El algoritmo se repite hasta que el criterio de convergencia γ es alcanzado. Dicho criterio se basa en la nueva matriz de datos imputados y la matriz imputada en el paso anterior. Si de las r variables que componen la matriz $\boldsymbol{X}$ existen N variables cuantitativas y F variables cualitativas, de modo que $r = N + F$, la diferencia entre los valores $x_{ij}^{*(k)}$ de la matriz $\boldsymbol{X}^{*(k)}$ imputada en el k-ésimo paso del proceso iterativo y los valores $x_{ij}^{*(k-1)}$ de la matriz $\boldsymbol{X}^{*(k-1)}$ imputada en el paso anterior se evalúa según las siguientes fórmulas:

- Para las N variables cuantitativas: $\Delta_N = \dfrac{\sum_{j \in N} \sum_{i \in \boldsymbol{i}_j^{miss}} \left(x_{ij}^{*(k)} - x_{ij}^{*(k-1)} \right)^2}{\sum_{j \in N} \sum_{i \in \boldsymbol{i}_j^{miss}} \left(x_{ij}^{*(k)} \right)^2}$.

- Para las F variables cualitativas: $\Delta_F = \dfrac{\sum_{j \in F} \sum_{i \in \boldsymbol{i}_j^{miss}} l_{ij}}{\#NA}$, siendo l_{ij} los elementos de una matriz $\boldsymbol{L}$ que toman el valor 1 cuando $x_{ij}^{*(k)} \neq x_{ij}^{*(k-1)}$ o 0 cuando $x_{ij}^{*(k)} = x_{ij}^{*(k-1)}$ y $\#NA$ el número de valores faltantes en dichas variables.

El criterio γ establece parar el procedimiento cuando Δ_N o Δ_F se incrementen por primera vez.

El pseudo-código correspondiente al Algoritmo 2, incluído en el Anexo I, representa el algoritmo utilizado por la técnica missForest.

3.3.4. Imputación múltiple por ecuaciones encadenadas

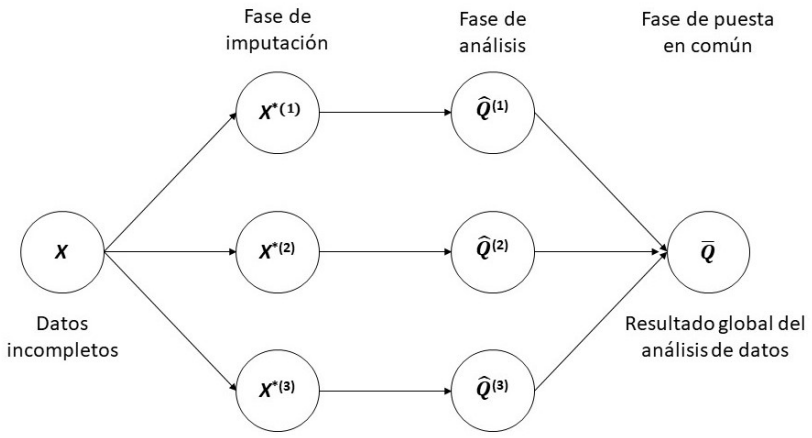
3.3.4.a. Imputación múltiple

La *imputación múltiple* (*multiple imputation*, MI) ha incrementado su popularidad debido al desarrollo computacional ocurrido en los últimos años [57, 76]. La clave de MI es la utilización de la distribución de los datos observados para estimar valores razonables para imputar los datos faltantes. Se incorporan componentes aleatorios a la estimación de los valores a fin de reflejar la falta de certeza de los mismos, mediante la generación de múltiples conjuntos de datos completos ($m \geq 1$). Se sugiere que el valor de m no sea pequeño para evitar errores de imputación importantes, especialmente si la proporción de datos faltantes es alta. El proceso de imputación requiere de la estimación de modelos de relación entre la variable a imputar y las variables conocidas. Los modelos habitualmente considerados se describen en la sección *3.3.4.c*.

Sea, nuevamente, X_j una variable con datos faltantes de la matriz de datos $\boldsymbol{X}$ y $\boldsymbol{x}_j^{obs}$, el sub-vector de $\boldsymbol{x}_j$ con valores completos y $\boldsymbol{x}_j^{miss}$ el sub-vector de $\boldsymbol{x}_j$ con valores faltantes. Los conjuntos $\boldsymbol{Z}^{obs} = \left\{\boldsymbol{x}_1^{obs}, \cdots, \boldsymbol{x}_r^{obs}\right\}$ y $\boldsymbol{Z}^{miss} = \{\boldsymbol{x}_1^{miss}, \cdots, \boldsymbol{x}_r^{miss}\}$ representan a los datos observados y perdidos, respectivamente, de la matriz de datos $\boldsymbol{X}$. La matriz de datos imputada con el k-ésimo conjunto de datos generados, con $1 \leq k \leq m$, se simboliza con $\boldsymbol{X}^{*(k)}$. Sea $\boldsymbol{X}_{-j} = (X_1, \cdots, X_{j-1}, X_{j+1}, \cdots, X_r)$ la matriz de datos compuesta por las variables de $\boldsymbol{X}$ salvo X_j. Supóngase que $\boldsymbol{Q}$ denota un vector de parámetros de interés, a estimar mediante los datos de $\boldsymbol{X}$ (por ejemplo, el vector de coeficientes de un modelo de regresión). El método de imputación múltiple está dividido en tres fases: fase de imputación, fase de análisis y fase de puesta en común (Figura 3.1).

En la fase de imputación, MI produce múltiples matrices completas (m) donde cada dato faltante es reemplazado por un valor obtenido aleatoriamente de la distribución de probabilidad de la variable correspondiente. La Figura 3.1 representa $m = 3$ conjuntos de datos imputados idénticos en los datos observados, pero diferentes en los valores imputados. La magnitud de esas diferencias refleja la incerteza respecto a los datos faltantes.

Figura 3.1: Proceso de análisis de datos mediante imputación múltiple.



La fase de análisis consiste en estimar $\boldsymbol{Q}$ mediante cada uno de los conjuntos de datos imputados, considerando que cada uno de ellos es un conjunto de datos completo. La estimación se realiza mediante el mismo mecanismo, de manera idéntica, para $\boldsymbol{X}^{*(1)}, \cdots, \boldsymbol{X}^{*(m)}$, obteniéndose las estimaciones $\hat{\boldsymbol{Q}}^{(1)}, \cdots, \hat{\boldsymbol{Q}}^{(m)}$. Finalmente, la fase de puesta en común consiste en amalgamar las m estimaciones $\hat{\boldsymbol{Q}}^{(1)}, \cdots, \hat{\boldsymbol{Q}}^{(m)}$ en una estimación única, $\bar{\boldsymbol{Q}}$, y obtener su matriz de varianzas y covarianzas, la cual incorpora tanto la variabilidad intra-imputación como la variabilidad entre-imputaciones (referencia). Si la estimación combinada $\bar{\boldsymbol{Q}}$ se obtiene como el promedio de las m estimaciones:

$$\bar{\boldsymbol{Q}} = \frac{1}{m} \sum_{k=1}^{m} \hat{\boldsymbol{Q}}^{(k)}, \tag{3.21}$$

la varianza total de $\bar{\boldsymbol{Q}}$ es:

$$var(\bar{\boldsymbol{Q}}) = \boldsymbol{W} + \left(1 + \frac{1}{m}\right) \boldsymbol{B}, \tag{3.22}$$

donde $\boldsymbol{W} = \frac{1}{m} \sum_{k=1}^{m} \boldsymbol{W}^{(k)}$ corresponde a la matriz de varianza intra-imputación y $\boldsymbol{B} = \frac{1}{(m-1)} \sum_{k=1}^{m} (\hat{\boldsymbol{Q}}^{(k)} - \bar{\boldsymbol{Q}})^2$, a la varianza entre-imputaciones.

La propuesta del estimador amalgamado supone estimaciones de $\boldsymbol{Q}$ con distribución asintótica Normal [57]. En casos particulares, por ejemplo, estimaciones de razones de odds, razones de riesgo, riesgo basal, probabilidad de supervivencia, desvio estándar, correlación, proporción de varianza explicada, coeficientes de asimetría y curtosis, etc. puede requerirse realizar transformaciones previas a la propuesta de estimación global [28].

Para generar los m conjuntos de valores imputados existen dos enfoques principales: *modelos conjuntos* (*joint modelling*, JM) y *especificación totalmente condicional* (*fully conditional specification*, FCS), también conocido como *imputación múltiple por ecuaciones encadenadas* (*multiple imputation by chained equations*, MICE).

Schafer propuso, utilizando el enfoque JM, considerar la normal multivariada como distribución marginal para las variables con pérdidas [77]. Las imputaciones se basan en distribuciones condicionales de esta Normal dadas las variables sin pérdidas. Si el supuesto de normalidad conjunta no es adecuado es preferible utilizar otros métodos más flexibles, como por ejemplo, FCS, que permite definir modelos específicos para cada variable. FCS ha sido propuesta por diferentes autores, utilizando diferentes nombres: *relajación estocástica* [78], *imputación variable por variable* [79], *cambio de regresión* [31], *regresiones secuenciales* [80], *muestreo pseudo-Gibbs ordenado* [81], *Markov Chain Monte Carlo* (MCMC) *parcialmente incompatible* [82], *imputación univariada iterativa* [83], MICE [84] y FCS [85]. Una de las ventajas de FCS es que puede imputar simultáneamente variables mixtas, ya que para cada variable se utiliza su propio modelo. En particular, se detalla el Algoritmo MICE, que es uno de los métodos comparados en esta investigación.

3.3.4.b. Algoritmo MICE

Este método supone que $\boldsymbol{X}$ corresponde a una muestra aleatoria multidimensional proveniente de una distribución $P(\boldsymbol{X}|\boldsymbol{\theta})$. Se asume que esta ley está completamente especificada por $\boldsymbol{\theta}$, un conjunto de parámetros desconocidos. El algoritmo MICE obtiene la distribución posterior de $\boldsymbol{\theta}$ muestreando iterativamente sobre las distribuciones condicionales de la forma:

$$\begin{aligned} &P(X_1|\boldsymbol{X}_{-1}, \boldsymbol{\theta}_1) \\ &\vdots \\ &P(X_r|\boldsymbol{X}_{-r}, \boldsymbol{\theta}_r), \end{aligned} \tag{3.23}$$

donde $\boldsymbol{X}_{-j}$ es la matriz definida en la sección *3.3.4.a* y $\boldsymbol{\theta}_j$ es el vector de parámetros que describe la distribución condicional de X_j.

Inicialmente, cada dato faltante en una variable, es imputado por un valor seleccionado al azar y con reposición entre los datos observados, o mediante algún método simple de imputación. La t-ésima iteración de las ecuaciones encadenadas es un muestreador de

Gibbs que genera sucesivamente:

$$
\begin{aligned}
\boldsymbol{\theta}_1^{*(t)} &\sim P\left(\boldsymbol{\theta}_1|\boldsymbol{x}_1^{obs}, \boldsymbol{x}_2^{(t-1)}, \cdots, \boldsymbol{x}_r^{(t-1)}\right) \\
\boldsymbol{x}_1^{*(t)} &\sim P\left(\boldsymbol{x}_1|\boldsymbol{x}_1^{obs}, \boldsymbol{x}_2^{(t-1)}, \cdots, \boldsymbol{x}_r^{(t-1)}, \boldsymbol{\theta}_1^{*(t)}\right) \\
\boldsymbol{\theta}_2^{*(t)} &\sim P\left(\boldsymbol{\theta}_2|\boldsymbol{x}_2^{obs}, \boldsymbol{x}_1^{(t)}, \boldsymbol{x}_3^{(t-1)}, \cdots, \boldsymbol{x}_r^{(t-1)}\right) \\
\boldsymbol{x}_2^{*(t)} &\sim P\left(\boldsymbol{x}_2|\boldsymbol{x}_2^{obs}, \boldsymbol{x}_1^{(t)}, \boldsymbol{x}_3^{(t-1)}, \cdots, \boldsymbol{x}_r^{(t-1)}, \boldsymbol{\theta}_2^{*(t)}\right) \\
&\vdots \\
\boldsymbol{\theta}_r^{*(t)} &\sim P\left(\boldsymbol{\theta}_r|\boldsymbol{x}_r^{obs}, \boldsymbol{x}_1^{(t)}, \cdots, \boldsymbol{x}_{r-1}^{(t)}\right) \\
\boldsymbol{x}_r^{*(t)} &\sim P\left(\boldsymbol{x}_r|\boldsymbol{x}_r^{obs}, \boldsymbol{x}_1^{(t)}, \cdots, \boldsymbol{x}_{r-1}^{(t)}, \boldsymbol{\theta}_r^{*(t)}\right)
\end{aligned}
$$

donde $\boldsymbol{x}_j^{(t)} = \left(\boldsymbol{x}_j^{obs}, \boldsymbol{x}_j^{*(t)}\right)$ es la j-ésima variable imputada en la iteración t. Obsérvese que $\boldsymbol{x}_j^{*(t-1)}$ solo se ingresa a $\boldsymbol{x}_j^{*(t)}$ mediante su relación con las otras variables y no en forma directa, por lo tanto, la convergencia puede ser bastante rápida.

El método MICE supone que las pérdidas respondan a un mecanismo MAR o MNAR. En este último caso se requieren supuestos de modelado adicionales que influyen en las imputaciones [24]. Este método puede utilizar modelos específicos en la etapa de imputación para cada variable incompleta en los datos, por ejemplo, regresión múltiple, modelo log-lineal, regresión logística, etc. Para ello, se deben especificar tanto la parte estructural del modelo como la distribución asumida para el error. Respecto de las variables explicativas en estos modelos, se sugiere incluir las variables relevantes y sus interacciones, no superando una cantidad de 15 a 25. Las selección de las variables apropiadas, especialmente en casos de grandes matrices de datos, puede realizarse siguiendo la estrategias propuestas por Van Buuren *et al.* [24].

El orden en que las variables son imputadas puede afectar a la convergencia del algoritmo. En particular, para datos que faltan monotónicamente, la convergencia es inmediata si las variables están ordenadas de acuerdo al número de casos faltantes, pero puede no darse en otros casos. El número de iteraciones sugeridas es entre 10 y 20 para cada conjunto de datos imputados, a fin de estabilizar los resultados del procedimiento [24, 28].

El pseudo-código correspondiente a MICE (Algoritmo 3), se presenta en el Anexo I y describe el proceso en el caso general donde todas las variables de $\boldsymbol{X}$ son considera-

das como explicativas al momento del ajuste de modelos para imputación, y sin incluir interacciones ni transformaciones de las mismas.

3.3.4.c. Modelos para imputar variables según su clasificación

Entre los métodos de imputación múltiple descriptos frecuentemente se hace mención a la utilización de distintos modelos para predecir los valores faltantes en función de algunos de los valores observados de la matriz de datos. En este apartado, se indican algunos de los modelos comúnmente considerados para imputar variables con datos faltantes, de acuerdo a su clasificación.

A modo general, supóngase que se quieren imputar los datos faltantes correspondientes a la variable X_j, la cual será considerada variable respuesta en un modelo adecuado. Por simplicidad, se considera el caso en el que la matriz de variables explicativas es $\boldsymbol{X}_{-j} = (X_1, \cdots, X_{j-1}, X_{j+1}, \cdots, X_r)$, de dimensión $n^{obs} \times (r-1)$, siendo n^{obs} el número de datos observados para la variable X_j y r la cantidad de parámetros a estimar, incluyendo el intercepto. El vector de parámetros del modelo se simboliza con $\boldsymbol{\beta}$.

Variable respuesta continua: El modelo de regresión lineal $X_j = \boldsymbol{\beta}\boldsymbol{X}_{-j} + e$, es la elección típica para imputar variables continuas con distribución Normal:

$$(X_j|\boldsymbol{X}_{-j};\boldsymbol{\beta}) \sim N(\boldsymbol{\beta}\boldsymbol{X}_{-j}, \sigma^2). \tag{3.24}$$

Como paso inicial, se completa la matriz $\boldsymbol{X}_{-j}$ y se calcula $\hat{\boldsymbol{\beta}}$. Este estimador tiene distribución asintótica Normal con matriz de covarianzas estimada $\hat{\boldsymbol{V}}$. σ^2 se estima con la raiz del error cuadrático medio. A continuación, se seleccionan los parámetros de imputación σ^* y $\boldsymbol{\beta}^*$ a partir de las distribuciones posteriores de σ y $\boldsymbol{\beta}$. En un primer paso, σ^* se obtiene como:

$$\sigma^* = \hat{\sigma}\sqrt{\frac{(n^{obs} - r)}{g}}, \tag{3.25}$$

donde g es un valor aleatorio de una distribución χ^2 de $n^{obs} - r$ grados de libertad. Luego, $\boldsymbol{\beta}^*$ se obtiene como:

$$\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}} + \frac{\sigma^*}{\hat{\sigma}}\boldsymbol{u}_1\hat{\boldsymbol{V}}^{\frac{1}{2}}, \tag{3.26}$$

donde $\boldsymbol{u}_1$ es un vector fila de r valores independientes seleccionados al azar a partir de

una distribución Normal estándar y $\hat{\boldsymbol{V}}^{\frac{1}{2}}$ es la descomposición de Cholesky de $\hat{\boldsymbol{V}}$. Luego, el valor imputado para la posición i-ésima, x^*_{ij}, se obtiene como:

$$x^*_{ij} = \boldsymbol{\beta}^* \boldsymbol{X}_{-j,i} + u_{2i}\sigma^*, \tag{3.27}$$

donde u_{2i} es un valor obtenido al azar a partir de una distribución Normal estándar.

El supuesto de distribución Normal para las variables a imputar no siempre puede asumirse. Si aún así se trabaja bajo este supuesto, la distribución de los valores imputados no representará a la de los valores observados, es decir, existirá falta de validez aparente. Si, por ejemplo, X_j es intrínsecamente positiva y aún así es imputada por una regresión lineal en $\boldsymbol{X}_{-j}$, puede ocurrir que el proceso de imputación complete un dato faltante mediante un valor negativo, lo cual implicaría un error. Para evitar este inconveniente, White *et al.* proponen dos alternativas: realizar una transformación de la variable a imputar, de modo que la variable transformada tenga distribución Normal o utilizar el método de *coincidencia media predictiva* (*predictive mean matching*, PMM) [28]. Este método imputa valores perdidos con la propiedad de que los valores imputados se muestrean solamente a partir de los valores observados para la misma variable. En consecuencia, su distribución a menudo coincide con la de la variable observada. Esta estrategia también es útil cuando la relación entre la variable a imputar y sus predictoras no es lineal. Sin embargo, no es adecuada cuando la imputación involucra la extrapolación más allá del rango de los valores observados o cuando el tamaño de la muestra es pequeño.

PMM se utiliza en la última etapa del método de imputación descripto, identificando las q unidades con los menores valores de $|\hat{\boldsymbol{\beta}}\boldsymbol{X}_{-j,h} - \boldsymbol{\beta}^*\boldsymbol{X}_{-j,i}|$, con $(h = 1, \cdots, n^{obs})$, es decir, identifica un conjunto de q casos con valor observado en la variable a imputar cuyos valores predichos se encuentran cerca del valor predicho para el dato faltante. La elección del valor a imputar se hace por seleccionan al azar entre los q casos cercanos. De este modo, el valor imputado es un valor observado cuya predicción con los datos observados coincide estrechamente con la predicción perturbada [86]. Respecto del número de casos q entre los cuáles seleccionar el valor imputado, no hay acuerdo acerca de cuál es su valor óptimo. Distintos autores han realizado trabajos por simulación encontrando que el mejor valor depende, en gran medida, del tamaño de la muestra [87, 88]. Con muestras grandes, probablemente la mejor opción es $q = 10$, pero con muestras más pequeñas, este valor probablemente llevaría a elegir entre casos muy diferentes. Esta falta de acuerdo lleva a

utilizar con frecuencia un valor de compromiso $q = 5$.

Variable respuesta binaria: Para imputar una variable binaria X_j, el modelo usualmente utilizado es el de regresión logística,

$$logit\ P(X_j = 1|\boldsymbol{X}_{-j}; \boldsymbol{\beta}) = \boldsymbol{\beta}\boldsymbol{X}_{-j}. \tag{3.28}$$

Sea $\hat{\boldsymbol{\beta}}$ el vector de parámetros estimados mediante el ajuste del modelo con las unidades con valores observados de X_j y $\hat{\boldsymbol{V}}$ su matriz de varianzas y covarianzas estimada. Sea $\boldsymbol{\beta}^*$ un valor seleccionado en base a la distribución posterior aproximada de $\boldsymbol{\beta}$, una Normal multivariada con parámetros $\hat{\boldsymbol{\beta}}$ y $\hat{\boldsymbol{V}}$. Para cada observación faltante x_{ij}^{miss}, sea $p_i^* = [1 + \exp(-\boldsymbol{\beta}^* \boldsymbol{X}_{-j,i})]^{-1}$, entonces se selecciona un valor de imputación x_{ij}^* como:

$$x_{ij}^* = \begin{cases} 1 & \text{si } u_i < p_i^* \\ 0 & \text{en otro caso,} \end{cases} \tag{3.29}$$

donde u_i es un valor aleatorio de una distribución Uniforme $(0, 1)$. Este procedimiento resulta sencillo de implementar. Sin embargo, pueden surgir problemas en la identificación de $\boldsymbol{\beta}^*$ debido a la *predicción perfecta*, que se produce cuando una o más observaciones tienen una probabilidad ajustada exactamente 0 o exactamente 1. La misma dificultad surge para las variables cualitativas con más de dos categorías.

Variable respuesta cualitativa nominal: Una variable cualitativa X_j, con $L > 2$ categorías o niveles, puede modelarse mediante un modelo de regresión logística multinomial en el cual, a cada una de las categorías l, le corresponde un modelo de regresión logística que compara la probabilidad de ocurrencia de la misma frente a una categoría basal (por ejemplo, 1):

$$P(X_j = l|\boldsymbol{X}_{-j}; \boldsymbol{\beta}) = \left[\sum_{l'=1}^{L} \exp(\boldsymbol{\beta}_{l'} \boldsymbol{X}_{-j})\right]^{-1} \exp(\boldsymbol{\beta}_l \boldsymbol{X}_{-j}), \tag{3.30}$$

donde $\boldsymbol{\beta}_l$ es un vector de dimensión r y $\boldsymbol{\beta}_1 = 0$. Sea $\boldsymbol{\beta}^*$ un vector aleatorio seleccionado en base a la aproximación Normal de la distribución posterior de $\boldsymbol{\beta} = (\beta_2, \cdots, \beta_L)$, de dimensión $r(L-1)$. Para cada observación faltante x_{ij}^{miss}, sea $p_{il}^* = P(x_{ij}^{miss} = l|\boldsymbol{X}_{-j,i}; \boldsymbol{\beta}^*)$,

con $l = 1, \cdots, L$ y $c_{il} = \sum_{l'=1}^{l} p^*_{il'}$. Cada valor imputado x^*_{ij} resulta:

$$x^*_{ij} = 1 + \sum_{l=1}^{L-1} I(u_i > c_{il}), \tag{3.31}$$

donde u_i es un valor aleatorio de una distribución Uniforme $(0, 1)$ y $I(u_i > c_{il}) = 1$ si $u_i > c_{il}$ o $I(u_i > c_{il}) = 0$ en otro caso.

Variable respuesta cualitativa ordinal: Una variable cualitativa ordinal X_j, con $L > 2$ categorías o niveles, puede modelarse mediante un modelo de regresión logística multinomial o bien un modelo de odds proporcionales, que extiende el modelo logístico binario al restringir las probabilidades de pertenencia a una clase para X_j de acuerdo con un supuesto de odds proporcionales entre las categorías ordenadas. A diferencia del modelo logístico multinomial, este modelo tiene un solo predictor lineal, $\boldsymbol{\beta X}_{-j}$. El modelo es:

$$logit\ P(X_j \leq l|\boldsymbol{X}_{-j}; \boldsymbol{\beta}, \boldsymbol{\zeta}) = \zeta_l - \boldsymbol{\beta X}_{-j}, \tag{3.32}$$

donde $\boldsymbol{\zeta} = (\zeta_1, \cdots, \zeta_{L-1})$ con $-\infty < \zeta_l < \cdots < \zeta_L < \infty$. $\boldsymbol{\beta}$ y $\boldsymbol{\zeta}$ son estimados por máxima verosimilitud y los valores $\boldsymbol{\beta}^*$ y $\boldsymbol{\zeta}^*$ se obtienen aleatoriamente de la distribución Normal aproximada de sus distribuciones posteriores. La probabilidad de pertenencia de la i-ésima unidad a la categoría l, con $l = 1, \cdots, L$, está dada por $p^*_{il} = P(x^{miss}_{ij} \leq l|\boldsymbol{X}_{-j,i}; \boldsymbol{\beta}^*, \boldsymbol{\zeta}^*) - P(x^{miss}_{ij} \leq l-1|\boldsymbol{X}_{-j,i}; \boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)$. Luego,el proceso de imputación se realiza del mismo modo que en el caso de variables cualitativas nominales.

3.4. Antecedentes en el tratamiento de datos faltantes en modelos de regresión de Cox

Uno de los trabajos pioneros en evaluar la eficacia de distintos métodos de imputación para estimar modelos de regresión de Cox es el publicado por Van Buuren *et al.* en 1999, quienes presentan un caso estudio particular en el cual se pretende estimar el riesgo de mortalidad de individuos mayores de 85 años según edad, sexo, presión sanguínea y un conjunto de 29 covariables relacionadas con diferentes aspectos de la salud. Las únicas variables con datos perdidos son medidas en escala continua (presión sanguínea sistólica y diastólica). Los autores proponen utilizar imputación múltiple y solo focalizan la comparación en la estimación del riesgo de muerte con el subgrupo de casos completos y con

la matriz con datos imputados [31]. El trabajo plantea aspectos prácticos relacionados con la especificación de los modelos utilizados para realizar las imputaciones mediante la información observada en las otras variables cuando la cantidad de datos completos es grande y se derivan pautas generales para la elección de los modelos de imputación múltiple. Los resultados no muestran diferencias entre las estimaciones obtenidas con los casos completos e imputación múltiple, por lo cual los autores concluyen que, en el caso estudiado, la estimación del riesgo es robusta ante la presencia de datos faltantes.

Clark & Altman en 2003 [37] utilizan las recomendaciones establecidas en [31] para realizar imputaciones múltiples sobre un conjunto de variables explicativas mixtas en otro estudio particular para predecir el riesgo de muerte en pacientes con cáncer de ovario. Sin embargo, en este trabajo los autores evalúan la eficiencia de la imputación múltiple, en comparación con el uso de CCA, mediante un estudio de la calibración y la capacidad discriminante de los modelos de Cox estimados para predecir mortalidad. Los resultados obtenidos resultan más favorables cuando se utiliza la matriz imputada y los autores concluyen que, debido al uso de imputación múltiple, aumenta la potencia y la precisión de las estimaciones.

En 2004, fue presentado otro caso real por Barzi & Woodward, el cual considera que las pérdidas se presentan en una única variable cuantitativa [35]. Se trata de un estudio de cohorte multicéntrico que estudia los determinantes de enfermedades cardiovasculares en países asiáticos y se consideran distintos métodos de imputación para asignar valores a los datos perdidos. Si bien las pérdidas solo se producen en la misma variable a través de todos los centros intervinientes en el estudio, el porcentaje de pérdidas en cada centro es diferente. La comparación se establece sobre las razones de riesgo de mortalidad por causas cardiovasculares estimadas solamente con los individuos con datos completos y con la muestra total con datos imputados. El estudio permite concluir que, en los centros donde hubo porcentajes de pérdidas inferiores al 10 % todos los métodos produjeron resultados similares, pero cuando porcentajes están entre el 10 % y el 60 %, existen claras diferencias entre los métodos, por lo que se recomienda utilizar imputación múltiple siguiendo recomendaciones de otros autores. Para porcentajes superiores, se encuentra que ninguna técnica funciona de manera aceptable. Cuando se analiza el conjunto de los centros a través de un metanálisis, se encuentran resultados similares al aplicar el método de casos completos y diferentes modelos de imputación, sugiriendo una ventaja del metanáli-

sis frente a la presencia de valores faltantes, situación muy común en la epidemiología observacional.

Un trabajo posterior, publicado en 2009, evalúa la importancia de agregar al conjunto de covariables del modelo utilizado para realizar imputaciones múltiples al riesgo basal y el estado (censura o evento) de los individuos [30]. En este trabajo se considera solo una variable con datos faltantes que responde a un mecanismo de pérdida MCAR o MAR. Esta variable puede ser binaria con probabilidad de éxito $\pi = 0.50$, o continua con distribución Normal estándar. Mediante simulación, se generan diferentes conjuntos de datos en base a modelos de Cox prefijados. Sobre ellos se producen pérdidas en dos proporciones distintas (p=0.20 o 0.50) y se utiliza solo imputación múltiple comparando resultados a través de distintos modelos. Las variables explicativas que se utilizan incluyen el riesgo basal, el estado, el tiempo hasta el evento o censura o alguna función de ellas, en diferentes combinaciones. Las covariables se generan a partir de distribuciónes Normales estándar. Para las comparaciones de los métodos se evalúa el sesgo que se produce en la estimación de los parámetros del modelo de Cox respecto del valor conocido que da origen a los datos, la diferencia entre el error estándar teórico y el observado para dichas estimaciones y la cobertura de los intervalos de confianza. Este trabajo también presenta un caso de aplicación donde se pretende estimar el riesgo de mortalidad de pacientes con carcinoma renal por medio de seis variables explicativas, donde una de ellas presenta más de la mitad de sus datos perdidos. Si bien no en todos los casos se observa una ventaja al utilizar imputación múltiple respecto de analizar los casos completos, los autores recomiendan el uso de métodos de imputación múltiple que consideren modelos que incluyan como covariables el estado y una estimación del riesgo acumulado. Este trabajo es el único que sugiere la incorporación de la estimación de un riesgo como covariable.

Recién en el año 2010 se publica uno de los trabajos más completos en relación al efecto de diferentes métodos de imputación sobre la estimación de parámetros de un modelo de regresión de Cox [34]. Marshall *et al.* simulan un conjunto de datos a partir de un modelo basado en los resultados de un estudio de sobrevida en pacientes con cáncer de mamas. Las covariables a incluir en el conjunto de datos se simulan siguiendo la misma distribución de probabilidad que la observada en las variables explicativas consideradas en dicho estudio, incluyendo tanto variables categóricas como cuantitativas, algunas de las cuales presentan distribución asimétrica. También se simula la ocurrencia de censura

y el tiempo hasta el evento. La simulación considera solo un tamaño de muestra y este es muy grande ($n = 1000$). Los autores consideran diferentes mecanismos de pérdidas (MCAR, MAR, MNAR y una combinación de ellos) sobre cuatro covariables mixtas y las imputaciones se realizaron solo utilizando MICE, en sus versiones de imputación simple y múltiple, considerando diferentes configuraciones dentro del algoritmo que son comparadas con CCA. En todos los casos, se incluye en los modelos de imputación el estado y el logaritmo del tiempo hasta el evento o censura como variables explicativas. Las medidas de eficiencia evaluadas son sesgo de las estimaciones, sus desvios estándar, significación de las covariables, bondad de ajuste, habilidad discriminante y capacidad predictiva de los modelos resultantes. Los autores concluyen que CCA produce estimaciones menos sesgadas y con mejor cobertura que imputación múltiple. Sin embargo, cuando el porcentaje de pérdidas es superior al 25 % y se observa la significancia de las covariables, el método CCA pierde eficiencia. Cuando el porcentaje de pérdidas es inferior al 50 % y el mecanismo de las mismas no es MNAR, recomiendan el uso de MICE con PMM como mecanismo de imputación, debido a sus ventajas cuando alguna covariable es asimétrica. Aunque los autores no lo mencionan, la buena *performance* del método CCA en los casos señalados podría estar influenciada por el importante tamaño de muestra considerado en las simulaciones.

En [34], los autores manifiestan preferir los estudios por simulación a partir de datos reales por considerar que así se evita elegir arbitrariamente un conjunto de parámetros que dificulten a posteriori la obtención de conclusiones generales. Sin embargo, critican esta estrategia en un trabajo posterior [89] considerándola simplista, prefiriendo estudiar los comportamientos aleatorios a través de remuestreo para reflejar el nivel apropiado de diversidad y variabilidad encontrado en poblaciones realistas. El conjunto de datos elegido corresponde a un estudio de sobrevida de pacientes con cáncer colorrectal, del cual se seleccionan 500 muestras de tamaño $n = 1000$ por remuestreo. Se producen pérdidas suponiendo un mecanismo MAR sobre tres de ocho covariables, con porcentajes de datos perdidos que van del 5 % al 75 %. Con este tranajo, los autores ratifican las conclusiones obtenidas en [34].

Otros autores que también emplean remuestreo son Ali *et al.* [17]. Los mecanismos de pérdida considerados son MCAR y MAR, comparando CCA, RVI y métodos de imputación múltiple, tanto incluyendo como no el estado (evento o censura) como variable

explicativa en el modelo de imputación. Como particularidad, en este estudio las muestras por bootstrap mantienen siempre el tamaño de la población original ($n = 5443$) así como la proporción de datos faltantes en cada variable (entre un 6 % y un 30 %). Las variables con pérdidas son solo categóricas y, como medidas de eficiencia, se evalúa el sesgo y desvío estándar en las estimaciones de los parámetros del modelo de Cox. Se recomienda imputación múltiple incluyendo el estado como variable informativa para obtener estimaciones con menos sesgo.

Solo algunos autores proponen un método denominado *probabilidad inversa aumentada ponderada* (*augmented inverse probability weighted*, AIPW) [32, 33, 90]. Este método no se trata de un método de imputación, sino que implica una modificación en la forma de estimar los parámetros del modelo de interés. Lo recomiendan como opción atractiva para el tratamiento de datos faltantes en estudios de sobrevida y proponen un método no paramétrico que consideran robustos frente a las pérdidasy eficiente para reducir el sesgo en las estimaciones de interés. Estudian escasas medidas de eficiencia y consideran a lo sumo tres variables explicativas con pérdidas en una sola de ellas.

En general, los trabajos analizados utilizan imputaciones múltiples bajo diferentes configuraciones y comparan su eficiencia en relación al uso de CCA, cuyas desventajas son conocidas. El uso de imputaciones múltiples implica el ajuste de modelos, los cuales requieren del cumplimiento de supuestos que no siempre son tenidos en cuenta por los analistas de datos. Por ende, resulta importante incluir en las comparaciones, técnicas de imputación más flexibles, como lo son MF y KNN. En ese sentido, solo se ha encontrado un trabajo donde se evalúa el uso de MICE, MF y un nuevo método que combina MICE con Random Forest, en el contexto de modelos de regresión de Cox, a través de remuestreo, a partir de un conjunto de datos reales, sobre los que se generan datos perdidos según un mecanismo MCAR sobre una variable cualitativa y pérdidas MAR con un patrón similar al observado en el conjunto de datos reales [91]. Si bien los resultados obtenidos mediante el uso de MICE y del nuevo método propuesto muestran ventajas sobre MF, fundamentalmente en relación al sesgo y cobertura, los autores concluyen que MF es un método interesante en situaciones con relaciones complejas entre las variables y que debe ser estudiado en mayor profundidad para obtener resultados generales.

Los antecedentes evaluados evidencian la falta de investigaciones que tengan en cuenta situaciones generales que incluyan tanto métodos modelo-dependientes como métodos

no paramétricos más flexibles. Además, en el espectro de escenarios considerados hay escasos trabajos con variables explicativas mixtas y variantes del grado de incidencia de los valores perdidos sobre el total de datos, la cantidad de variables donde se producen pérdidas y que incluyan la consideración de la influencia de distintos tamaños muestrales sobre las propiedades de los estimadores. A más de esto, hay propiedades importantes de los estimadores que no han sido estudiadas al evaluar comparativamente distintas alternativas de métodos de imputación. En los dos capítulos siguientes se proponen y desarrollan estudios por simulación (Capítulo 4) y por remuestreo a partir de un caso de la vida real (Capítulo 5) que pretenden llenar este vacío.

4. Impacto de la imputación de datos perdidos en la estimación de modelos de Cox. Estudio por simulación

En este capítulo se presentan los resultados de un estudio por simulación de tiempos de sobrevida a partir de un modelo especificado. También se describen los escenarios considerados, las medidas de eficiencia definidas para evaluar el impacto de las pérdidas así como los aspectos computacionales para su implementación.

4.1. Diseño del estudio comparativo

Se plantea un modelo de regresión de Cox para tiempos de sobrevida con variables explicativas mixtas que son generadas a partir de distribuciones preestablecidas. A partir de este modelo se generan muestras de tamaño $n = 100$, 200, 500 y 1000 y, en ellas, pérdidas según mecanismos MCAR, MAR y MNAR con distintas proporciones de datos faltantes ($p = 0.10$, 0.20, 0.30 y 0.40) sobre todas o sobre algunas de las variables, según el mecanismo de pérdida. En el caso del mecanismo MCAR con $n = 100$ y 200, solo se consideran los porcentajes de pérdida $p = 0.10$, 0.20 y 0.30, a fin de mantener una proporción suficiente de información conocida. Cada escenario (combinación de mecanismo de pérdida, proporción de datos faltantes y tamaño de muestra) se replica 5000 veces y en cada una se utilizan distintos métodos de tratamiento de datos faltantes: CCA, RVI, KNN, MF y MICE. El modelo planteado es:

$$h(t) = \exp(-0.10 \cdot X_1 + 0.25 \cdot X_2 - 0.13 \cdot X_3 + 0.50 \cdot X_4 + 0.35 \cdot X_5 - 0.15 \cdot X_6) h_0(t), \quad (4.1)$$

donde X_1, X_3 y X_4 se suponen variables cuantitativas absolutamente continuas, X_2 es dicotómica y X_5 y X_6 son variables *dummies* que, conjuntamente, representan el resultado de una variable cualitativa de tres categorías.

Los modelos utilizados para generar las variables aleatorias continuas son: para X_1 Normal estándar, para X_3 Uniforme en el intervalo (0,10) y para X_4 Exponencial con tasa de falla igual a 1. La variable dicotómica X_2 se genera a partir de una distribución Binomial con probabilidad de éxito igual a 0.30, y para las variables X_5 y X_6 se generan valores para una variable Y con distribución Multinomial con 3 resultados posibles, con probabilidades $p_1 = 0.30$, $p_2 = 0.20$ y $p_3 = 0.50$ y, a partir de ella, si $y = 1$ entonces $(x_5, x_6) = (0, 0)$, si $y = 2$ entonces $(x_5, x_6) = (0, 1)$ y si $y = 3$, $(x_5, x_6) = (1, 0)$.

La aleatoriedad de la información de las muestras se incorpora al modelo mediante la simulación de los valores de las covariables. El riesgo basal se genera de acuerdo a una función log-normal con media 50 y desvío estándar 10. Los valores del tiempo hasta el evento se simulan por Monte Carlo eligiendo un número al azar entre 0 y 1 e identificando el tiempo correspondiente sobre curva de sobrevida $S(t) = S_0(t)^{exp(\boldsymbol{x}'\boldsymbol{\beta})}$. Se establece un tiempo máximo de seguimiento de $T = 1000$ y se impone un porcentaje de censura por derecha del 20 %.

Para generar datos faltantes según un mecanismo MCAR, se eliminan aleatoriamente $n \times p$ observaciones para cada variable explicativa en cada conjunto de datos completos.

Para generar pérdidas de acuerdo a un mecanismo MAR, se eliminan aleatoriamente $n \times p$ observaciones de la variable X_1 para las unidades en las que $x_2 = 0$ y $n \times p$ observaciones de las variables X_5 y X_6 cuando el valor observado en la variable X_3 es superior a su mediana.

Para generar datos faltantes según un mecanismo MNAR, se eliminan aleatoriamente $n \times p$ observaciones del valor 0 de la variable X_2 y de los valores de X_4 inferiores a su mediana.

En cada iteración del procedimiento y para cada escenario simulado, se estiman los parámetros del modelo considerando cinco métodos para el tratamiento de datos faltantes, uno de ellos que descarta las observaciones con valores perdidos (CCA) y cuatro que completan la matriz de información previamente a la estimación de los parámetros del modelo 4.1. Los cuatro métodos considerados pueden tratar simultáneamente pérdidas en

variables cuantitativas continuas, cualitativas o dicotómicas y corresponden a alternativas de imputación simple y múltiple, dos parámetricas (RVI y MICE) y dos no paramétricas (KNN y MF).

El método RVI, para valores faltantes en variables cualitativas, asigna el valor modal de los datos observados mientras que, para variables cuantitativas, reemplaza el valor faltante por la mediana de los valores observados.

Con KNN se utilizan diferentes cantidades de donantes en el proceso de imputación: $k = 3$, 5 y 9. Para cada una de las opciones de k, se consideran dos alternativas para evaluar cercanías utilizando la distancia de Gower. Una de ellas considera solamente las variables sin pérdida y la otra agrega la información del *estado* y el *tiempo hasta el evento o censura*.

En el método MF se ajustan 100 árboles de decisión en cada bosque aleatorio, realizando un máximo de 10 iteraciones del proceso de imputación. También con esta técnica se repite el proceso considerando o no el *estado* y el *tiempo* como variables informativas.

Con la metodología MICE las variables cuantitativas se imputan con el método PMM, descripto en la sección *3.3.4.b*. Para imputar las variables cualitativas se consideran modelos de regresión logística. Nuevamente, se repite el proceso considerando el *estado* y el *tiempo* como variables informativas y sin tenerlas en cuenta. El número de variantes para la imputación es $m = 5$.

4.1.1. Medidas de eficiencia

Se define un conjunto de medidas que son evaluadas comparativamente para identificar la *performance* de cada método para tratar la información faltante. Estas medidas son calculadas en cada iteración ($s = 5000$) y resumidas a través de promedios, desvíos y porcentajes para cada escenario. Los resultados se muestran en forma tabular o gráfica de modo de facilitar la comparación. Los aspectos distribucionales de los estimadores de los parámetros del modelo se reflejan a través de la significatividad del test de bondad de ajuste de Anderson-Darling. En las comparaciones se enfocan, para los estimadores de los coeficientes del modelo, aspectos como sesgo, error cuadrático medio, cobertura y distribución en el muestreo; para evaluar globalmente el modelo se comparan probabilidades de sobrevida para un tiempo mayor a 100 unidades y, para la precisión de las imputaciones,

se estudia el error cuadrático medio y la coincidencia de las imputaciones con los valores originales.

4.1.1.a. Medidas que enfocan las estimaciones de los parámetros

El **error cuadrático medio** (MSE) de los estimadores de β_j, para $j = 1, \cdots, p$, se calcula promediando las desviaciones cuadráticas de los $\hat{\beta}_{jl}$ respecto de β_j, a través de las repeticiones de cada escenario, donde l indica la iteración, $l = 1, \cdots, 5000$. Es decir:

$$\text{MSE}_j = \frac{\sum_{l=1}^{5000} \left(\hat{\beta}_{jl} - \beta_j\right)^2}{5000}. \tag{4.2}$$

El análisis del MSE se realiza individualmente para cada parámetro en cada escenario. Este engloba propiedades sobre sesgo y variabilidad. Estos dos aspectos también son estudiados por separado en una escala relativa a la magnitud del parámetro que resultan comparables.

El **sesgo de los estimadores** de β_j para $j = 1, \cdots, p$ se evalúa por medio de las diferencias relativas en valor absoluto del promedio de las estimaciones $\hat{\beta}_{jl}$ en cada escenario a través de las 5000 réplicas, respecto del verdadero valor de β_j en el modelo 4.1. Las desviaciones relativas se promedian para los seis parámetros del modelo a fin de obtener una medida global del sesgo relativo (SR):

$$\text{SR} = \frac{1}{6} \sum_{j=1}^{6} \frac{|\hat{\overline{\beta}}_j - \beta_j|}{|\beta_j|}, \tag{4.3}$$

donde $\hat{\overline{\beta}}_j = \frac{1}{5000} \sum_{l=1}^{5000} \hat{\beta}_{jl}$.

El indicador de variabilidad de los estimadores compara la variancia empírica (promedio de los desvíos estándar estimados de las 5000 réplicas) con la teórica esperada de las distribuciones asintóticas de los $\hat{\beta}_j$, estimados sin datos faltantes. Esta diferencia se considera relativa al valor del desvío teórico. Nuevamente se obtiene una medida global de variabilidad para los seis parámetros estimados. La diferencia de desvíos relativa (DDR) se obtiene como:

$$\text{DDR} = \frac{1}{6} \sum_{j=1}^{6} \frac{|\hat{\sigma}_j - \sigma_j|}{\sigma_j}, \tag{4.4}$$

donde $\hat{\sigma}_j = \sqrt{\frac{\sum_{l=1}^{5000}(\hat{\beta}_{jl} - \hat{\bar{\beta}}_j)^2}{4999}}$ y $\sigma_j = \frac{\sum_{l=1}^{5000} \boldsymbol{I}^{-1}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{jl}})}{5000}$.

La **cobertura**, C_j, se define como el porcentaje de los 5000 intervalos de confianza del 95 % que cubren al verdadero valor de β_j:

$$\mathrm{C}_j = \frac{\sum_{l=1}^{5000} \mathrm{I}_{\mathrm{IC}_{jl}}}{5000} \times 100\,\%, \tag{4.5}$$

con $\mathrm{I}_{\mathrm{IC}_{jl}} = 0$ si $\beta_j \notin \mathrm{IC}_{jl,95\,\%}$ y $\mathrm{I}_{\mathrm{IC}_{jl}} = 1$ si $\beta_j \in \mathrm{IC}_{jl,95\,\%}$.

Para evaluar la **aproximación de la distribución de los estimadores** $\hat{\beta}_j$ a la distribución teórica Normal conocida cuando se dispone de datos completos, se realizan pruebas de Anderson-Darling y se calcula el porcentaje de coeficientes para los que se rechaza la hipótesis de normalidad, considerando un nivel de significación del 5 %, es decir:

$$\%\text{ de rechazo} = \frac{\sum_{l=1}^{5000} \mathrm{I}_{ADp<0.05}}{5000} \times 100\,\%, \tag{4.6}$$

siendo $\mathrm{I}_{ADp<0.05} = 1$ si la probabilidad asociada al test de Anderson-Darling (ADp) es inferior a 0.05 y $\mathrm{I}_{ADp<0.05} = 0$ si ADp es mayor a 0.05. Este resultado se acompaña por los *box-plots* de las distribuciones empíricas obtenidas en cada escenario. A modo de ejemplo, se muestran para un parámetro en particular.

Como medida global de evaluación del modelo se define la consistencia del mismo para estimar la *probabilidad media de sobrevida* (PMS) después de $t = 100$, es decir:

$$\mathrm{PMS} = \frac{\sum_{l=1}^{5000} P_l(T > 100)}{5000}, \tag{4.7}$$

siendo $P_l(T > 100) = S_l(100|\boldsymbol{x}) = \left[\left(-\int_0^t h_0(u)du\right)\right]^{exp(\boldsymbol{x}'\hat{\boldsymbol{\beta}}_j)}$, para una unidad con $\boldsymbol{x} = (0, 1, 5, 1, 0, 1)$. La probabilidad real, calculada con los valores de $\boldsymbol{\beta}$, es, aproximadamente, 0.40.

4.1.1.b. Medidas que enfocan las imputaciones individuales

Para la comparación de los valores imputados con el valor real de la observación se definen dos medidas según el tipo de variable: el promedio de la raíz del error cuadrático medio normalizado ($\overline{\mathrm{NRMSE}}$) para las variables continuas y el porcentaje promedio de no coincidencias para las variables categóricas. Cada medida se calcula por variable imputada

en cada escenario. El $\overline{\text{NRMSE}}$ se define, para cada variable, como:

$$\overline{\text{NRMSE}}_j = \frac{\sum_{l=1}^{5000} \text{NRMSE}_{jl}}{5000}, \tag{4.8}$$

donde:

$$\text{NRMSE}_{jl} = \sqrt{\frac{\sum_{i=1}^{n} \left(x^*_{ijl} - x_{ijl}\right)^2}{\text{Var}\left(\boldsymbol{x}^{obs}_{jl}\right)}}, \tag{4.9}$$

siendo x^*_{ijl} los valores imputados y x_{ijl} los valores reales en el conjunto de datos completos, con i el indicador del individuo, j el indicador de la variable imputada y l el número de iteración correspondiente.

En el caso de MICE, dado que dicha técnica genera cinco conjuntos de datos imputados, se estableció el valor de NRMSE_{jl} como su promedio en los cinco casos a fin de obtener un único valor por iteración.

El porcentaje promedio de no coincidencia para las variables categóricas se calcula como el porcentaje de datos imputados que no coinciden con el valor correspondiente en el conjunto de datos completos, dividido la cantidad de valores imputados, promediado para las 5000 repeticiones. En el caso de MICE, el porcentaje de datos imputados en forma incorrecta en cada iteración se promedia para obtener un único valor por iteración.

4.1.2. Implementación computacional

El proceso de simulación, imputación y análisis de datos se realiza completamente en el lenguaje de programación R, mediante su entorno RStudio [92], incluyendo en cada caso un valor inicial para la generación de valores pseudo-aleatorios (*semilla*) a fin de garantizar la reproducibilidad de los resultados. El código completo utilizado en este capítulo se encuentra disponible en el Anexo II.

Para generar un tiempo hasta el evento o censura individual y un indicador de dicho estado, se utiliza el paquete de R *Coxed* [93], el cual permite generar conjuntos de datos simulados para modelos de Cox en base a los lineamientos establecidos por Harden & Kropko [94].

Para imputar los datos mediante KNN, se utiliza la opción *kNN* del paquete *VIM* [95]. Para hacerlo mediante MF, se utiliza el paquete *missForest* con todas las opciones

establecidas de forma predeterminada [96]. Y para realizar imputaciones mediante MICE, se emplea el paquete *MICE* [24].

Mediante la función *proc.time* se registra el tiempo demandado por cada método de imputación para arribar al conjunto de datos completos en cada iteración, obteniéndose luego el tiempo promedio requerido por cada escenario evaluado. Para la ejecución de los procedimientos, se utiliza una computadora de escritorio con sistema operativo Windows 10 Home x64, procesador Intel(R) Core(TM) i3-3227U, 1.90 GHz y 4.00 GB de memoria RAM.

4.2. Resultados

Los resultados se muestran separadamente según el mecanismo de pérdida. La sección 1, corresponde a los datos perdidos completamente al azar, la sección 2 corresponde a los datos perdidos al azar y la sección 3, a los perdidos no al azar. Dentro de cada sección, se muestran los resultados para las propiedades de los estimadores en forma individual y global, un apartado sobre la reproducibilidad de los datos perdidos y otro correspondiente al estudio de los tiempos computacionales requeridos para la imputación.

Se simboliza con MF+ al uso de *missForest* incluyendo el *estado* (evento o censura) y el *tiempo hasta el evento o censura* como variables informativas adicionales y, con MF-, cuando se utiliza dicha tenga sin incluir las mencionadas variables. De manera similar, se simboliza MICE+ y MICE- cuando se realizan *imputaciones múltiples por ecuaciones encadenadas* según se incluyan o no las variables adicionales. Para el método KNN, se utiliza KNN+ y KNN- de manera general y, por ejemplo, K3+ cuando se emplea KNN con 3 donantes e incluyendo las variables adicionales.

Los hallazgos sobre los errores cuadráticos medios (MSE) se presentan en forma tabular y no gráfica porque, aún estando estandarizados, presentan rangos muy diferentes a través de los distintos escenarios. En los Cuadros que muestran los resultados del MSE según el tamaño muestral y la proporción de datos faltantes, p, los valores sombreados en tonos rosados corresponden a los MSE más grandes para cada parámetro, siendo estos los resultados más desfavorables, mientras que los MSE sombreados en tonos verdes son aquellos de menor valor, indicando resultados más favorables en la estimación del correspondiente parámetro (Cuadros 4.1 a 4.4, 4.5 a 4.8 y 4.9 a 4.12).

4.2.1. Datos perdidos completamente al azar

4.2.1.a. Propiedades distribucionales de los estimadores de los coeficientes del modelo de regresión de Cox

Error cuadrático medio

Los resultados correspondientes al MSE para cada parámetro del modelo 4.1 se muestran en los Cuadros 4.1 a 4.4, según tamaño muestral y proporción de datos faltantes.

Para $n = 100$ y $p = 0.10$, los resultados más desfavorables se observan con el uso de CCA y MF+, mientras que MICE- aporta los menores valores de MSE para la mayoría de los parámetros. Cuando $p = 0.20$, los mayores MSE se encuentran con CCA y MF+, mientras que los menores se observan con K5- y MICE-, aunque este último resulta desfavorable para dos de los seis parámetros estimados, correspondientes a las variables cuantitativas no Normales del modelo (X_3 y X_4). Para $p = 0.30$, CCA, MF+ y MICE+ presentan los mayores MSE, y MICE- y K3+, los menores. En general, para este tamaño de muestra, se observan, dentro de cada técnica de imputación, mayores MSE al incluir el *tiempo* y el *estado* como variables informativas que al no incluirlas (Cuadro 4.1).

Para $n = 200$, también se encuentran resultados desfavorables al utilizar CCA y MF+. Para $p = 0.10$, K3- genera valores de MSE bajos. MICE- resulta favorable para cuatro de los seis parámetros, aunque su MSE es de los más altos para los otros dos parámetros estimados, correspondientes a las variables cuantitativas no Normales del modelo. Cuando $p = 0.20$ y 0.30, K5- se encuentra entre los mejores resultados y MICE- repite el comportamiento descripto para $p = 0.10$. No se observan tendencias en los MSE según se incluyan o no el *tiempo* y el *estado* como variables informativas (Cuadro 4.2).

Cuando $n = 500$ y $p = 0.10$, CCA y K3- presentan los mayores MSE. K5+, K9- y MICE+ presentan los MSE más bajos para tres parámetros y en ningún caso se ubican entre los mayores MSE para los otros parámetros. Para $p \geq 0.20$, CCA y MF+ se relacionan con los mayores MSE y MICE+ y K5+ tienen los menores MSE (Cuadro 4.3).

Para $n = 1000$ y $p \leq 0.30$, CCA, K3- y MICE- presentan los resultados menos favorables, mientras que K5+ y K9+ resultan los más favorables. Para $p = 0.40$, CCA, MICE- y MF- se asocian a los MSE más grandes en tres de los seis parámetros estimados y K5+, K3+ y MICE+ lo hacen a los MSE más chicos (Cuadro 4.4).

Para los dos tamaños de muestra más grandes se observan menores MSE al incluir el *tiempo* y el *estado* como variables informativas, de manera inversa a lo observado para $n = 100$. Para todos los tamaños de muestra, se destaca que el comportamiento de los métodos difiere según el parámetro estimado, con un comportamiento común para $\beta_1, \beta_2, \beta_5$ y β_6 y otro para β_3 y β_4.

Cuadro 4.1: Error cuadrático medio de los parámetros del modelo de regresión de Cox para n=100, según método de imputación y proporción de datos faltantes (p).

p	Método	B1	B2	B3	B4	B5	B6
0.1	CCA	0.02827	0.13620	0.00366	0.03803	0.24578	0.14085
	RVI	0.01568	0.07384	0.00409	0.02406	0.13672	0.07269
	K3-	0.01542	0.07159	0.00258	0.02350	0.12751	0.07018
	K3+	0.01743	0.07818	0.00241	0.02152	0.13833	0.08112
	K5-	0.01546	0.07215	0.00250	0.02196	0.12966	0.06972
	K5+	0.01740	0.07743	0.00238	0.02158	0.13775	0.08274
	K9-	0.01568	0.07248	0.00244	0.02117	0.13209	0.07104
	K9+	0.01717	0.07693	0.00235	0.02153	0.13884	0.08401
	MF-	0.01594	0.07252	0.00243	0.02193	0.12827	0.07075
	MF+	0.01832	0.08782	0.00246	0.02289	0.15828	0.09265
	MICE-	0.01308	0.06298	0.00252	0.02441	0.11134	0.06272
	MICE+	0.01544	0.07687	0.00240	0.02166	0.13631	0.07794
0.2	CCA	0.07890	0.49375	0.00992	0.11006	1.33516	0.47468
	RVI	0.01779	0.08107	0.00620	0.03355	0.15982	0.07845
	K3-	0.01720	0.07575	0.00334	0.03207	0.13915	0.07397
	K3+	0.02218	0.08990	0.00287	0.02569	0.16077	0.09930
	K5-	0.01771	0.07674	0.00313	0.02756	0.14093	0.07350
	K5+	0.02219	0.08877	0.00287	0.02573	0.16342	0.10227
	K9-	0.01819	0.07858	0.00288	0.02498	0.14506	0.07378
	K9+	0.02134	0.08537	0.00284	0.02572	0.16420	0.10576
	MF-	0.01867	0.07964	0.00287	0.02670	0.14314	0.07759
	MF+	0.02597	0.11902	0.00322	0.02982	0.21098	0.12788
	MICE-	0.01263	0.06054	0.00337	0.03862	0.10947	0.05936
	MICE+	0.01838	0.09083	0.00282	0.02725	0.16400	0.09530
0.3	CCA	0.02154	1.90797	1.21051	0.41058	6.58671	1.51687
	RVI	0.01081	0.05948	1.20984	0.36021	0.05208	0.03127
	K3-	0.01076	0.05883	1.20981	0.36315	0.04854	0.02977
	K3+	0.01076	0.05880	1.20976	0.36314	0.04834	0.02984
	K5-	0.01080	0.05896	1.20977	0.36347	0.04915	0.02991
	K5+	0.01081	0.05899	1.20974	0.36344	0.04928	0.02995
	K9-	0.01082	0.05934	1.20976	0.36406	0.05012	0.03016
	K9+	0.01082	0.05933	1.20974	0.36411	0.05016	0.03027
	MF-	0.01087	0.05972	1.20979	0.36358	0.04845	0.03088
	MF+	0.01142	0.06616	1.20967	0.36747	0.05980	0.03877
	MICE-	0.01048	0.05732	1.20979	0.36279	0.04486	0.02835
	MICE+	0.01094	0.06066	1.20976	0.36574	0.05610	0.03536

Ref.: Bj=β_j; CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales. En escala de rosas: mayores errores cuadráticos medios, por parámetro. En escala de verdes: menores errores cuadráticos medios, por parámetro.

Cuadro 4.2: Error cuadrático medio de los parámetros del modelo de regresión de Cox para n=200, según método de imputación y proporción de datos faltantes (p).

p	Método	B1	B2	B3	B4	B5	B6
0.1	CCA	0.01131	0.05409	0.00172	0.01517	0.09896	0.05722
	RVI	0.00708	0.03341	0.00363	0.01592	0.06248	0.03391
	K3-	0.00692	0.03243	0.00170	0.01430	0.05973	0.03198
	K3+	0.00774	0.03445	0.00119	0.00990	0.06390	0.03686
	K5-	0.00694	0.03250	0.00158	0.01237	0.06060	0.03221
	K5+	0.00776	0.03441	0.00114	0.00973	0.06392	0.03728
	K9-	0.00697	0.03255	0.00147	0.01129	0.06036	0.03242
	K9+	0.00771	0.03422	0.00111	0.00961	0.06394	0.03814
	MF-	0.00705	0.03297	0.00141	0.01161	0.05981	0.03221
	MF+	0.00815	0.03920	0.00114	0.00964	0.07236	0.04287
	MICE-	0.00606	0.02965	0.00188	0.02022	0.05328	0.02858
	MICE+	0.00697	0.03374	0.00138	0.01251	0.06187	0.03476
0.2	CCA	0.02492	0.12172	0.00324	0.03227	0.22286	0.12143
	RVI	0.00828	0.03844	0.00582	0.02477	0.07225	0.03767
	K3-	0.00810	0.03774	0.00254	0.02448	0.06888	0.03411
	K3+	0.01010	0.04223	0.00138	0.01131	0.07524	0.04432
	K5-	0.00816	0.03721	0.00223	0.01878	0.06844	0.03411
	K5+	0.01023	0.04162	0.00131	0.01093	0.07473	0.04603
	K9-	0.00823	0.03757	0.00197	0.01513	0.06906	0.03459
	K9+	0.01013	0.04058	0.00128	0.01078	0.07609	0.04843
	MF-	0.00844	0.03869	0.00180	0.01612	0.07043	0.03530
	MF+	0.01162	0.05518	0.00142	0.01215	0.09835	0.05988
	MICE-	0.00645	0.03153	0.00294	0.03854	0.05734	0.02767
	MICE+	0.00826	0.04008	0.00164	0.01739	0.07245	0.04075
0.3	CCA	0.07768	0.38638	0.00976	0.10893	1.83320	0.43095
	RVI	0.00929	0.04294	0.00756	0.03541	0.08103	0.04135
	K3-	0.00910	0.04304	0.00368	0.04014	0.07712	0.03596
	K3+	0.01329	0.04967	0.00174	0.01426	0.08685	0.05415
	K5-	0.00934	0.04209	0.00323	0.02985	0.07572	0.03632
	K5+	0.01357	0.04966	0.00170	0.01380	0.08842	0.05727
	K9-	0.00962	0.04273	0.00276	0.02214	0.07510	0.03697
	K9+	0.01315	0.04769	0.00176	0.01404	0.08982	0.06233
	MF-	0.00982	0.04533	0.00241	0.02398	0.08286	0.03982
	MF+	0.01674	0.07556	0.00207	0.01839	0.12691	0.08153
	MICE-	0.00675	0.03436	0.00423	0.06112	0.06133	0.02663
	MICE+	0.00964	0.04810	0.00199	0.02441	0.08385	0.04868

Ref.: Bj=β_j; CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales. En escala de rosas: mayores errores cuadráticos medios, por parámetro. En escala de verdes: menores errores cuadráticos medios, por parámetro.

Cuadro 4.3: Error cuadrático medio de los parámetros del modelo de regresión de Cox para n=500, según método de imputación y proporción de datos faltantes (p).

p	Método	B1	B2	B3	B4	B5	B6
0.1	CCA	0.00408	0.02028	0.00093	0.00657	0.03613	0.02026
	RVI	0.00288	0.01503	0.00347	0.01216	0.02554	0.01461
	K3-	0.00290	0.01524	0.00136	0.01073	0.02675	0.01368
	K3+	0.00299	0.01437	0.00066	0.00478	0.02554	0.01468
	K5-	0.00285	0.01501	0.00120	0.00846	0.02576	0.01355
	K5+	0.00297	0.01423	0.00059	0.00448	0.02529	0.01463
	K9-	0.00282	0.01467	0.00108	0.00724	0.02489	0.01345
	K9+	0.00296	0.01412	0.00055	0.00434	0.02509	0.01489
	MF-	0.00282	0.01523	0.00098	0.00727	0.02682	0.01342
	MF+	0.00304	0.01569	0.00055	0.00382	0.02831	0.01716
	MICE-	0.00279	0.01465	0.00165	0.01979	0.02502	0.01226
	MICE+	0.00281	0.01427	0.00093	0.00839	0.02474	0.01362
0.2	CCA	0.00776	0.03777	0.00129	0.01056	0.06877	0.03856
	RVI	0.00334	0.01745	0.00574	0.02113	0.02917	0.01744
	K3-	0.00354	0.01912	0.00224	0.02257	0.03376	0.01505
	K3+	0.00384	0.01713	0.00066	0.00519	0.03018	0.01684
	K5-	0.00342	0.01863	0.00192	0.01583	0.03178	0.01512
	K5+	0.00390	0.01675	0.00054	0.00455	0.02981	0.01757
	K9-	0.00336	0.01765	0.00159	0.01157	0.02951	0.01551
	K9+	0.00392	0.01643	0.00049	0.00432	0.03008	0.01858
	MF-	0.00342	0.01914	0.00133	0.01160	0.03435	0.01515
	MF+	0.00433	0.02105	0.00053	0.00443	0.03854	0.02397
	MICE-	0.00336	0.01800	0.00282	0.04140	0.03176	0.01264
	MICE+	0.00335	0.01666	0.00112	0.01325	0.02896	0.01556
0.3	CCA	0.01733	0.08707	0.00252	0.02230	0.15316	0.08535
	RVI	0.00374	0.02101	0.00740	0.03129	0.03375	0.01988
	K3-	0.00420	0.02469	0.00339	0.04034	0.04279	0.01670
	K3+	0.00502	0.02102	0.00074	0.00638	0.03718	0.02046
	K5-	0.00407	0.02366	0.00289	0.02797	0.04022	0.01705
	K5+	0.00531	0.02042	0.00063	0.00533	0.03649	0.02150
	K9-	0.00397	0.02257	0.00236	0.01925	0.03596	0.01745
	K9+	0.00546	0.02013	0.00065	0.00515	0.03734	0.02321
	MF-	0.00407	0.02566	0.00185	0.01917	0.04608	0.01768
	MF+	0.00663	0.03122	0.00079	0.00743	0.05456	0.03307
	MICE-	0.00389	0.02280	0.00414	0.06487	0.04103	0.01339
	MICE+	0.00391	0.02047	0.00131	0.01924	0.03400	0.01781
0.4	CCA	0.05876	0.27850	0.00736	0.07987	0.80926	0.37001
	RVI	0.00449	0.02360	0.00871	0.04149	0.03850	0.02451
	K3-	0.00517	0.02958	0.00475	0.06335	0.05228	0.01983
	K3+	0.00757	0.02587	0.00097	0.00820	0.04604	0.02751
	K5-	0.00503	0.02840	0.00419	0.04644	0.04901	0.02013
	K5+	0.00828	0.02589	0.00097	0.00718	0.04578	0.02914
	K9-	0.00500	0.02660	0.00346	0.03209	0.04394	0.02036
	K9+	0.00856	0.02511	0.00115	0.00732	0.04706	0.03341
	MF-	0.00524	0.03222	0.00265	0.03265	0.05834	0.02097
	MF+	0.01165	0.04612	0.00160	0.01396	0.07510	0.04856
	MICE-	0.00477	0.02795	0.00563	0.09076	0.05045	0.01477
	MICE+	0.00489	0.02488	0.00153	0.02572	0.04203	0.02286

Ref.: Bj=β_j; CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales. En escala de rosas: mayores errores cuadráticos medios, por parámetro. En escala de verdes: menores errores cuadráticos medios, por parámetro.

Cuadro 4.4: Error cuadrático medio de los parámetros del modelo de regresión de Cox para n=1000, según método de imputación y proporción de datos faltantes (p).

p	Método	B1	B2	B3	B4	B5	B6
0.1	CCA	0.00209	0.01078	0.00070	0.00456	0.01851	0.01018
	RVI	0.00167	0.00927	0.00342	0.01128	0.01443	0.00856
	K3-	0.00175	0.00991	0.00126	0.01017	0.01684	0.00768
	K3+	0.00157	0.00818	0.00050	0.00334	0.01407	0.00753
	K5-	0.00167	0.00949	0.00110	0.00760	0.01583	0.00778
	K5+	0.00155	0.00796	0.00042	0.00301	0.01346	0.00748
	K9-	0.00163	0.00910	0.00097	0.00622	0.01470	0.00772
	K9+	0.00155	0.00789	0.00037	0.00286	0.01354	0.00763
	MF-	0.00162	0.00958	0.00084	0.00605	0.01672	0.00762
	MF+	0.00159	0.00822	0.00036	0.00217	0.01460	0.00885
	MICE-	0.00182	0.00993	0.00158	0.02038	0.01696	0.00711
	MICE+	0.00161	0.00832	0.00079	0.00729	0.01375	0.00723
0.2	CCA	0.00378	0.01884	0.00087	0.00635	0.03190	0.01752
	RVI	0.00193	0.01136	0.00567	0.01986	0.01682	0.01123
	K3-	0.00223	0.01349	0.00218	0.02231	0.02400	0.00937
	K3+	0.00192	0.00953	0.00044	0.00326	0.01667	0.00890
	K5-	0.00208	0.01263	0.00183	0.01502	0.02154	0.00933
	K5+	0.00194	0.00912	0.00032	0.00260	0.01601	0.00896
	K9-	0.00196	0.01170	0.00149	0.01044	0.01870	0.00964
	K9+	0.00195	0.00900	0.00026	0.00230	0.01594	0.00935
	MF-	0.00195	0.01345	0.00117	0.00958	0.02458	0.00913
	MF+	0.00216	0.01102	0.00027	0.00215	0.01960	0.01263
	MICE-	0.00244	0.01378	0.00279	0.04261	0.02488	0.00807
	MICE+	0.00193	0.00990	0.00095	0.01157	0.01643	0.00830
0.3	CCA	0.00738	0.03533	0.00127	0.01052	0.06553	0.03815
	RVI	0.00225	0.01361	0.00741	0.02945	0.01918	0.01352
	K3-	0.00284	0.01851	0.00340	0.04151	0.03331	0.01125
	K3+	0.00260	0.01149	0.00046	0.00384	0.02045	0.01045
	K5-	0.00264	0.01703	0.00288	0.02851	0.02940	0.01143
	K5+	0.00269	0.01094	0.00033	0.00279	0.01928	0.01101
	K9-	0.00242	0.01521	0.00229	0.01867	0.02442	0.01169
	K9+	0.00284	0.01052	0.00032	0.00252	0.01955	0.01233
	MF-	0.00252	0.01925	0.00172	0.01710	0.03520	0.01086
	MF+	0.00349	0.01611	0.00043	0.00472	0.02850	0.01906
	MICE-	0.00322	0.01867	0.00419	0.06735	0.03513	0.00921
	MICE+	0.00228	0.01168	0.00114	0.01655	0.01942	0.00955
0.4	CCA	0.02025	0.09222	0.00276	0.02556	0.17654	0.09600
	RVI	0.00255	0.01567	0.00869	0.03993	0.02252	0.01645
	K3-	0.00358	0.02372	0.00477	0.06642	0.04467	0.01358
	K3+	0.00381	0.01408	0.00057	0.00513	0.02583	0.01370
	K5-	0.00330	0.02180	0.00416	0.04931	0.03928	0.01390
	K5+	0.00428	0.01373	0.00048	0.00368	0.02460	0.01438
	K9-	0.00306	0.01944	0.00344	0.03364	0.03248	0.01448
	K9+	0.00474	0.01318	0.00062	0.00374	0.02553	0.01646
	MF-	0.00326	0.02560	0.00261	0.03162	0.04650	0.01403
	MF+	0.00621	0.02587	0.00104	0.01020	0.04278	0.02822
	MICE-	0.00401	0.02391	0.00566	0.09286	0.04572	0.01102
	MICE+	0.00278	0.01419	0.00131	0.02269	0.02418	0.01159

Ref.: Bj=β_j; CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales. En escala de rosas: mayores errores cuadráticos medios, por parámetro. En escala de verdes: menores errores cuadráticos medios, por parámetro.

Sesgo

El sesgo se estudia en forma global para los seis parámetros, promediando la diferencia relativa entre la estimación promedio y el valor real (Figura 4.1).

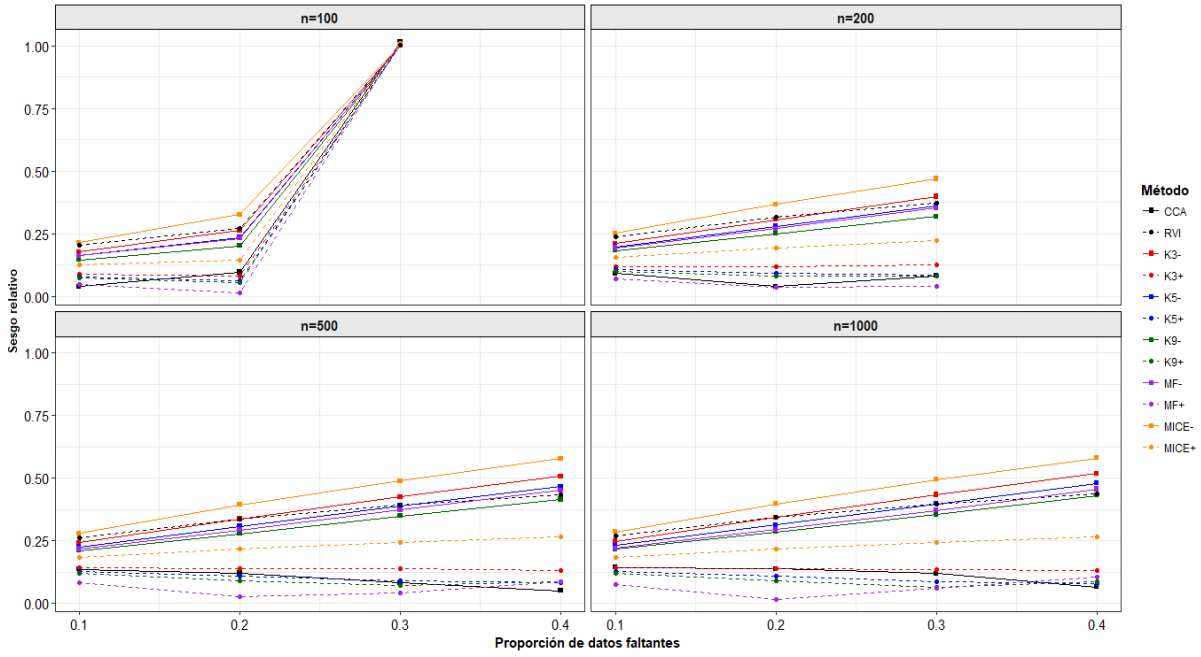
El efecto del incremento en la proporción de datos faltantes, presente en todos los métodos estudiados, es más pronunciado con el tamaño de muestra más pequeño ($n = 100$). Para los otros tamaños de muestra, este incremento en el sesgo se atenúa o desaparece al incluir como variables explicativas al *tiempo* y *estado*.

Para todos los tamaños de muestra, los perfiles de sesgo se agrupan según si son o no utilizadas las variables informativas *tiempo* y *estado*, mostrándose con menor sesgo el grupo de métodos que las incluyen. MICE evidencia los resultados más desfavorables en ambos grupos, es decir, considerando o no las variables informativas.

Al emplear KNN, los resultados mejoran al aumentar el número de donantes, k.

MF muestra siempre los mejores resultados, por lo que MF+ se presenta como el método con menor diferencia promedio relativa entre la media teórica y empírica de los estimadores de los parámetros del modelo, en la mayoría de las situaciones.

Figura 4.1: Diferencia relativa media, en valor absoluto, entre el promedio de las estimaciones de los parámetros y sus valores reales, según método de imputación, tamaño de muestra y proporción de datos faltantes.



Ref.: CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

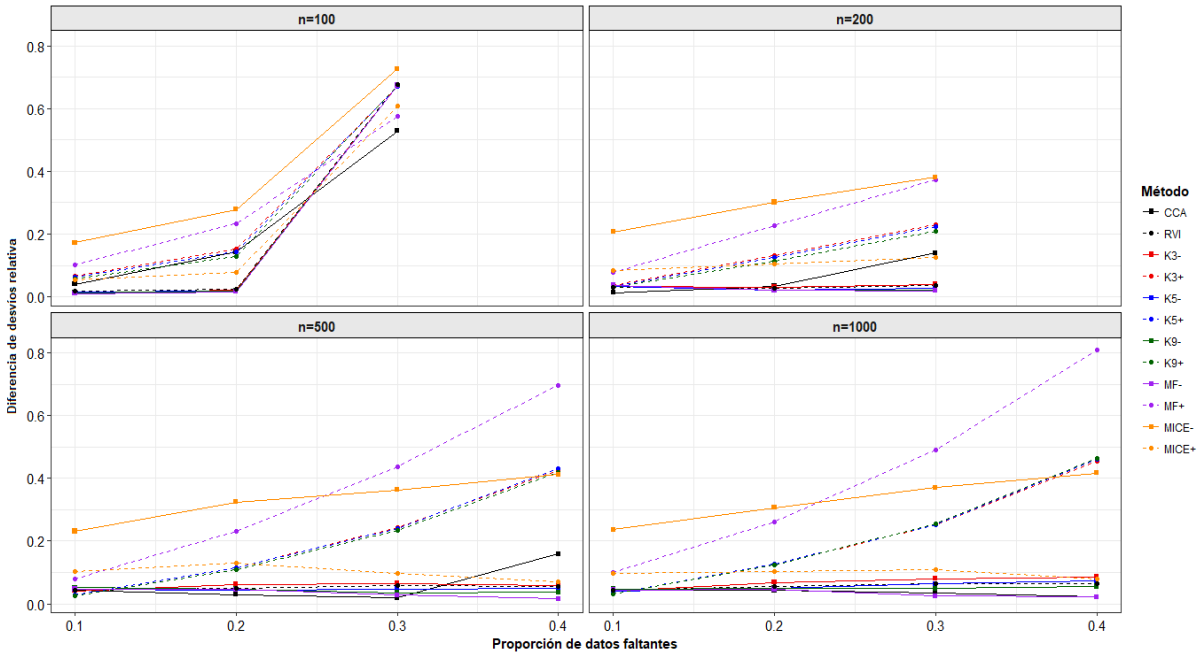
Variabilidad de los estimadores respecto de la variabilidad esperada

Aquí también el escenario que corresponde al menor tamaño muestral es el que evidencia mayor impacto de los porcentajes altos de pérdida, ahora sobre la capacidad de reproducir la variabilidad esperada de los estimadores a partir de la calculada con datos imputados (Figura 4.2). Si bien este efecto disminuye cuando el tamaño de muestra crece, aún hay métodos de imputación que no logran corregirlo (KNN+, MICE- y MF+). MICE- presenta los mayores valores para el indicador en la mayor parte de los escenarios, salvo cuando $n = 500$ y $n = 1000$ para $p = 0.30$ y 0.40, donde MF+ muestra valores superiores.

Para todas las combinaciones de n y p, la no inclusión del *tiempo* y el *estado* como variables informativas provee indicadores menores, cercanos a cero y similares para todas las técnicas, a excepción de MICE. Estos valores se incrementan muy levemente a medida que aumenta p, para cada tamaño de muestra. Cuando se incluyen las variables informativas adicionales, el incremento del indicador en cada técnica es más notorio.

KNN muestra resultados muy similares para los tres valores de k, ya sea cuando se tienen en cuenta las variables informativas adicionales o cuando no se las considera.

Figura 4.2: Diferencia relativa media, en valor absoluto, entre el promedio de los desvíos estándar teóricos de los estimadores y el desvío estándar empírico, según método de imputación, tamaño de muestra y proporción de datos faltantes.



Ref.: CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

Cobertura

Esta propiedad se analiza, separadamente, para cada parámetro del modelo de regresión de Cox 4.1 (Figura 4.3).

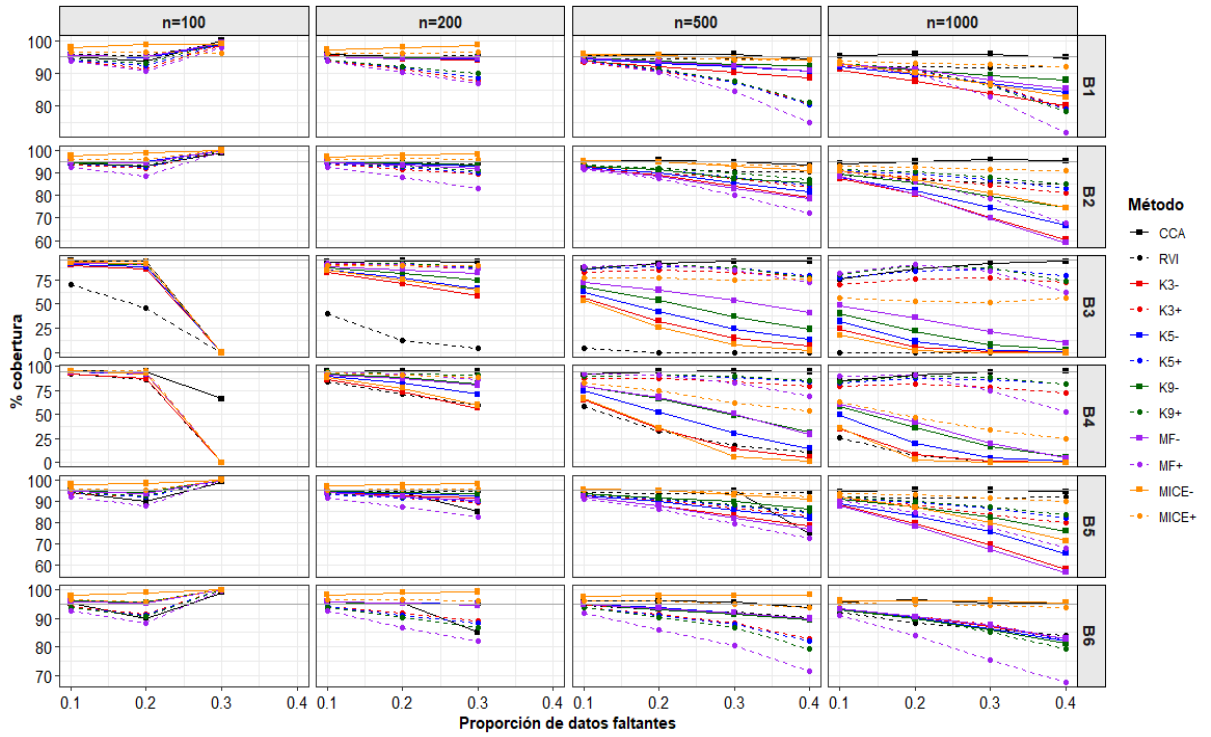
El método CCA evidencia altos porcentajes de cobertura, salvo en algunos escenarios y cuando el porcentaje de pérdidas es alto.

Salvo algunas excepciones, para todos los parámetros y valores de n, la diferencia observada entre los métodos de imputación respecto a la cobertura aumenta a medida que aumenta la proporción de pérdidas.

Para β_1, β_2, β_5 y β_6, que corresponden a la variable con distribución Normal y a las cualitativas, el comportamiento de los métodos es muy similar para cada tamaño de muestra. Con bajo porcentaje de pérdida, la cobertura de esos cuatro parámetros es muy alta, cercana al 95 %, para todos los métodos. En estos casos, el método de imputación que deriva en menores coberturas es MF+, mientras que CCA y MICE muestran los valores más altos.

Al analizar los resultados de cobertura para los parámetros β_3 y β_4, correspondientes a las variables continuas con distribución no Normal, para tamaños de muestras grandes ($n = 500$ y 1000) hay uniformidad de cada método a través de los distintos porcentajes de pérdidas. Para tamaños de muestra $n = 100$ y 200, la influencia de los porcentajes de pérdida produce efectos diferentes aún para los mismos métodos.

Figura 4.3: Porcentaje de cobertura para los intervalos de confianza del 95 %, según método de imputación, tamaño de muestra y proporción de datos faltantes, según método de imputación, tamaño de muestra y proporción de datos faltantes.



Ref.: Bj=β_j; CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

Distribuciones en el muestreo

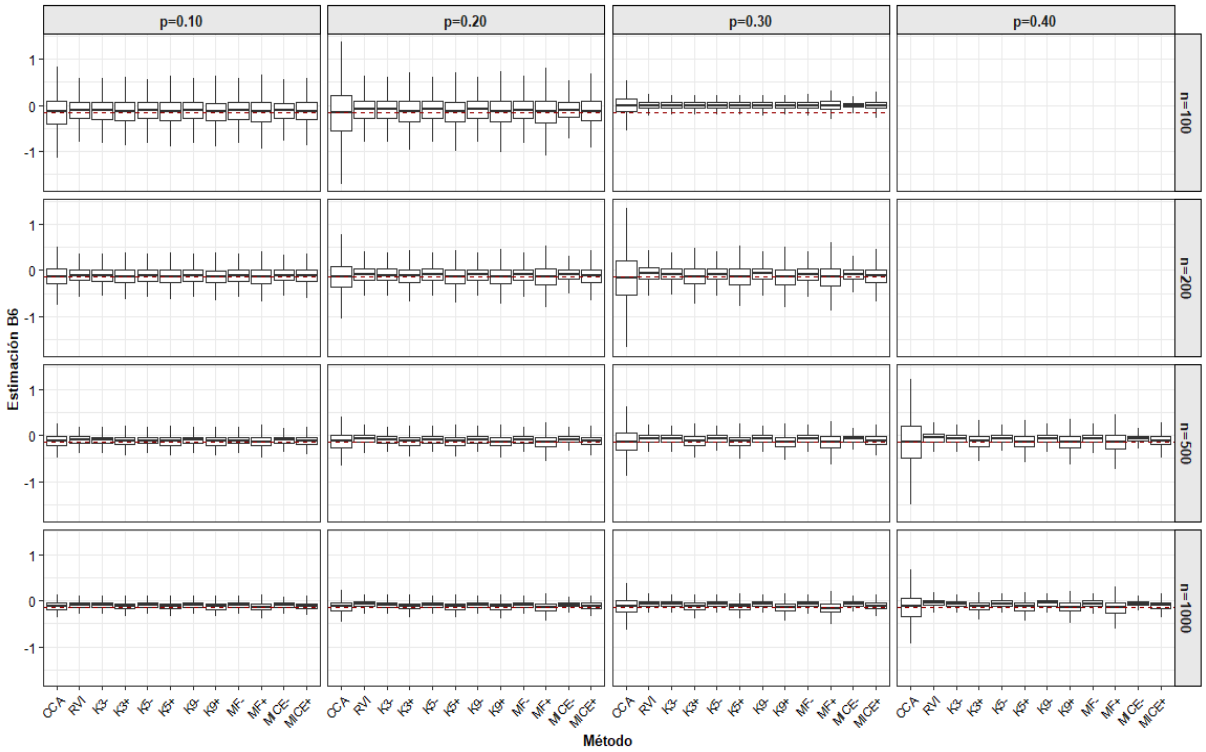
Los box-plots de la Figura 4.4 aproximan algunas características de la distribución de los estimadores a partir de las 5000 repeticiones del proceso de simulación. Cada panel compara las distribuciones según método de tratamiento de la información faltante. Se muestran los resultados correspondientes a uno de los seis parámetros del modelo 4.1 (β_6). Se omiten las distribuciones restantes por presentar características similares.

Las distribuciones no son homogéneas en cuanto a variabilidad, tanto si se comparan los escenarios como si se comparan los métodos para un escenario fijo. Lógicamente, a mayor tamaño de muestra, las distribuciones resultan más concentradas, salvo para el método CCA, que presenta la mayor dispersión frente a cualquier escenario. Ninguna de las distribuciones evidencia asimetrías.

Se observa que, dentro de cada método de imputación, existe menor variabilidad en las estimaciones cuando no se incluyen el *tiempo* y el *estado* como variables informativas,

aunque su mediana presenta un alejamiento respecto al verdadero valor del parámetro.

Figura 4.4: Distribución de los estimadores de β_6, según método de imputación, tamaño de muestra y proporción de datos faltantes, sin incluir valores extremos.



Ref.: CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

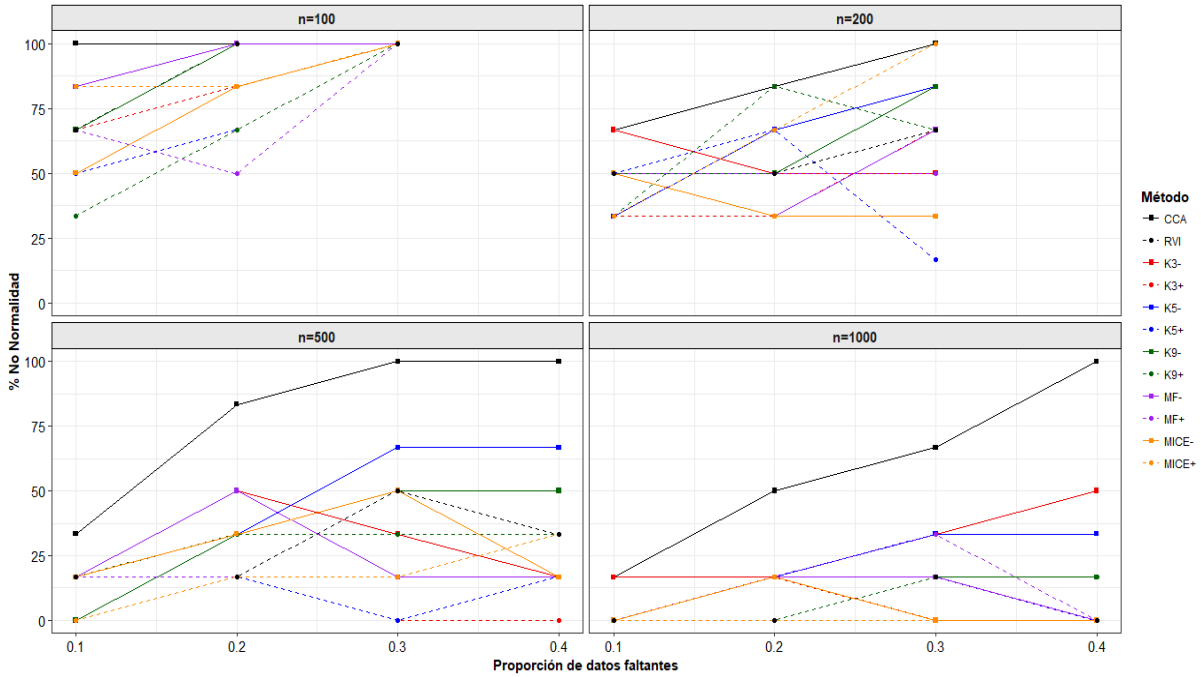
Para complementar la visión descriptiva de las distribuciones de los estimadores, se presenta los resultados de un test de bondad de ajuste a la distribución Normal esperada. En la Figura 4.5 se muestra el porcentaje de parámetros, entre los seis correspondientes al modelo 4.1, en los que se rechaza la hipótesis de distribución Normal de acuerdo al test de Anderson-Darling.

Si bien no se aprecia una tendencia al comparar los métodos en los distintos escenarios, se distingue que, a medida que aumenta el tamaño de muestra, disminuye el porcentaje en estudio, es decir, aumenta el porcentaje de parámetros para los cuales se acepta que su estimador presenta distribución Normal.

Dado que el test de Anderson-Darling se aplica sobre un gran número de datos (5000 estimaciones para cada método y escenario), el mismo se vuelve muy sensible, provocando un rechazo de la hipótesis de normalidad ante pequeños alejamientos de dicha

distribución. Sin embargo, no se observan características en las distribuciones que sugieran que el supuesto de normalidad no es aceptable.

Figura 4.5: Porcentaje de parámetros para los cuales se rechaza la hipótesis del test de Anderson-Darling, según método de imputación, tamaño de la muestra y proporción de datos faltantes.



Ref.: CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

Estimación de la probabilidad de sobrevida

Con cada modelo estimado, se predice la probabilidad de sobrevida después de un tiempo particular, $t = 100$. Según el modelo poblacional, dicha probabilidad es de 0.40. Se estima este valor en cada simulación, promediando los resultados para cada método y escenario (Figura 4.6).

El método RVI muestra estimaciones muy alejadas del verdadero valor de la probabilidad de supervivencia de interés, en cualquier escenario.

Salvo cuando $n = 100$, la diferencia entre las distintas técnicas en la estimación de la probabilidad de sobrevida, se amplia a medida que aumenta p.

En general, dentro de cada técnica, las estimaciones promedio de la probabilidad de supervivencia resultan más cercanas al valor real cuando se incluyen el *tiempo* y el *estado* como variables informativas. Además, para los niveles más altos de pérdidas ($p =$

0.30 y 0.40), cuando se incluyen dichas variables la probabilidad de supervivencia resulta sobreestimada y, cuando no se incluyen, es subestimada. El efecto del incremento de la proporción de pérdidas resulta sobre las estimaciones promedio resulta mayor cuando se incluyen las variables adicionales, en comparación con la variación observada en las estimaciones cuando las mismas no son consideradas en el proceso de imputación.

Cuando $n = 100$ y $p = 0.10$, K9+ y K5+ muestran los resultados más favorables, mientras que KNN- presenta los más desfavorables, para los tres valores de k. Con $p = 0.20$, K5+ presenta el promedio de estimaciones más cercano al real y KNN-, para los tres valores de k, se ubica a mayor distancia de dicho valor. Para $p = 0.30$, CCA es quien presenta la estimación media más cercana al verdadero valor, mientras que todas las otras técnicas se comportan muy similares entre sí.

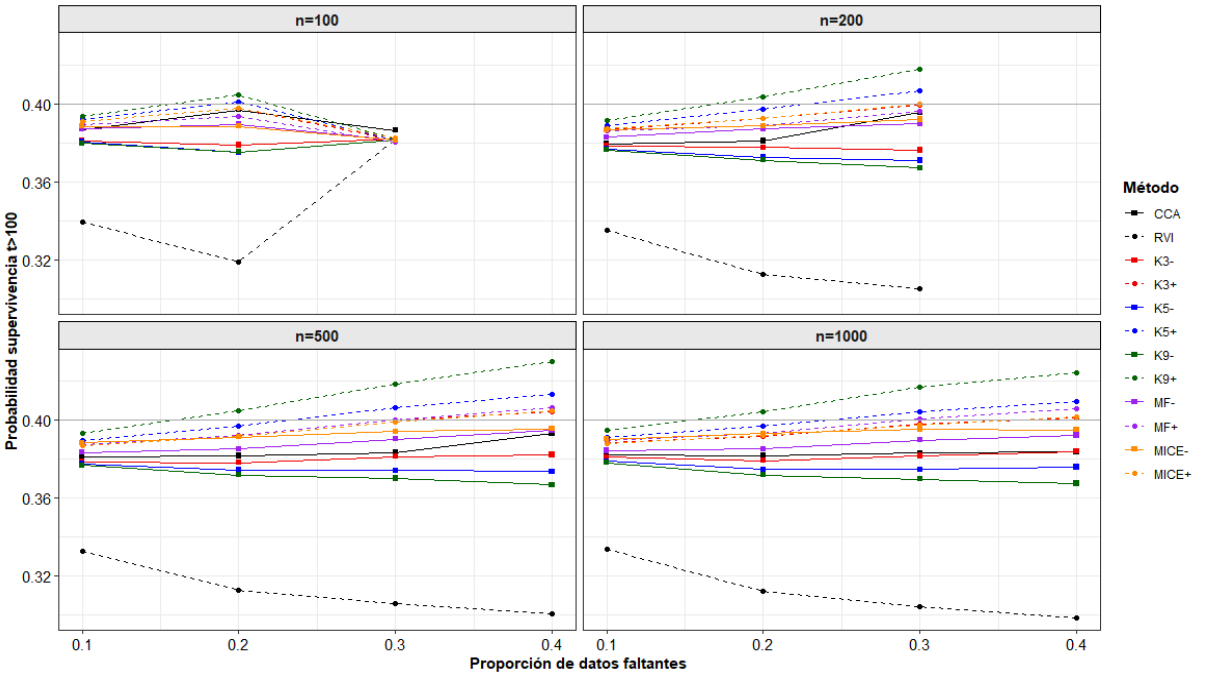
Cuando $n = 200$ y $p = 0.30$, K3+ y MICE+ están más cercanos al valor real y K9- es quien presenta el valor más alejado.

Cuando $n = 200$, 500 y 1000 y $p = 0.10$, K9+ presenta el promedio de estimaciones más cercano al real y KNN-, para los tres valores de k, se ubica a mayor distancia de dicho valor. Para $p = 0.20$, K5+ y K9+ presentan los mejores resultados y K5- y K9-, los peores. Cuando $p = 0.30$, K3+ y MICE+ muestran los resultados más cercanos al verdadero mientras que K9- presenta el más alejado. En particular, cuando $n = 200$, MF+ se incluye entre los resultados más cercanos al verdadero.

En el caso de $n = 500$ y $p = 0.40$, K3+, MF+, MICE+, MICE-, MF- y CCA muestran los mejores resultados y K9- y K9+, los peores.

Cuando $n = 1000$ y $p = 0.40$, MICE+ y K3+ presentan los mejores resultados y K9- y K9+, los peores.

Figura 4.6: Promedio de la probabilidad de supervivencia estimada para $t > 100$, según método de estimación, tamaño de la muestra y proporción de datos faltantes.



Ref.: CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

4.2.1.b. Reproducibilidad de los datos perdidos

En este apartado, se realiza una evaluación de la eficiencia de los métodos de imputación para asignar valores a los datos faltantes, según se trate de observaciones correspondientes a las variables cuantitativas X_1, X_3 y X_4 o a la variables cualitativas X_2 y (X_5, X_6), cuando se han generado datos perdidos de acuerdo a un mecanismo MCAR.

Raíz del error cuadrático medio normalizado

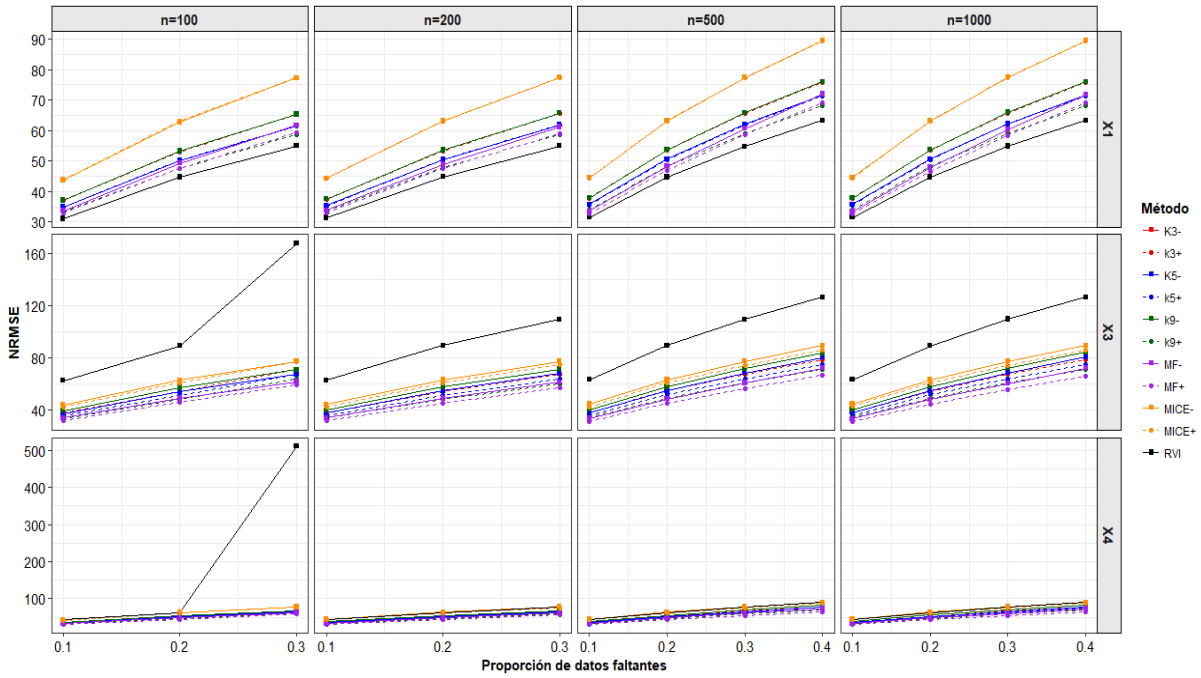
En relación a la imputación de datos para X_1, de acuerdo con la Figura 4.7, se encuentran los mayores valores para $\overline{\text{NRMSE}}$ al utilizar MICE y los menores valores corresponden al uso de RVI, seguidos por MF. En el caso del empleo de KNN, los resultados más favorables se encuentran con $k = 5$.

En cuanto a la imputación para X_3 y X_4, se encuentran los mayores valores para $\overline{\text{NRMSE}}$ al utilizar RVI, seguidos por MICE, y los menores valores corresponden al uso de MF. Para KNN, nuevamente, los resultados más favorables ocurren con $k = 5$.

Para ninguna de las tres variables se distinguen diferencias importantes en los valores de $\overline{\text{NRMSE}}$ según se considere el *tiempo* y el *estado* como variables informativas.

El comportamiento de $\overline{\text{NRMSE}}$ es muy similar para todos los tamaños de muestra considerados. En todos los casos, $\overline{\text{NRMSE}}$ aumenta a medida que aumenta la proporción de datos faltantes.

Figura 4.7: Error cuadrático medio normalizado (NRMSE) para las variables explicativas cuantitativas, según método de imputación, tamaño de la muestra y proporción de datos faltantes.



Ref.: CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

Porcentaje de no coincidencia

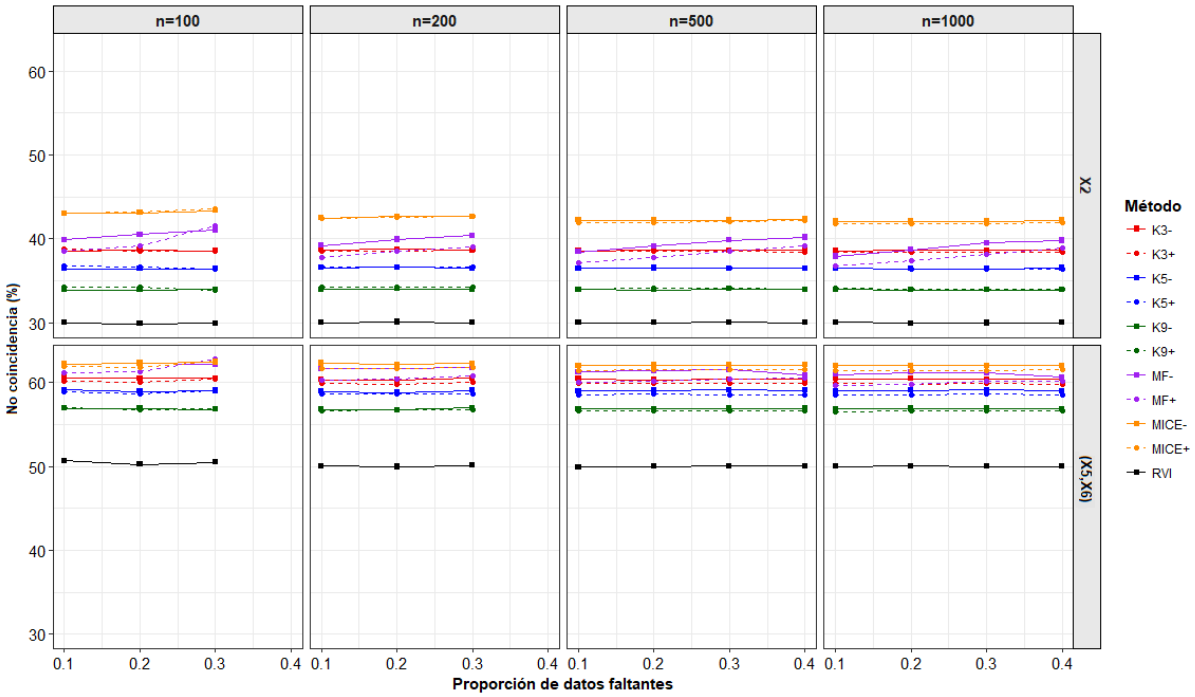
En la Figura 4.8 se muestran los porcentajes medios de datos imputados en forma incorrecta para las variables categóricas X_2 y (X_5, X_6).

Si bien en magnitudes diferentes, el comportamiento es muy similar para ambas variables. Los porcentajes medios de datos imputados en forma incorrecta por cada método son muy similares para todos los tamaños de muestra y para todas las proporciones de datos faltantes, aunque al utilizar MF se distingue un leve incremento a medida que aumenta el porcentaje de valores faltantes. Los valores más pequeños se encuentran bajo RVI, mientras que los más desfavorables corresponden a MICE. Haciendo uso de KNN,

se observan resultados más favorables a medida que aumenta el número de donantes.

En ningún caso se distinguen mejores resultados al considerar el *tiempo* y el *estado* como variables informativas.

Figura 4.8: Porcentaje de datos imputados no coincidentes con los reales para las variables explicativas cualitativas, según método de imputación, tamaño de la muestra y proporción de datos faltantes.



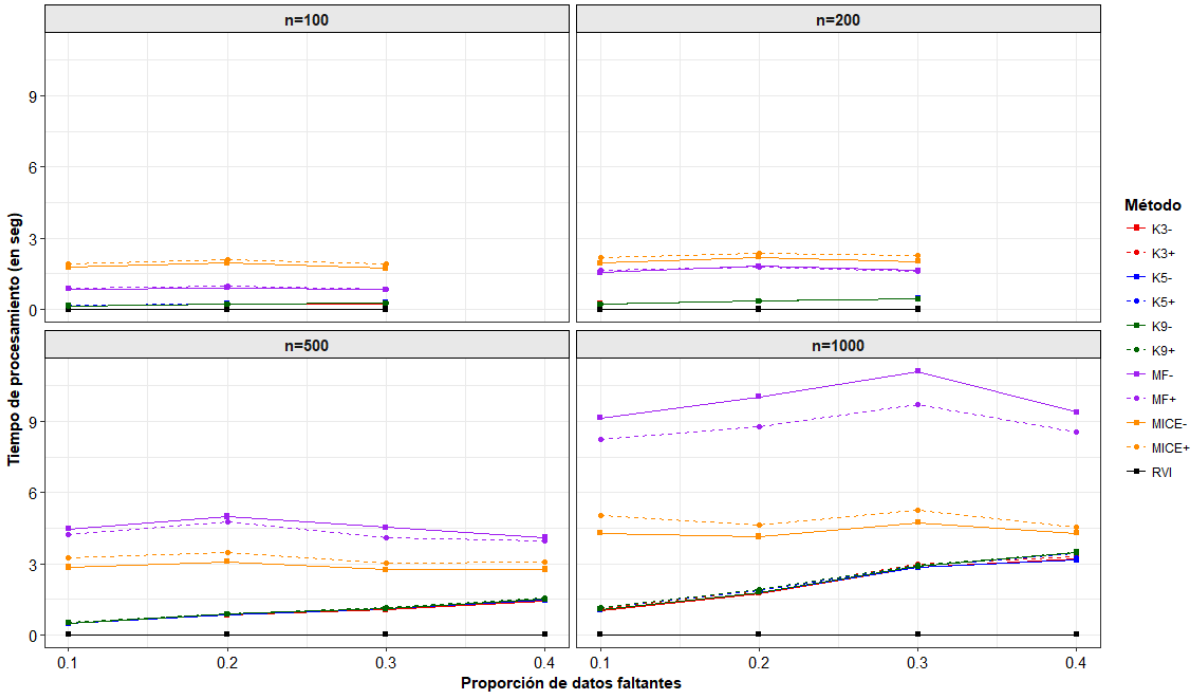
Ref.: CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

4.2.1.c. Tiempos computacionales de imputación

En relación a los tiempos que demora el software utilizado para realizar las imputaciones, se observa que RVI demanda muy poco tiempo para realizar la imputación completa del conjunto de datos, seguido por KNN (Figura 4.9). Para esta técnica, no hay diferencias destacables entre las distintas configuraciones empleadas.

Cuando $n = 100$ o $n = 200$, MF demanda menor tiempo para la imputación que MICE. Esta relación se invierte cuando $n = 500$ o $n = 1000$. Para este último tamaño del conjunto de datos, el tiempo de procesamiento bajo MF es más de dos veces el utilizado por MICE.

Figura 4.9: Tiempo promedio, en segundos, demandado para la imputación completa de los datos, según método de imputación, tamaño de la muestra y proporción de datos faltantes.



Ref.: CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

4.2.2. Datos perdidos al azar

4.2.2.a. Propiedades distribucionales de los estimadores de los coeficientes del modelo de regresión de Cox

Error cuadrático medio

Los resultados obtenidos al calcular el MSE para cada uno de los seis parámetros del modelo 4.1 se muestran en los Cuadros 4.5 a 4.8, según el tamaño de la muestra, n, y la proporción de datos faltantes, p.

En el caso de $n = 100$, los resultados más desfavorables se observan bajo el uso de CCA y MF+, mientras que MICE- aporta los menores valores de MSE. El comportamiento es similar para los distintos valores de p (Cuadro 4.5).

Para $n = 200$, se encuentran resultados desfavorables al utilizar CCA, MF+ y K9+. MICE- es el método que mejores resultados presenta (Cuadro 4.6).

Cuando $n = 500$, CCA y K9+ se asocian a los mayores valores de MSE, mientras que MICE presenta los menores valores. Tanto para este tamaño de muestra como para $n = 100$ y $n = 200$, se observan menores MSE cuando no se consideran el *tiempo* y el *estado* como variables informativas (Cuadro 4.7).

No se distingue una regularidad en los resultados de MSE cuando $n = 1000$ (Cuadro 4.8).

Cuadro 4.5: Error cuadrático medio de los parámetros del modelo de regresión de Cox para n=100, según proporción de datos faltantes (p).

p	Método	B1	B2	B3	B4	B5	B6
0.1	CCA	0.01796	0.08350	0.00251	0.02263	0.16489	0.09363
	RVI	0.01508	0.06760	0.00213	0.01809	0.14020	0.07554
	K3-	0.01475	0.06744	0.00203	0.01825	0.12860	0.07093
	K3+	0.01679	0.06758	0.00209	0.01842	0.14545	0.08503
	K5-	0.01502	0.06732	0.00206	0.01826	0.13089	0.07200
	K5+	0.01687	0.06801	0.00211	0.01839	0.14783	0.08767
	K9-	0.01515	0.06777	0.00209	0.01816	0.13467	0.07306
	K9+	0.01675	0.06804	0.00211	0.01828	0.14655	0.08864
	MF-	0.01530	0.06782	0.00203	0.01826	0.12951	0.07296
	MF+	0.01745	0.06909	0.00210	0.01845	0.16227	0.09237
	MICE-	0.01290	0.06692	0.00200	0.01804	0.11490	0.06631
	MICE+	0.01487	0.06803	0.00204	0.01818	0.14215	0.08086
0.2	CCA	0.02560	0.10761	0.00362	0.03233	0.22352	0.13092
	RVI	0.01708	0.06654	0.00236	0.01793	0.15568	0.08439
	K3-	0.01607	0.06647	0.00218	0.01828	0.13892	0.07669
	K3+	0.02099	0.06708	0.00232	0.01873	0.17230	0.10666
	K5-	0.01673	0.06642	0.00221	0.01825	0.13974	0.07771
	K5+	0.02154	0.06719	0.00235	0.01870	0.17500	0.11344
	K9-	0.01697	0.06621	0.00228	0.01830	0.14450	0.07879
	K9+	0.02146	0.06806	0.00236	0.01854	0.17490	0.11870
	MF-	0.01751	0.06772	0.00215	0.01831	0.13439	0.07776
	MF+	0.02346	0.07060	0.00231	0.01888	0.20454	0.12289
	MICE-	0.01252	0.06616	0.00207	0.01786	0.10804	0.06473
	MICE+	0.01689	0.06909	0.00215	0.01826	0.16181	0.09600
0.3	CCA	0.04030	0.15591	0.00571	0.05156	0.43336	0.19043
	RVI	0.02060	0.06613	0.00255	0.01881	0.18061	0.09086
	K3-	0.01851	0.06634	0.00217	0.01931	0.14367	0.07583
	K3+	0.02784	0.06724	0.00242	0.02006	0.20578	0.13944
	K5-	0.01910	0.06633	0.00226	0.01934	0.15258	0.07867
	K5+	0.02892	0.06766	0.00246	0.02001	0.21172	0.14896
	K9-	0.02005	0.06596	0.00236	0.01925	0.16219	0.08186
	K9+	0.03028	0.06834	0.00249	0.02010	0.21859	0.15883
	MF-	0.02149	0.06823	0.00214	0.01939	0.14531	0.07929
	MF+	0.03384	0.07300	0.00245	0.02051	0.26968	0.16021
	MICE-	0.01262	0.06548	0.00200	0.01872	0.10633	0.05842
	MICE+	0.02049	0.06947	0.00211	0.01945	0.19342	0.11315
0.4	CCA	0.07286	0.27516	0.01262	0.09720	1.03153	0.43179
	RVI	0.02477	0.06930	0.00280	0.01851	0.21935	0.11639
	K3-	0.02095	0.06964	0.00226	0.01943	0.17118	0.08764
	K3+	0.03700	0.07165	0.00268	0.02093	0.27144	0.21988
	K5-	0.02233	0.06966	0.00238	0.01925	0.17914	0.08983
	K5+	0.04043	0.07140	0.00273	0.02099	0.27111	0.22894
	K9-	0.02365	0.06930	0.00250	0.01922	0.19270	0.09829
	K9+	0.04144	0.07182	0.00281	0.02109	0.26970	0.23405
	MF-	0.02702	0.07256	0.00221	0.01984	0.17310	0.09678
	MF+	0.05157	0.07943	0.00272	0.02161	0.35135	0.22596
	MICE-	0.01316	0.06933	0.00200	0.01851	0.10739	0.06161
	MICE+	0.02584	0.07639	0.00216	0.01988	0.24666	0.15716

Ref.: Bj=β_j; CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales. En escala de rosas: mayores errores cuadráticos medios, por parámetro. En escala de verdes: menores errores cuadráticos medios, por parámetro.

Cuadro 4.6: Error cuadrático medio de los parámetros del modelo de regresión de Cox para n=200, según método de imputación y proporción de datos faltantes (p).

p	Método	B1	B2	B3	B4	B5	B6
0.1	CCA	0.00792	0.03552	0.00130	0.01020	0.06941	0.04070
	RVI	0.00699	0.03008	0.00129	0.00880	0.05982	0.03433
	K3-	0.00686	0.03010	0.00117	0.00886	0.05626	0.03268
	K3+	0.00758	0.03021	0.00119	0.00883	0.06170	0.03726
	K5-	0.00690	0.02995	0.00119	0.00886	0.05642	0.03250
	K5+	0.00764	0.03035	0.00121	0.00884	0.06310	0.03842
	K9-	0.00701	0.03010	0.00122	0.00883	0.05758	0.03293
	K9+	0.00769	0.03030	0.00124	0.00884	0.06355	0.03969
	MF-	0.00703	0.03013	0.00115	0.00888	0.05632	0.03276
	MF+	0.00793	0.03078	0.00120	0.00882	0.06962	0.04236
	MICE-	0.00614	0.02990	0.00114	0.00881	0.05073	0.03025
	MICE+	0.00681	0.03033	0.00115	0.00880	0.05989	0.03554
0.2	CCA	0.01051	0.04538	0.00166	0.01330	0.09311	0.05221
	RVI	0.00777	0.03070	0.00143	0.00896	0.06971	0.03815
	K3-	0.00738	0.03070	0.00117	0.00913	0.06432	0.03371
	K3+	0.00942	0.03129	0.00124	0.00919	0.07733	0.04606
	K5-	0.00756	0.03078	0.00121	0.00909	0.06456	0.03454
	K5+	0.00970	0.03129	0.00129	0.00918	0.07968	0.04884
	K9-	0.00770	0.03080	0.00127	0.00910	0.06499	0.03485
	K9+	0.00992	0.03114	0.00135	0.00920	0.08002	0.05299
	MF-	0.00785	0.03115	0.00114	0.00913	0.06440	0.03441
	MF+	0.01066	0.03221	0.00125	0.00916	0.09550	0.05575
	MICE-	0.00609	0.03062	0.00112	0.00898	0.05270	0.02872
	MICE+	0.00760	0.03128	0.00115	0.00897	0.07081	0.03985
0.3	CCA	0.01489	0.05997	0.00226	0.01772	0.12450	0.07400
	RVI	0.00924	0.03055	0.00163	0.00888	0.08106	0.04339
	K3-	0.00858	0.03045	0.00119	0.00909	0.07021	0.03558
	K3+	0.01233	0.03099	0.00134	0.00926	0.09727	0.06262
	K5-	0.00886	0.03053	0.00126	0.00904	0.07040	0.03623
	K5+	0.01321	0.03109	0.00142	0.00931	0.10104	0.06941
	K9-	0.00905	0.03050	0.00137	0.00908	0.07486	0.03789
	K9+	0.01368	0.03131	0.00153	0.00930	0.10261	0.07771
	MF-	0.00944	0.03093	0.00115	0.00918	0.07263	0.03734
	MF+	0.01548	0.03287	0.00136	0.00938	0.12544	0.07590
	MICE-	0.00645	0.03021	0.00111	0.00900	0.05618	0.02831
	MICE+	0.00889	0.03141	0.00116	0.00897	0.08512	0.04790
0.4	CCA	0.02321	0.09126	0.00413	0.02834	0.19813	0.11875
	RVI	0.01043	0.03060	0.00200	0.00916	0.09328	0.05125
	K3-	0.00929	0.03050	0.00133	0.00956	0.08001	0.03937
	K3+	0.01635	0.03186	0.00166	0.01015	0.13231	0.10296
	K5-	0.00983	0.03057	0.00143	0.00961	0.08243	0.04000
	K5+	0.01810	0.03196	0.00178	0.01023	0.13955	0.11394
	K9-	0.01006	0.03085	0.00160	0.00948	0.08415	0.04279
	K9+	0.01908	0.03233	0.00195	0.01020	0.13642	0.12773
	MF-	0.01112	0.03166	0.00128	0.00961	0.08341	0.04097
	MF+	0.02327	0.03572	0.00166	0.01023	0.17299	0.10767
	MICE-	0.00665	0.03033	0.00119	0.00920	0.06094	0.02830
	MICE+	0.01073	0.03243	0.00124	0.00936	0.10252	0.06088

Ref.: Bj=β_j; CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales. En escala de rosas: mayores errores cuadráticos medios, por parámetro. En escala de verdes: menores errores cuadráticos medios, por parámetro.

Cuadro 4.7: Error cuadrático medio de los parámetros del modelo de regresión de Cox para n=500, según método de imputación y proporción de datos faltantes (p).

p	Método	B1	B2	B3	B4	B5	B6
0.1	CCA	0.00306	0.01381	0.00079	0.00512	0.02628	0.01533
	RVI	0.00276	0.01189	0.00089	0.00489	0.02365	0.01425
	K3-	0.00280	0.01191	0.00075	0.00494	0.02485	0.01319
	K3+	0.00292	0.01197	0.00077	0.00485	0.02497	0.01438
	K5-	0.00281	0.01193	0.00078	0.00494	0.02428	0.01326
	K5+	0.00293	0.01198	0.00079	0.00484	0.02484	0.01468
	K9-	0.00278	0.01190	0.00081	0.00492	0.02385	0.01336
	K9+	0.00293	0.01199	0.00082	0.00485	0.02500	0.01515
	MF-	0.00279	0.01193	0.00074	0.00495	0.02512	0.01306
	MF+	0.00303	0.01210	0.00078	0.00478	0.02735	0.01674
	MICE-	0.00268	0.01184	0.00073	0.00495	0.02304	0.01221
	MICE+	0.00279	0.01195	0.00075	0.00487	0.02389	0.01360
0.2	CCA	0.00386	0.01717	0.00086	0.00577	0.03289	0.01911
	RVI	0.00307	0.01235	0.00105	0.00502	0.02656	0.01609
	K3-	0.00319	0.01239	0.00078	0.00512	0.02925	0.01398
	K3+	0.00353	0.01263	0.00081	0.00500	0.02911	0.01752
	K5-	0.00316	0.01235	0.00081	0.00510	0.02820	0.01433
	K5+	0.00361	0.01266	0.00086	0.00499	0.03005	0.01829
	K9-	0.00314	0.01239	0.00088	0.00509	0.02649	0.01456
	K9+	0.00362	0.01261	0.00093	0.00501	0.03052	0.02009
	MF-	0.00320	0.01246	0.00075	0.00517	0.02977	0.01403
	MF+	0.00402	0.01290	0.00083	0.00486	0.03774	0.02280
	MICE-	0.00304	0.01232	0.00074	0.00514	0.02665	0.01222
	MICE+	0.00316	0.01251	0.00075	0.00500	0.02721	0.01507
0.3	CCA	0.00525	0.02156	0.00105	0.00693	0.04481	0.02614
	RVI	0.00351	0.01237	0.00126	0.00524	0.03078	0.01866
	K3-	0.00367	0.01238	0.00080	0.00538	0.03531	0.01560
	K3+	0.00457	0.01281	0.00086	0.00526	0.03808	0.02409
	K5-	0.00363	0.01244	0.00086	0.00535	0.03292	0.01576
	K5+	0.00487	0.01284	0.00095	0.00526	0.04092	0.02681
	K9-	0.00359	0.01240	0.00096	0.00530	0.03067	0.01662
	K9+	0.00504	0.01281	0.00109	0.00527	0.04264	0.03245
	MF-	0.00370	0.01250	0.00076	0.00540	0.03785	0.01557
	MF+	0.00597	0.01335	0.00090	0.00511	0.05666	0.03373
	MICE-	0.00349	0.01229	0.00074	0.00540	0.03324	0.01279
	MICE+	0.00364	0.01260	0.00076	0.00524	0.03193	0.01776
0.4	CCA	0.00718	0.02968	0.00159	0.00903	0.06480	0.03603
	RVI	0.00385	0.01247	0.00147	0.00541	0.03804	0.02127
	K3-	0.00413	0.01258	0.00081	0.00564	0.04409	0.01702
	K3+	0.00596	0.01322	0.00095	0.00573	0.05901	0.03892
	K5-	0.00404	0.01262	0.00089	0.00560	0.04067	0.01744
	K5+	0.00680	0.01337	0.00112	0.00581	0.06613	0.04818
	K9-	0.00401	0.01252	0.00104	0.00559	0.03883	0.01824
	K9+	0.00740	0.01333	0.00133	0.00582	0.06803	0.06095
	MF-	0.00449	0.01297	0.00076	0.00565	0.04762	0.01741
	MF+	0.00960	0.01454	0.00102	0.00555	0.08554	0.04938
	MICE-	0.00395	0.01250	0.00073	0.00562	0.04190	0.01320
	MICE+	0.00415	0.01295	0.00075	0.00547	0.04059	0.02082

Ref.: Bj=β_j; CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales. En escala de rosas: mayores errores cuadráticos medios, por parámetro. En escala de verdes: menores errores cuadráticos medios, por parámetro.

Cuadro 4.8: Error cuadrático medio de los parámetros del modelo de regresión de Cox para n=1000, según método de imputación y proporción de datos faltantes (p).

p	Método	B1	B2	B3	B4	B5	B6
0.1	CCA	0.00156	0.00753	0.00063	0.00360	0.01363	0.00748
	RVI	0.00146	0.00687	0.00077	0.00369	0.01246	0.00776
	K3-	0.00156	0.00692	0.00063	0.00377	0.01473	0.00707
	K3+	0.00149	0.00698	0.00064	0.00365	0.01318	0.00715
	K5-	0.00152	0.00692	0.00065	0.00375	0.01414	0.00720
	K5+	0.00147	0.00697	0.00066	0.00363	0.01279	0.00726
	K9-	0.00150	0.00690	0.00069	0.00372	0.01333	0.00718
	K9+	0.00146	0.00697	0.00069	0.00363	0.01287	0.00739
	MF-	0.00149	0.00692	0.00062	0.00377	0.01492	0.00699
	MF+	0.00150	0.00693	0.00065	0.00352	0.01424	0.00837
	MICE-	0.00160	0.00689	0.00061	0.00377	0.01445	0.00658
	MICE+	0.00152	0.00692	0.00062	0.00368	0.01281	0.00676
0.2	CCA	0.00198	0.00892	0.00068	0.00381	0.01680	0.00971
	RVI	0.00166	0.00708	0.00095	0.00383	0.01382	0.00945
	K3-	0.00190	0.00710	0.00066	0.00393	0.01953	0.00806
	K3+	0.00181	0.00726	0.00068	0.00373	0.01544	0.00861
	K5-	0.00181	0.00714	0.00070	0.00390	0.01778	0.00836
	K5+	0.00182	0.00728	0.00073	0.00372	0.01540	0.00907
	K9-	0.00175	0.00713	0.00077	0.00388	0.01574	0.00859
	K9+	0.00184	0.00727	0.00081	0.00372	0.01580	0.01001
	MF-	0.00175	0.00719	0.00064	0.00395	0.02012	0.00794
	MF+	0.00201	0.00730	0.00071	0.00353	0.02007	0.01219
	MICE-	0.00209	0.00710	0.00062	0.00396	0.01978	0.00713
	MICE+	0.00180	0.00716	0.00064	0.00381	0.01441	0.00778
0.3	CCA	0.00245	0.01072	0.00072	0.00416	0.02161	0.01255
	RVI	0.00183	0.00692	0.00112	0.00403	0.01550	0.01106
	K3-	0.00225	0.00699	0.00067	0.00419	0.02604	0.00955
	K3+	0.00218	0.00724	0.00071	0.00396	0.01950	0.01147
	K5-	0.00210	0.00697	0.00072	0.00416	0.02258	0.00977
	K5+	0.00232	0.00725	0.00080	0.00396	0.02026	0.01277
	K9-	0.00198	0.00696	0.00082	0.00412	0.01901	0.01025
	K9+	0.00247	0.00729	0.00093	0.00394	0.02118	0.01546
	MF-	0.00206	0.00709	0.00064	0.00421	0.02700	0.00941
	MF+	0.00304	0.00747	0.00076	0.00369	0.03087	0.01799
	MICE-	0.00261	0.00692	0.00061	0.00423	0.02676	0.00801
	MICE+	0.00208	0.00706	0.00063	0.00404	0.01627	0.00868
0.4	CCA	0.00341	0.01409	0.00094	0.00497	0.03013	0.01792
	RVI	0.00206	0.00699	0.00136	0.00417	0.01788	0.01184
	K3-	0.00265	0.00709	0.00069	0.00441	0.03353	0.01104
	K3+	0.00297	0.00745	0.00077	0.00422	0.02942	0.01858
	K5-	0.00245	0.00706	0.00076	0.00438	0.02907	0.01153
	K5+	0.00340	0.00757	0.00093	0.00424	0.03291	0.02321
	K9-	0.00229	0.00706	0.00090	0.00433	0.02337	0.01174
	K9+	0.00386	0.00761	0.00115	0.00427	0.03521	0.03132
	MF-	0.00266	0.00726	0.00065	0.00443	0.03530	0.01089
	MF+	0.00559	0.00801	0.00085	0.00398	0.04985	0.02832
	MICE-	0.00323	0.00703	0.00062	0.00442	0.03539	0.00893
	MICE+	0.00242	0.00721	0.00064	0.00425	0.01947	0.01022

Ref.: Bj=β_j; CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales. En escala de rosas: mayores errores cuadráticos medios, por parámetro. En escala de verdes: menores errores cuadráticos medios, por parámetro.

Sesgo

El sesgo se estudia en forma global para los seis parámetros, promediando la diferencia relativa entre la estimación promedio y el valor real (Figura 4.10).

Cuando el *tiempo* y el *estado* son incluidas como variables informativas, el indicador presenta los valores más pequeños, debido a diferencias relativas pequeñas entre la media teórica y empírica de los estimadores. Cuando dichas variables no son incluidas, el efecto de la proporción de pérdidas sobre el sesgo es más evidente, ya que el mismo se incrementa de manera notoria al aumentar p y MICE- presenta los resultados más desfavorables, en todas las situaciones.

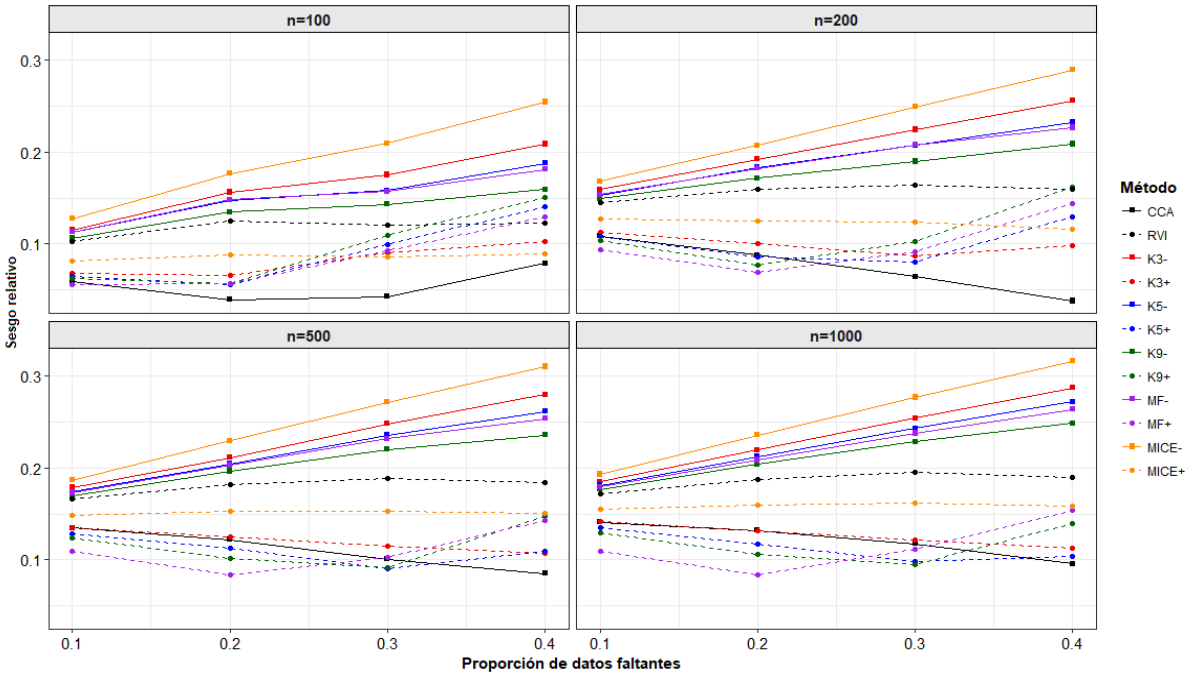
Al emplear KNN, los resultados mejoran al aumentar el número de vecinos, k. El indicador correspondiente a MF- resulta intermedio a los obtenidos en K5- y K9-, y los valores correspondientes a RVI se encuentran en la zona intermedia de los resultados globales.

En el caso de $n = 100$, CCA produce los resultados más favorables para todos los valores de p, salvo en el caso de $p = 0.10$, donde es levemente mejorado por MF+. Cuando se incluyen las variables informativas adicionales, MICE presenta el mayor valor del indicador para $p = 0.10$ y 0.20, y MF muestra el menor valor del mismo. Para $p = 0.30$ y 0.40, esto se revierte y MICE+ presenta el valor más chico, mejorado solamente por CCA.

Al tener en cuenta las variables informativas adicionales para $n = 200$, MICE muestra indicadores menos favorables cuando $p \leq 0.30$. Esto se revierte cuando $p = 0.40$, siendo mejorado solo por CCA y K3+. Para $p = 0.10$ y 0.20, MF+ muestra los mejores valores del indicador, mientras que para $p = 0.30$, CCA y K5+ resultan los mejores.

Para $n = 500$ y 1000, MICE se ubica como el indicador con mayor valor dentro de los casos en los que se tienen en cuenta las variables informativas adicionales. MF+ resulta la técnica con mejor indicador para $p = 0.10$ y 0.20, K5+ y K9+ lo hacen cuando $p = 0.30$ y CCA presenta el mejor indicador para $p = 0.40$, seguido por K3+ y K5+. Para esta proporción de datos faltantes, MF+ y MICE+ muestran resultados muy similares.

Figura 4.10: Diferencia relativa media, en valor absoluto, entre el promedio de las estimaciones de los parámetros y sus valores reales, según método de imputación, tamaño de muestra y proporción de datos faltantes.



Ref.: CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

Variabilidad de los estimadores respecto de la variabilidad esperada

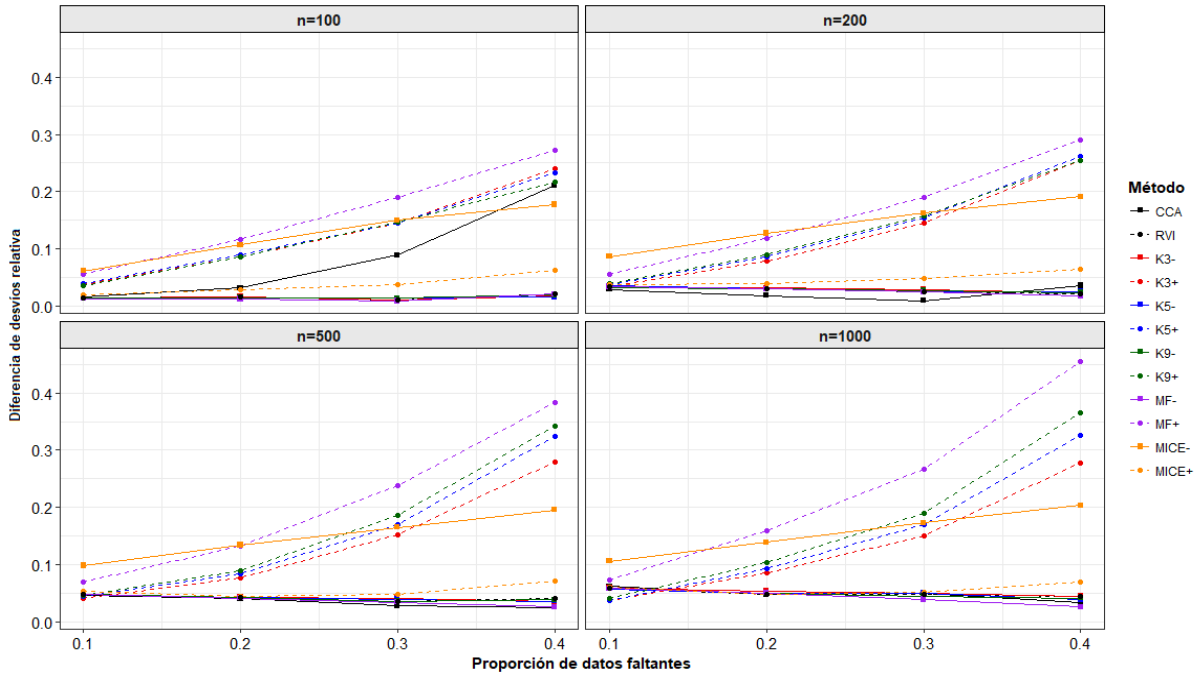
En la Figura 4.11 se muestran los resultados obtenidos al calcular el indicador de la diferencia relativa en valor absoluto, promediada para los seis parámetros, entre el desvío estándar empírico y teórico.

Para todas las combinaciones de n y p, la no inclusión del *tiempo* y el *estado* como variables informativas proveen indicadores menores, muy cercanos a cero y similares para todas las técnicas, incluyendo CCA y RVI y a excepción de MICE. Estos valores se incrementan muy levemente a medida que aumenta p, para cada tamaño de muestra. Cuando se incluyen las variables informativas adicionales, el incremento en el indicador para cada técnica es más notorio.

MICE- presenta los mayores valores para el indicador cuando $p = 0.10$ y 0.20, mientras que para los p mayores, MF+ y KNN+ muestran valores superiores.

MICE es la única técnica bajo la cual se obtienen indicadores menores cuando se incluyen las variables informativas adicionales.

Figura 4.11: Diferencia relativa media, en valor absoluto, entre el promedio de los desvíos estándar teóricos de los estimadores y el desvío estándar empírico, según método de imputación, tamaño de muestra y proporción de datos faltantes.



Ref.: CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

Cobertura

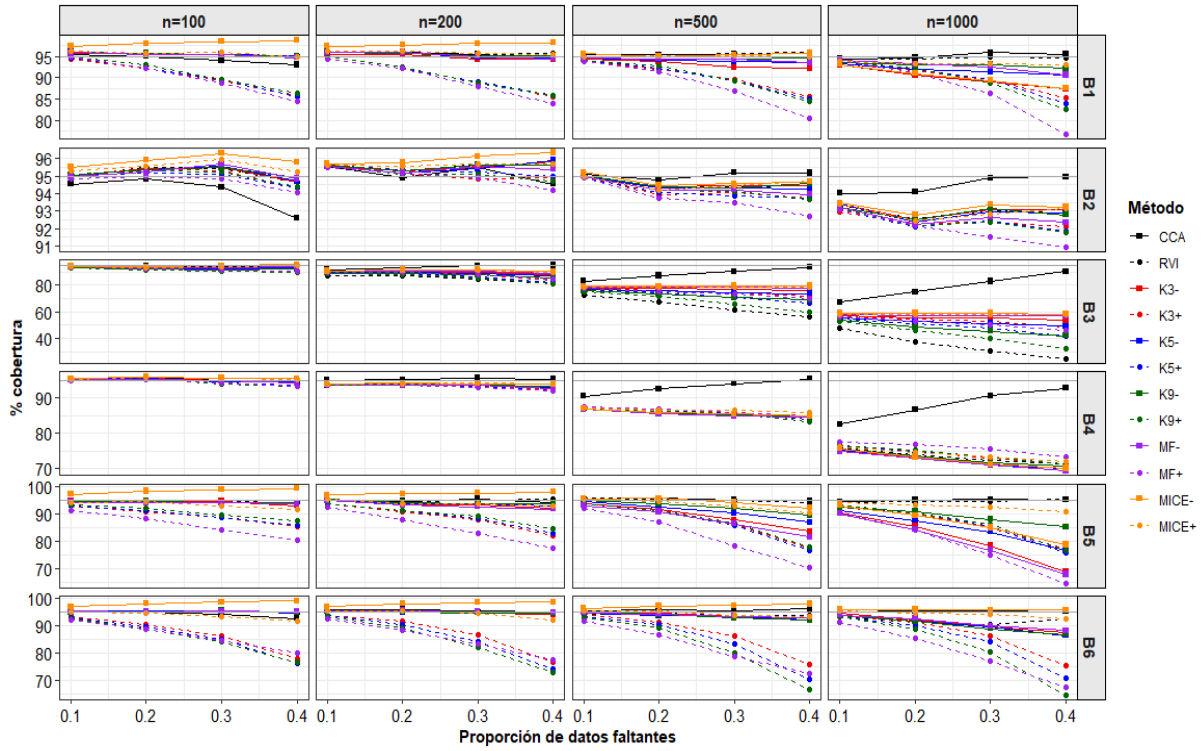
Esta propiedad se analiza, separadamente, para cada parámetro del modelo de regresión de Cox 4.1 (Figura 4.12). En general, para todos los parámetros y valores de n, la diferencia observada entre los métodos de imputación aumenta a medida que aumenta la proporción de pérdidas.

Para los parámetros β_1, β_2, β_5 y β_6, correspondientes a la variable con distribución Normal y a las cualitativas, los métodos de imputación que incluyen como variables informativas adicionales al *tiempo* y *estado* presentan menores porcentajes de cobertura, siendo MF+ el que muestra los valores más desfavorables en la mayoría de los escenarios. Para $n = 100$ y 200, MICE- presenta la mayor cobertura. Para $n = 500$ y 1000, también CCA se ubica como la técnica con resultados más favorables. En todos los casos, el porcentaje de cobertura es cercano al 95 % para todos los métodos excepto MF+ y KN+, quienes se alejan de dicho valor a medida que aumenta p.

Para los parámetros β_3 y β_4, correspondientes a las variables continuas con distri-

bución no Normal, no se observan diferencias en los resultados cuando $n = 100$. Para $n = 200$, 500 y 1000, CCA muestra el mayor porcentaje de cobertura. En particular, para β_3 el porcentaje de cobertura de RVI resulta el más bajo, mientras que para β_4 no se observan diferencias en los resultados de los métodos de imputación. Para estos parámetros, la cobertura con $n = 100$ y 200 es cercana al 95 % para todos los métodos, pero resulta inferior a dicho valor para tamaños muestrales iguales a 500 y 1000.

Figura 4.12: Porcentaje de cobertura para los intervalos de confianza del 95 %, según método de imputación, tamaño de muestra y proporción de datos faltantes.



Ref.: Bj: β_j; CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

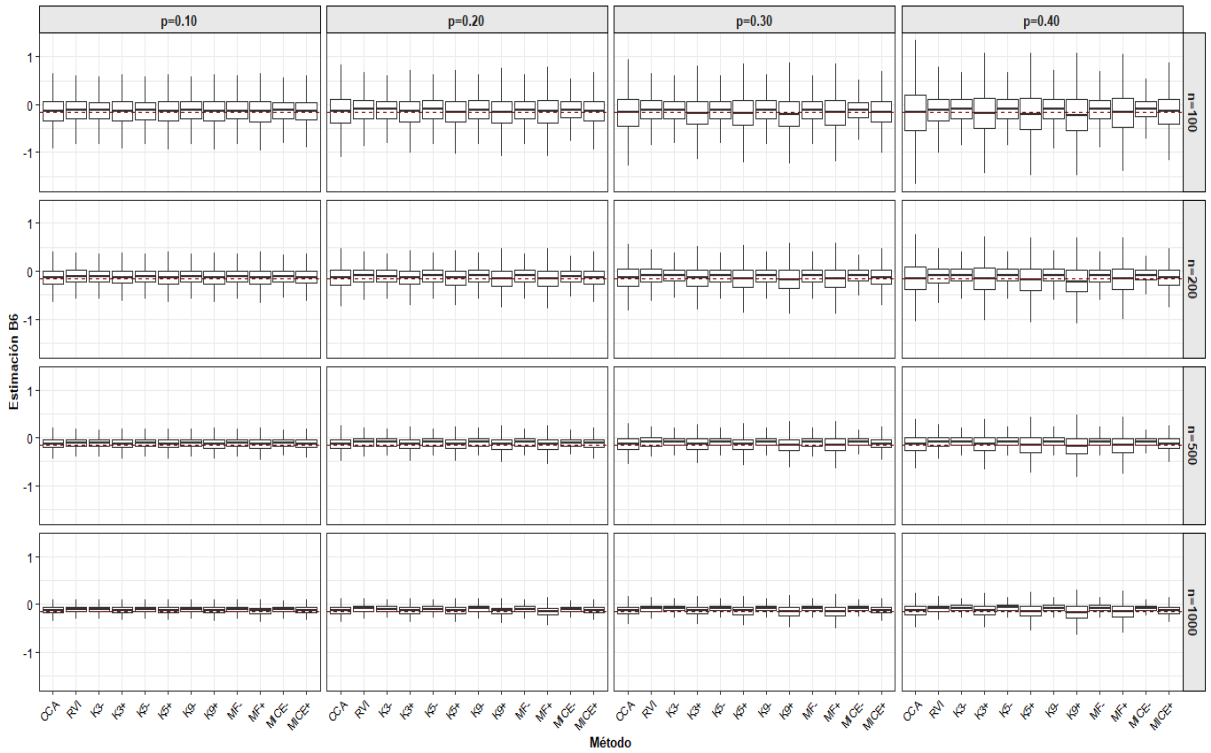
Distribuciones en el muestreo

Los box-plots de la Figura 4.13 aproximan la distribución de los estimadores a partir de los valores obtenidos en las 5000 iteraciones del proceso de simulación. Se presentan los resultados para el parámetro β_6 del modelo 4.1, según método de tratamiento de la información faltante, en diferentes paneles según tamaño muestral y proporción de pérdidas.

Se observa que, dentro de cada método de imputación, existe menor variabilidad en

las estimaciones cuando no se incluyen el *tiempo* y el *estado* como variables informativas, aunque su mediana presenta un alejamiento respecto al verdadero valor del parámetro. Además, se evidencia que, a mayor tamaño de muestra, menor en la variabilidad en las estimaciones. En todos los escenarios, CCA es la técnica donde se presenta mayor variación de las estimaciones.

Figura 4.13: Distribución de los estimadores de β_6, según método de imputación, tamaño de muestra y proporción de datos faltantes, sin incluir valores extremos(1).



(1): La presencia de valores extremos no está asociada con ningún efecto medido. Ref.: CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

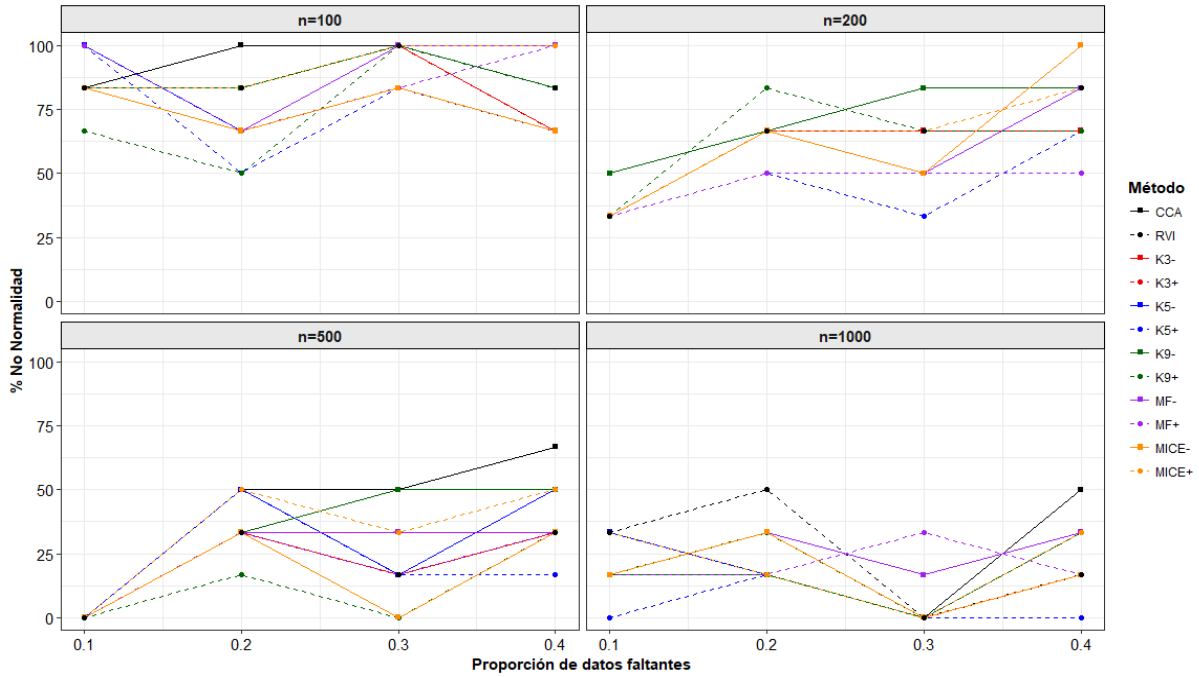
En la Figura 4.14 se muestra el porcentaje de parámetros, entre los seis correspondientes al modelo 4.1, en los que se rechaza la hipótesis de distribución Normal de acuerdo al test de Anderson-Darling aplicado sobre las 5000 estimaciones para cada uno de ellos.

Si bien no se aprecia una tendencia al comparar los métodos en los distintos escenarios, se distingue que, a medida que aumenta el tamaño de muestra, disminuye el porcentaje en estudio, es decir, aumenta el porcentaje de parámetros para los cuales se acepta que su estimador presenta distribución Normal.

Dado que el test de Anderson-Darling se aplica sobre un gran número de datos

(5000 estimaciones para cada método y escenario), el mismo se vuelve muy sensible, provocando un rechazo de la hipótesis de normalidad ante pequeños alejamientos de dicha distribución. Sin embargo, no se observan características en las distribuciones que sugieran que el supuesto de normalidad no es aceptable.

Figura 4.14: Porcentaje de parámetros para los cuales se rechaza la hipótesis del test de Anderson-Darling, según método de imputación, tamaño de la muestra y proporción de datos faltantes.



Ref.: CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

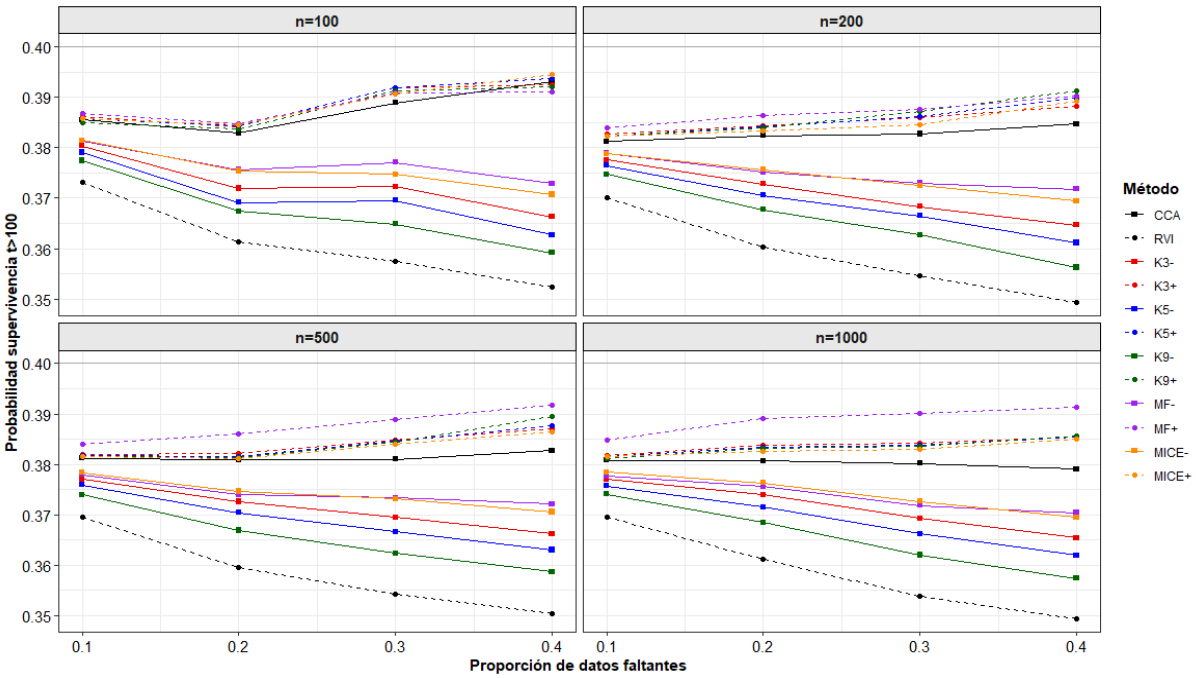
Estimación de la probabilidad de sobrevida

En la Figura 4.15 se muestra el promedio de la probabilidad de supervivencia estimada para $t > 100$, a partir de las 5000 estimaciones de los parámetros del modelo 4.1, para una unidad con vector de variables explicativas $\boldsymbol{x} = (0, 1, 5, 1, 0, 1)$.

Los resultados son muy similares para los cuatro valores de n considerados y la diferencia entre los resultados de las diferentes técnicas se amplía a medida que aumenta p. En todos los casos, la probabilidad de supervivencia resulta subestimada. RVI muestra los resultados más alejados del valor real, seguido por KNN-, quien presenta valores más cercanos al real para los valores de k más chicos. MICE- y MF- muestran resultados muy similares entre sí. Cuando se incluyen las variables informativas adicionales, las estima-

ciones medias con muy similares para todas las técnicas, incluyendo CCA. MF+ es quien presenta valores más cercanos a la probabilidad de supervivencia verdadera es la mayoría de las situaciones.

Figura 4.15: Promedio de la probabilidad de supervivencia estimada para $t > 100$, según método de imputación, tamaño de la muestra y proporción de datos faltantes.



Ref.: CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

4.2.2.b. Reproducibilidad de los datos perdidos

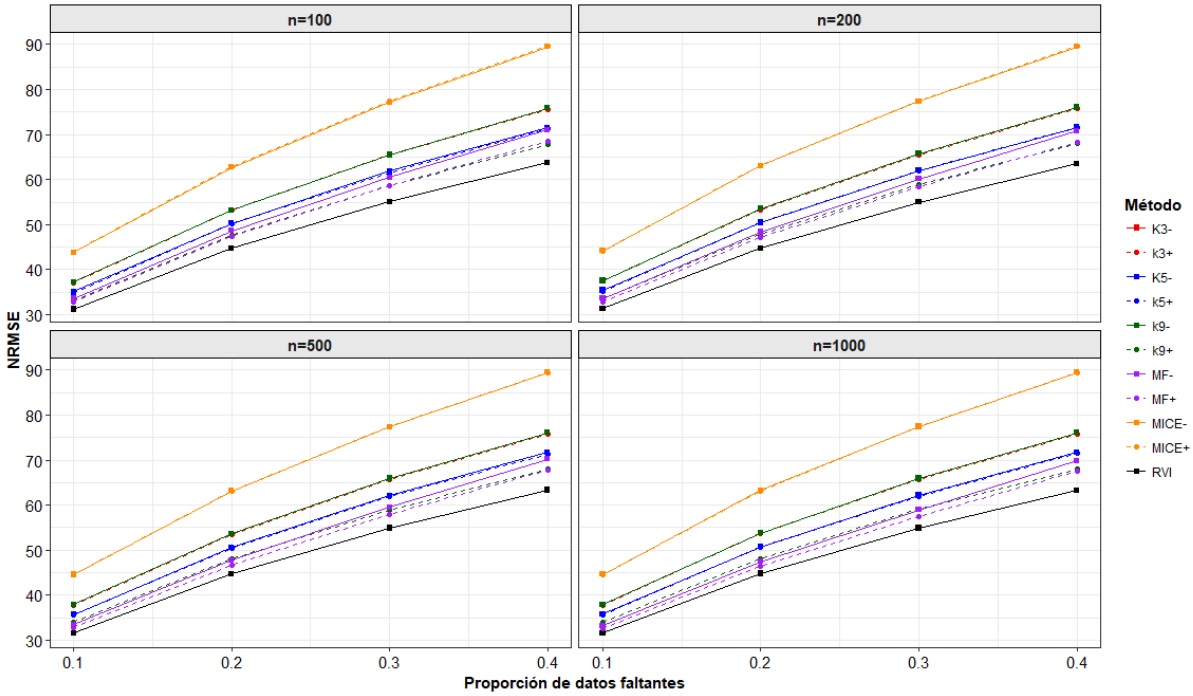
En este apartado, se realiza una evaluación de la eficiencia de los métodos de imputación para asignar valores a los datos faltantes, según se trate de observaciones correspondientes a la variable cuantitativa X_1 o a las variables *dummies* (X_6, X_7), cuando se han generado datos perdidos de acuerdo a un mecanismo MAR.

Raíz del error cuadrático medio normalizado

En relación a la imputación de datos para X_1, de acuerdo con la Figura 4.16, se encuentran los mayores valores para $\overline{\text{NRMSE}}$ al utilizar MICE y los menores valores corresponden al uso de RVI, seguidos por MF. En el caso del empleo de KNN, los resultados más favorables se encuentran con $k = 5$. No se distinguen diferencias importantes en los

valores de $\overline{\text{NRMSE}}$ según se considere el *tiempo* y el *estado* como variables informativas. El comportamiento de $\overline{\text{NRMSE}}$ es muy similar para todos los tamaños de muestra considerados. En todos los casos, $\overline{\text{NRMSE}}$ aumenta a medida que aumenta la proporción de datos faltantes.

Figura 4.16: Error cuadrático medio normalizado (NRMSE) para X_1, según método de imputación, tamaño de la muestra y proporción de datos faltantes.

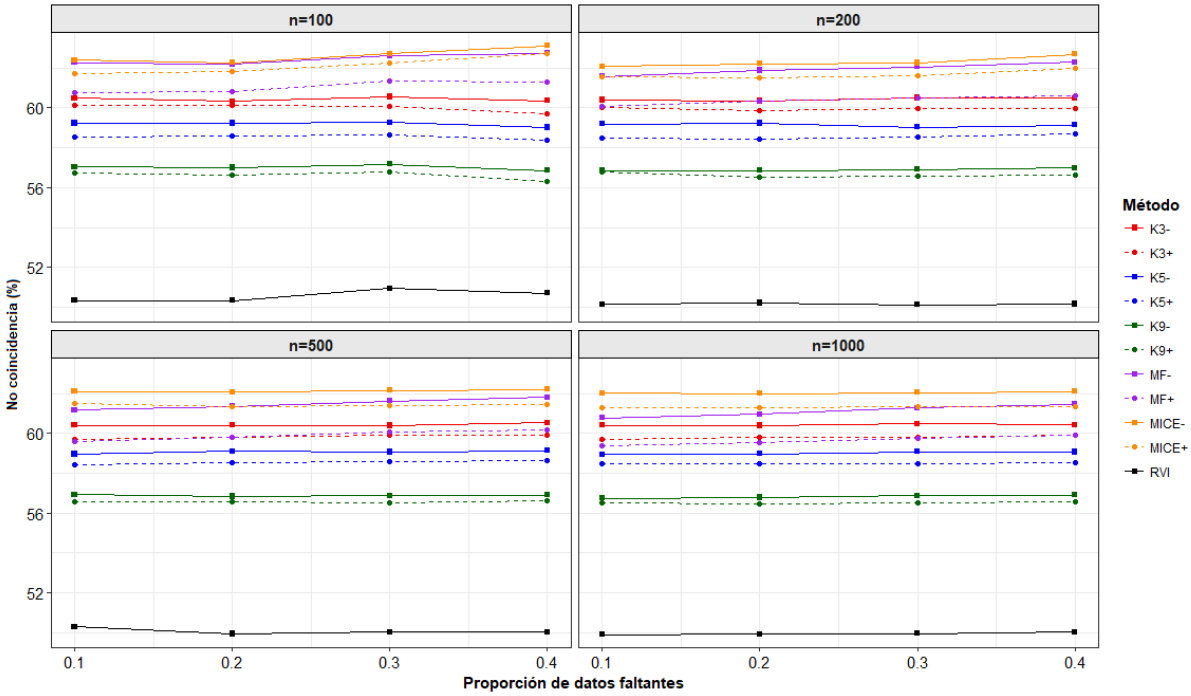


Ref.: CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

Porcentaje de no coincidencia

En la Figura 4.17 se muestran los porcentajes medios de datos imputados en forma incorrecta para las variables *dummies* (X_6, X_7). Los valores más pequeños se encuentran bajo RVI, mientras que los más desfavorables corresponden a MICE-. Haciendo uso de KNN, se observan resultados más favorables a medida que aumenta el número de donantes. En todos los casos, se distinguen mejores resultados al considerar el *tiempo* y el *estado* como variables informativas. El comportamiento de los porcentajes medios de datos imputados en forma incorrecta es muy similar para todos los tamaños de muestra y para todas las proporciones de datos faltantes.

Figura 4.17: Porcentaje de datos imputados no coincidentes con los reales para (X_6, X_7), según método de imputación, tamaño de la muestra y proporción de datos faltantes.



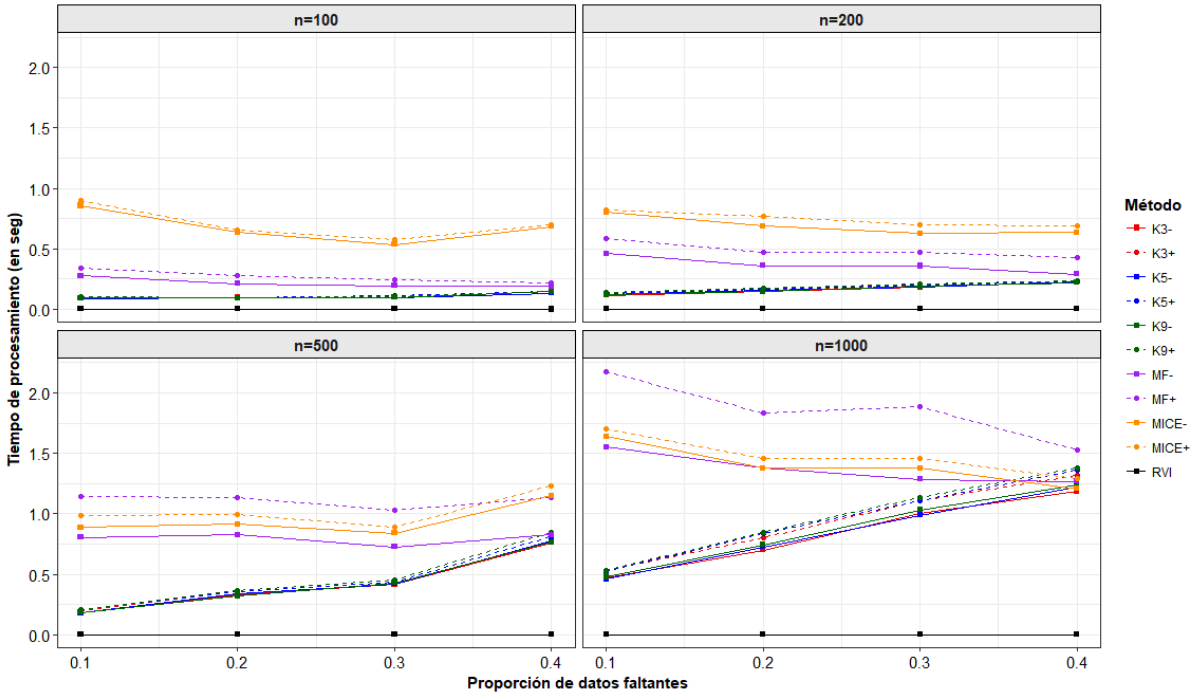
Ref.: CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

4.2.2.c. Tiempos computacionales de imputación

En relación a los tiempos que demora el software utilizado para realizar las imputaciones, de acuerdo a los datos presentados en la figura 4.18, se observa que RVI demanda muy poco tiempo para realizar la imputación completa del conjunto de datos, seguido por KNN. Para esta técnica, no se observan diferencias destacables para las distintas configuraciones empleadas.

Cuando $n = 100$ o $n = 200$, MF demanda menor tiempo para la imputación que MICE. Cuando $n = 500$ o $n = 1000$, MF+ es la técnica que demanda mayor tiempo de procesamiento, aunque MF- se mantiene por debajo de MICE. Cabe destacar, además, que para $n = 1000$ y $p = 0.40$, los tiempos de procesamiento requeridos por todas las técnicas son muy similares entre sí.

Figura 4.18: Tiempo promedio, en segundos, demandado para la imputación completa de los datos, según método de imputación, tamaño de la muestra y proporción de datos faltantes.



Ref.: CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

4.2.3. Datos perdidos no al azar

4.2.3.a. Propiedades distribucionales de los estimadores de los coeficientes del modelo de regresión de Cox

Error cuadrático medio

Los resultados obtenidos al calcular el MSE para cada uno de los seis parámetros del modelo 4.1 se muestran en los Cuadros 4.9 a 4.12, según el tamaño de la muestra y la proporción de datos faltantes p.

En el caso de $n = 100$, los resultados más desfavorables se observan bajo el uso de CCA y MF+, mientras que RVI aporta los menores valores de MSE. En el caso de $p = 0.10$, K3+ y K5- también se relacionan con los MSE menores (Cuadro 4.9).

Para $n = 200$ y $p = 0.10$, se encuentran los resultados más desfavorables al utilizar CCA, mientras que RVI, K5+ y K9+ se encuentran entre los más favorables. Para $p = 0.20$ y $p = 0.30$, CCA y MF presentan los resultados más desfavorables y RVI y K5+, los más favorables. Se observan menores MSE cuando se consideran el *tiempo* y el *estado* como variables informativas (Cuadro 4.10).

Cuando $n = 500$ y $p = 0.10$, CCA, K3- y MICE- presentan los mayores MSE y K9+ junto con RVI presentan los menores MSE. Para $p = 0.20$, MF- y K3- se asocian a los mayores MSE, siendo K9+, K5+ y RVI quienes tienen menores MSE. Cuando $p = 0.30$, nuevamente MF- y K3- se asocian a los mayores MSE y KNN+, en general, presentan los MSE más chicos. Y para $p = 0.40$, MF tiene los MSE mayores, mientras que RVI, K3+ y K5+ resultan los más favorables para tres de los seis parámetros estimados (Cuadro 4.11).

Para $n = 1000$ y $p = 0.10$, MICE- y K3- tienen los MSE más grandes y K9+, K5+ y MF+, los más pequeños. En el caso de $p = 0.20$, MICE- y K3- también presentan los MSE más grandes y K9+ junto con K5+, los más pequeños. Para $p = 0.30$, MICE-, K3- y MF- se ubican entre los más desfavorables respecto a sus MSE, mientras que RVI, K5+ y K9+ presentan los valores más favorables. Para $p = 0.40$, MF- tiene los MSE superiores y K5+ y K9+ se asocian a los MSE inferiores (Cuadro 4.12).

Cuadro 4.9: Error cuadrático medio de los parámetros del modelo de regresión de Cox para n=100, según método de imputación y proporción de datos faltantes (p).

p	Método	B1	B2	B3	B4	B5	B6
0.1	CCA	0.01782	0.08636	0.00254	0.02228	0.16237	0.09279
	RVI	0.01354	0.06874	0.00200	0.01813	0.12092	0.07094
	K3-	0.01369	0.06883	0.00207	0.02148	0.12249	0.07165
	K3+	0.01374	0.07576	0.00203	0.01964	0.11945	0.07030
	K5-	0.01359	0.06844	0.00205	0.01995	0.12175	0.07143
	K5+	0.01367	0.07531	0.00203	0.01982	0.11916	0.07029
	K9-	0.01361	0.06914	0.00204	0.01912	0.12116	0.07109
	K9+	0.01364	0.07520	0.00202	0.01994	0.11967	0.07080
	MF-	0.01373	0.06918	0.00208	0.02030	0.12281	0.07178
	MF+	0.01384	0.08342	0.00203	0.02098	0.12257	0.07234
	MICE-	0.01334	0.06065	0.00205	0.02485	0.11975	0.06999
	MICE+	0.01355	0.07260	0.00203	0.02018	0.12151	0.07114
0.2	CCA	0.02682	0.10608	0.00338	0.02986	0.22629	0.12574
	RVI	0.01433	0.06512	0.00214	0.01899	0.12676	0.06786
	K3-	0.01485	0.06713	0.00233	0.02740	0.12925	0.07018
	K3+	0.01509	0.08073	0.00225	0.02273	0.12545	0.06803
	K5-	0.01490	0.06743	0.00229	0.02266	0.12898	0.06989
	K5+	0.01496	0.08230	0.00222	0.02357	0.12566	0.06795
	K9-	0.01485	0.06737	0.00226	0.02095	0.12854	0.06917
	K9+	0.01475	0.08136	0.00221	0.02415	0.12585	0.06802
	MF-	0.01515	0.06824	0.00235	0.02416	0.13288	0.07200
	MF+	0.01545	0.09965	0.00226	0.02654	0.13322	0.07271
	MICE-	0.01409	0.05364	0.00229	0.04248	0.12744	0.06809
	MICE+	0.01473	0.07493	0.00222	0.02498	0.13030	0.07060
0.3	CCA	0.04132	0.16153	0.00505	0.04482	0.33575	0.19503
	RVI	0.01383	0.07001	0.00207	0.02005	0.11962	0.06979
	K3-	0.01471	0.07268	0.00234	0.03429	0.12559	0.07272
	K3+	0.01526	0.09903	0.00230	0.02928	0.11972	0.06922
	K5-	0.01474	0.07286	0.00232	0.02632	0.12403	0.07227
	K5+	0.01486	0.10170	0.00229	0.03224	0.11897	0.06940
	K9-	0.01455	0.07466	0.00229	0.02304	0.12317	0.07136
	K9+	0.01447	0.10163	0.00224	0.03408	0.11889	0.06924
	MF-	0.01517	0.07589	0.00238	0.02968	0.12873	0.07413
	MF+	0.01607	0.13709	0.00232	0.03891	0.13102	0.07859
	MICE-	0.01371	0.05320	0.00228	0.06292	0.12207	0.06930
	MICE+	0.01465	0.08934	0.00223	0.02925	0.12811	0.07353
0.4	CCA	0.07057	0.25765	0.00884	0.08557	1.46489	0.34102
	RVI	0.01391	0.08438	0.00200	0.02108	0.12648	0.06921
	K3-	0.01495	0.08051	0.00238	0.04635	0.13107	0.07178
	K3+	0.01567	0.11822	0.00239	0.04243	0.12480	0.06849
	K5-	0.01515	0.08345	0.00236	0.03262	0.13023	0.07055
	K5+	0.01546	0.12846	0.00236	0.04943	0.12474	0.06817
	K9-	0.01502	0.08726	0.00238	0.02746	0.12882	0.06933
	K9+	0.01503	0.12733	0.00231	0.05258	0.12490	0.06782
	MF-	0.01562	0.08952	0.00247	0.04003	0.13487	0.07446
	MF+	0.01738	0.17934	0.00244	0.06562	0.14417	0.08260
	MICE-	0.01385	0.05498	0.00227	0.09020	0.12592	0.06788
	MICE+	0.01498	0.10849	0.00226	0.03769	0.13706	0.07558

Ref.: Bj=β_j; CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales. En escala de rosas: mayores errores cuadráticos medios, por parámetro. En escala de verdes: menores errores cuadráticos medios, por parámetro.

Cuadro 4.10: Error cuadrático medio de los parámetros del modelo de regresión de Cox para n=200, según método de imputación y proporción de datos faltantes (p).

p	Método	B1	B2	B3	B4	B5	B6
0.1	CCA	0.00806	0.03612	0.00126	0.01082	0.06861	0.04069
	RVI	0.00627	0.03077	0.00113	0.01038	0.05404	0.03200
	K3-	0.00638	0.03131	0.00121	0.01316	0.05474	0.03268
	K3+	0.00638	0.03356	0.00113	0.00929	0.05408	0.03206
	K5-	0.00634	0.03133	0.00119	0.01117	0.05475	0.03239
	K5+	0.00636	0.03353	0.00112	0.00905	0.05392	0.03209
	K9-	0.00629	0.03099	0.00118	0.01026	0.05444	0.03232
	K9+	0.00634	0.03281	0.00112	0.00897	0.05383	0.03211
	MF-	0.00639	0.03077	0.00122	0.01159	0.05509	0.03265
	MF+	0.00640	0.03714	0.00113	0.00940	0.05501	0.03284
	MICE-	0.00628	0.02869	0.00126	0.02235	0.05406	0.03192
	MICE+	0.00633	0.03236	0.00117	0.01208	0.05437	0.03247
0.2	CCA	0.01039	0.04332	0.00150	0.01265	0.09377	0.05271
	RVI	0.00616	0.02912	0.00116	0.01129	0.05470	0.03011
	K3-	0.00645	0.03314	0.00130	0.01831	0.05673	0.03138
	K3+	0.00659	0.03570	0.00115	0.00968	0.05536	0.03074
	K5-	0.00642	0.03218	0.00126	0.01293	0.05631	0.03108
	K5+	0.00653	0.03632	0.00114	0.01012	0.05495	0.03053
	K9-	0.00640	0.03147	0.00124	0.01054	0.05585	0.03093
	K9+	0.00646	0.03700	0.00114	0.01050	0.05450	0.03060
	MF-	0.00655	0.03340	0.00130	0.01388	0.05696	0.03214
	MF+	0.00670	0.04616	0.00116	0.01163	0.05691	0.03261
	MICE-	0.00621	0.02806	0.00136	0.04248	0.05563	0.03048
	MICE+	0.00639	0.03333	0.00122	0.01429	0.05644	0.03159
0.3	CCA	0.01524	0.06146	0.00207	0.01628	0.12776	0.07432
	RVI	0.00639	0.03133	0.00115	0.01232	0.05526	0.03127
	K3-	0.00671	0.03818	0.00143	0.02740	0.05948	0.03224
	K3+	0.00713	0.04403	0.00122	0.01248	0.05650	0.03086
	K5-	0.00675	0.03677	0.00138	0.01691	0.05842	0.03176
	K5+	0.00701	0.04707	0.00120	0.01449	0.05590	0.03087
	K9-	0.00672	0.03613	0.00135	0.01203	0.05745	0.03142
	K9+	0.00689	0.05023	0.00121	0.01600	0.05540	0.03088
	MF-	0.00692	0.03862	0.00145	0.01987	0.06009	0.03296
	MF+	0.00739	0.06518	0.00123	0.01889	0.06082	0.03458
	MICE-	0.00636	0.03005	0.00147	0.06849	0.05672	0.03080
	MICE+	0.00672	0.03867	0.00129	0.01866	0.05832	0.03219
0.4	CCA	0.02289	0.08519	0.00297	0.02477	0.19187	0.11305
	RVI	0.00605	0.04378	0.00117	0.01375	0.05472	0.03071
	K3-	0.00655	0.04665	0.00153	0.03691	0.05730	0.03116
	K3+	0.00718	0.05475	0.00130	0.01968	0.05569	0.03011
	K5-	0.00660	0.04579	0.00150	0.02088	0.05672	0.03093
	K5+	0.00706	0.06189	0.00130	0.02499	0.05479	0.03007
	K9-	0.00661	0.04681	0.00148	0.01356	0.05597	0.03056
	K9+	0.00672	0.07110	0.00132	0.02860	0.05458	0.02997
	MF-	0.00681	0.05133	0.00157	0.02813	0.05920	0.03249
	MF+	0.00767	0.09133	0.00132	0.03647	0.06384	0.03646
	MICE-	0.00614	0.03438	0.00155	0.09665	0.05586	0.02995
	MICE+	0.00653	0.04663	0.00139	0.02371	0.05922	0.03241

Ref.: Bj=β_j; CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales. En escala de rosas: mayores errores cuadráticos medios, por parámetro. En escala de verdes: menores errores cuadráticos medios, por parámetro.

Cuadro 4.11: Error cuadrático medio de los parámetros del modelo de regresión de Cox para n=500, según método de imputación y proporción de datos faltantes (p).

p	Método	B1	B2	B3	B4	B5	B6
0.1	CCA	0.00308	0.01442	0.00075	0.00519	0.02652	0.01452
	RVI	0.00251	0.01259	0.00074	0.00619	0.02179	0.01186
	K3-	0.00260	0.01394	0.00082	0.00871	0.02262	0.01217
	K3+	0.00256	0.01326	0.00070	0.00368	0.02181	0.01188
	K5-	0.00257	0.01350	0.00080	0.00643	0.02223	0.01197
	K5+	0.00254	0.01337	0.00070	0.00346	0.02182	0.01189
	K9-	0.00254	0.01314	0.00078	0.00532	0.02200	0.01186
	K9+	0.00252	0.01307	0.00070	0.00334	0.02178	0.01188
	MF-	0.00260	0.01406	0.00081	0.00654	0.02240	0.01198
	MF+	0.00255	0.01463	0.00070	0.00335	0.02211	0.01211
	MICE-	0.00261	0.01393	0.00090	0.02167	0.02262	0.01194
	MICE+	0.00254	0.01336	0.00077	0.00704	0.02209	0.01198
0.2	CCA	0.00381	0.01754	0.00077	0.00567	0.03276	0.01890
	RVI	0.00251	0.01267	0.00074	0.00750	0.02140	0.01200
	K3-	0.00268	0.01809	0.00094	0.01470	0.02344	0.01234
	K3+	0.00268	0.01527	0.00069	0.00346	0.02204	0.01210
	K5-	0.00263	0.01700	0.00089	0.00885	0.02297	0.01224
	K5+	0.00266	0.01518	0.00068	0.00358	0.02187	0.01203
	K9-	0.00261	0.01592	0.00086	0.00607	0.02246	0.01211
	K9+	0.00264	0.01505	0.00068	0.00380	0.02173	0.01200
	MF-	0.00269	0.01778	0.00093	0.00939	0.02335	0.01247
	MF+	0.00267	0.01927	0.00069	0.00444	0.02239	0.01267
	MICE-	0.00263	0.01734	0.00105	0.04650	0.02329	0.01216
	MICE+	0.00260	0.01505	0.00082	0.00968	0.02253	0.01227
0.3	CCA	0.00513	0.02034	0.00088	0.00656	0.04433	0.02624
	RVI	0.00248	0.01206	0.00075	0.00882	0.02192	0.01188
	K3-	0.00271	0.02138	0.00106	0.02426	0.02405	0.01230
	K3+	0.00274	0.01692	0.00069	0.00492	0.02297	0.01188
	K5-	0.00271	0.02089	0.00100	0.01287	0.02365	0.01208
	K5+	0.00272	0.01770	0.00068	0.00644	0.02263	0.01192
	K9-	0.00266	0.01994	0.00095	0.00714	0.02307	0.01196
	K9+	0.00270	0.01907	0.00068	0.00766	0.02229	0.01182
	MF-	0.00276	0.02290	0.00105	0.01421	0.02429	0.01255
	MF+	0.00280	0.02744	0.00069	0.00996	0.02354	0.01322
	MICE-	0.00268	0.01981	0.00117	0.07441	0.02395	0.01186
	MICE+	0.00263	0.01587	0.00088	0.01232	0.02351	0.01222
0.4	CCA	0.00709	0.02839	0.00105	0.00868	0.06278	0.03816
	RVI	0.00246	0.02453	0.00075	0.01048	0.02202	0.01201
	K3-	0.00277	0.02832	0.00113	0.03596	0.02505	0.01253
	K3+	0.00298	0.02098	0.00070	0.00920	0.02409	0.01215
	K5-	0.00278	0.02841	0.00108	0.01698	0.02440	0.01237
	K5+	0.00290	0.02325	0.00069	0.01381	0.02362	0.01213
	K9-	0.00278	0.02861	0.00103	0.00817	0.02386	0.01215
	K9+	0.00285	0.02859	0.00071	0.01688	0.02335	0.01211
	MF-	0.00288	0.03042	0.00115	0.02334	0.02587	0.01304
	MF+	0.00314	0.04254	0.00070	0.02342	0.02661	0.01493
	MICE-	0.00268	0.02458	0.00122	0.10377	0.02450	0.01212
	MICE+	0.00272	0.01869	0.00094	0.01543	0.02503	0.01303

Ref.: Bj=β_j; CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales. En escala de rosas: mayores errores cuadráticos medios, por parámetro. En escala de verdes: menores errores cuadráticos medios, por parámetro.

Cuadro 4.12: Error cuadrático medio de los parámetros del modelo de regresión de Cox para n=1000, según método de imputación y proporción de datos faltantes (p).

p	Método	B1	B2	B3	B4	B5	B6
0.1	CCA	0.00164	0.00745	0.00060	0.00375	0.01363	0.00770
	RVI	0.00139	0.00671	0.00063	0.00497	0.01150	0.00640
	K3-	0.00147	0.00888	0.00072	0.00742	0.01215	0.00655
	K3+	0.00140	0.00700	0.00059	0.00220	0.01173	0.00639
	K5-	0.00144	0.00833	0.00069	0.00513	0.01190	0.00649
	K5+	0.00139	0.00675	0.00058	0.00197	0.01165	0.00638
	K9-	0.00142	0.00782	0.00067	0.00404	0.01174	0.00643
	K9+	0.00139	0.00647	0.00058	0.00186	0.01157	0.00637
	MF-	0.00145	0.00864	0.00071	0.00508	0.01206	0.00652
	MF+	0.00139	0.00714	0.00058	0.00176	0.01154	0.00644
	MICE-	0.00149	0.00932	0.00082	0.02270	0.01251	0.00657
	MICE+	0.00141	0.00747	0.00065	0.00561	0.01171	0.00645
0.2	CCA	0.00196	0.00863	0.00056	0.00397	0.01679	0.00969
	RVI	0.00136	0.00670	0.00063	0.00661	0.01150	0.00649
	K3-	0.00150	0.01189	0.00082	0.01424	0.01280	0.00691
	K3+	0.00140	0.00767	0.00054	0.00170	0.01202	0.00664
	K5-	0.00146	0.01099	0.00078	0.00796	0.01248	0.00677
	K5+	0.00139	0.00748	0.00053	0.00173	0.01189	0.00660
	K9-	0.00144	0.00986	0.00074	0.00494	0.01219	0.00668
	K9+	0.00139	0.00736	0.00052	0.00188	0.01174	0.00658
	MF-	0.00148	0.01183	0.00080	0.00779	0.01273	0.00695
	MF+	0.00139	0.00997	0.00053	0.00240	0.01171	0.00691
	MICE-	0.00152	0.01260	0.00096	0.04871	0.01316	0.00685
	MICE+	0.00140	0.00828	0.00068	0.00804	0.01209	0.00677
0.3	CCA	0.00257	0.01055	0.00057	0.00435	0.02152	0.01274
	RVI	0.00139	0.00682	0.00063	0.00810	0.01163	0.00624
	K3-	0.00159	0.01726	0.00094	0.02355	0.01387	0.00684
	K3+	0.00150	0.00888	0.00052	0.00263	0.01313	0.00657
	K5-	0.00155	0.01611	0.00088	0.01158	0.01333	0.00665
	K5+	0.00149	0.00888	0.00050	0.00392	0.01280	0.00646
	K9-	0.00153	0.01509	0.00083	0.00592	0.01304	0.00652
	K9+	0.00147	0.00973	0.00049	0.00501	0.01256	0.00643
	MF-	0.00160	0.01751	0.00092	0.01226	0.01383	0.00700
	MF+	0.00151	0.01647	0.00050	0.00750	0.01268	0.00707
	MICE-	0.00160	0.01722	0.00107	0.07721	0.01412	0.00681
	MICE+	0.00150	0.00979	0.00073	0.01045	0.01303	0.00672
0.4	CCA	0.00337	0.01328	0.00063	0.00510	0.02929	0.01706
	RVI	0.00140	0.01964	0.00066	0.00989	0.01186	0.00668
	K3-	0.00169	0.02386	0.00108	0.03614	0.01440	0.00732
	K3+	0.00166	0.01073	0.00055	0.00630	0.01389	0.00701
	K5-	0.00168	0.02371	0.00102	0.01630	0.01401	0.00714
	K5+	0.00162	0.01144	0.00052	0.01040	0.01344	0.00693
	K9-	0.00165	0.02345	0.00096	0.00686	0.01358	0.00703
	K9+	0.00160	0.01485	0.00053	0.01321	0.01320	0.00690
	MF-	0.00173	0.02525	0.00108	0.02083	0.01506	0.00767
	MF+	0.00170	0.02874	0.00052	0.02063	0.01407	0.00823
	MICE-	0.00167	0.02229	0.00118	0.10627	0.01450	0.00719
	MICE+	0.00158	0.01141	0.00083	0.01269	0.01400	0.00742

Ref.: Bj=β_j; CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales. En escala de rosas: mayores errores cuadráticos medios, por parámetro. En escala de verdes: menores errores cuadráticos medios, por parámetro.

Sesgo

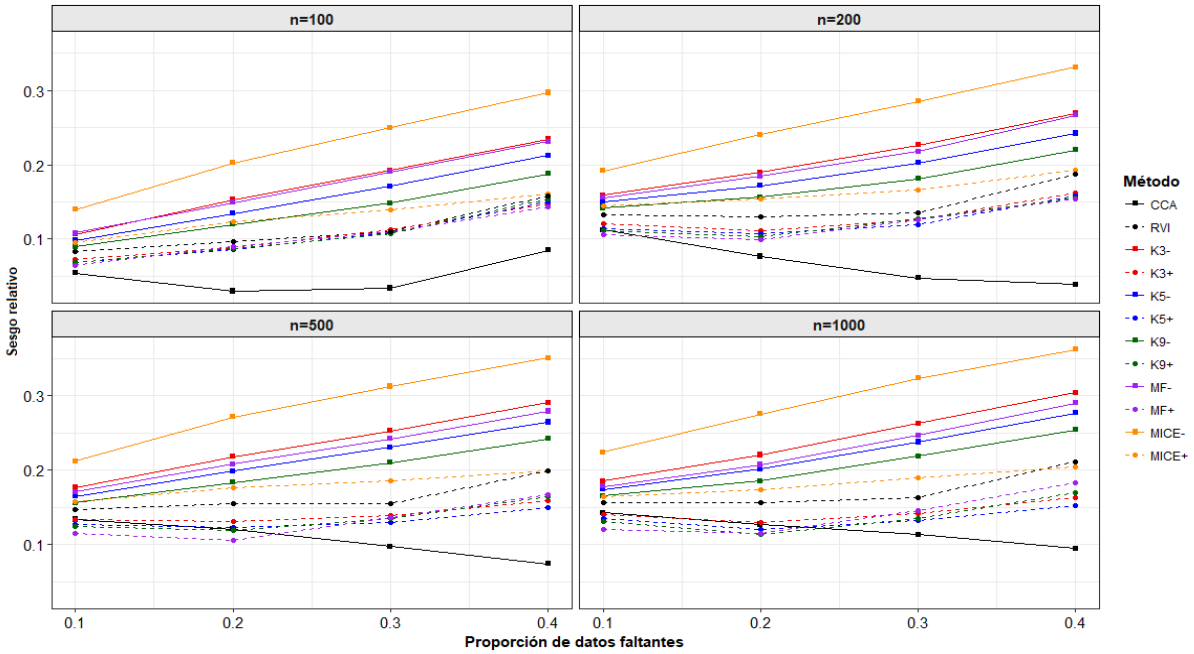
En la Figura 4.19 se muestran los resultados obtenidos al calcular el indicador del sesgo de estimación global para los parámetros del modelo 4.1.

Para todos los tamaños de muestra, los perfiles de sesgo se agrupan según sean o no utilizadas las variables informativas *tiempo* y *estado*, mostrándose con menor sesgo el grupo de métodos que las incluyen. MICE- evidencia los resultados más desfavorables en todos los escenarios.

Para KNN, los resultados mejoran al aumentar el número de vecinos, k. MF- presenta valores del indicador intermedios a K3- y K5-. Entre las técnicas que incluyen a las variables informativas adicionales, MICE presenta los valores más grandes para el indicador. MF y KNN presentan valores muy similares en todos los escenarios. RVI muestra valores intermedios.

Los valores más pequeños del indicador se encuentran bajo el uso de CCA en la mayoría de los escenarios, siendo mejorados por MF+ cuando $n = 200$ y $p = 0.10$, y por MF+ y KNN+ cuando $n = 500$ y $p = 0.20$, y $n = 1000$ y $p = 0.10$ y 0.20.

Figura 4.19: Diferencia relativa media, en valor absoluto, entre el promedio de las estimaciones de los parámetros y sus valores reales, según método de imputación, tamaño de muestra y proporción de datos faltantes.



Ref.: CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

Variabilidad de los estimadores respecto de la variabilidad esperada

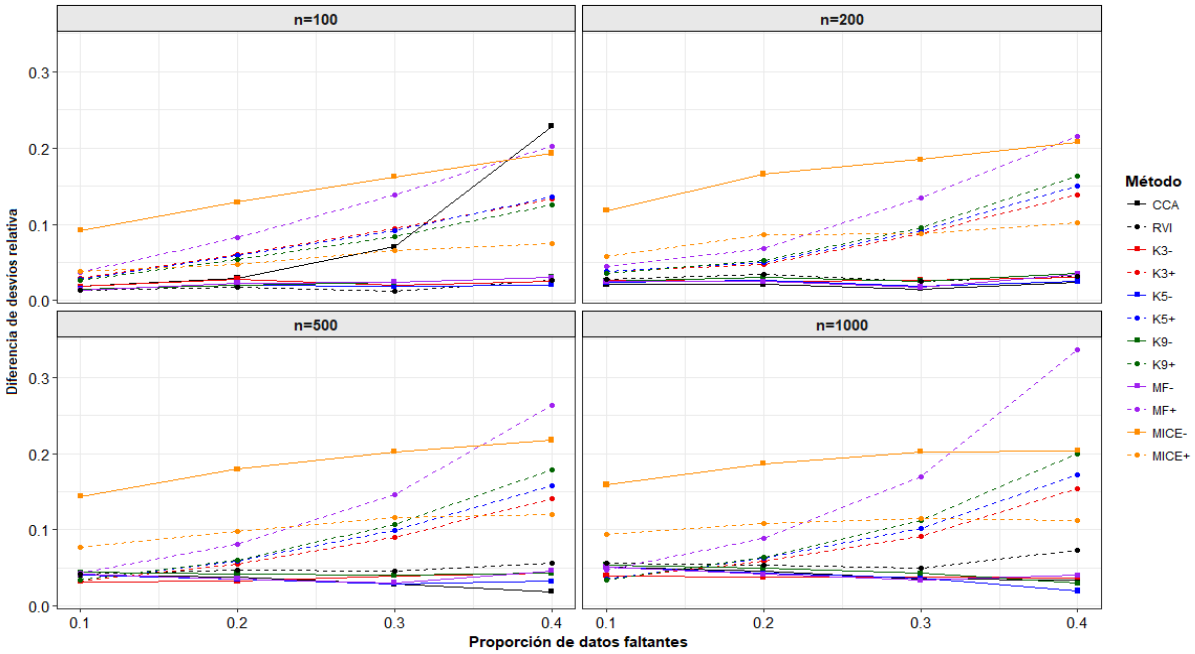
En la Figura 4.20 se muestran los resultados obtenidos al calcular el indicador global de la diferencia relativa en valor absoluto entre el desvío estándar empírico y teórico de los estimadores los parámetros del modelo 4.1.

MICE- presenta los mayores valores para el indicador cuando $p \leq 0.30$ mientras que, para $p = 0.40$, MF+ muestra valores superiores para todos los n, y CCA también lo hace cuando $n = 100$.

Para todas las combinaciones de n y p, la no inclusión del *tiempo* y el *estado* como variables informativas proveen indicadores menores, cercanos a cero y similares para todas las técnicas, incluyendo CCA y RVI y a excepción de MICE. Estos valores varían muy levemente a medida que aumenta p, para cada tamaño de muestra. Cuando se incluyen las variables informativas adicionales, el incremento en el indicador para cada técnica es notorio.

MICE es la única técnica bajo la cual se obtienen indicadores menores cuando se incluyen las variables informativas adicionales .

Figura 4.20: Diferencia relativa media, en valor absoluto, entre el promedio de los desvíos estándar teóricos de los estimadores y el desvío estándar empírico, según método de imputación, tamaño de muestra y proporción de datos faltantes.



Ref.: CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

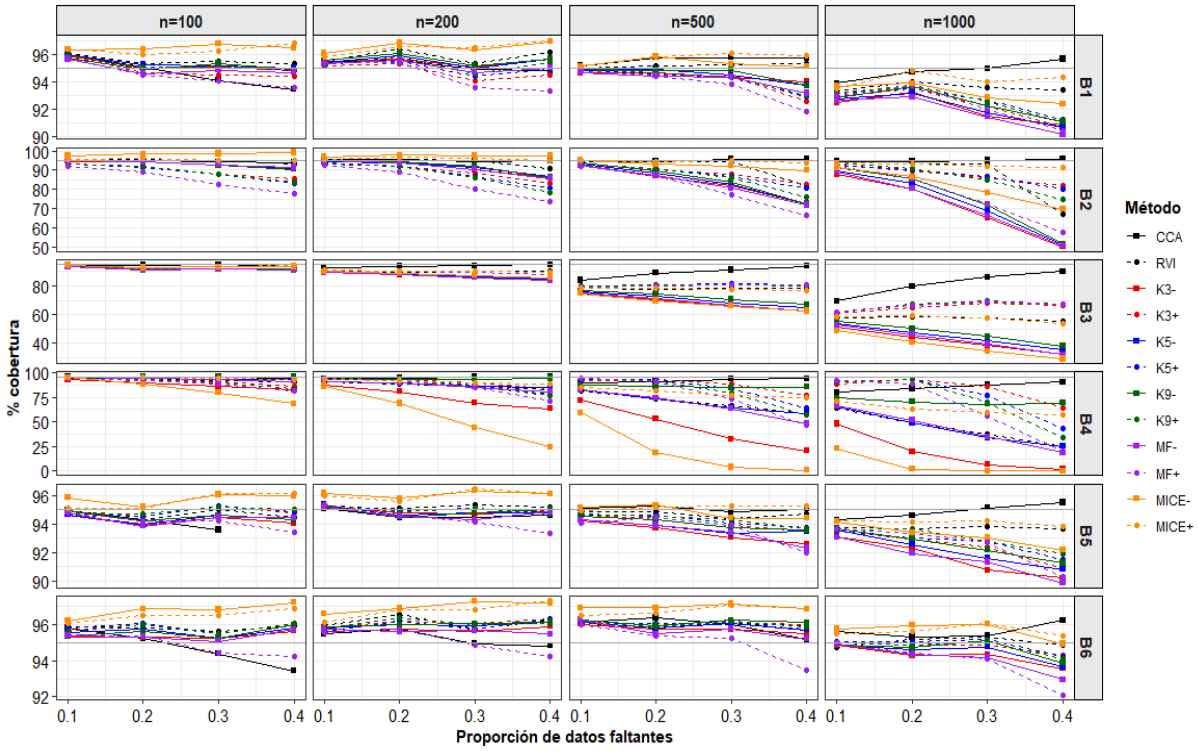
Cobertura

En la Figura 4.21 se presentan los porcentajes de intervalos de confianza que cubren al verdadero valor, para cada parámetro del modelo y según el tamaño muestral y el porcentaje de pérdida. En general, para todos los parámetros y valores de n, la diferencia observada entre los métodos de imputación aumenta a medida que aumenta la proporción de pérdidas.

Para los parámetros $\beta_1, \beta_2, \beta_5$ y β_6, correspondientes a la variable con distribución Normal y a las cualitativas, para $n = 100$ y 200 MICE presenta la mayor cobertura. Para $n = 500$ y 1000, también CCA se ubica como la técnica con resultados más favorables. Para estos coeficientes, MF+ muestra la cobertura más baja cuando $p = 0.30$ y 0.40.

Para los parámetros β_3 y β_4, correspondientes a las variables continuas con distribución no Normal, MICE- y K3- se corresponden con los menores porcentajes de cobertura. CCA muestra un buen desempeño, seguido por todas las técnicas que no incluyen a las variables informativas adicionales, con resultados similares entre sí.

Figura 4.21: Porcentaje de cobertura para los intervalos de confianza del 95 %, según método de imputación, tamaño de muestra y proporción de datos faltantes.



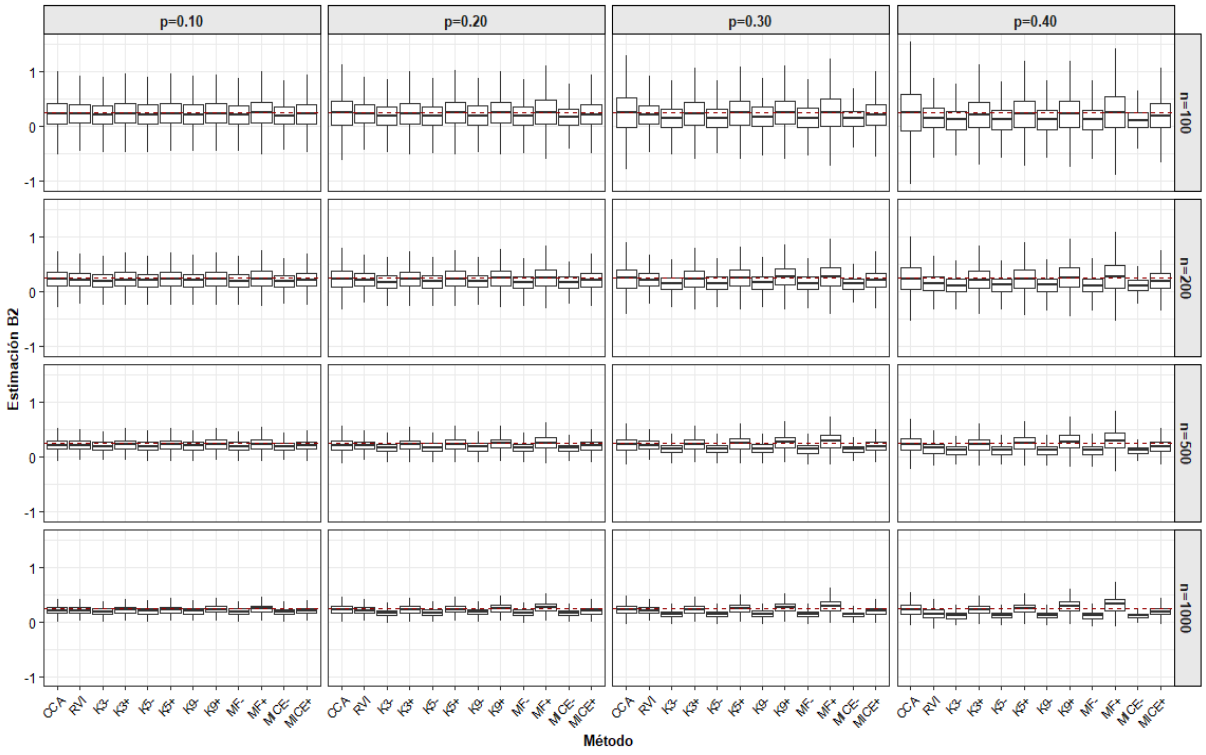
Ref.: Bj: β_j; CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

Distribuciones en el muestreo

En la Figura 4.22 se muestra la distribución de las estimaciones obtenidas en las 5000 iteraciones del proceso de simulación para el parámetro β_2 del modelo 4.1.

Se observa que, dentro de cada método de imputación, existe menor variabilidad en las estimaciones cuando no se incluyen el *tiempo* y el *estado* como variables informativas, aunque su mediana presenta un alejamiento respecto al verdadero valor del parámetro. Además, se evidencia que, a mayor tamaño de muestra, menor en la variabilidad en las estimaciones.

Figura 4.22: Distribución de los estimadores de β_2, según método de imputación, tamaño de muestra y proporción de datos faltantes, sin incluir valores extremos(1).



(1): La presencia de valores extremos no está asociada con ningún efecto medido. Ref.: CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

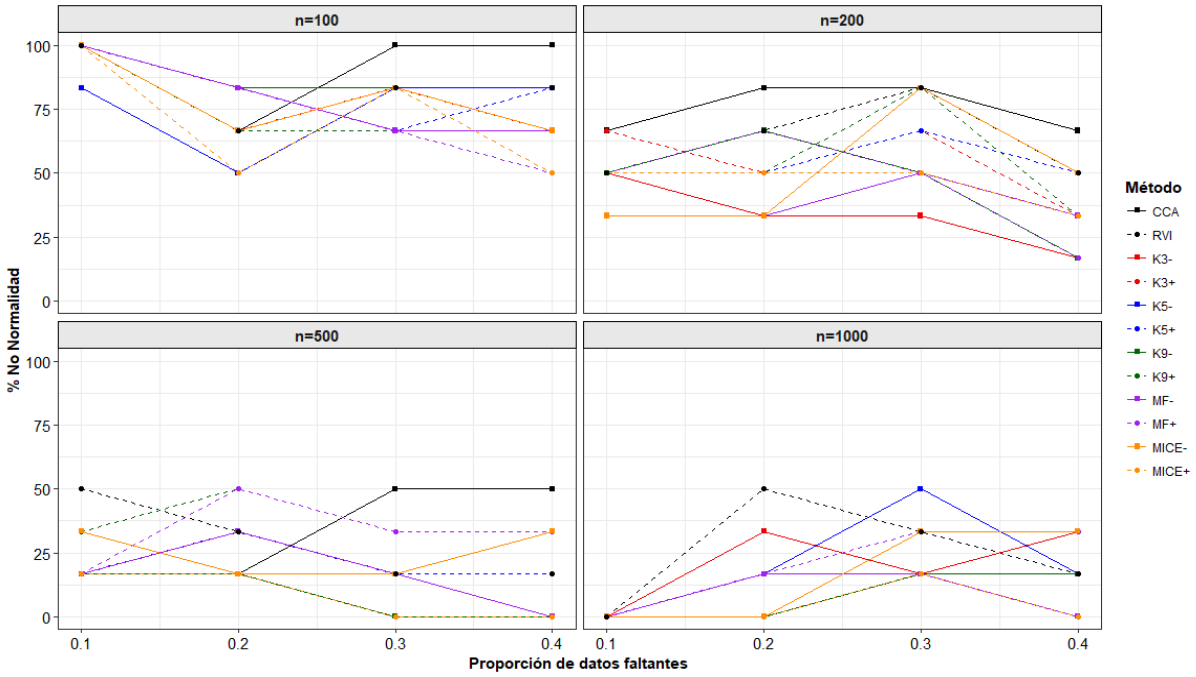
En la Figura 4.23 se muestra el porcentaje de parámetros, entre los seis correspondientes al modelo 4.1, en los que se rechaza la hipótesis de distribución Normal de acuerdo al test de Anderson-Darling aplicado sobre las 5000 estimaciones para cada uno de ellos.

Si bien no se aprecia una tendencia al comparar los métodos en los distintos escenarios, se distingue que, a medida que aumenta el tamaño de muestra, disminuye el

porcentaje en estudio, es decir, aumenta el porcentaje de parámetros para los cuales se acepta que su estimador presenta distribución Normal.

Dado que el test de Anderson-Darling se aplica sobre un gran número de datos (5000 estimaciones para cada método y escenario), el mismo se vuelve muy sensible, provocando un rechazo de la hipótesis de normalidad ante pequeños alejamientos de dicha distribución. Sin embargo, no se observan características en las distribuciones que sugieran que el supuesto de normalidad no es aceptable.

Figura 4.23: Porcentaje de parámetros para los cuales se rechaza la hipótesis del test de Anderson-Darling, según método de imputación, tamaño de la muestra y proporción de datos faltantes(1).



(1): La presencia de valores extremos no está asociada con ningún efecto medido. Ref.: CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

Estimación de la probabilidad de sobrevida

En la Figura 4.24 se muestra el promedio de la probabilidad de supervivencia estimada para $t > 100$, a partir de las 5000 estimaciones de los parámetros del modelo 4.1, para una unidad con vector de variables explicativas $\boldsymbol{x} = (0, 1, 5, 1, 0, 1)$. Los resultados son muy similares para los cuatro valores de n considerados.

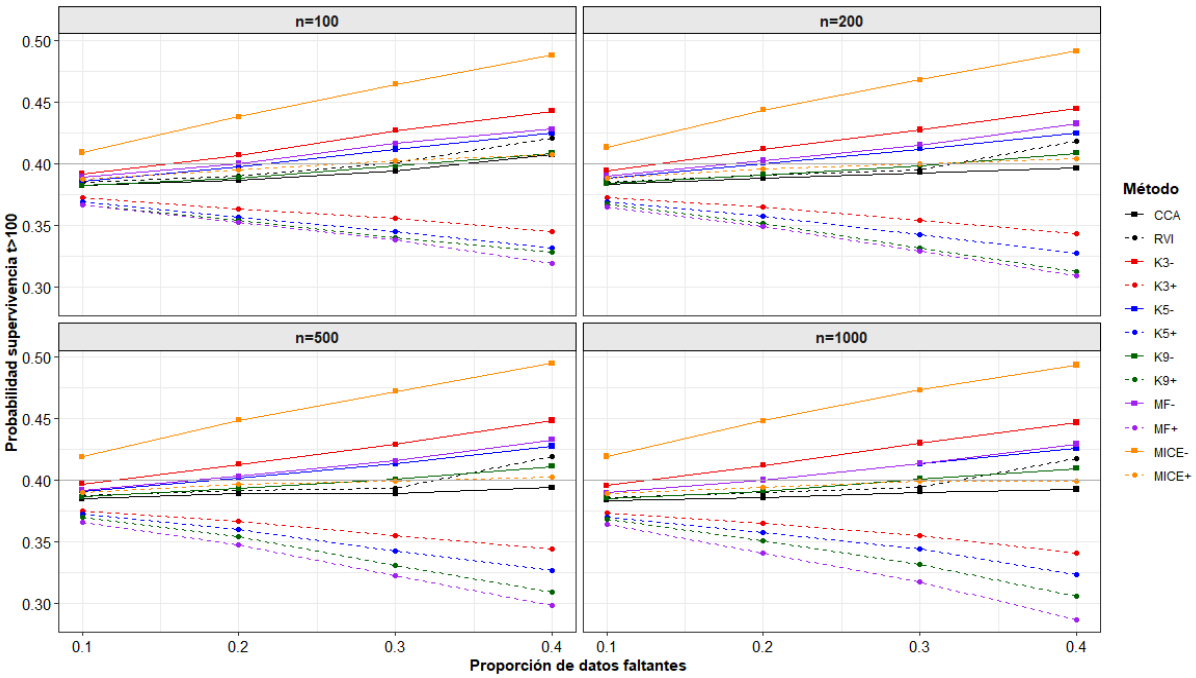
Para todos los tamaños de muestra considerados, la diferencia entre los resultados

de las diferentes técnicas se amplía a medida que aumenta p.

En general, cuando se incluyen las variables informativas adicionales, las técnicas de imputación generan estimaciones de la probabilidad de sobrevida inferiores a la real, mientras que, cuando no se incluyen dichas variables, la probabilidad resulta sobreestimada. En el primer caso, MF+ es la técnica cuyos valores se encuentran más alejados de la probabilidad real, mientras que en el segundo caso, es MICE- quien tiene los resultados menos favorables.

Para todos los n establecidos resulta que, cuando $p = 0.10$, los mejores resultados se encuentran bajo K3-. Cuando $p = 0.20$, se encuentran con MF- y K5-. Cuando, $p = 0.30$, MICE+ y K9- presentan estimaciones cercanas a la real y, cuando $p = 0.40$, MICE+ muestra los mejores resultados en este aspecto.

Figura 4.24: Promedio de la probabilidad de supervivencia estimada para $t > 100$, según método de imputación, tamaño de la muestra y proporción de datos faltantes.



Ref.: CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

4.2.3.b. Reproducibilidad de los datos perdidos

En este apartado, se realiza una evaluación de la eficiencia de los métodos de imputación para asignar valores a los datos faltantes, según se trate de observaciones corres-

pondientes a la variable cuantitativa X_4 o a la variable cualitativa X_2, cuando se han generado datos perdidos de acuerdo a un mecanismo MNAR.

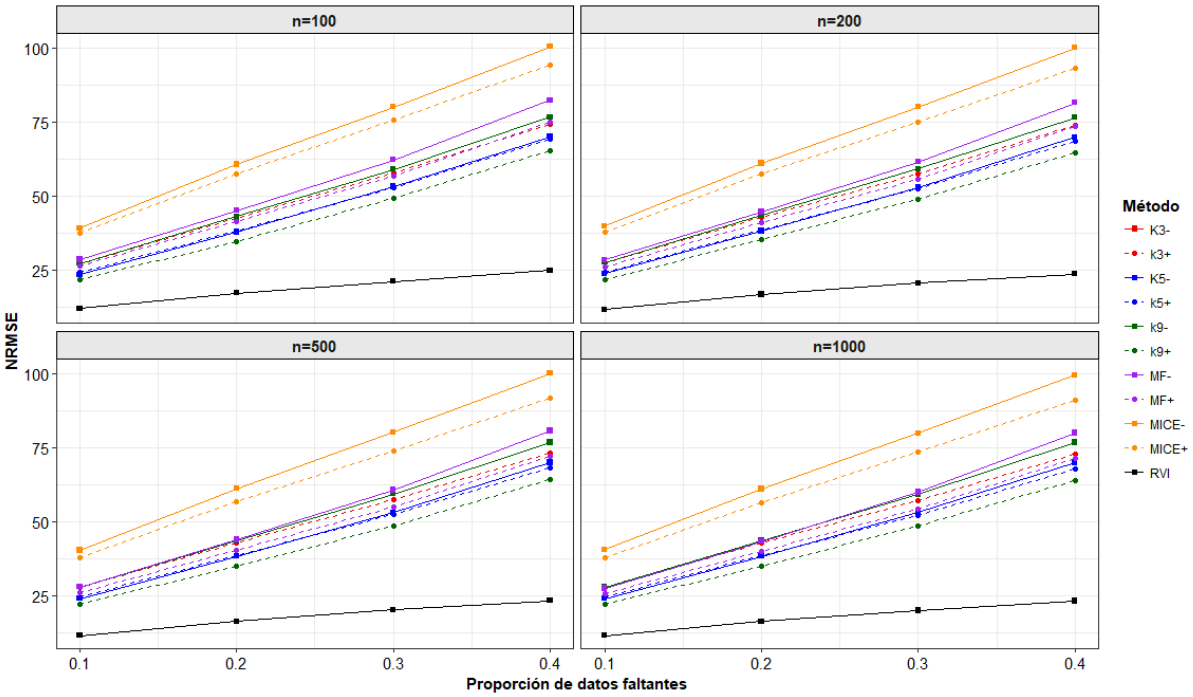
Raíz del error cuadrático medio normalizado

En relación a la imputación de datos para X_4, de acuerdo con la Figura 4.25, se encuentran los mayores valores para $\overline{\text{NRMSE}}$ al utilizar MICE y los menores valores corresponden al uso de RVI, seguidos por K9+.

En general, los valores de $\overline{\text{NRMSE}}$ para cada técnica resultan menores cuando se consideran el *tiempo* y el *estado* como variables informativas.

El comportamiento de $\overline{\text{NRMSE}}$ es muy similar para todos los tamaños de muestra considerados. En todos los casos, $\overline{\text{NRMSE}}$ aumenta a medida que aumenta la proporción de datos faltantes.

Figura 4.25: Error cuadrático medio normalizado (NRMSE) para X_4, según método de imputación, tamaño de la muestra y proporción de datos faltantes.



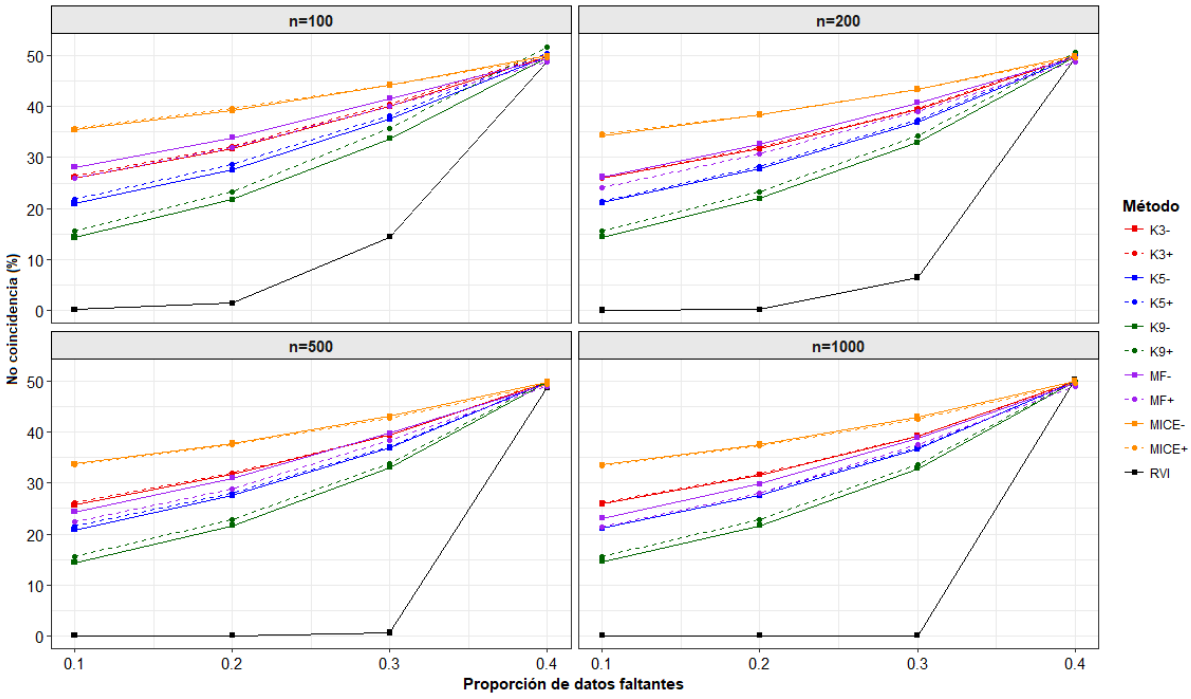
Ref.: CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

Porcentaje de no coincidencia

En la Figura 4.26 se muestran los porcentajes medios de datos imputados en forma incorrecta para la variable categórica X_2. Los valores más pequeños se encuentran bajo RVI, mientras que los más desfavorables corresponden a MICE.

Haciendo uso de KNN, se observan resultados más favorables a medida que aumenta el número de donantes. No se distinguen diferencias en los resultados al considerar el *tiempo* y el *estado* como variables informativas. El comportamiento de los porcentajes medios de datos imputados en forma incorrecta es muy similar para todos los tamaños de muestra y se incrementa a medida que aumenta el porcentaje de datos faltantes.

Figura 4.26: Porcentaje de datos imputados no coincidentes con los reales para X_2, según método de imputación, tamaño de la muestra y proporción de datos faltantes.



Ref.: CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

4.2.3.c. Tiempos computacionales de imputación

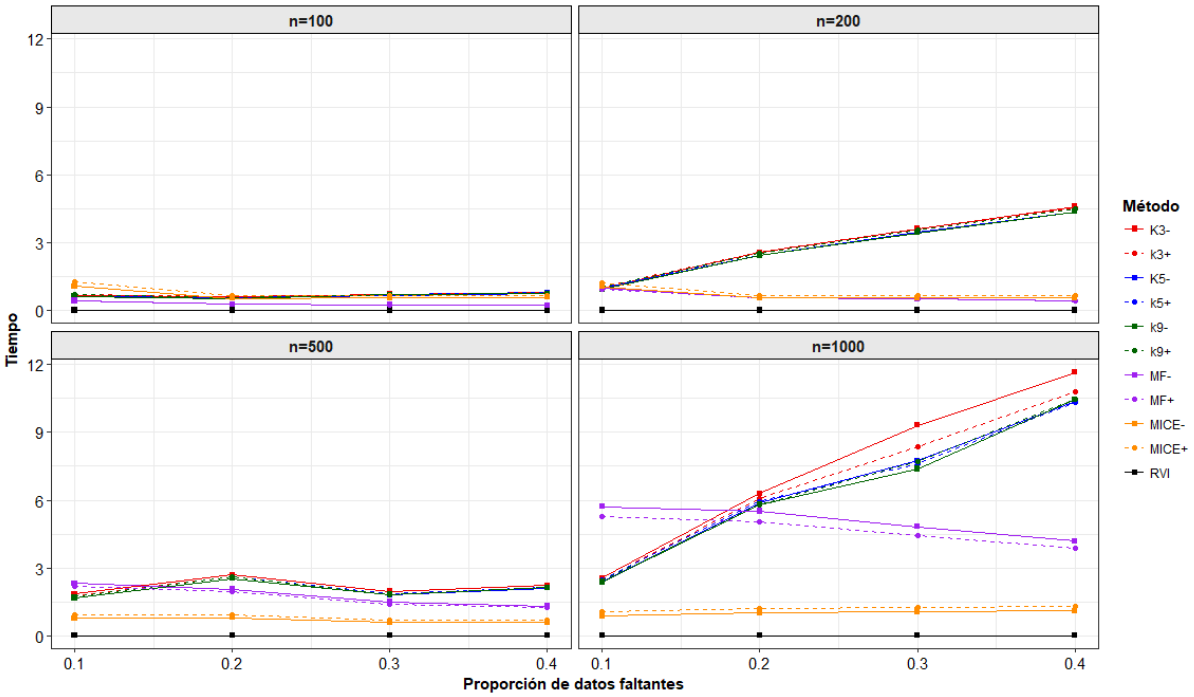
En relación a los tiempos que demora el software utilizado para realizar las imputaciones, de acuerdo a los datos presentados en la figura 4.27, se observa que RVI demanda muy poco tiempo para realizar la imputación completa del conjunto de datos.

Cuando $n = 100$, MF+ presenta tiempos de imputación levemente superiores a RVI, siendo superados por las otras técnicas, que muestras tiempos muy similares entre sí. Los tiempos requeridos por todos los métodos no presentan variaciones para los distintos valores de p.

Para $n = 200$, MICE y MF muestran tiempos computacionales muy parecidos entre sí, estables para los valores de p. KNN no presenta diferencias en sus distintas configuraciones. Cuando $p = 0.10$ requiere tiempos similares al resto de los métodos, que aumentan notoriamente a medida que la proporción de datos faltantes crece.

En el caso de $n = 500$ y 1000, MICE es el método que menor tiempo demanda, luego de RVI, y es seguido por MF, salvo cuando $p = 0.10$ donde resultan menores los tiempos requeridos por KNN. Para este último método, los tiempos se incrementan cuando p aumenta.

Figura 4.27: Tiempo promedio, en segundos, demandado para la imputación completa de los datos, según método de imputación, tamaño de la muestra y proporción de datos faltantes.



Ref.: CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

5. Estudio por remuestreo a partir de un caso real

A fin de complementar al estudio por simulación presentado en el capítulo anterior y que permitió obtener resultados originales sobre la eficiencia de distintos métodos de imputación en la estimación de parámetros en análisis de supervivencia frente a datos perdidos, se presenta un estudio por remuestreo a partir de los datos de la investigación psiquiátrica referida en el capítulo introductorio, que permitirá evaluar la estabilidad de los hallazgos en el contexto de los datos reales de interés.

El capítulo se organiza en cuatro secciones, comenzando por la descripción del contexto de la investigación, variables recolectadas y características principales de la misma. A continuación, se resumen los aspectos de diseño, el cual mantiene los escenarios y algunas medidas de eficiencia utilizadas en el capítulo anterior. La tercer sección detalla los resultados del estudio. Por último, se agrega un análisis de los datos reales surgidos del trabajo que incorpora el uso de los métodos de imputación recomendados a partir de los estudios por simulación en un paso previo a la estimación del modelo de regresión para explicar el tiempo hasta el reintento de suicidio.

5.1. Presentación del caso real. Un estudio en el área de psiquiatría

Los datos que se utilizan en este capítulo fueron obtenidos por investigadores pertenecientes al Instituto de Farmacología de la Facultad de Medicina de la Universidad de Buenos Aires, en el marco de un proyecto de investigación cuyo objetivo principal es evaluar las características personales y clínicas de pacientes con intento de suicidio que influyen sobre el tiempo transcurrido hasta un nuevo intento de suicidio. Los primeros resultados de este estudio fueron publicados en el año 2019 [7, 97].

Entre los años 2012 y 2016 se reclutaron pacientes que ingresaron al área de emergencias de tres hospitales de la ciudad de Buenos Aires, por intento o ideación suicida. Las instituciones involucradas fueron el Hospital Neuropsiquiátrico “Braulio A. Moyano”, el Hospital Interdisciplinario Psicoasistencial “José Tiburcio Borda” y el Hospital de Clínicas “José de San Martín”. Para participar del estudio, los pacientes debían tener entre 18

y 65 años y estar en condiciones de entendimiento, por lo que se excluyeron a quienes no comprendían el idioma español o bien se encontraban bajo efectos sedativos. Todos los participantes debieron firmar un consentimiento informado.

Inicialmente, se indagó a los pacientes acerca de características socio-demográficas basales y se les realizó una batería de cuestionarios validados a fin de definir sus características clínicas y psiquiátricas, así como para detectar patologías mentales que pudieran presentar. Todos los cuestionarios fueron realizados por médicos psiquiatras capacitados para tal fin.

Posteriormente, dos médicos psiquiatras pertenecientes al equipo de investigación fueron los encargados de contactar telefónicamente a los pacientes cada seis meses, durante un periodo de dos años posteriores al intento de suicidio que les dió ingreso al estudio, es decir, a los seis, 12, 18 y 24 meses de tal evento. Si no pudo establecerse contacto después de tres intentos, el participante fue declarado como "pérdida de seguimiento".

La información de interés principal durante el seguimiento fue si el paciente había realizado un reintento de suicidio o bien había tenido un suicidio consumado. Para ambos casos, se registró la fecha de ocurrencia y se estableció la cantidad de días transcurridos desde el ingreso al estudio hasta el evento de interés. Para quienes no presentaron dicho evento, se calculó la cantidad de días transcurridos desde el ingreso al estudio hasta la pérdida de seguimiento o el final de seguimiento a los dos años.

Uno de los estudios publicados por el grupo de investigación se centró en aquellos pacientes en los que se había detectado Trastorno Límite de Personalidad (TLP) [7], según las mediciones del indicador *Structured Clinical Interview for DSM-IV Axis II Disorders* (SCID-II, versión en español) [98]. Se identificaron 161 pacientes con TLP, de los cuales 146 tuvieron al menos un seguimiento. Entre estos, 46 (31.51 %) reintentaron suicidarse y dos (1.37 %) cometieron suicidio, por lo que 98 (67.12 %) resultaron censurados debido a pérdida de seguimiento o final de seguimiento a los dos años. Se ajustó un modelo de regresión de Cox considerando los 114 individuos sin datos faltantes (78.08 %), utilizando un grupo de variables mixtas correspondientes a características socio-demográficas y clínicas. Mediante el *método de eliminación hacia atrás* (backward) se seleccionaron las variables estadísticamente significativas ($p < 0.05$) para estimar el riesgo instantáneo de reintento de suicidio. Ellas fueron *hostilidad, funcionamiento psicosocial, número de intentos de suicidio previos, presencia de antecedentes de abuso sexual infantil, edad al*

momento de la admisión al estudio y *edad al momento del primer intento de suicidio.*

El concepto de *hostilidad* fue determinado mediante la *Escala de Hostilidad de Buss-Durkee* (BDHS, versión en español) [99, 100], instrumento diseñado para evaluar los distintos aspectos de la hostilidad y los sentimientos de culpabilidad asociados. Es un instrumento autoaplicado que consta de 75 ítems agrupados en siete subescalas de hostilidad y una de culpabilidad. Proporciona una puntuación numérica total, que fue utilizada para clasificar a los pacientes en dos niveles de hostilidad, bajo y alto, según la puntuación estuviera por debajo o por encima del valor de la mediana.

El *funcionamiento psicosocial* se define como el conjunto de capacidades y habilidades que presenta una determinada persona para desenvolverse de forma correcta en el ámbito social y personal y fue valorado mediante la *Social Adaptation Self-evaluation Scale* (SASS, versión en español) [101, 102]. Los pacientes se clasificaron con *bajo funcionamiento psicosocial* si su valor estaba por debajo de la puntuación mediana o con *alto funcionamiento psicosocial* en caso contrario.

El *número de intentos de suicidio previos* corresponde a una variable cualitativa ordinal con tres categorías: *sin intentos previos*, *uno o dos intentos previos* o *tres o más intentos previos.*

Los *antecedentes de abuso sexual infantil* fueron recogidos por formularios autoadministrados debido al riesgo de subregistro. En caso de que la respuesta fuese afirmativa, se indagó también a qué edad sufrió el abuso y se consideró como *infantil* si ocurrió cuando el paciente tenía menos de 18 años de edad.

Se muestra un análisis exploratorio de las variables explicativas en el Anexo III.

5.2. Diseño de un estudio comparativo por remuestreo

Se realiza un muestreo aleatorio con reemplazo sobre el conjunto de datos completos ($n = 114$) generando una muestra de tamaño 2000. Sobre esta muestra, se realiza un análisis de regresión de Cox a fin de determinar el efecto de cada covariable sobre el tiempo hasta el reintento de suicidio, estimando los parámetros del siguiente modelo:

$$h_i(t) = \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7}) h_0(t), \qquad (5.1)$$

donde x_{i1}: hostilidad (nivel bajo = 0, nivel alto = 1), x_{i2}: funcionamiento psicosocial (nivel bajo = 0, nivel alto = 1), x_{i3}: tres o más intentos de suicidio anteriores (no = 0, sí = 1), x_{i4}: sin intentos de suicidio previo (no = 0, sí = 1), x_{i5}: historial de abuso sexual infantil (no = 0, sí = 1), x_{i6}: edad en el momento de la admisión al estudio, y x_{i7}: edad de primer intento de suicidio, para el i-ésimo individuo. Las estimaciones puntuales obtenidas son consideradas como los *valores reales* para los parámetros del modelo.

Se consideran escenarios que combinan cuatro tamaños de muestra (n = 100, 200, 500 y 1000), cuatro proporciones de datos faltantes (p = 0.10, 0.20, 0.30 y 0.40) y tres mecanismos de pérdida (MCAR, MAR, MNAR). Cada escenario (combinación de mecanismo de pérdida, proporción de datos faltantes y tamaño de muestra) se replica 100 veces y en cada una se utilizan distintos métodos de tratamiento de datos faltantes: CCA, RVI, KNN, MF y MICE. En el caso del mecanismo MCAR con n = 100 y 200 y para MNAR con n = 100, solo se consideran los porcentajes de pérdida p = 0.10, 0.20 y 0.30, a fin de mantener una proporción suficiente de información conocida.

Para generar datos faltantes según un mecanismo MCAR en los conjuntos de datos completos, se eliminan aleatoriamente $n \times p$ observaciones para cada variable explicativa en cada conjunto de datos completo.

Para simular datos faltantes según un mecanismo MAR, se elimina una proporción p de los datos de *edad del primer intento de suicidio* considerando los pacientes con tres o más intentos de suicidios previos, ya que se supone que aquellos individuos con mayor frecuencia de esta conducta son más propensos a no recordar cuándo la presentaron por primera vez. Del mismo modo se procede con la variable *abuso sexual infantil*, eliminando datos entre los pacientes de mayor edad al momento del ingreso al estudio, dado que ellos podrían ser quienes no recordaran dicho antecendente.

Para simular datos faltantes según un mecanismo MNAR, se elimina una proporción p de las observaciones de *abuso sexual infantil* sobre los casos en los que se presenta dicho antecedente, ya que estos pacientes podrían ser más proclives a no brindar información. También se eliminan registros para la *edad del primer intento de suicidio* cuando la misma presenta valores pequeños, dado que es posible suponer que quienes presentan esa conducta a edades tempranas no recuerdan con precisión dicha información.

Se estiman los parámetros del modelo 5.1 mediante CCA, es decir, considerando solamente los individuos sin pérdidas y se imputan los datos faltantes mediante RVI,

KNN, MF y MICE, considerando las mismas opciones descriptas en el capítulo anterior.

Con cada conjunto de datos completos se estima el modelo 5.1 mediante cada método. En el caso de las variables que se incluyen en el modelo como dicotómicas según su mediana (*hostilidad* y *funcionamiento psicosocial*), se consideran como cuantitativas durante el proceso de imputación y luego son dicotomizadas según la mediana obtenida en cada conjunto de datos imputados.

5.2.1. Medidas de eficiencia

Se define un conjunto de medidas que son evaluadas comparativamente para identificar la *performance* de cada método para tratar la información faltante. Estas medidas son calculadas en cada iteración ($s = 100$) y resumidas a través de promedios, desvíos y porcentajes para cada escenario. Los resultados se muestran en forma tabular o gráfica de modo de facilitar la comparación. En las comparaciones se enfocan, para los estimadores de los coeficientes del modelo, el error cuadrático medio y la cobertura y, para la precisión de las imputaciones, se estudia el error cuadrático medio y la coincidencia de las imputaciones con los valores originales. Estas medidas corresponden a las descriptas en el capítulo anterior en la sección 4.1.1.

5.2.2. Implementación computacional

El proceso de simulación, imputación y análisis de datos se realiza completamente en el lenguaje de programación R, mediante su entorno RStudio [92], incluyendo en cada caso un valor inicial para la generación de valores pseudo-aleatorios (*semilla*) a fin de garantizar la reproducibilidad de los resultados.

Los paquetes de R utilizados para las imputaciones son los ya mencionados en el capítulo anterior, en la sección 4.1.2. También se registran los tiempos demandados por cada método de imputación para arribar al conjunto de datos completos en cada iteración. El equipo computacional es el mismo que el mencionado en la sección 4.1.2.

5.3. Resultados

Los resultados se muestran separadamente según el mecanismo de pérdida. La sección 1, corresponde a los datos perdidos completamente al azar, la sección 2 corresponde a los datos perdidos al azar y la sección 3, a los perdidos no al azar. Dentro de cada sección, se muestran los resultados para las propiedades de los estimadores en forma global, un apartado sobre la reproducibilidad de los datos perdidos y otro correspondiente al estudio de los tiempos computacionales requeridos para la imputación.

Para mencionar los métodos de tratamiento de datos faltantes y sus configuraciones, se utiliza la nomenclatura definida en la sección 4.2.

En los Cuadros que muestran los resultados del error cuadrático medio (MSE) de acuerdo a la cantidad de unidades del conjunto de datos y la proporción de datos faltantes en cada uno de ellos, p, los valores sombreados en tonos rosados corresponden a los MSE más grandes para cada parámetro, siendo estos los resultados más desfavorables, mientras que los MSE sombreados en tonos verdes son aquellos de menor valor, indicando resultados más favorables en la estimación del correspondiente parámetro (Cuadros 5.1 a 5.4, 5.5 a 5.8 y 5.9 a 5.12).

Los resultados en relación al tiempo computacional demandado por cada método para realizar las imputaciones no se presenta debido a resultar muy similares a lo observado en el Capítulo 4, secciones *4.2.1.c*, *4.2.2.c* y *4.2.3.c* para los mecanismos MCAR, MAR y MNAR, respectivamente.

5.3.1. Datos perdidos completamente al azar

5.3.1.a. Propiedades distribucionales de los estimadores de los coeficientes del modelo de regresión de Cox

Error cuadrático medio

Los resultados obtenidos al calcular el MSE para cada uno de los siete parámetros del modelo 5.1 se muestran en los Cuadros 5.1 a 5.4, de acuerdo a la cantidad de unidades del conjunto de datos y la proporción de datos faltantes en cada uno de ellos, p.

Cuando $n = 100$, para todos los valores de p, se observa que utilizando CCA se

obtienen los resultados más desfavorables. Para $p = 0.10$, las mejores estimaciones se obtienen utilizando imputación mediante el método MF, con leves diferencias según se tenga en cuenta o no el *estado* y el *tiempo hasta el evento o censura* como variables informativas. También se observan MSE dentro de los más bajos al utilizar MICE+ para realizar las imputaciones. En el caso de $p = 0.20$, MICE es la técnica de imputación que muestra mejores resultados, estando los MSE correspondientes a MICE- entre los menores para todos los parámetros del modelo. El uso de MF+ provee MSE grandes en relación a la estimación de cinco de los siete parámetros. Cuando $p = 0.30$, los mejores resultados se alcanzan con MICE+, encontrándose también MSE bajos con MF-, MF+ y MICE- (Cuadro 5.1).

En el caso de $n = 200$, CCA presenta los mayores valores de MSE para todos los p propuestos. Se observan buenos resultados al utilizar MF para la imputación de datos, al igual que con MICE+. En general, salvo para el caso de MF, se reconocen MSE menores cuando se consideran el *estado* y el *tiempo hasta el evento o censura* como variables informativas (Cuadro 5.2).

Para $n = 500$ y $p = 0.10$, los resultados más desfavorables se encuentran bajo CCA, RVI y KNN considerando 9 observaciones donantes. Los mejores resultados se obtienen con MF+, seguido por MF-. En el caso de $p = 0.20$, además del uso de CCA y RVI, el empleo de MICE- provee las estimaciones con mayores MSE, mientras que MF+ y MF- presentan MSE ubicados entre los más pequeños para todos los parámetros, acompañados por MICE+. Estos resultados son similares a lo detectado cuando la proporción de datos faltantes se establece en $p = 0.30$ o $p = 0.40$. Además, en general resulta que para una misma técnica, los MSE obtenidos son menores cuando se consideran el *estado* y el *tiempo hasta el evento o censura* como variables informativas (Cuadro 5.3).

Finalmente, para el caso de $n = 1000$, se encuentra que CCA, RVI y MICE- presentan los mayores MSE, mientras que MF se destaca como la técnica que provee menores MSE. En los casos de $p = 0.10$ y $p = 0.20$, se observa que los MSE correspondientes al uso de K3+ se encuentran entre los menores, indicando resultados favorables de la estimación de los parámetros cuando los datos son imputados bajo este método. Para $p = 0.30$ y $p = 0.40$, MICE+ provee MSE pequeños que la ubican como una de las técnicas con resultados más favorables. En este caso, también resulta que para una misma técnica, los MSE obtenidos son menores cuando se consideran el *estado* y el *tiempo hasta el evento o*

censura como variables informativas (Cuadro 5.4).

Cuadro 5.1: Error cuadrático medio de los parámetros del modelo de regresión de Cox para n=100, según método de imputación y proporción de datos faltantes (p).

p	Método	B1	B2	B3	B4	B5	B6	B7
0.1	CCA	0.34731	0.32806	36.11298	11.97721	0.29563	0.00215	0.00231
	RVI	0.04932	0.06528	2.98815	1.01919	0.09810	0.00115	0.00169
	K3-	0.06430	0.06776	0.17649	0.03540	0.12704	0.00136	0.00186
	K3+	0.06456	0.04907	2.87701	1.01081	0.11635	0.00131	0.00184
	K5-	0.08115	0.06359	2.91821	1.01523	0.11595	0.00103	0.00157
	K5+	0.06695	0.06209	2.95037	1.01650	0.14428	0.00121	0.00173
	K9-	0.07537	0.07461	2.89375	1.04237	0.12477	0.00104	0.00155
	K9+	0.05904	0.05189	2.87058	1.04835	0.14308	0.00108	0.00152
	MF-	0.03460	0.04551	0.10773	0.03054	0.03926	0.00037	0.00049
	MF+	0.03541	0.04650	0.10904	0.02589	0.04522	0.00036	0.00047
	MICE-	0.03692	0.05811	1.20606	0.41318	0.07708	0.00079	0.00112
	MICE+	0.04645	0.04595	0.38907	0.13263	0.06498	0.00041	0.00055
0.2	CCA	6314.11987	11291.08672	2590.65655	3071.74228	6809.30933	36.74150	44.60573
	RVI	0.18833	0.13924	4.66345	1.59470	0.10046	0.00042	0.00039
	K3-	0.24982	0.14633	4.78401	1.70673	0.10943	0.00047	0.00043
	K3+	0.18458	0.15964	3.21731	1.12820	0.12285	0.00041	0.00045
	K5-	0.27152	0.14575	4.71724	1.76961	0.10166	0.00047	0.00047
	K5+	0.16819	0.15472	3.17340	1.13163	0.14576	0.00046	0.00042
	K9-	0.20561	0.13396	4.74744	1.71456	0.11772	0.00037	0.00040
	K9+	0.19217	0.16352	4.71628	1.69826	0.10983	0.00041	0.00040
	MF-	0.11633	0.10077	4.66975	1.64399	0.09553	0.00043	0.00039
	MF+	0.16062	0.24296	4.93184	1.64759	0.16884	0.00066	0.00047
	MICE-	0.11479	0.05433	1.24876	0.47044	0.09867	0.00023	0.00024
	MICE+	0.11862	0.10974	2.59157	0.84338	0.12705	0.00052	0.00047
0.3	CCA	28794.04547	17129.41851	34211.62322	8538.28621	45053.57563	267.83572	293.18686
	RVI	0.20405	0.13378	13.02663	4.46130	0.24389	0.00040	0.00122
	K3-	0.24855	0.19924	8.90331	3.03703	0.31436	0.00069	0.00116
	K3+	0.14636	0.10929	8.83061	3.10653	0.22153	0.00070	0.00152
	K5-	0.24977	0.19941	11.67009	4.00270	0.28447	0.00068	0.00102
	K5+	0.13722	0.12849	11.46026	3.99028	0.23829	0.00060	0.00134
	K9-	0.23552	0.21399	13.06941	4.60328	0.32278	0.00045	0.00087
	K9+	0.17675	0.13194	13.16226	4.60728	0.23614	0.00046	0.00100
	MF-	0.20774	0.18531	1.71705	0.61657	0.18551	0.00049	0.00063
	MF+	0.15393	0.16007	6.48171	2.09846	0.17090	0.00056	0.00073
	MICE-	0.21998	0.15607	0.36313	0.11845	0.28538	0.00053	0.00096
	MICE+	0.09042	0.09343	4.06437	1.34283	0.12301	0.00042	0.00066

Ref.: Bj=β_j; CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales. En escala de rosas: mayores errores cuadráticos medios, por parámetro. En escala de verdes: menores errores cuadráticos medios, por parámetro.

Cuadro 5.2: Error cuadrático medio de los parámetros del modelo de regresión de Cox para n=200, según método de imputación y proporción de datos faltantes (p).

p	Método	B1	B2	B3	B4	B5	B6	B7
0.1	CCA	0.13575	0.09484	7.98290	2.64732	0.11642	0.00032	0.00032
	RVI	0.03166	0.02712	0.07864	0.01166	0.02993	0.00027	0.00043
	K3-	0.02773	0.02422	0.07716	0.01478	0.03202	0.00031	0.00045
	K3+	0.02706	0.01819	0.06178	0.01690	0.02378	0.00020	0.00030
	K5-	0.03079	0.02606	0.08035	0.01321	0.04414	0.00034	0.00052
	K5+	0.03129	0.02032	0.06211	0.01809	0.03406	0.00025	0.00040
	K9-	0.03319	0.02975	0.07207	0.01462	0.04680	0.00022	0.00042
	K9+	0.03323	0.01878	0.05854	0.01742	0.04158	0.00021	0.00034
	MF-	0.01587	0.01287	0.03496	0.01126	0.00615	0.00003	0.00004
	MF+	0.01608	0.01678	0.02742	0.01229	0.00799	0.00004	0.00005
	MICE-	0.01320	0.03130	0.08010	0.01210	0.02593	0.00022	0.00028
	MICE+	0.02643	0.02088	0.03747	0.01439	0.02010	0.00007	0.00007
0.2	CCA	0.45801	0.50393	85.00816	27.65681	0.72045	0.00184	0.00174
	RVI	0.10792	0.13255	1.44428	0.50350	0.12084	0.00057	0.00074
	K3-	0.12760	0.11336	0.22898	0.04111	0.11149	0.00055	0.00069
	K3+	0.07359	0.08069	1.51285	0.50919	0.07614	0.00049	0.00059
	K5-	0.14267	0.13214	1.50423	0.52053	0.12716	0.00058	0.00073
	K5+	0.08179	0.06818	1.44799	0.52832	0.07159	0.00044	0.00056
	K9-	0.17722	0.12889	1.44752	0.53264	0.12864	0.00046	0.00060
	K9+	0.06815	0.06814	1.41820	0.51679	0.07457	0.00045	0.00048
	MF-	0.06771	0.07407	0.18020	0.04154	0.04202	0.00025	0.00028
	MF+	0.04600	0.04743	1.41604	0.49783	0.04184	0.00017	0.00018
	MICE-	0.09296	0.15221	0.28741	0.03016	0.13183	0.00053	0.00075
	MICE+	0.02830	0.06571	0.95972	0.35046	0.06446	0.00027	0.00031
0.3	CCA	5365.96080	4090.49973	14159.71028	1167.67341	55.31849	25.17230	34.94213
	RVI	0.14658	0.14425	4.06220	1.35452	0.11559	0.00038	0.00135
	K3-	0.18643	0.14575	1.74943	0.51514	0.18230	0.00061	0.00145
	K3+	0.15516	0.06194	2.91645	0.98167	0.09542	0.00055	0.00138
	K5-	0.21221	0.15426	4.12230	1.39729	0.19782	0.00059	0.00146
	K5+	0.15791	0.06541	4.18843	1.38883	0.12073	0.00048	0.00138
	K9-	0.18322	0.17243	4.06885	1.37382	0.17960	0.00048	0.00116
	K9+	0.16598	0.05534	4.13318	1.39460	0.13145	0.00047	0.00130
	MF-	0.12273	0.14631	0.28362	0.07368	0.08566	0.00020	0.00029
	MF+	0.04961	0.06563	2.99404	1.00606	0.04841	0.00028	0.00045
	MICE-	0.16827	0.17048	0.81040	0.09221	0.18123	0.00050	0.00105
	MICE+	0.06279	0.07298	1.19289	0.40372	0.08760	0.00025	0.00038

Ref.: Bj=β_j; CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales. En escala de rosas: mayores errores cuadráticos medios, por parámetro. En escala de verdes: menores errores cuadráticos medios, por parámetro.

Cuadro 5.3: Error cuadrático medio de los parámetros del modelo de regresión de Cox para n=500, según método de imputación y proporción de datos faltantes (p).

p	Método	B1	B2	B3	B4	B5	B6	B7
0.1	CCA	0.06594	0.02803	0.08440	0.04474	0.02173	0.00011	0.00013
	RVI	0.11331	0.01102	0.06540	0.00656	0.03627	0.00050	0.00078
	K3-	0.03902	0.00551	0.02431	0.00310	0.00744	0.00015	0.00021
	K3+	0.04267	0.00451	0.01619	0.00189	0.00434	0.00013	0.00016
	K5-	0.04902	0.00602	0.03027	0.00454	0.01493	0.00020	0.00031
	K5+	0.04463	0.00594	0.02785	0.00325	0.00678	0.00020	0.00024
	K9-	0.05878	0.01538	0.05884	0.00679	0.03455	0.00037	0.00055
	K9+	0.07712	0.00737	0.06559	0.00687	0.01696	0.00030	0.00045
	MF-	0.05561	0.00453	0.00367	0.00132	0.00167	0.00001	0.00002
	MF+	0.05344	0.00698	0.00144	0.00053	0.00098	0.00001	0.00002
	MICE-	0.08050	0.01357	0.05780	0.00471	0.03276	0.00028	0.00043
	MICE+	0.03700	0.00676	0.01769	0.00590	0.00857	0.00006	0.00009
0.2	CCA	0.14304	0.07940	14.92787	4.92489	0.10166	0.00036	0.00025
	RVI	0.11190	0.06076	0.09149	0.01830	0.09514	0.00044	0.00069
	K3-	0.07730	0.03216	0.12964	0.01351	0.06409	0.00040	0.00049
	K3+	0.04599	0.01045	0.08075	0.01284	0.03341	0.00033	0.00033
	K5-	0.10140	0.04347	0.14565	0.01406	0.08339	0.00047	0.00057
	K5+	0.05408	0.01701	0.10261	0.01247	0.04250	0.00041	0.00042
	K9-	0.09874	0.06034	0.10619	0.01617	0.08582	0.00046	0.00062
	K9+	0.04883	0.01844	0.08536	0.01655	0.04370	0.00045	0.00056
	MF-	0.04115	0.01003	0.04003	0.01367	0.01240	0.00004	0.00005
	MF+	0.01199	0.01288	0.01205	0.00733	0.00693	0.00003	0.00003
	MICE-	0.16833	0.08146	0.16707	0.00744	0.12369	0.00046	0.00064
	MICE+	0.03355	0.02410	0.04027	0.01337	0.03643	0.00011	0.00013
0.3	CCA	0.35693	0.56152	47.34951	15.65709	0.54500	0.00163	0.00181
	RVI	0.07476	0.14241	0.17528	0.01526	0.09281	0.00050	0.00129
	K3-	0.10334	0.06913	0.27815	0.01937	0.08780	0.00079	0.00124
	K3+	0.07820	0.02697	0.18851	0.01759	0.03315	0.00062	0.00101
	K5-	0.11780	0.08093	0.26175	0.01640	0.10330	0.00082	0.00143
	K5+	0.09054	0.02351	0.21855	0.01887	0.03641	0.00060	0.00117
	K9-	0.12629	0.08037	0.21318	0.01479	0.10687	0.00072	0.00141
	K9+	0.07903	0.02355	0.22576	0.02592	0.03288	0.00064	0.00129
	MF-	0.05111	0.01932	0.08789	0.02112	0.03351	0.00014	0.00020
	MF+	0.01002	0.06684	0.02305	0.01346	0.01821	0.00006	0.00006
	MICE-	0.07971	0.17047	0.32921	0.01017	0.16777	0.00072	0.00117
	MICE+	0.02016	0.02510	0.07895	0.01864	0.03644	0.00020	0.00025
0.4	CCA	2328.30624	4351.49543	5885.45375	1507.50765	5492.59777	103.84535	59.09104
	RVI	0.03488	0.16973	0.27373	0.03035	0.16323	0.00083	0.00201
	K3-	0.04174	0.13440	0.50843	0.03787	0.20395	0.00124	0.00225
	K3+	0.08299	0.05719	0.32498	0.03338	0.12549	0.00093	0.00162
	K5-	0.05757	0.12684	0.41587	0.02559	0.19916	0.00123	0.00219
	K5+	0.09586	0.05555	0.32052	0.03663	0.10985	0.00076	0.00156
	K9-	0.07001	0.12777	0.30068	0.03081	0.21466	0.00103	0.00203
	K9+	0.07775	0.04208	0.28953	0.04073	0.09562	0.00062	0.00145
	MF-	0.05034	0.08418	0.18371	0.03796	0.08472	0.00011	0.00022
	MF+	0.03839	0.08881	0.07199	0.03231	0.03838	0.00072	0.00066
	MICE-	0.02739	0.22959	0.63626	0.03579	0.22436	0.00156	0.00211
	MICE+	0.02556	0.05854	0.08359	0.03544	0.05677	0.00014	0.00022

Ref.: Bj=β_j; CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales. En escala de rosas: mayores errores cuadráticos medios, por parámetro. En escala de verdes: menores errores cuadráticos medios, por parámetro.

Cuadro 5.4: Error cuadrático medio de los parámetros del modelo de regresión de Cox para n=1000, según método de imputación y proporción de datos faltantes (p).

p	Método	B1	B2	B3	B4	B5	B6	B7
0.1	CCA	0.04780	0.01498	0.03215	0.02033	0.01183	0.00006	0.00006
	RVI	0.07410	0.01237	0.05048	0.00234	0.03387	0.00043	0.00064
	K3-	0.03027	0.00306	0.00922	0.00096	0.00398	0.00007	0.00009
	K3+	0.03142	0.00292	0.00443	0.00045	0.00271	0.00006	0.00008
	K5-	0.03462	0.00290	0.01403	0.00130	0.00519	0.00010	0.00013
	K5+	0.03750	0.00291	0.00551	0.00053	0.00317	0.00007	0.00009
	K9-	0.04216	0.00530	0.03092	0.00257	0.00989	0.00018	0.00026
	K9+	0.05092	0.00484	0.01415	0.00171	0.00349	0.00010	0.00013
	MF-	0.06733	0.00479	0.00135	0.00062	0.00073	0.00002	0.00003
	MF+	0.06196	0.00977	0.00043	0.00019	0.00010	0.00001	0.00002
	MICE-	0.05951	0.01778	0.05219	0.00324	0.03916	0.00033	0.00043
	MICE+	0.02552	0.00512	0.01090	0.00275	0.00892	0.00007	0.00008
0.2	CCA	0.04419	0.03384	0.14515	0.05831	0.04024	0.00025	0.00020
	RVI	0.10593	0.04941	0.12916	0.00977	0.06343	0.00074	0.00094
	K3-	0.04449	0.01616	0.05994	0.00847	0.02744	0.00028	0.00031
	K3+	0.02355	0.00475	0.02812	0.00368	0.01476	0.00020	0.00023
	K5-	0.05903	0.02041	0.08810	0.01277	0.03496	0.00041	0.00048
	K5+	0.03329	0.00570	0.03834	0.00463	0.01631	0.00027	0.00028
	K9-	0.08470	0.02666	0.16881	0.01857	0.05548	0.00058	0.00075
	K9+	0.04916	0.00677	0.07257	0.01033	0.02003	0.00042	0.00046
	MF-	0.05782	0.00227	0.00957	0.00426	0.00493	0.00004	0.00003
	MF+	0.03564	0.01007	0.00106	0.00056	0.00058	0.00002	0.00001
	MICE-	0.13195	0.07186	0.27747	0.01465	0.10398	0.00069	0.00082
	MICE+	0.03784	0.02120	0.03970	0.00819	0.02898	0.00013	0.00012
0.3	CCA	0.18604	0.14315	19.88650	6.68748	0.14089	0.00078	0.00058
	RVI	0.07625	0.12978	0.24805	0.00876	0.07134	0.00080	0.00158
	K3-	0.06689	0.05208	0.18837	0.01645	0.05031	0.00065	0.00088
	K3+	0.05704	0.01857	0.12221	0.01060	0.02362	0.00049	0.00068
	K5-	0.09286	0.05967	0.20861	0.01620	0.06686	0.00075	0.00113
	K5+	0.07900	0.01814	0.13039	0.01596	0.03457	0.00053	0.00077
	K9-	0.10428	0.05859	0.22428	0.00949	0.08887	0.00084	0.00135
	K9+	0.09197	0.01447	0.13364	0.01394	0.03519	0.00052	0.00085
	MF-	0.06458	0.00734	0.03006	0.01262	0.01790	0.00003	0.00006
	MF+	0.00935	0.03130	0.00546	0.00301	0.00411	0.00003	0.00004
	MICE-	0.08183	0.16729	0.48446	0.02226	0.13275	0.00098	0.00141
	MICE+	0.02280	0.02731	0.06219	0.01245	0.02543	0.00015	0.00018
0.4	CCA	0.59315	0.80832	74.06247	24.66238	1.04419	0.00406	0.00291
	RVI	0.08581	0.11597	0.25339	0.01657	0.12798	0.00099	0.00198
	K3-	0.09648	0.09476	0.42250	0.03198	0.13938	0.00117	0.00175
	K3+	0.10134	0.02460	0.24097	0.01712	0.07591	0.00077	0.00114
	K5-	0.09851	0.09795	0.37165	0.01940	0.17235	0.00119	0.00192
	K5+	0.12683	0.02312	0.26795	0.01782	0.07780	0.00075	0.00126
	K9-	0.09400	0.09591	0.30629	0.01378	0.17895	0.00125	0.00205
	K9+	0.13018	0.01943	0.25521	0.01969	0.06942	0.00071	0.00132
	MF-	0.06034	0.03282	0.06893	0.02309	0.04236	0.00004	0.00008
	MF+	0.01054	0.03571	0.02150	0.00475	0.01113	0.00022	0.00018
	MICE-	0.07803	0.18114	0.64854	0.03535	0.16856	0.00158	0.00196
	MICE+	0.01765	0.03875	0.06582	0.02010	0.02817	0.00014	0.00020

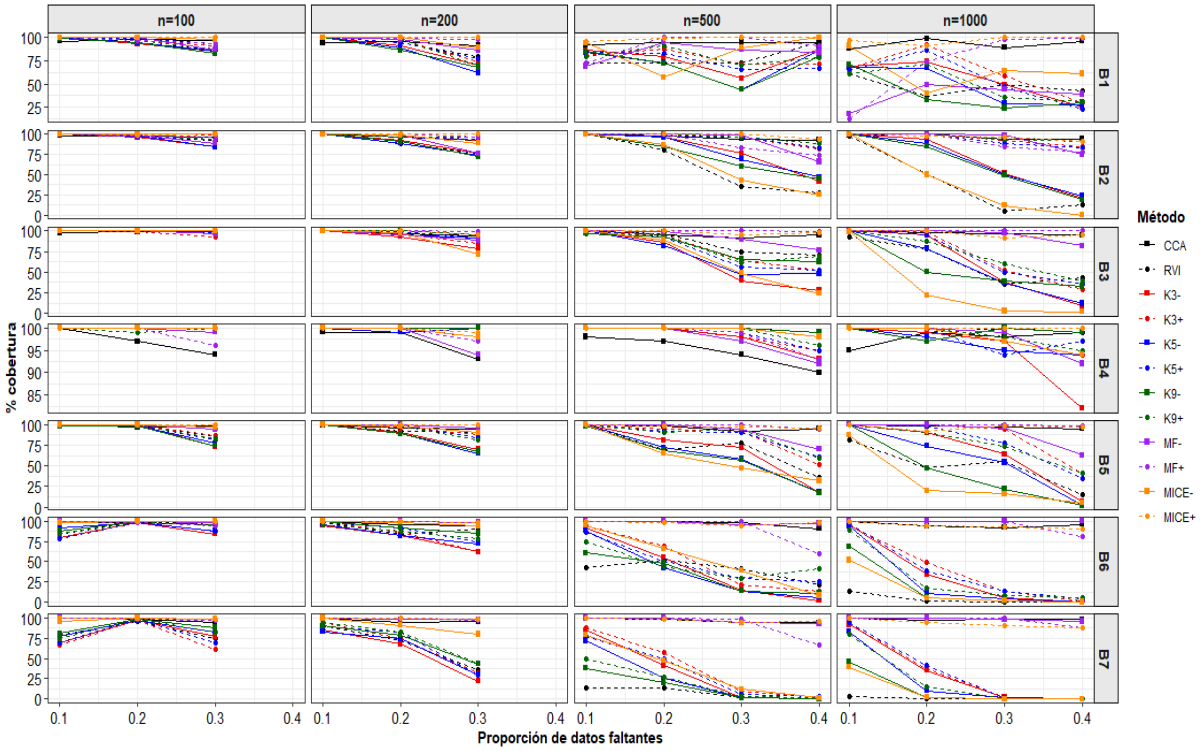
Ref.: Bj=β_j; CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales. En escala de rosas: mayores errores cuadráticos medios, por parámetro. En escala de verdes: menores errores cuadráticos medios, por parámetro.

Cobertura

En la Figura 5.1 se presentan los porcentajes de intervalos de confianza que cubren al verdadero valor, para cada parámetro del modelo y según el tamaño muestral y el porcentaje de pérdida.

Cuando $n = 100$, la cobertura resulta alta y muy similar para todos los parámetros y proporciones de datos faltantes. Para los tamaños de muestra más grandes, para todos los parámetros se observa que el porcentaje de cobertura presenta mayores variaciones entre las técnicas a medida que aumenta la proporción de pérdida. Sin embargo, no se observa un patrón claro que permita establecer mejores resultados para un determinado método.

Figura 5.1: Porcentaje de cobertura para los intervalos de confianza del 95 %, según método de imputación, tamaño de muestra y proporción de datos faltantes.



Ref.: Bj=β_j; CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

5.3.1.b. Reproducibilidad de los datos perdidos

En esta sección, se realiza una evaluación de la eficiencia de los métodos de imputación para asignar valores a los datos faltantes, según se trate de observaciones corres-

pondientes a variables cuantitativas o cualitativas, cuando se han generado datos perdidos de acuerdo a un mecanismo MCAR.

Raíz del error cuadrático medio normalizado

Respecto a la imputación de datos para variables cuantitativas, los valores de $\overline{\text{NRMSE}}$ obtenidos para los distintos escenarios se presentan en la figura 5.2. Para las cuatro variables estudiadas se observa que, para un n fijo, el $\overline{\text{NRMSE}}$ aumenta a medida que se incrementa la proporción de datos faltantes, p. Además, se encuentra que a medida que aumenta el tamaño de muestra n se incrementa la diferencia observada en los $\overline{\text{NRMSE}}$ correspondientes a los distintos métodos de imputación.

En todos los casos, los mejores resultados (es decir, los menores valores de $\overline{\text{NRMSE}}$) se obtienen utilizando el método MF, mejorando cuando se tienen en cuenta el *estado* y el *tiempo hasta el evento o censura* como variables informativas (MF+). Por el contrario, el método MICE genera imputaciones que implican los mayores valores de $\overline{\text{NRMSE}}$, siendo estos los resultados más desfavorables para todos los tamaños de muestra y proporciones de datos perdidos.

En el caso de $n = 100$ y para todos los valores de p, el método KNN presenta mejores resultados en relación a $\overline{\text{NRMSE}}$ cuando se consideran 9 observaciones donantes, con valores similares a los obtenidos bajo RVI.

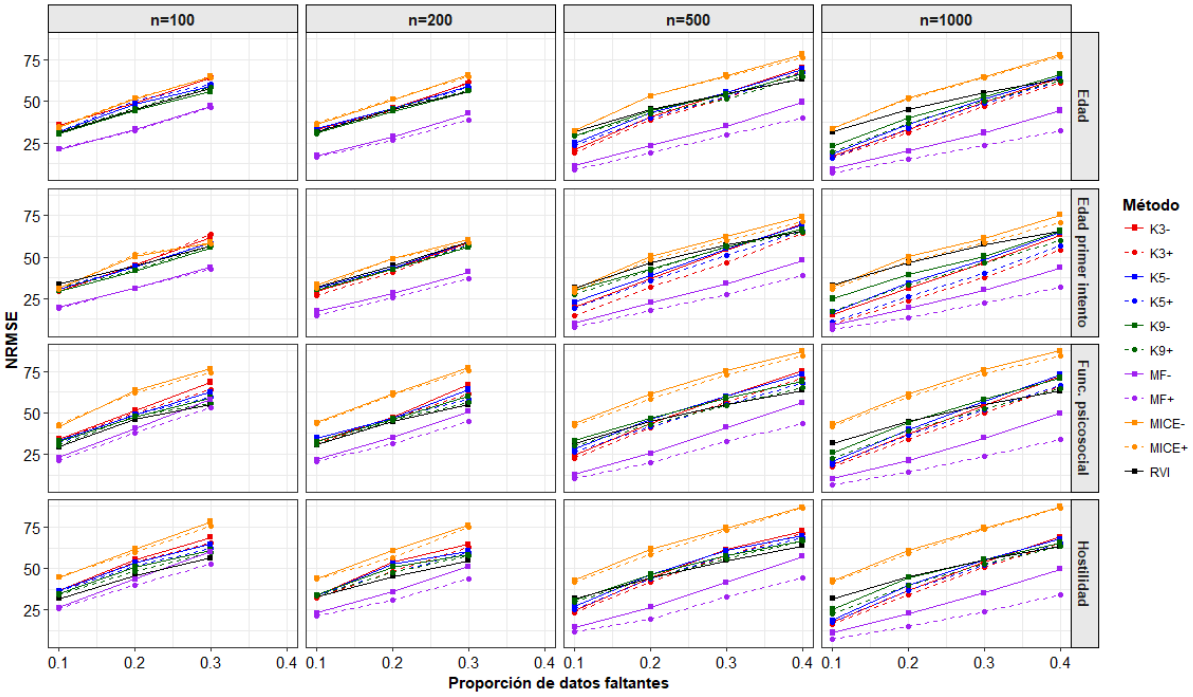
Para $n = 200$, para los dos valores más pequeños de p, se observan resultados muy similares para todos los posibles valores de k evaluados y para RVI, mientras que cuando $p = 0.30$ los resultados son mejores para KNN con 9 observaciones donantes y para RVI y empeoran cuando se consideran solo 3 observaciones donantes.

Cuando $n = 500$ y $p = 0.10$, los menores valores de $\overline{\text{NRMSE}}$ se encuentran con $k = 3$ y los mayores, con $k = 9$. Para el mismo tamaño muestral y $p = 0.20$ o $p = 0.30$ los resultados son muy similares para todos los valores de k y para RVI, mientras que cuando $p = 0.40$ los menores $\overline{\text{NRMSE}}$ ocurren bajo RVI y cuando $k = 9$ y los mayores $\overline{\text{NRMSE}}$ se presentan con $k = 3$.

Finalmente, para $n = 1000$ se encuentran resultados levemente favorables con K3+ en relación a las otras configuraciones probadas para KNN. Respecto a RVI, muestra mayores $\overline{\text{NRMSE}}$ que KNN, aunque menores que bajo MICE, salvo cuando $p = 0.40$ en

el que se comporta de manera similar a KNN.

Figura 5.2: Error cuadrático medio normalizado (NRMSE) para las variables explicativas cuantitativas, según método de imputación, tamaño de la muestra y proporción de datos faltantes.



Ref.: CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

Porcentaje de no coincidencia

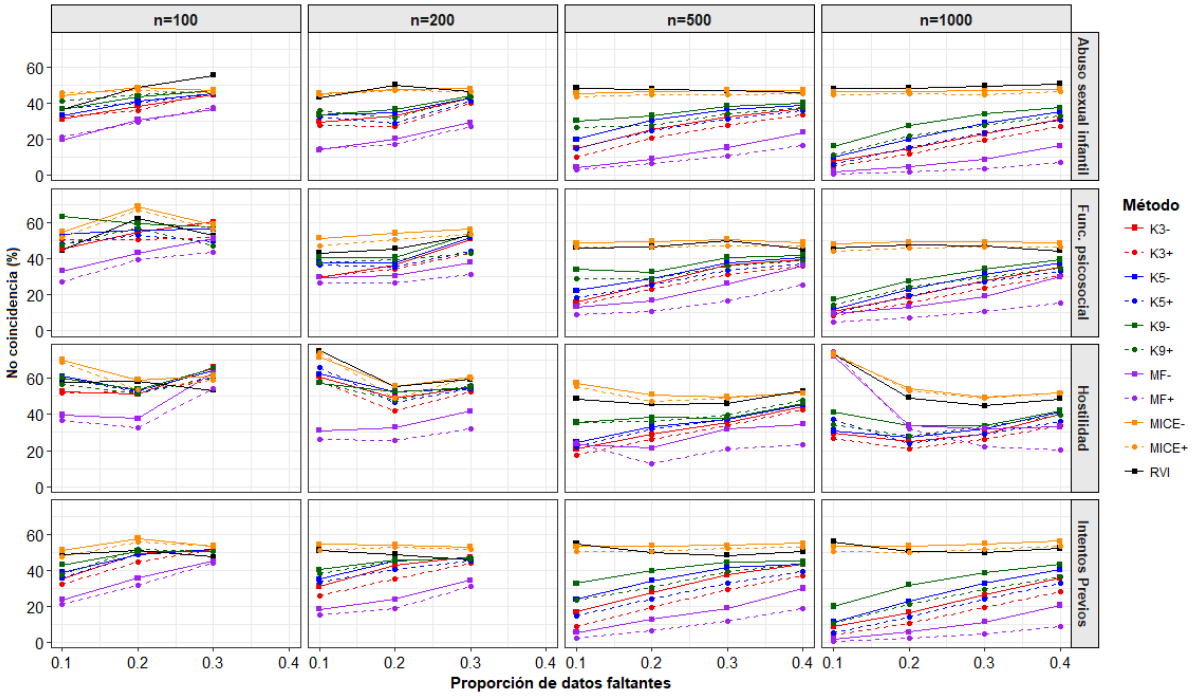
En el caso de las variables cualitativas, los porcentajes promedio de clasificación inadecuada se muestran en la figura 5.3, de acuerdo al tamaño del conjunto de datos y a la proporción de datos faltantes en cada uno de ellos. Se observa que, en la gran mayoría de las situaciones, los menores porcentajes de clasificación incorrecta se obtienen utilizando el método MF, mejorando cuando se tienen en cuenta el *estado* y el *tiempo hasta el evento o censura* como variables informativas (MF+).

Las técnicas MICE y RVI presentan, en general, los mayores porcentajes de datos imputados en forma incorrecta, incluso cuando la proporción de datos faltantes es pequeña.

Dentro de las configuraciones establecidas para el método KNN, se observa que el uso de 3 datos vecinos y la inclusión del *estado* y el *tiempo hasta el evento o censura* como variables informativas, es decir K3+, genera los datos con menor porcentaje de

clasificación incorrecta.

Figura 5.3: Porcentaje de datos imputados no coincidentes con los reales para las variables explicativas cualitativas, según método de imputación, tamaño de la muestra y proporción de datos faltantes.



Ref.: CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

5.3.2. Datos perdidos al azar

5.3.2.a. Propiedades distribucionales de los estimadores de los coeficientes del modelo de regresión de Cox

Error cuadrático medio

Los resultados obtenidos al calcular el MSE para cada uno de los siete parámetros del modelo 5.1 se muestran en los Cuadros 5.5 a 5.8, de acuerdo a la cantidad de unidades del conjunto de datos y la proporción de datos faltantes en cada uno de ellos, p.

Al considerar conjuntos de 100 unidades, con una proporción de datos faltantes de $p = 0.10$, se encuentran MSE altos al emplear K3+, K5- o CCA, mientras que MF y MICE brindan los resultados más favorables. Cuando $p = 0.20$, CCA, K9+ y MF+ son las técnicas con MSE más altos, mientras que KNN con 3 datos donantes y MICE- presentan los menores MSE. Con $p = 0.30$, CCA y K3+ se relacionan con los mayores MSE, obteniéndose los mejores resultados con MICE+. Y para $p = 0.40$, CCA es quien muestra los mayores valores de MSE y MICE la que implica los menores MSE. En general, no se observa un patrón al comparar para cada configuración los MSE obtenidos al considerar o no el *estado* y el *tiempo hasta el evento o censura* como variables informativas (Cuadro 5.5).

En el caso de $n = 200$, para $p = 0.10$, los mayores MSE se observan al emplear CCA o KNN, mientras que los valores más bajos se encuentran con MF, seguidos por MICE+. Para $p = 0.20$, CCA y RVI implican los mayores MSE, encontrándose los menores valores al utilizar K3+, MF o MICE+. Cuando $p = 0.30$, los resultados más desfavorables se encuentran con CCA y K5-. Si bien con MF se obtienen MSE que se ubican entre los más pequeños para la mayoría de los parámetros, también se observan valores entre los más grandes para algunos de ellos. El uso de MICE para la imputación de datos también implica MSE que se ubican entre los más chicos. Finalmente, al considerar $p = 0.40$, los resultados más desfavorables se encuentran con CCA, seguido de K3- y K5-. Los MSE inferiores se obtienen con MICE y MF (Cuadro 5.6).

Cuando $n = 500$ y $p = 0.10$, 0.20 o 0.30, el uso de CCA, RVI o KNN con 9 datos donantes brindan los MSE más grandes. Mediante la imputación con MF se encuentran los resultados más favorables en relación al MSE. Para la técnica KNN, se observa una

tendencia a obtener menores MSE cuando se consideran menos datos vecinos. Sin embargo, cuando $p = 0.40$, junto a CCA los valores más grandes de MSE ocurren bajo el uso de K5-. Los MSE se encuentran ante el uso de MICE o MF como métodos de imputación (Cuadro 5.7).

Para $n = 1000$, se observan comportamientos similares para todos los valores de p propuestos. Los MSE más grandes se encuentran bajo CCA y utilizando RVI o KNN con 9 datos donantes. MF es la técnica que muestra MSE más pequeños, así como MICE+ y KNN con 3 datos donantes. En este caso, se detecta una tendencia en los MSE a resultar más pequeños cuando se tienen en cuenta el *estado* y el *tiempo hasta el evento o censura* como variables informativas en comparación a los obtenidos cuando no se los incluye (Cuadro 5.8).

Cuadro 5.5: Error cuadrático medio de los parámetros del modelo de regresión de Cox para n=100, según método de imputación y proporción de datos faltantes (p).

p	Método	B1	B2	B3	B4	B5	B6	B7
0.1	CCA	0.05381	0.04828	3.42064	1.15268	0.04789	0.00035	0.00042
	RVI	0.00742	0.00688	0.02444	0.00463	0.05978	0.00044	0.00053
	K3-	0.00660	0.01359	0.03289	0.00188	0.09432	0.00055	0.00073
	K3+	0.00387	0.02186	0.06282	0.00220	0.11895	0.00077	0.00112
	K5-	0.01350	0.01856	0.03024	0.00288	0.10430	0.00056	0.00072
	K5+	0.00838	0.01046	0.02770	0.00301	0.11121	0.00043	0.00050
	K9-	0.01120	0.01597	0.03042	0.00349	0.08323	0.00041	0.00056
	K9+	0.01553	0.00891	0.03200	0.00447	0.07494	0.00024	0.00032
	MF-	0.00324	0.00533	0.01289	0.00071	0.03621	0.00017	0.00022
	MF+	0.00446	0.00585	0.01737	0.00070	0.03201	0.00014	0.00023
	MICE-	0.00401	0.01023	0.01879	0.00116	0.05890	0.00029	0.00043
	MICE+	0.00524	0.00809	0.01476	0.00144	0.04307	0.00016	0.00023
0.2	CCA	0.36569	0.47282	0.13212	0.10850	0.33963	0.00086	0.00127
	RVI	0.01410	0.01929	0.00922	0.00404	0.05143	0.00010	0.00035
	K3-	0.01451	0.01443	0.00852	0.00645	0.05863	0.00013	0.00031
	K3+	0.00929	0.01762	0.01207	0.00504	0.05298	0.00020	0.00035
	K5-	0.01680	0.01851	0.01014	0.00901	0.07605	0.00010	0.00030
	K5+	0.01362	0.03339	0.00895	0.01545	0.09629	0.00012	0.00024
	K9-	0.01860	0.01940	0.01003	0.01152	0.06118	0.00012	0.00035
	K9+	0.03834	0.05399	0.01837	0.03053	0.14250	0.00026	0.00094
	MF-	0.01316	0.01176	0.01177	0.01101	0.05511	0.00016	0.00035
	MF+	0.01767	0.02913	0.01777	0.02266	0.14121	0.00024	0.00056
	MICE-	0.00651	0.02470	0.00809	0.00808	0.03999	0.00011	0.00021
	MICE+	0.00790	0.04370	0.01256	0.01606	0.09679	0.00017	0.00037
0.3	CCA	0.11648	0.32335	60.97074	20.04295	0.14626	0.00270	0.00154
	RVI	0.02423	0.00653	0.02688	0.00645	0.09791	0.00008	0.00047
	K3-	0.03013	0.00587	0.03437	0.00633	0.22874	0.00025	0.00083
	K3+	0.02754	0.01313	0.08566	0.02130	0.18990	0.00047	0.00197
	K5-	0.02608	0.00490	0.02963	0.00684	0.22427	0.00021	0.00059
	K5+	0.01499	0.00844	0.03839	0.00886	0.13708	0.00023	0.00064
	K9-	0.03598	0.00312	0.03775	0.00693	0.29225	0.00025	0.00089
	K9+	0.01714	0.00857	0.04936	0.01238	0.13006	0.00016	0.00040
	MF-	0.01775	0.00622	0.02469	0.00525	0.14841	0.00025	0.00049
	MF+	0.01501	0.00790	0.03084	0.00737	0.11028	0.00031	0.00048
	MICE-	0.02391	0.00436	0.02958	0.00689	0.17369	0.00023	0.00063
	MICE+	0.01293	0.00640	0.02829	0.00415	0.09182	0.00016	0.00036
0.4	CCA	6.92327	6.08939	10.28545	8.30016	6.85319	0.00272	0.00150
	RVI	0.05277	0.00556	0.03112	0.00624	0.11609	0.00102	0.00109
	K3-	0.05006	0.00422	0.35634	0.09757	0.24943	0.00033	0.00138
	K3+	0.01818	0.00864	0.30585	0.08607	0.12753	0.00053	0.00197
	K5-	0.05419	0.01989	0.29558	0.10845	0.16048	0.00042	0.00159
	K5+	0.01608	0.02542	0.13799	0.05275	0.04840	0.00063	0.00185
	K9-	0.08939	0.04276	0.21852	0.09991	0.15059	0.00029	0.00148
	K9+	0.01755	0.02560	0.20554	0.06615	0.06428	0.00081	0.00192
	MF-	0.03185	0.02251	0.62004	0.17917	0.09483	0.00021	0.00065
	MF+	0.01885	0.02155	0.49441	0.15130	0.08981	0.00025	0.00163
	MICE-	0.03363	0.00945	0.11941	0.04167	0.05746	0.00016	0.00027
	MICE+	0.01736	0.00894	0.14346	0.05790	0.05212	0.00018	0.00058

Ref.: Bj=β_j; CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales. En escala de rosas: mayores errores cuadráticos medios, por parámetro. En escala de verdes: menores errores cuadráticos medios, por parámetro.

Cuadro 5.6: Error cuadrático medio de los parámetros del modelo de regresión de Cox para n=200, según método de imputación y proporción de datos faltantes (p).

p	Método	B1	B2	B3	B4	B5	B6	B7
0.1	CCA	0.02663	0.02278	0.02714	0.01356	0.01897	0.00011	0.00011
	RVI	0.00721	0.00414	0.00736	0.00108	0.02339	0.00009	0.00017
	K3-	0.00236	0.00513	0.01032	0.00098	0.02062	0.00012	0.00026
	K3+	0.00196	0.00536	0.01564	0.00092	0.02381	0.00017	0.00031
	K5-	0.00347	0.00874	0.01440	0.00162	0.03731	0.00018	0.00037
	K5+	0.00268	0.00906	0.02691	0.00119	0.03752	0.00027	0.00051
	K9-	0.00324	0.01054	0.00890	0.00085	0.05640	0.00009	0.00022
	K9+	0.00228	0.00446	0.00613	0.00074	0.05203	0.00010	0.00016
	MF-	0.00092	0.00130	0.00297	0.00020	0.00687	0.00003	0.00006
	MF+	0.00108	0.00093	0.00363	0.00018	0.00626	0.00002	0.00006
	MICE-	0.00296	0.00509	0.00634	0.00062	0.02370	0.00006	0.00014
	MICE+	0.00113	0.00330	0.00388	0.00080	0.00896	0.00003	0.00006
0.2	CCA	0.04023	0.04237	0.05847	0.01865	0.03792	0.00046	0.00028
	RVI	0.01031	0.03112	0.01309	0.00422	0.09044	0.00018	0.00031
	K3-	0.00714	0.02029	0.00843	0.00141	0.06993	0.00015	0.00029
	K3+	0.00521	0.01084	0.00793	0.00155	0.02798	0.00017	0.00026
	K5-	0.00803	0.02058	0.00898	0.00164	0.06414	0.00018	0.00033
	K5+	0.00530	0.01216	0.01048	0.00148	0.02293	0.00024	0.00038
	K9-	0.00794	0.02088	0.00726	0.00407	0.08105	0.00014	0.00021
	K9+	0.00454	0.01228	0.00850	0.00175	0.04567	0.00018	0.00025
	MF-	0.00782	0.00746	0.00250	0.00387	0.03424	0.00004	0.00007
	MF+	0.01070	0.00979	0.00304	0.00361	0.05633	0.00004	0.00009
	MICE-	0.00546	0.01921	0.00780	0.00420	0.05297	0.00012	0.00022
	MICE+	0.00508	0.01246	0.00643	0.00443	0.02852	0.00009	0.00016
0.3	CCA	0.06440	0.11273	65.82509	22.60590	0.12623	0.00077	0.00065
	RVI	0.03126	0.00212	0.01604	0.00200	0.08918	0.00004	0.00025
	K3-	0.03146	0.00370	0.03735	0.00119	0.18106	0.00022	0.00086
	K3+	0.01207	0.00297	0.03252	0.00366	0.05666	0.00030	0.00104
	K5-	0.03390	0.00406	0.05651	0.00125	0.19792	0.00029	0.00131
	K5+	0.02621	0.00215	0.07024	0.00240	0.10722	0.00049	0.00227
	K9-	0.02986	0.00534	0.03796	0.00178	0.17964	0.00020	0.00081
	K9+	0.01202	0.00368	0.01703	0.00605	0.08708	0.00020	0.00048
	MF-	0.00644	0.00567	0.00806	0.00257	0.03595	0.00007	0.00015
	MF+	0.01008	0.00771	0.01580	0.00407	0.04157	0.00009	0.00025
	MICE-	0.03014	0.00412	0.02329	0.00043	0.13778	0.00012	0.00047
	MICE+	0.01394	0.00342	0.00703	0.00107	0.05686	0.00005	0.00012
0.4	CCA	0.16750	1.67604	113.65606	46.75720	0.14798	0.00237	0.00090
	RVI	0.02582	0.05514	0.04813	0.00662	0.18194	0.00154	0.00137
	K3-	0.04795	0.08021	0.09611	0.02894	0.63092	0.00065	0.00353
	K3+	0.02989	0.07785	0.08248	0.01878	0.52400	0.00066	0.00275
	K5-	0.04242	0.07741	0.09498	0.02392	0.60084	0.00069	0.00326
	K5+	0.02028	0.05318	0.08181	0.01078	0.33420	0.00079	0.00266
	K9-	0.01852	0.05998	0.05567	0.00462	0.32268	0.00049	0.00128
	K9+	0.02434	0.03645	0.09308	0.01013	0.27894	0.00082	0.00268
	MF-	0.01319	0.02800	0.01628	0.00864	0.22504	0.00029	0.00054
	MF+	0.01336	0.01788	0.02205	0.00927	0.13676	0.00028	0.00054
	MICE-	0.01806	0.05982	0.01498	0.00451	0.22160	0.00015	0.00043
	MICE+	0.00930	0.05085	0.00931	0.00499	0.11388	0.00015	0.00026

Ref.: Bj=β_j; CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales. En escala de rosas: mayores errores cuadráticos medios, por parámetro. En escala de verdes: menores errores cuadráticos medios, por parámetro.

Cuadro 5.7: Error cuadrático medio de los parámetros del modelo de regresión de Cox para n=500, según método de imputación y proporción de datos faltantes (p).

p	Método	B1	B2	B3	B4	B5	B6	B7
0.1	CCA	0.006754	0.013057	0.030496	0.026080	0.005125	0.000040	0.000030
	RVI	0.012450	0.002513	0.011700	0.000938	0.023682	0.000222	0.000278
	K3-	0.001711	0.000589	0.003450	0.000149	0.001952	0.000027	0.000054
	K3+	0.001243	0.000722	0.001735	0.000145	0.001443	0.000018	0.000028
	K5-	0.002519	0.001263	0.006708	0.000282	0.006881	0.000068	0.000106
	K5+	0.001551	0.000714	0.003210	0.000395	0.002818	0.000039	0.000050
	K9-	0.005500	0.005824	0.020444	0.000663	0.035259	0.000302	0.000458
	K9+	0.002632	0.003369	0.007245	0.001426	0.019475	0.000153	0.000208
	MF-	0.000250	0.000211	0.000361	0.000046	0.001249	0.000003	0.000007
	MF+	0.000244	0.000169	0.000302	0.000024	0.000666	0.000003	0.000007
	MICE-	0.004303	0.002615	0.003301	0.000339	0.012385	0.000042	0.000088
	MICE+	0.001491	0.001043	0.000848	0.000388	0.003318	0.000012	0.000018
0.2	CCA	0.034050	0.045180	0.135600	0.110500	0.056420	0.000100	0.000060
	RVI	0.027830	0.010630	0.013480	0.003120	0.085020	0.000080	0.000190
	K3-	0.002260	0.002870	0.002030	0.001090	0.022720	0.000030	0.000040
	K3+	0.004190	0.002780	0.003420	0.001740	0.013130	0.000030	0.000030
	K5-	0.004290	0.006420	0.001830	0.001370	0.065440	0.000040	0.000070
	K5+	0.003740	0.003910	0.003810	0.002470	0.022840	0.000050	0.000080
	K9-	0.009930	0.019180	0.005360	0.001470	0.114620	0.000100	0.000190
	K9+	0.007290	0.005700	0.007190	0.001060	0.056900	0.000140	0.000210
	MF-	0.001230	0.001190	0.000620	0.000510	0.005650	0.000010	0.000010
	MF+	0.001310	0.001440	0.000650	0.000480	0.004660	0.000010	0.000010
	MICE-	0.021370	0.007150	0.006810	0.002500	0.077380	0.000060	0.000180
	MICE+	0.008830	0.002980	0.001820	0.002560	0.029880	0.000020	0.000040
0.3	CCA	0.038110	0.134560	0.408680	0.224700	0.062570	0.000210	0.000210
	RVI	0.038310	0.020580	0.023920	0.001430	0.107830	0.000050	0.000340
	K3-	0.014150	0.012040	0.023440	0.000590	0.099730	0.000240	0.000670
	K3+	0.009220	0.006660	0.016520	0.001170	0.042570	0.000240	0.000510
	K5-	0.021380	0.017720	0.019460	0.000660	0.193990	0.000190	0.000590
	K5+	0.011040	0.007500	0.018860	0.001830	0.058330	0.000240	0.000590
	K9-	0.041060	0.026060	0.032280	0.000410	0.286660	0.000190	0.000810
	K9+	0.017710	0.006640	0.025170	0.001030	0.098290	0.000280	0.000660
	MF-	0.002980	0.003330	0.004220	0.000580	0.018380	0.000060	0.000080
	MF+	0.002720	0.002910	0.002710	0.000620	0.016060	0.000040	0.000050
	MICE-	0.034370	0.021020	0.017520	0.000440	0.197880	0.000100	0.000450
	MICE+	0.017950	0.006650	0.004960	0.001330	0.055510	0.000040	0.000110
0.4	CCA	0.186030	1.199940	21.993120	8.502680	0.027390	0.000400	0.000370
	RVI	0.013520	0.019570	0.047120	0.003690	0.140280	0.001100	0.000670
	K3-	0.018290	0.047610	0.106320	0.005530	0.410680	0.000720	0.002520
	K3+	0.008410	0.037990	0.085630	0.002980	0.253750	0.000780	0.002100
	K5-	0.019280	0.048360	0.121180	0.007400	0.468090	0.000780	0.002820
	K5+	0.009310	0.045280	0.108470	0.003440	0.303030	0.000860	0.002440
	K9-	0.014390	0.046090	0.107680	0.004320	0.459240	0.000670	0.002280
	K9+	0.007980	0.047820	0.131970	0.003740	0.294030	0.000960	0.002800
	MF-	0.004500	0.010850	0.011450	0.003920	0.070240	0.000160	0.000230
	MF+	0.008720	0.008420	0.026990	0.006110	0.057110	0.000160	0.000480
	MICE-	0.008860	0.022400	0.021980	0.000850	0.094620	0.000160	0.000510
	MICE+	0.001950	0.017120	0.005190	0.007280	0.020170	0.000060	0.000110

Ref.: Bj=β_j; CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales. En escala de rosas: mayores errores cuadráticos medios, por parámetro. En escala de verdes: menores errores cuadráticos medios, por parámetro.

Cuadro 5.8: Error cuadrático medio de los parámetros del modelo de regresión de Cox para n=1000, según método de imputación y proporción de datos faltantes (p).

p	Método	B1	B2	B3	B4	B5	B6	B7
0.1	CCA	0.0035656	0.0110360	0.0169635	0.0185893	0.0037627	0.0000250	0.0000189
	RVI	0.0080783	0.0017810	0.0068836	0.0003284	0.0166010	0.0001055	0.0001260
	K3-	0.0000397	0.0000338	0.0000297	0.0000119	0.0003147	0.0000003	0.0000006
	K3+	0.0000965	0.0000402	0.0000418	0.0000173	0.0002281	0.0000004	0.0000009
	K5-	0.0003805	0.0003237	0.0006407	0.0000598	0.0007945	0.0000053	0.0000128
	K5+	0.0004017	0.0002467	0.0005277	0.0000644	0.0007797	0.0000060	0.0000114
	K9-	0.0020303	0.0014139	0.0051630	0.0002521	0.0031378	0.0000493	0.0001133
	K9+	0.0008874	0.0005492	0.0016458	0.0002546	0.0028492	0.0000329	0.0000493
	MF-	0.0000458	0.0000271	0.0000691	0.0000075	0.0000445	0.0000003	0.0000019
	MF+	0.0000086	0.0000037	0.0000119	0.0000008	0.0000059	0.0000001	0.0000003
	MICE-	0.0027599	0.0019832	0.0021712	0.0001991	0.0125006	0.0000257	0.0000568
	MICE+	0.0009964	0.0005872	0.0004240	0.0001377	0.0025117	0.0000110	0.0000103
0.2	CCA	0.0165490	0.0343800	0.1204200	0.0997650	0.0277910	0.0000570	0.0000370
	RVI	0.0222720	0.0048790	0.0207610	0.0005930	0.0604100	0.0002150	0.0002510
	K3-	0.0005190	0.0005020	0.0009300	0.0001710	0.0032180	0.0000070	0.0000160
	K3+	0.0004860	0.0004410	0.0007860	0.0002110	0.0026300	0.0000080	0.0000090
	K5-	0.0017360	0.0017160	0.0034080	0.0003960	0.0077910	0.0000330	0.0000680
	K5+	0.0014300	0.0015980	0.0020400	0.0004840	0.0062130	0.0000300	0.0000390
	K9-	0.0066890	0.0054630	0.0098700	0.0008510	0.0297080	0.0001460	0.0002330
	K9+	0.0023600	0.0022910	0.0029700	0.0012140	0.0109110	0.0000970	0.0000950
	MF-	0.0001290	0.0001020	0.0002990	0.0000570	0.0003860	0.0000010	0.0000110
	MF+	0.0000890	0.0000450	0.0001350	0.0000190	0.0001800	0.0000010	0.0000030
	MICE-	0.0137930	0.0042970	0.0057610	0.0004760	0.0542290	0.0000740	0.0001550
	MICE+	0.0034700	0.0011230	0.0014610	0.0005190	0.0162990	0.0000110	0.0000260
0.3	CCA	0.0175700	0.2167600	0.2914000	0.2780100	0.0461700	0.0001200	0.0000900
	RVI	0.0434500	0.0096700	0.0331400	0.0001900	0.1042700	0.0001500	0.0003100
	K3-	0.0035700	0.0035700	0.0087000	0.0005300	0.0232400	0.0001000	0.0002100
	K3+	0.0029800	0.0031800	0.0083800	0.0005900	0.0198900	0.0001100	0.0002000
	K5-	0.0092100	0.0098700	0.0251200	0.0009700	0.0767600	0.0002700	0.0006100
	K5+	0.0053800	0.0048700	0.0143500	0.0010200	0.0385700	0.0002300	0.0003600
	K9-	0.0225200	0.0190800	0.0493600	0.0019400	0.1845800	0.0004200	0.0011200
	K9+	0.0059100	0.0091400	0.0233100	0.0010800	0.0565200	0.0003500	0.0005000
	MF-	0.0006700	0.0006500	0.0007600	0.0002100	0.0036100	0.0000100	0.0000200
	MF+	0.0007300	0.0004300	0.0007400	0.0001700	0.0024200	0.0000100	0.0000200
	MICE-	0.0301700	0.0080800	0.0150800	0.0005200	0.1219200	0.0001000	0.0003600
	MICE+	0.0119600	0.0016500	0.0020100	0.0006600	0.0319300	0.0000200	0.0000400
0.4	CCA	0.0187300	1.2243500	8.6993000	3.8931000	0.0408300	0.0002100	0.0002700
	RVI	0.0330400	0.0129800	0.0631100	0.0024800	0.1633300	0.0010800	0.0008800
	K3-	0.0259000	0.0224900	0.0677400	0.0011000	0.2012300	0.0006400	0.0016700
	K3+	0.0161800	0.0147800	0.0512500	0.0015400	0.1408900	0.0006400	0.0012800
	K5-	0.0382000	0.0269500	0.0920800	0.0016100	0.3001200	0.0008100	0.0023000
	K5+	0.0273300	0.0202000	0.0869600	0.0011700	0.1818500	0.0009300	0.0021300
	K9-	0.0483500	0.0273600	0.1163100	0.0023200	0.3531100	0.0009400	0.0029200
	K9+	0.0420400	0.0264400	0.1249200	0.0014100	0.2154600	0.0010900	0.0031500
	MF-	0.0038500	0.0038500	0.0074100	0.0008900	0.0298900	0.0001000	0.0001400
	MF+	0.0022400	0.0032900	0.0099900	0.0010300	0.0167300	0.0000600	0.0002200
	MICE-	0.0285300	0.0134900	0.0351800	0.0003200	0.1325400	0.0002900	0.0008800
	MICE+	0.0120800	0.0093100	0.0074500	0.0039000	0.0501600	0.0000600	0.0001100

Ref.: Bj=β_j; CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales. En escala de rosas: mayores errores cuadráticos medios, por parámetro. En escala de verdes: menores errores cuadráticos medios, por parámetro.

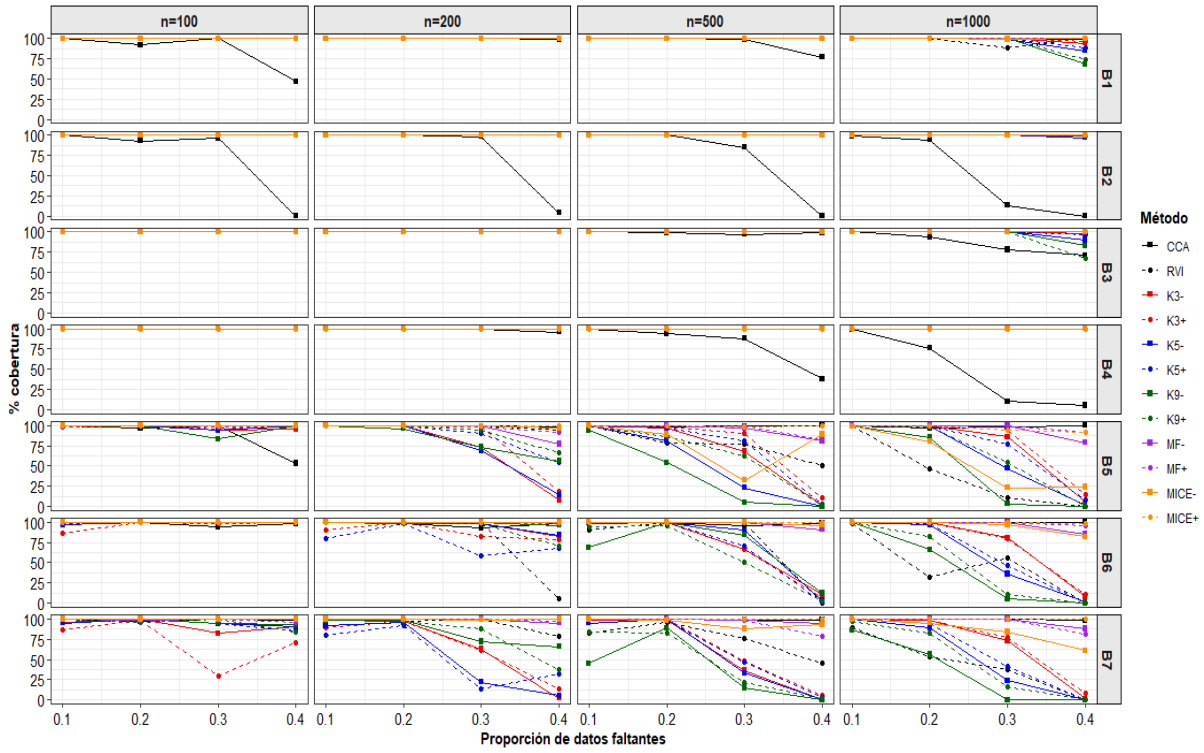
Cobertura

En la Figura 5.4 se presentan los porcentajes de intervalos de confianza que cubren al verdadero valor, para cada parámetro del modelo y según el tamaño muestral y el porcentaje de pérdida.

Para los coeficientes $\beta_1, \beta_2, \beta_3$ y β_4, no se observan diferencias en el porcentaje de cobertura de acuerdo al método usado, siendo cercano al 100 % para todas las proporciones de pérdida p, a excepción de CCA cuya cobertura disminuye a medida que aumenta p.

En el caso de β_5, β_6 y β_7, las diferencias en la cobertura para los distintos métodos aumenta a medida que aumenta la proporción de datos faltantes. RVI y KNN muestran porcentajes de cobertura inferiores al resto de los métodos, siendo menor la cobertura cuando mayor es el valor de k. Sin embargo, no se observa un patrón claro que permita indicar que un determinado método presenta mejores resultados que el resto.

Figura 5.4: Porcentaje de cobertura para los intervalos de confianza del 95 %, según método de imputación, tamaño de muestra y proporción de datos faltantes.



Ref.: Bj=β_j; CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

5.3.2.b. Reproducibilidad de los datos perdidos

En este apartado, se realiza una evaluación de la eficiencia de los métodos de imputación para asignar valores a los datos faltantes, según se trate de observaciones correspondientes a la variable cuantitativa *edad del primer intento de suicidio* o a la variable dicotómica *abuso sexual infantil*, cuando se han generado datos perdidos de acuerdo a un mecanismo MAR.

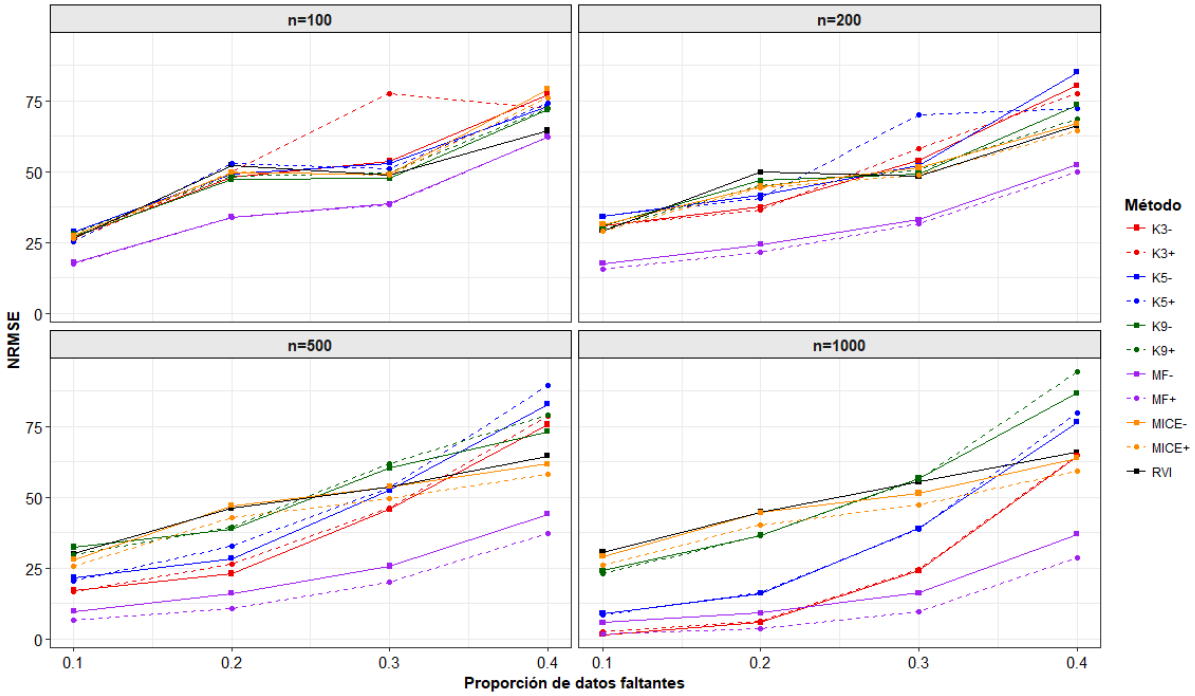
Raíz del error cuadrático medio normalizado

Respecto a la imputación de datos para la variable cuantitativa *edad del primer intento de suicidio*, los valores de $\overline{\text{NRMSE}}$ obtenidos para los distintos escenarios se presentan en la Figura 5.5. Se observa que, en todos los casos, los mejores resultados (es decir, los menores valores de $\overline{\text{NRMSE}}$) se obtienen utilizando el método MF, mejorando cuando se tienen en cuenta el *estado* y el *tiempo hasta el evento o censura* como variables informativas (MF+). Para $n = 100$ y $n = 200$, se encuentran $\overline{\text{NRMSE}}$ similares para todas las técnicas restantes, dentro de los posibles valores de p.

Cuando $n = 500$ y $p = 0.10$, 0.20 o 0.30, KNN presenta menores $\overline{\text{NRMSE}}$ cuanto menos vecinos más cercanos considera como observaciones donantes. Para dichos valores de p, los $\overline{\text{NRMSE}}$ obtenidos al considerar 9 observaciones donantes se encuentran entre los mayores, junto con los obtenidos al utilizar RVI y MICE. Para el caso de $p = 0.40$, MICE y RVI presentan valores de $\overline{\text{NRMSE}}$ más favorables a los obtenidos con KNN, si bien MF se mantiene como la técnica con menor valor de dicho indicador.

Para el caso de $n = 1000$, para $p = 0.10$, 0.20 o 0.30, el uso de KNN con 9 observaciones donantes, RVI y MICE se asocia con los valores menos favorables de $\overline{\text{NRMSE}}$, si bien estas últimas dos técnicas mejoran comparativamente su *performance* cuando $p = 0.40$, con valores similares a los obtenidos por KNN con 3 observaciones donantes. Así, KNN con 5 o 9 vecinos presenta los $\overline{\text{NRMSE}}$ más altos.

Figura 5.5: Error cuadrático medio normalizado (NRMSE) para *Edad del primer intento de suicidio*, según método de imputación, tamaño de la muestra y proporción de datos faltantes.



Ref.: CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

Porcentaje de no coincidencia

En la Figura 5.6 se muestran los porcentajes medios de datos imputados en forma incorrecta para la variable dicotómica *abuso sexual infantil*.

Cuando $n = 100$ y $p = 0.10$ o 0.20, MF brinda los menores porcentajes promedio de datos asigandos de manera errónea, mientras que MICE genera porcentajes de imputaciones incorrectas que se encuentran entre las más desfavorables. Para $p = 0.30$, los resultados de MF se ven levemente mejorados al utilizar K3+, quien a su vez en mejorado por K5+ en el caso de $p = 0.40$. Cuando $p = 0.30$, RVI muestra el mayores porcentajes promedio de datos imputados de manera incorrecta, si bien para $p = 0.40$ son K3- y K5- quienes implican los resultados menores favorables.

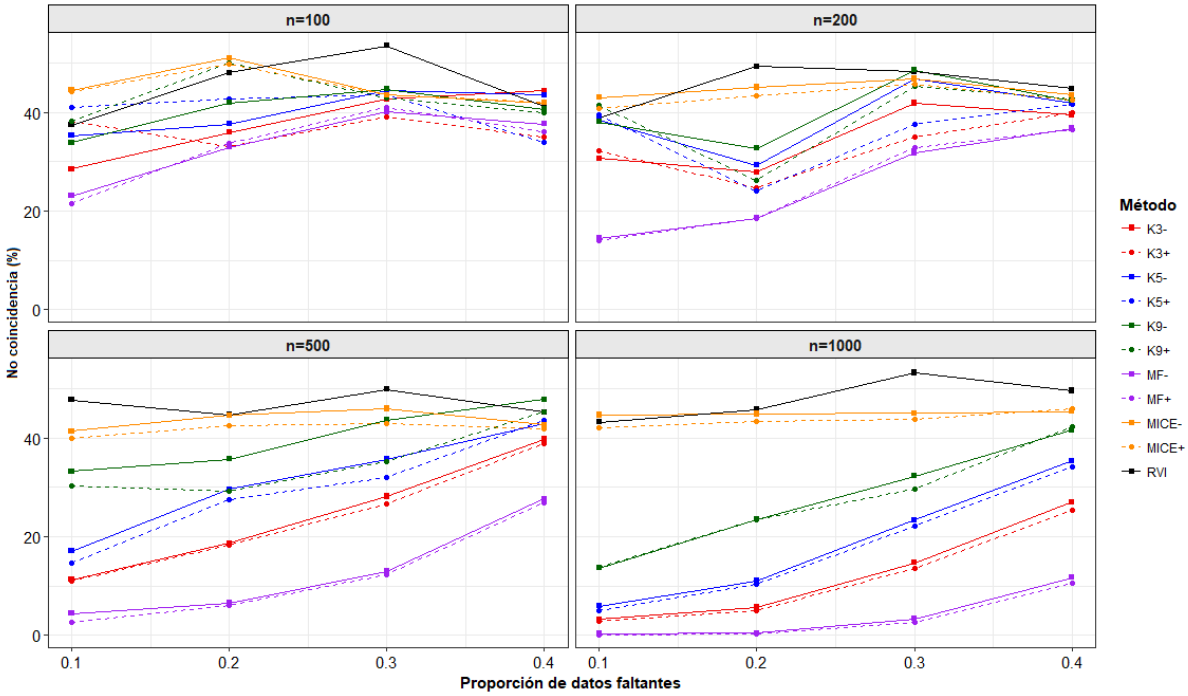
Al considerar $n = 200$, para todas las proporciones de datos faltantes consideradas, se detectan menores porcentajes de datos imputados en forma incorrecta que al utilizar MF, con ínfimas diferencias según se consideren o no el *estado* y el *tiempo hasta el evento o censura* como variables informativas. Dentro de KNN, cuando $p = 0.1$ se obtienen los

mejores resultados con K3-, aunque muy similares a los obtenidos por K3+. Al considerar 5 o 9 valores donantes, se obtienen porcentajes de datos imputados en forma incorrecta muy similares y similares, a su vez, a los obtenidos con RVI. Los resultados con MICE- resultan los menos favorables. Cuando $p = 0.20$, luego de MF, los menores porcentajes de datos asignados en forma errónea surgen con K3+ y K5+, mientras que RVI muestra los valores menos favorables. Un análisis similar corresponde al caso de $p = 0.30$, aunque el porcentaje de imputaciones incorrectas es similar cuando se utiliza RVI y K9-. Finalmente, cuando $p = 0.40$, los valores obtenidos resultan similares al compararlos entre técnicas, siendo MF+ la que permite alcanzar el porcentaje más pequeño y RVI quien implica el mayor porcentaje de observaciones asignadas incorrectamente.

Cuando $n = 500$ y para las proporciones de datos perdidos 0.1, 0.2 y 0.3, se observan menores porcentajes de datos imputados en forma incorrecta que al utilizar MF, seguidos por los obtenidos por KNN (a mayor número de datos donantes, mayor porcentaje de imputaciones erróneas), luego por MICE y siendo RVI el método que implica mayor porcentaje de datos asignados en forma errónea. Además, dentro de cada técnica, se obtienen mejores resultados cuando se consideran el *estado* y el *tiempo hasta el evento o censura* como variables informativas. Sin embargo, cuando $p = 0.40$, MICE presenta resultados levemente más favorables en comparación con KNN con 5 o 9 datos vecinos y con RVI. Además, el porcentaje de datos imputados en forma incorrecta al utilizar RVI es inferior al obtenido con KNN con 9 datos vecinos. Dentro de cada técnica, se obtienen mejores resultados cuando se consideran el *estado* y el *tiempo hasta el evento o censura* como variables informativas.

En el caso de $n = 1000$, para todos las proporciones de datos faltantes generadas, se encuentra que MF es la técnica con menor porcentajes de datos imputados en forma incorrecta, seguido por KNN (a mayor número de datos donantes, mayor porcentaje de imputaciones erróneas), luego por MICE y siendo RVI el método con resultados menos favorables. Además, dentro de cada técnica, se obtienen mejores resultados cuando se consideran el *estado* y el *tiempo hasta el evento o censura* como variables informativas.

Figura 5.6: Porcentaje de datos imputados no coincidentes con los reales para *Abuso Sexual Infantil*, según método de imputación, tamaño de la muestra y proporción de datos faltantes.



Ref.: CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

5.3.3. Datos perdidos no al azar

5.3.3.a. Propiedades distribucionales de los estimadores de los coeficientes del modelo de regresión de Cox

Error cuadrático medio

Los resultados obtenidos al calcular el MSE para cada uno de los siete parámetros del modelo 5.1 se muestran en los Cuadros 5.9 a 5.12, de acuerdo a la cantidad de unidades del conjunto de datos y la proporción de datos faltantes en cada uno de ellos, p.

Para el caso de $n = 100$, cuando $p = 0.10$, los resultados menos favorables se encuentran bajo el uso de CCA, K3+ y K5+, mientras que con MF y MICE se encuentran los MSE más chicos. Para $p = 0.20$ o 0.30, no se observa un patrón definido, si bien pueden identificarse resultados desfavorables para CCA y buenos resultados con MICE- en general y con RVI en la estimación de algunos parámetros (Cuadro 5.9).

Cuando $n = 200$, para una proporción de pérdidas de 0.10, los menores MSE se observan al utilizar MF, MICE+ y RVI, encontrándose resultados pocos favorables al emplear KNN con 3 o 5 datos donantes o CCA. Cuando $p = 0.20$, MF y MICE+ permiten la obtención de los menores MSE, mientras que con CCA, RVI o KNN con 3 datos donantes se encuentran los resultados menos favorables. Al aumentar p a 0.30, MF- brinda los menores valores de MSE, mientras que RVI presenta MSE bajos en la estimación de cinco de los siete parámetros aunque resultan MSE muy altos para los otros dos parámetros estimados. En este escenario, CCA y KNN en general se asocian a los mayores MSE. Para la máxima proporción de datos faltantes evaluada, es decir $p = 0.40$, MICE presenta los MSE más favorables y CCA junto con K3+, los más desfavorables (Cuadro 5.10).

En el escenario con 500 unidades, para los dos valores más bajos de p se observan resultados similares: los menores MSE corresponden a MF, K3- y MICE+, mientras que los mayores MSE se encuentran bajo KNN con 9 vecinos. Cuando $p = 0.30$ o 0.40, MF y MICE+ se asocian a los MSE inferiores, mientras que CCA y KNN con 5 datos donantes presentan los MSE superiores (Cuadro 5.11).

Finalmente, cuando $n = 1000$, para $p = 0.10$ o 0.20, MF y KNN con 3 datos donantes tienen menores MSE en comparación con las otras técnicas, mientras que CCA, RVI y KNN con 9 datos donantes se asocian con los MSE más grandes. Los resultados observados

cuando $p = 0.30$ o $p = 0.40$ son similares a los descriptos para valores de p inferiores, aunque MICE+ se ubica entre las técnicas con menores MSE (Cuadro 5.12).

Cuadro 5.9: Error cuadrático medio de los parámetros del modelo de regresión de Cox para n=100, según método de imputación y proporción de datos faltantes (p).

p	Método	B1	B2	B3	B4	B5	B6	B7
0.1	CCA	0.05322	0.04557	0.10360	0.03997	0.04141	0.00036	0.00044
	RVI	0.00993	0.00809	0.02468	0.00191	0.03198	0.00048	0.00054
	K3-	0.00945	0.01304	0.02711	0.00633	0.06603	0.00034	0.00048
	K3+	0.00607	0.05320	0.15490	0.01328	0.19182	0.00187	0.00288
	K5-	0.01735	0.03023	0.04431	0.00695	0.10292	0.00054	0.00069
	K5+	0.00984	0.03648	0.05507	0.00772	0.20534	0.00105	0.00138
	K9-	0.01146	0.02338	0.04332	0.00762	0.11202	0.00057	0.00078
	K9+	0.01582	0.02209	0.02531	0.00508	0.12412	0.00052	0.00066
	MF-	0.00426	0.00398	0.01122	0.00570	0.02903	0.00012	0.00013
	MF+	0.00563	0.00390	0.01407	0.00553	0.01939	0.00013	0.00019
	MICE-	0.00678	0.01053	0.02010	0.00425	0.07052	0.00024	0.00047
	MICE+	0.00618	0.00800	0.00997	0.00293	0.04281	0.00015	0.00025
0.2	CCA	0.24657	0.33519	24.53688	8.06474	0.13056	0.00128	0.00083
	RVI	0.02267	0.00632	0.00842	0.00501	0.24601	0.00015	0.00031
	K3-	0.01049	0.00529	0.01551	0.00717	0.11837	0.00009	0.00013
	K3+	0.02037	0.01102	0.01311	0.00667	0.10801	0.00012	0.00022
	K5-	0.01357	0.00921	0.01485	0.00660	0.10933	0.00010	0.00016
	K5+	0.02224	0.01201	0.01721	0.00677	0.10879	0.00012	0.00022
	K9-	0.01728	0.01248	0.01197	0.00667	0.14698	0.00012	0.00020
	K9+	0.02351	0.01145	0.01297	0.00571	0.19713	0.00018	0.00033
	MF-	0.00777	0.00633	0.01370	0.00699	0.12956	0.00011	0.00018
	MF+	0.01186	0.00739	0.01601	0.01101	0.07870	0.00023	0.00037
	MICE-	0.00859	0.00555	0.01093	0.00453	0.10448	0.00012	0.00018
	MICE+	0.01075	0.00717	0.01500	0.00743	0.05788	0.00021	0.00029
0.3	CCA	0.19727	0.20008	0.40623	0.09316	0.13921	0.00288	0.00390
	RVI	0.08647	0.00788	0.01973	0.01317	0.34224	0.00011	0.00041
	K3-	0.03313	0.00984	0.03928	0.01611	0.19393	0.00025	0.00074
	K3+	0.05434	0.01344	0.10186	0.02306	0.14943	0.00038	0.00172
	K5-	0.04279	0.00804	0.03614	0.02087	0.25574	0.00026	0.00078
	K5+	0.04231	0.01874	0.03871	0.01758	0.19476	0.00021	0.00062
	K9-	0.05976	0.00801	0.02822	0.01955	0.28873	0.00021	0.00066
	K9+	0.05377	0.01832	0.02922	0.01457	0.21284	0.00011	0.00041
	MF-	0.03413	0.01945	0.03871	0.01398	0.12905	0.00015	0.00037
	MF+	0.04363	0.03054	0.07443	0.02784	0.14604	0.00025	0.00067
	MICE-	0.02961	0.00882	0.02158	0.01260	0.12138	0.00010	0.00035
	MICE+	0.02872	0.01782	0.03632	0.01479	0.09742	0.00025	0.00051

Ref.: Bj=β_j; CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales. En escala de rosas: mayores errores cuadráticos medios, por parámetro. En escala de verdes: menores errores cuadráticos medios, por parámetro.

Cuadro 5.10: Error cuadrático medio de los parámetros del modelo de regresión de Cox para n=200, según método de imputación y proporción de datos faltantes (p).

p	Método	B1	B2	B3	B4	B5	B6	B7
0.1	CCA	0.02023	0.02957	0.03814	0.01999	0.01669	0.00009	0.00008
	RVI	0.00392	0.00150	0.01067	0.00079	0.00510	0.00010	0.00012
	K3-	0.00438	0.00376	0.00844	0.00210	0.03726	0.00009	0.00015
	K3+	0.00305	0.00911	0.02962	0.00496	0.03860	0.00034	0.00056
	K5-	0.00475	0.00800	0.01784	0.00238	0.06313	0.00019	0.00033
	K5+	0.00265	0.01129	0.04035	0.00528	0.05703	0.00040	0.00069
	K9-	0.00259	0.00750	0.00997	0.00172	0.04453	0.00010	0.00015
	K9+	0.00414	0.00700	0.01272	0.00325	0.08300	0.00016	0.00024
	MF-	0.00064	0.00070	0.00102	0.00039	0.00873	0.00001	0.00002
	MF+	0.00080	0.00062	0.00154	0.00033	0.00731	0.00002	0.00004
	MICE-	0.00311	0.00370	0.00355	0.00088	0.02404	0.00003	0.00008
	MICE+	0.00237	0.00218	0.00250	0.00087	0.00980	0.00004	0.00006
0.2	CCA	0.07279	0.11369	21.19642	6.79272	0.05810	0.00053	0.00038
	RVI	0.01755	0.03746	0.01431	0.00197	0.20237	0.00019	0.00027
	K3-	0.00893	0.02413	0.02174	0.00362	0.08010	0.00016	0.00031
	K3+	0.00548	0.01993	0.02037	0.00328	0.06637	0.00019	0.00029
	K5-	0.00530	0.00990	0.00896	0.00278	0.06124	0.00005	0.00009
	K5+	0.00635	0.01109	0.01147	0.00346	0.04906	0.00008	0.00014
	K9-	0.00579	0.01754	0.01979	0.00209	0.07104	0.00011	0.00020
	K9+	0.00679	0.01976	0.02036	0.00167	0.10722	0.00014	0.00023
	MF-	0.00347	0.00480	0.00652	0.00219	0.02446	0.00004	0.00008
	MF+	0.00411	0.00487	0.00697	0.00218	0.02219	0.00005	0.00009
	MICE-	0.00557	0.02385	0.01003	0.00155	0.09658	0.00008	0.00019
	MICE+	0.00460	0.01648	0.00875	0.00204	0.04702	0.00009	0.00017
0.3	CCA	0.16856	0.17049	0.08982	0.05777	0.08367	0.00181	0.00218
	RVI	0.05256	0.00396	0.01147	0.00275	0.30370	0.00005	0.00017
	K3-	0.03492	0.00470	0.05112	0.00325	0.16549	0.00029	0.00111
	K3+	0.01764	0.00616	0.09751	0.01132	0.07868	0.00031	0.00107
	K5-	0.02953	0.00628	0.05763	0.00417	0.15280	0.00031	0.00118
	K5+	0.03113	0.00559	0.13644	0.00831	0.07023	0.00041	0.00174
	K9-	0.02363	0.00785	0.06189	0.00577	0.10982	0.00037	0.00123
	K9+	0.02404	0.00784	0.05198	0.00729	0.16685	0.00023	0.00052
	MF-	0.00862	0.00558	0.01666	0.00408	0.03845	0.00008	0.00021
	MF+	0.00996	0.00628	0.02449	0.00636	0.03768	0.00013	0.00034
	MICE-	0.02712	0.00769	0.01932	0.00460	0.11862	0.00005	0.00024
	MICE+	0.01459	0.00910	0.01182	0.00522	0.05045	0.00021	0.00031
0.4	CCA	0.24438	0.25331	138.72582	38.95994	34.17464	0.01196	0.01359
	RVI	0.06449	0.14975	0.02610	0.03876	1.08978	0.00020	0.00057
	K3-	0.06267	0.16641	0.05123	0.05190	1.03371	0.00031	0.00180
	K3+	0.06957	0.17336	0.06040	0.05503	1.33990	0.00040	0.00223
	K5-	0.06220	0.16475	0.03760	0.05148	1.13292	0.00023	0.00135
	K5+	0.06435	0.16373	0.03930	0.05243	1.35086	0.00027	0.00153
	K9-	0.06940	0.16437	0.03298	0.04942	1.16447	0.00019	0.00109
	K9+	0.06931	0.13594	0.01587	0.05015	1.07739	0.00012	0.00056
	MF-	0.04732	0.15574	0.02542	0.03750	1.42475	0.00031	0.00066
	MF+	0.05615	0.14193	0.04262	0.03905	1.27441	0.00059	0.00126
	MICE-	0.04701	0.15265	0.02125	0.03163	0.86495	0.00011	0.00071
	MICE+	0.02914	0.14067	0.02354	0.02069	0.52182	0.00028	0.00049

Ref.: Bj=β_j; CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales. En escala de rosas: mayores errores cuadráticos medios, por parámetro. En escala de verdes: menores errores cuadráticos medios, por parámetro.

Cuadro 5.11: Error cuadrático medio de los parámetros del modelo de regresión de Cox para n=500, según método de imputación y proporción de datos faltantes (p).

p	Método	B1	B2	B3	B4	B5	B6	B7
0.1	CCA	0.011674	0.018666	0.016624	0.009416	0.004569	0.000038	0.000043
	RVI	0.011159	0.003993	0.009144	0.000585	0.059053	0.000219	0.000207
	K3-	0.001754	0.000632	0.004069	0.000336	0.003027	0.000016	0.000045
	K3+	0.001027	0.001481	0.005842	0.000895	0.003634	0.000075	0.000108
	K5-	0.002064	0.001102	0.006078	0.000811	0.009643	0.000034	0.000074
	K5+	0.001477	0.003209	0.013973	0.002087	0.008320	0.000210	0.000284
	K9-	0.004743	0.005512	0.016492	0.002312	0.045880	0.000262	0.000406
	K9+	0.002788	0.005698	0.014022	0.002262	0.024831	0.000296	0.000416
	MF-	0.000159	0.000113	0.000322	0.000100	0.000877	0.000003	0.000005
	MF+	0.000172	0.000095	0.000239	0.000052	0.000476	0.000003	0.000004
	MICE-	0.002333	0.001770	0.002498	0.000647	0.013134	0.000025	0.000063
	MICE+	0.000818	0.000877	0.001455	0.000694	0.003290	0.000018	0.000021
0.2	CCA	0.024602	0.071881	0.041194	0.015046	0.015896	0.000111	0.000131
	RVI	0.028994	0.012017	0.009046	0.000808	0.174012	0.000109	0.000114
	K3-	0.006737	0.002339	0.005871	0.001324	0.033855	0.000048	0.000127
	K3+	0.004234	0.002443	0.007916	0.002599	0.015512	0.000078	0.000142
	K5-	0.009974	0.006384	0.008858	0.001799	0.056434	0.000068	0.000214
	K5+	0.005964	0.005078	0.017995	0.004003	0.022596	0.000170	0.000307
	K9-	0.015769	0.014535	0.018988	0.002543	0.103095	0.000145	0.000400
	K9+	0.007230	0.006852	0.026835	0.004430	0.045533	0.000204	0.000398
	MF-	0.000999	0.000871	0.000859	0.000684	0.005257	0.000004	0.000011
	MF+	0.001290	0.000816	0.000888	0.000575	0.005044	0.000006	0.000013
	MICE-	0.010532	0.006565	0.010269	0.000715	0.054413	0.000039	0.000164
	MICE+	0.003157	0.003240	0.004252	0.001009	0.012751	0.000024	0.000044
0.3	CCA	0.040280	0.177650	0.059950	0.020850	0.036720	0.000850	0.000920
	RVI	0.046440	0.034510	0.018930	0.001070	0.306500	0.000110	0.000210
	K3-	0.012710	0.017220	0.027960	0.002690	0.160050	0.000160	0.000600
	K3+	0.007180	0.009380	0.037230	0.004920	0.057820	0.000260	0.000670
	K5-	0.018030	0.021830	0.032640	0.003590	0.187950	0.000230	0.000760
	K5+	0.010070	0.009800	0.045220	0.008310	0.059070	0.000330	0.000840
	K9-	0.015400	0.019050	0.026300	0.004000	0.144010	0.000260	0.000680
	K9+	0.012890	0.008810	0.056170	0.014130	0.047530	0.000340	0.000730
	MF-	0.001620	0.002550	0.004630	0.001460	0.017480	0.000020	0.000030
	MF+	0.001320	0.002390	0.004680	0.000950	0.009060	0.000030	0.000040
	MICE-	0.016190	0.020770	0.012470	0.002390	0.164020	0.000040	0.000240
	MICE+	0.003180	0.009050	0.006820	0.004820	0.025670	0.000100	0.000110
0.4	CCA	0.098910	0.316700	0.131350	0.012300	0.231520	0.006100	0.004690
	RVI	0.017230	0.047410	0.045910	0.005910	0.345290	0.000310	0.000420
	K3-	0.017730	0.076060	0.091530	0.017580	0.814320	0.000510	0.002100
	K3+	0.009400	0.060170	0.095250	0.022550	0.476720	0.000630	0.001870
	K5-	0.022210	0.079050	0.125970	0.019200	0.810700	0.000730	0.003040
	K5+	0.014020	0.071920	0.120700	0.020030	0.572440	0.000780	0.002500
	K9-	0.018530	0.069420	0.080660	0.013610	0.462330	0.000440	0.001810
	K9+	0.013330	0.077190	0.111300	0.018270	0.491780	0.000780	0.002510
	MF-	0.007630	0.022660	0.008800	0.004520	0.306750	0.000130	0.000180
	MF+	0.009920	0.017810	0.014230	0.005940	0.200800	0.000240	0.000360
	MICE-	0.015350	0.062110	0.027000	0.006750	0.269580	0.000060	0.000570
	MICE+	0.009310	0.052720	0.007710	0.002230	0.123620	0.000300	0.000180

Ref.: Bj=β_j; CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales. En escala de rosas: mayores errores cuadráticos medios, por parámetro. En escala de verdes: menores errores cuadráticos medios, por parámetro.

Cuadro 5.12: Error cuadrático medio de los parámetros del modelo de regresión de Cox para n=1000, según método de imputación y proporción de datos faltantes (p).

p	Método	B1	B2	B3	B4	B5	B6	B7
0.1	CCA	0.0051417	0.0126120	0.0138996	0.0048945	0.0030414	0.0000243	0.0000237
	RVI	0.0092782	0.0063723	0.0099840	0.0007550	0.0699820	0.0002080	0.0001827
	K3-	0.0000911	0.0000673	0.0001127	0.0000133	0.0007779	0.0000010	0.0000025
	K3+	0.0001154	0.0002303	0.0004598	0.0000940	0.0004686	0.0000067	0.0000103
	K5-	0.0004328	0.0002221	0.0005282	0.0000470	0.0016048	0.0000034	0.0000092
	K5+	0.0005175	0.0009136	0.0018048	0.0003613	0.0018532	0.0000278	0.0000408
	K9-	0.0011225	0.0012151	0.0042686	0.0003657	0.0044983	0.0000126	0.0000446
	K9+	0.0007683	0.0048465	0.0123040	0.0026881	0.0090243	0.0001759	0.0002485
	MF-	0.0000270	0.0000132	0.0000305	0.0000055	0.0000953	0.0000006	0.0000017
	MF+	0.0000042	0.0000029	0.0000055	0.0000006	0.0000117	0.0000001	0.0000002
	MICE-	0.0019535	0.0020067	0.0020072	0.0006419	0.0166434	0.0000193	0.0000527
	MICE+	0.0004224	0.0006892	0.0007879	0.0004607	0.0023701	0.0000065	0.0000086
0.2	CCA	0.0123800	0.0685750	0.0368020	0.0074770	0.0054690	0.0000510	0.0000620
	RVI	0.0255840	0.0106410	0.0178900	0.0005740	0.1556260	0.0002740	0.0002020
	K3-	0.0009490	0.0007310	0.0013040	0.0003240	0.0078730	0.0000090	0.0000220
	K3+	0.0007530	0.0004860	0.0005560	0.0002930	0.0035690	0.0000060	0.0000110
	K5-	0.0026500	0.0018780	0.0037490	0.0007060	0.0183790	0.0000290	0.0000650
	K5+	0.0017600	0.0017140	0.0020600	0.0008000	0.0098150	0.0000290	0.0000480
	K9-	0.0123300	0.0049390	0.0228870	0.0023210	0.0293750	0.0001200	0.0003650
	K9+	0.0044360	0.0048370	0.0167820	0.0052610	0.0129010	0.0001870	0.0003160
	MF-	0.0001560	0.0000640	0.0001900	0.0001050	0.0004980	0.0000040	0.0000090
	MF+	0.0000340	0.0000170	0.0000600	0.0000120	0.0002090	0.0000010	0.0000020
	MICE-	0.0065100	0.0056430	0.0076110	0.0011440	0.0482300	0.0000400	0.0001420
	MICE+	0.0011960	0.0024180	0.0020060	0.0013570	0.0096090	0.0000170	0.0000240
0.3	CCA	0.0154100	0.1673500	0.0265800	0.0122800	0.0160700	0.0003800	0.0003100
	RVI	0.0587300	0.0172000	0.0474700	0.0078500	0.2563400	0.0003100	0.0005400
	K3-	0.0055700	0.0035500	0.0076700	0.0021700	0.0564300	0.0000400	0.0001600
	K3+	0.0038000	0.0027200	0.0077500	0.0024700	0.0250600	0.0000800	0.0001800
	K5-	0.0104700	0.0091700	0.0167100	0.0044900	0.1274600	0.0000900	0.0003600
	K5+	0.0068400	0.0051700	0.0180500	0.0062600	0.0437000	0.0002100	0.0004500
	K9-	0.0202400	0.0188900	0.0466700	0.0089200	0.1733300	0.0002800	0.0009800
	K9+	0.0087500	0.0084800	0.0258300	0.0120900	0.0568400	0.0003000	0.0006700
	MF-	0.0006300	0.0003400	0.0007500	0.0003200	0.0037100	0.0000200	0.0000400
	MF+	0.0004500	0.0002600	0.0005700	0.0001500	0.0014300	0.0000100	0.0000200
	MICE-	0.0189000	0.0128000	0.0162300	0.0052600	0.1251900	0.0000500	0.0002900
	MICE+	0.0065600	0.0073100	0.0054200	0.0031600	0.0184100	0.0000400	0.0000500
0.4	CCA	0.0111600	0.1640500	0.0854400	0.0085700	0.0609900	0.0017400	0.0020700
	RVI	0.0373200	0.0214500	0.0393600	0.0007200	0.2485300	0.0003400	0.0002200
	K3-	0.0197300	0.0239200	0.0453600	0.0035200	0.2418000	0.0003500	0.0009900
	K3+	0.0150100	0.0153500	0.0691700	0.0099300	0.0862700	0.0005500	0.0012300
	K5-	0.0288400	0.0270300	0.0833900	0.0058300	0.2788700	0.0006500	0.0019000
	K5+	0.0186600	0.0179800	0.1074200	0.0113200	0.0862700	0.0007700	0.0018600
	K9-	0.0401900	0.0292700	0.1267600	0.0084100	0.2319400	0.0009600	0.0029400
	K9+	0.0356600	0.0247900	0.1433700	0.0080300	0.1180300	0.0010000	0.0027900
	MF-	0.0024800	0.0034700	0.0049800	0.0017100	0.0401400	0.0000300	0.0000800
	MF+	0.0024500	0.0034600	0.0060000	0.0019500	0.0266500	0.0000300	0.0000700
	MICE-	0.0196300	0.0269300	0.0168900	0.0022300	0.1547500	0.0000400	0.0003300
	MICE+	0.0058500	0.0226700	0.0037900	0.0012400	0.0345500	0.0002300	0.0001500

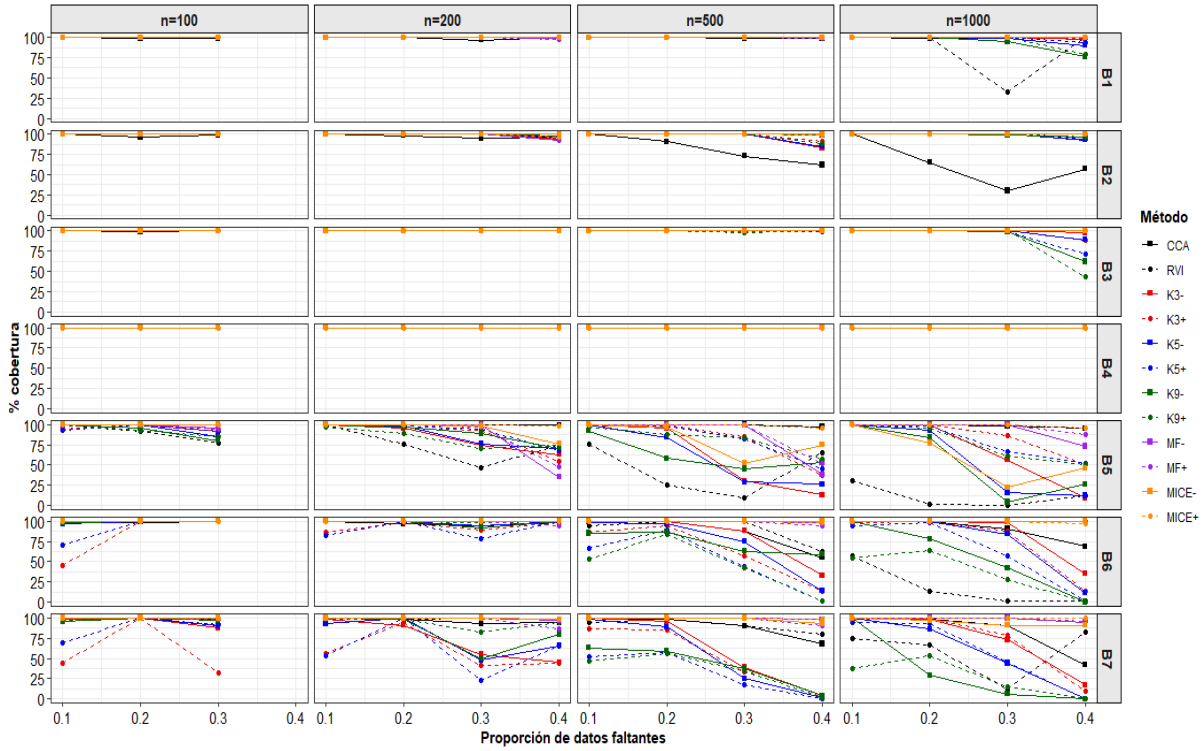
Ref.: Bj=β_j; CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales. En escala de rosas: mayores errores cuadráticos medios, por parámetro. En escala de verdes: menores errores cuadráticos medios, por parámetro.

Cobertura

En la Figura 5.7 se presentan los porcentajes de intervalos de confianza que cubren al verdadero valor, para cada parámetro del modelo y según el tamaño muestral y el porcentaje de pérdida. Para los coeficientes $\beta_1, \beta_2, \beta_3$ y β_4, no se observan diferencias en el porcentaje de cobertura de acuerdo al método usado, siendo cercano al 100 % para todas las proporciones de pérdida p, a excepción de CCA cuya cobertura disminuye a medida que aumenta p.

En el caso de β_5, β_6 y β_7, las diferencias en la cobertura para los distintos métodos aumenta a medida que aumenta la proporción de datos faltantes. RVI y KNN muestran porcentajes de cobertura inferiores al resto de los métodos, siendo menor la cobertura cuando mayor es el valor de k. Sin embargo, no se observa un patrón claro que permita indicar que un determinado método presenta mejores resultados que el resto.

Figura 5.7: Porcentaje de cobertura para los intervalos de confianza del 95 %, según método de imputación, tamaño de muestra y proporción de datos faltantes.



Ref.: Bj=β_j; CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

5.3.3.b. Reproducibilidad de los datos perdidos

En esta sección, se realiza una evaluación de la eficiencia de los métodos de imputación para asignar valores a los datos faltantes, según se trate de observaciones correspondientes a la variable cuantitativa *edad del primer intento de suicidio* o a la variable dicotómica *abuso sexual infantil*, cuando se han generado datos perdidos de acuerdo a un mecanismo MNAR.

Raíz del error cuadrático medio normalizado

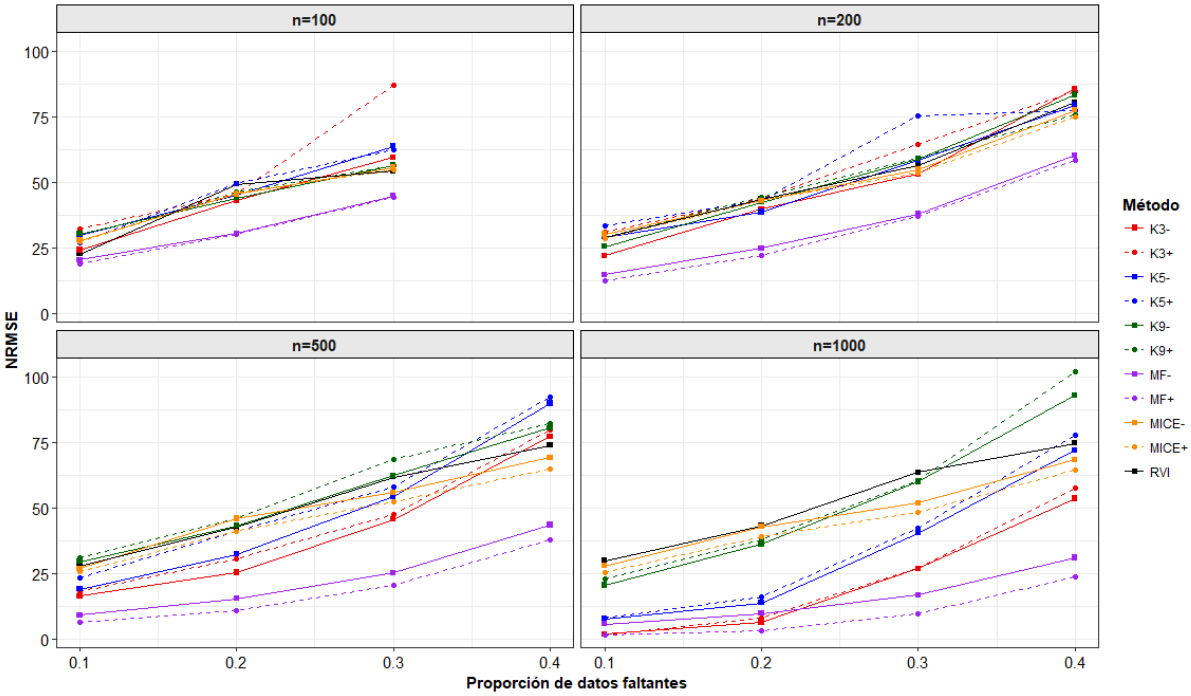
En relación a la imputación de datos para la variable cuantitativa *edad del primer intento de suicidio*, los $\overline{\text{NRMSE}}$ resultantes para los distintos escenarios se presentan en la Figura 5.8.

Se observa que, en general, los menores valores de $\overline{\text{NRMSE}}$ se obtienen utilizando el método MF, mejorando cuando se tienen en cuenta el *estado* y el *tiempo hasta el evento o censura* como variables informativas (MF+). Para $n = 100$ y $n = 200$, se encuentran $\overline{\text{NRMSE}}$ similares para todas las técnicas restantes, dentro de los posibles valores de p.

Cuando $n = 500$ y $p = 0.10$, 0.20 o 0.30, KNN presenta menores $\overline{\text{NRMSE}}$ cuanto menos vecinos más cercanos considera como observaciones donantes. Para dichos valores de p, los $\overline{\text{NRMSE}}$ obtenidos al considerar 9 observaciones donantes se encuentran entre los mayores, junto con los obtenidos al utilizar RVI y MICE. Para el caso de $p = 0.40$, MICE y RVI presentan valores de $\overline{\text{NRMSE}}$ más favorables a los obtenidos con KNN, si bien MF se mantiene como la técnica con menor valor de dicho indicador.

Para el caso de $n = 1000$, para $p = 0.10$, 0.20 o 0.30, el uso de KNN con 9 observaciones donantes, RVI y MICE se asocia con los valores menos favorables de $\overline{\text{NRMSE}}$, si bien estas últimas dos técnicas mejoran comparativamente su *performance* cuando $p = 0.40$, con valores similares a los obtenidos por KNN con 3 o 5 observaciones donantes. Así, KNN con 9 vecinos presenta los $\overline{\text{NRMSE}}$ más altos.

Figura 5.8: Error cuadrático medio normalizado (NRMSE) para *Edad del primer intento de suicidio*, según método de imputación, tamaño de la muestra y proporción de datos faltantes.



Ref.: CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

Porcentaje de no coincidencia

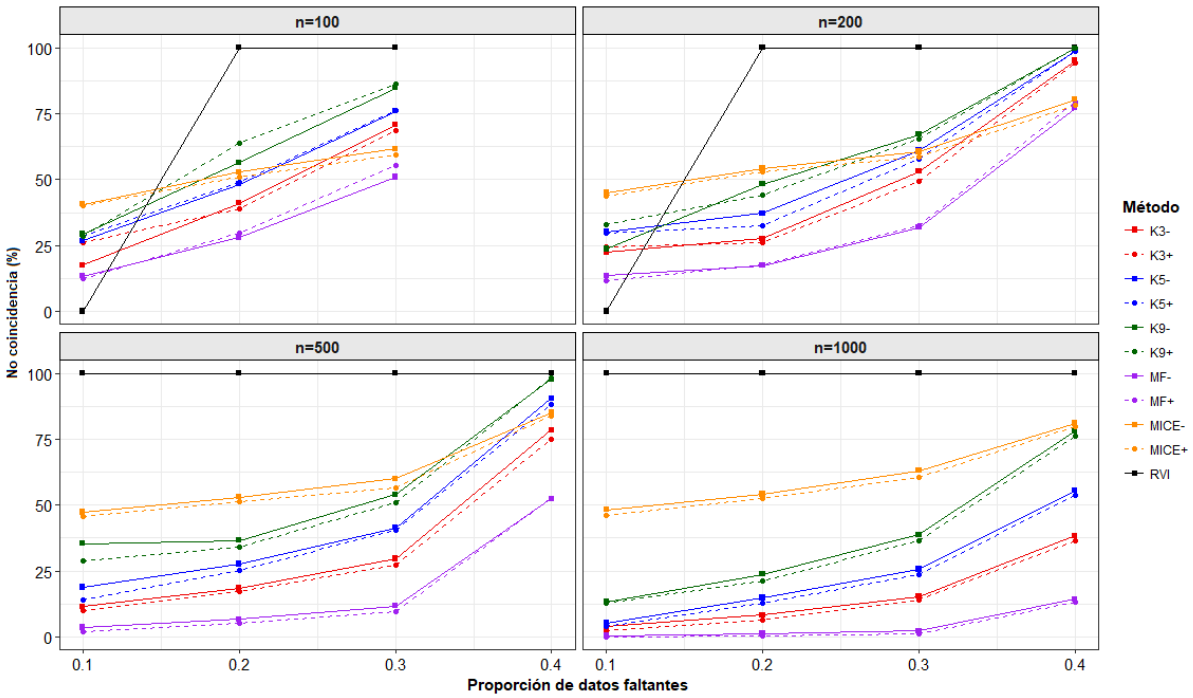
En la Figura 5.9 se muestran los porcentajes medios de datos imputados en forma incorrecta para la variable dicotómica *abuso sexual infantil*, cuando en la misma se provocaron datos perdidos dentro de los casos de respuesta afirmativa.

Se observa que, de forma general, los menores porcentajes de imputaciones realizadas en forma incorrecta se obtienen en el siguiente orden: MF, K3, K5, K9, MICE y RVI, siendo mejores los resultados cuando se tienen en cuenta el *estado* y el *tiempo hasta el evento o censura* como variables informativas.

Dado que RVI asigna a los datos faltantes el valor modal dentro de los observados, es lógico encontrar la totalidad de las imputaciones realizadas en forma incorrecta cuando el porcentaje de datos faltantes es alto. Solo en los casos donde los datos observados conservaron una mayoría de respuestas afirmativas en la variable *abuso sexual infantil* (tal es el caso de $n = 100$ o $n = 200$ y $p = 0.10$) las imputaciones se realizaron en forma completamente correcta.

Para $n = 100$, MICE muestra un menor porcentaje medio de datos imputados en forma incorrecta en comparación con KNN con 9 datos donantes cuando $p = 0.20$ y que KNN en todas sus configuraciones cuando $p = 0.30$, siendo solo mejorado por MF. Similar situación se presenta para $n = 200$, mientras que para $n = 500$ y $p = 0.40$ MICE resulta más favorable que KNN con 5 o 9 observaciones donantes.

Figura 5.9: Porcentaje de datos imputados no coincidentes con los reales para *Abuso Sexual Infantil*, según método de imputación, tamaño de la muestra y proporción de datos faltantes.



Ref.: CCA: análisis de casos completos; RVI: imputación por valores representativos; Kn: k-vecinos cercanos con n donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

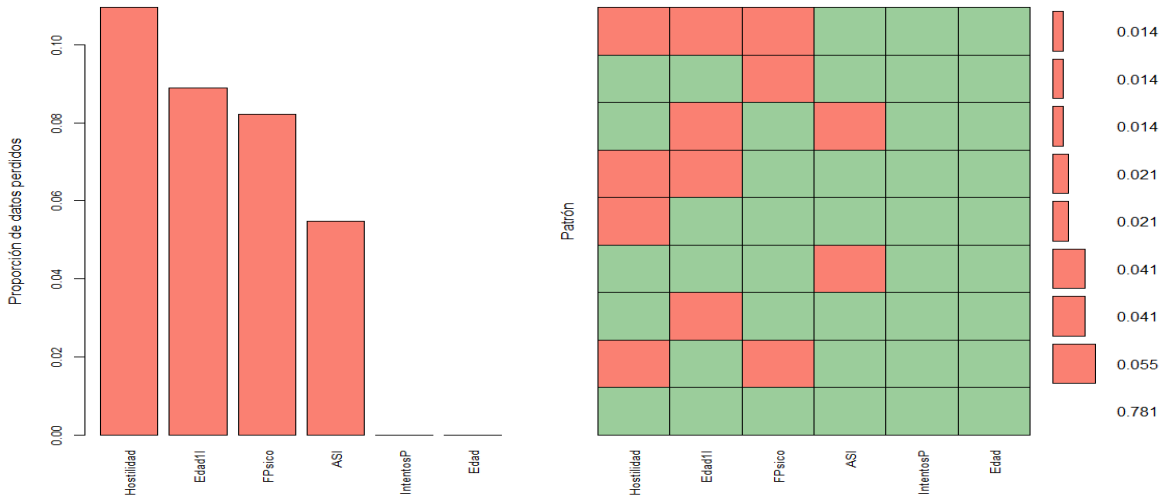
5.4. Análisis del caso estudio

Se retoman los datos correspondientes al caso presentado en la sección 5.1 sobre el estudio del tiempo hasta el reintento de suicidio de pacientes hospitalizados por intento de quitarse la vida. Se realiza un análisis descriptivo en relación a la pérdida de datos y se estima el modelo de regresión de Cox para evaluar la significancia estadística de las posibles variables explicativas del tiempo hasta el evento de interés, según se consideren solo los casos con datos completos o la matriz imputada mediante diversos mecanismos recomendados a partir de los estudios de simulación.

5.4.1. Esquema de datos perdidos

La base de datos consta de información sobre 146 individuos con TLP para los cuales se realizó al menos un seguimiento posterior al ingreso hospitalario y que cuentan con registro del tiempo hasta el evento o censura. De ellos, 114 (78.08 %) presentan datos completos para todas las covariables y 32 (21.92 %) tienen al menos un dato faltante. El porcentaje de pérdidas global para a matriz de información de dimensión 876 (seis variables explicativas medidas sobre 146 pacientes) es del 5.59 %. Las variables con pérdidas son: *hostilidad* (10.95 %), *edad del primer intento de suicidio* (8.90 %), *funcionamiento psicosocial* (8.21 %) y *presencia de antecedentes de abuso sexual infantil* (5.48 %). En la Figura 5.13 se muestra el esquema de pérdidas, indicando la proporción de datos faltantes para cada covariable en forma particular y para la combinación entre ellas.

Cuadro 5.13: Proporción de datos faltantes según covariable y combinación entre covariables con datos faltantes.



Ref.: Edad1I: *edad del primer intento de suicidio*; FPsico: *funcionamiento psicosocial*; ASI: *presencia de antecedentes de abuso sexual infantil*; IntentosP: *número de intentos previos.*

A fin de identificar si las pérdidas observadas en cada variable son dependientes de los valores de las covariables restantes, se presenta un análisis de subgrupos para cada una de ellas. Para cada variables se identifican los grupos de individuos con o sin pérdidas y se los compara respecto de las covariables restantes. Se evalúan medidas estadísticas de resumen y si existen diferencias significativas (Cuadro 5.14), a fin de observar si las pérdidas de una variable están relacionadas con los valores observados de otras. Para comparar

promedios de variables cuantitativas, se utiliza el test t-student y, para comparar frecuencias de variables cualitativas, el test chi-cuadrado. Todos los test se evalúan considerando significativa una probabilidad asociada inferior a 0.05.

Los resultados sugieren una posible relación de las pérdidas de datos para *edad del primer intento* con los valores observados de *funcionamiento psicosocial*, *presencia de antecedentes de abuso sexual infantil* y *número de intentos previos*, así como para las pérdidas en *presencia de antecedentes de abuso sexual infantil* según los valores de *hostilidad* y *funcionamiento psicosocial*. Por lo tanto, se supone un mecanismo de pérdida MAR, no siendo posible evaluar si se corresponde con un mecanismo MNAR.

5.4.2. Estimación del modelo de regresión de Cox

Previo a la estimación del modelo de regresión de Cox, se completa la información faltante en la matriz de datos, eligiendo el método según las recomendaciones derivadas de los capítulos 4 y 5 para las condiciones de este estudio, tamaño de muestra moderado, pérdidas según un mecanismo MAR en un porcentaje bajo, no superior al 11 %. También se comprueba la adecuación del supuesto de riesgos proporcionales utilizando solamente la información disponible en la base de datos original.

Bajo esas condiciones y en relación al error cuadrático medio, se han observado resultados favorables con MICE- y MF+. Si se analizan los resultados para el sesgo de las estimaciones, MF+ se muestra más favorable para las proporciones pequeñas de pérdida mientras que MICE- tiene valores desfavorables en este aspecto. Ambas técnicas son equivalentes a otros métodos en cuanto a la precisión de las estimaciones para porcentajes bajos de pérdida, 0.10 o 0.20. En cuanto a la reproducibilidad de los datos, MF+ ha mostrado buenos resultados, al igual que al momento de la estimación de la probabilidad de sobrevida. Estos aspectos conducen a elegir MF+ y MICE- como métodos de imputación adecuados para este estudio y se agregan, para evaluar estabilidad de los resultados, CCA y un método simple, de eficiencia intermedia, como K5+.

Los métodos CCA y MICE- presentan intervalos de confianza para los coeficientes que son más amplios que los obtenidos para MF+ y K5+ y similares entre sí (Figura 5.10). De este modo, las estimaciones obtenidas bajo MICE- resultan más conservadoras, ya que no se rechaza la hipótesis nula $\beta_i = 0$ en cuatro de los siete parámetros estimados.

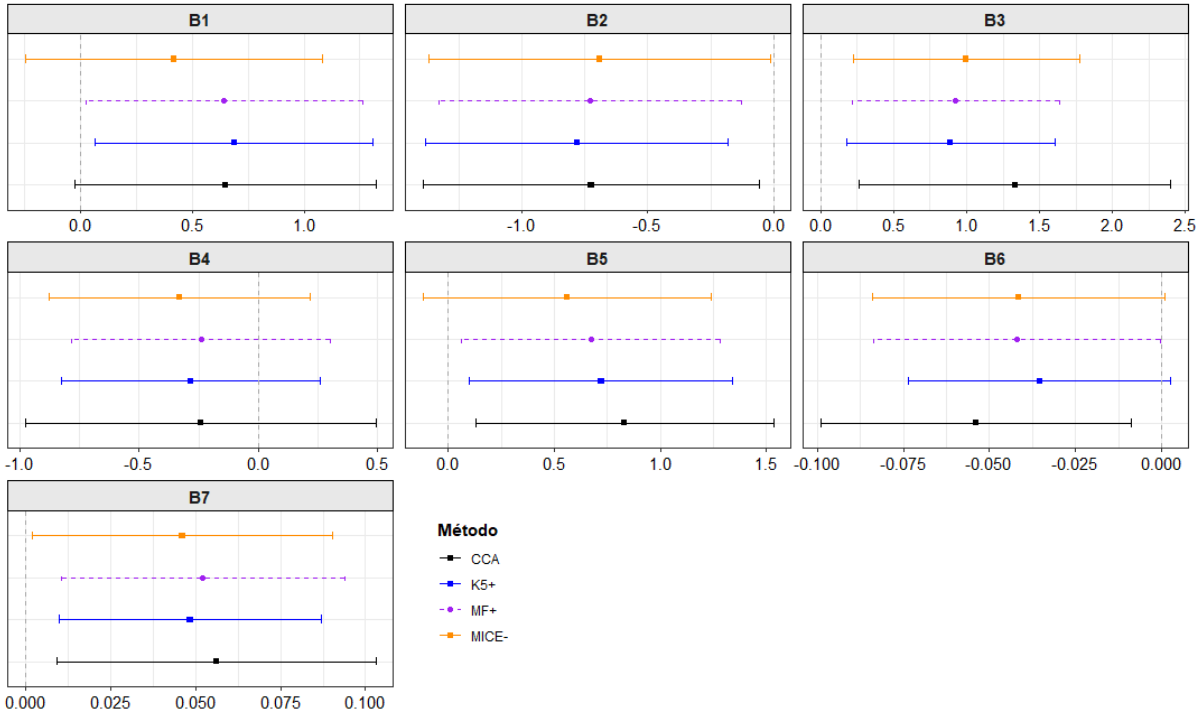
Cuadro 5.14: Variables explicativas según subgrupos de faltantes.

Variable	Hostilidad		Edad primer intento		Funcionamiento Psicosocial		Abuso sexual infantil	
	Observada (n=130, 89.04%)	Perdida (n=16, 10.95%)	Observada (n=133, 91.09%)	Perdida (n=13, 8.91%)	Observada (n=134, 91.78%)	Perdida (n=12, 8.22%)	Observada (n=138, 94.52%)	Perdida (n=8, 5.48%)
Hostilidad	-	-	46.89 +/- 9.54 (n=122)	47.50 +/- 8.57 (n=8)	46.82 +/- 9.49 (n=128)	54.00 +/- 0.00 (n=2)	46.58 +/- 9.58 (n=122)	52.25 +/- 5.26 (n=8)*
Edad	33.05 +/- 10.05 (n=130)	34.31 +/- 5.03 (n=16)	33.26 +/- 9.73 (n=130)	32.46 +/- 8.76 (n=13)	33.10 +/- 9.99 (n=130)	34.08 +/- 3.45 (n=12)	33.29 +/- 9.72 (n=138)	31.25 +/- 8.12 (n=8)
Edad primer intento	24.50 +/- 11.16 (n=122)	24.18 +/- 5.81 (n=11)	-	-	24.39 +/- 11.10 (n=123)	25.50 +/- 6.35 (n=10)	24.28 +/- 10.67 (n=127)	28.50 +/- 13.87 (n=6)
Func. Psicosocial	31.13 +/- 7.05 (n=128)	30.83 +/- 3.06 (n=6)	31.44 +/- 7.04 (n=123)	27.55 +/- 3.98 (n=11)*	-	-	31.44 +/- 6.86 (n=126)	26.00 +/- 6.09 (n=8)*
Abuso sexual infantil (sí)	63/122 (51.64%)	7/16 (43.75%)	61/127 (48.03%)	9/11 (81.82%)*	67/126 (53.17%)	3/12 (25.00%)	-	-
Intentos previos								
Ninguno	26/130 (20.00%)	3/16 (18.75%)	20/133 (15.04%)	9/13 (69.23%)*	26/134 (19.40%)	3/12 (25.00%)	26/138 (18.84%)	3/8 (37.50%)
1 o 2	48/130 (36.92%)	7/16 (47.75%)	52/133 (39.10%)	3/13 (23.08%)	52/134 (38.81%)	3/12 (25.00%)	51/138 (36.96%)	4/8 (50.00%)
3 o más	56/130 (43.08%)	6/16 (37.50%)	61/133 (45.86%)	1/13 (7.69%)	56/134 (41.79%)	6/12 (50.00%)	61/138 (44.20%)	1/8 (12.50%)

Ref.: Los bloques de columnas Observada/Perdida se identifican por subgrupos definidos por la variable correspondiente. Variables cuantitativas: media +/- desvío std. (n=datos observados). Variables cuantitativas: observaciones por categoría/datos observados (porcentaje).*:Diferencia significativa al 5 %.

Respecto a β_4, hay uniformidad en la decisión respecto al no rechazo de la hipótesis nula para todos los métodos. En general, hay mucho acuerdo en las estimaciones puntuales, salvo para β_3, β_5 y β_6 bajo CCA y para β_1, β_4 y β_5 bajo MICE-.

Figura 5.10: Intervalos de confianza del 95 % para los parámetros del modelo de regresión de Cox, según método para tratamiento de datos faltantes.



Ref.: CCA: análisis de casos completos; K5: k-vecinos cercanos con 5 donantes; MF: missForest; MICE: imputación múltiple por ecuaciones encadenadas; +: se incluye la información sobre el *estado* (censura o evento) y el *tiempo hasta alcanzar dicho estado*; -: no se incluyen variables adicionales.

Repasando las propiedades evaluadas en los estudios de los capítulos 4 y 5 y de acuerdo a los resultados encontrados en la presente sección, MF+ parece ser una alternativa razonable para imputar los datos en escenarios similares al del caso estudio presentado, es decir, con tamaño muestral moderado, baja proporción de datos perdidos al azar y con variables explicativas no normales. De todas formas, es recomendable, cuando existe la posibilidad, evaluar los resultados obtenidos bajo diferentes estrategias de manejo de datos faltantes y distinguir si las estimaciones se muestran sensibles a las mismas o, por el contrario, se mantienen estables. Un análisis de los valores obtenidos puede ser útil para distinguir la mejor estrategia a utilizar en cada caso particular. También se destaca como aspecto a tener en cuenta que los intervalos de confianza de la Figura 5.10 provienen de los resultados programados los cuales consideran, como desviación estándar, a los valores de las distribuciones asintóticas de los estimadores, cuando no hay datos faltantes y, por

los estudios realizados, se ha observado que la variabilidad de los estimadores cuando se basan en matrices con datos imputados puede resultar sensiblemente mayor.

6. Discusión y conclusiones

La existencia de datos faltantes es frecuente en investigaciones que involucran la recolección de información con fines estadísticos. Sin embargo, los métodos de análisis empleados requieren, generalmente, que la matriz de datos se encuentre completa. Según la estructura de las pérdidas, un bajo porcentaje global de datos perdidos puede impactar en un alto porcentaje de observaciones con algún dato no disponible. En estudios multivariados, el descartar observaciones con algún dato faltante, puede generar una reducción importante en el tamaño de la muestra que derive en pérdida de precisión en la etapa de inferencia. Estos aspectos han conducido al surgimiento de una línea de investigación en la que se han realizado numerosas propuestas metodológicas que incorporan, en los análisis estadísticos, etapas previas de imputación. Si bien se encuentran publicaciones que evalúan la capacidad de estos métodos a través de simulación computacional, es importante comprender en qué forma la imputación de información afecta a los resultados de los análisis estadísticos que se apliquen sobre ella. Sin embargo, aún quedan interrogantes sin responder, en particular en estudios que involucran mediciones a lo largo del tiempo.

Esta tesis se ha enfocado particularmente en el efecto de distintos métodos para la imputación de datos sobre la estimación de los parámetros de modelos de regresión de Cox con variables de naturaleza mixta, dado que, entre las publicaciones científicas revisadas, no se encontraron trabajos que aborden esta situación de manera completa, considerando diversos escenarios y enfocando las propiedades de los estimadores más allá de la capacidad de los métodos de reproducir los valores perdidos.

En el Capítulo de Antecedentes, se realiza una síntesis de los contenidos de las publicaciones consultadas, realizando una discusión crítica de algunos aspectos tratados por los autores, destacando situaciones no resueltas o abordadas en forma limitada, en relación al tema investigado.

Se plantearon estudios por simulación desde dos enfoques diferentes, a partir de modelos teóricos que consideras situaciones hipotéticas amplias y un estudio por remuestreo a partir de un modelo real surgido de un tema de interés en el área de investigación de la autora de esta tesis. De este modo, se pretende evaluar la consistencia de los resultados derivados según distintos enfoques en los que no hay acuerdo de superioridad.

En el Capítulo 4, mediante datos generados por simulación a partir de un modelo de regresión de Cox con parámetros prefijados, y provocando pérdidas según diferentes mecanismos y proporciones, se ha contrastado la distribución de probabilidad teórica de los estimadores de los parámetros con la distribución empírica obtenida tras la imputación de datos por diferentes mecanismos, no solo evaluando el sesgo de estimación, sino también comparando las desviaciones estándar y la forma de las distribuciones. Hasta el momento, no existían estudios que abordaran estas cuestiones, a pesar de ser aspectos fundamentales para realizar inferencia estadística válida para datos de supervivencia.

Otro aporte significativo de este capítulo es la evaluación de métodos de imputación no incluidos en trabajos anteriores. En general, se ha observado que las publicaciones que comparan técnicas de imputación en el contexto de modelos de Cox, lo hacen utilizando imputación múltiple y análisis de casos completos, coincidiendo con lo apuntado por Siri & Harel [26]. El primer método requiere de supuestos que condicionan su uso o lo vuelven más complejo, ya que requiere un tratamiento previo de los datos en caso de no cumplirse, y esta etapa, generalmente, no constituye el proceso seguido por los analistas. El método de casos completos presenta importantes desventajas, ampliamente señaladas en la bibliografía, aunque se utiliza com comparador. En la tesis, se propone la consideración de métodos de imputación más flexibles, que brinden resultados adecuados sin requerir de configuraciones sofisticadas para los analistas, incluyendo además las técnicas de k-vecinos cercanos (KNN) y *missForest* (MF), las cuales son libres de supuestos y considerando variables con distribución no Normal, lo cual implica el incumplimiento de uno de los supuestos de los modelos para realizar imputaciones múltiples.

Los resultados obtenidos por simulación, en cuanto a eficiencia de los métodos estudiados, son robustos frente a los distintos mecanismos de pérdidas. Respecto del error cuadrático medio, el uso de MICE considerando solo la información de las covariables restantes como explicativas para la imputación de cada variable con pérdidas (MICE-), muestra un buen desempeño en relación a los parámetros correspondientes a las variables

cualitativas y a la única cuantitativa con distribución Normal, mientras que MF permite una mejor estimación de los parámetros asociados a las variables cuantitativas no normales. Sin embargo, la eficiencia de MICE- es menor en relación al sesgo y reproducibilidad de la variabilidad teórica para todos los escenarios para proporciones de pérdidas inferiores a 0.30. Si bien MF, con la inclusión de las variables informativas adicionales (MF+) presenta buenos resultados en cuanto al sesgo de las estimaciones, la diferencia entre la desviación estándar teórica y la empíricas es la mayor cuando las pérdidas son importantes, del 30 % o del 40 %, pudiendo producir un efecto no deseado ya que el intervalo de confianza calculado automáticamente por los programas computacionales tendría amplitud diferente a la real. En general, las estimaciones obtenidas al incluir las variables adicionales resultan menos sesgadas pero con una variabilidad observada mayor a la teórica, provocando así una menor cobertura de los intervalos de confianza del 95 % calculados a partir de la desviación estándar teórica. Esta observación sugiere la necesidad de plantear modificaciones en las estimaciones de la desviación estándar de los estimadores de los parámetros de modelos de Cox, cuando se realizan imputaciones considerando las variables adicionales, para así obtener estimaciones más precisas y exactas. No se detectan alejamientos recurrentes del supuesto de distribución Normal. KNN presenta resultados moderados para los distintos aspectos evaluados, sin grandes diferencias según el número de donantes considerados.

Si bien con KNN, MF y MICE no hay diferencias en los indicadores de bondad de las imputaciones individuales, de acuerdo con la inclusión o no del *estado* y del *tiempo hasta el evento o censura*, la evaluación del impacto sobre las estimaciones de los parámetros sí muestra que la inclusión de las mismas se asocia a resultados más favorables. Este hecho coincide con los reportado en publicaciones donde se evalúa esta variante [17, 30]. Lo mismo se observa al estimar la sobrevida para un tiempo determinado, aspecto para el cual MF+ resultó el método más favorable en la mayoría de los escenarios. En cuanto a este punto, es importante destacar que la estimación puede realizarse para todos los métodos suponiendo conocida la función de riesgo basal, tomando en consideración la que da origen a los datos simulados. Sin embargo, para datos reales donde se desconoce dicha función no es posible llevar a cabo esta comparación utilizando MICE, ya que dicho método permite obtener una estimación amalgamada de los parámetros del modelo pero no del riesgo basal.

En el estudio por remuestreo, los resultados ratifican lo observado a través del proceso de simulación, ya que MF muestra resultados favorables en la mayoría de los caso, especialmente con tamaño de muestra grande ($n = 500$ y 1000).

De este modo, no se encontraron ventajas absolutas ante el uso de un método en particular, lo cual abre las puertas a la consideración de opciones al momento de la imputación de datos. Hasta el momento, la literatura especializada recomienda el uso de imputación múltiple, especialmente en el área de Medicina [25, 103, 104]. Sin embargo, se recomienda prestar especial atención al cumplimiento de los supuestos requeridos por los modelos elegidos para obtener las imputaciones y realizar análisis de sensibilidad en este aspecto [27, 105, 106]. Esto requiere de conocimientos por parte del analista de datos, un trabajo adicional previo a la imputación y mayor dificultad para el tratamiento correspondiente en caso de encontrar que los supuestos no pueden ser aceptados. Por estas razones, es importante considerar métodos alternativos como los propuestos en esta tesis que sean libres de supuestos y resulten más flexibles cuando se pretenden imputar datos en escenarios complejos. Se ha mostrado que MF y KNN proveen resultados aceptables, por lo que se recomienda su consideración en casos similares a los planteados en esta tesis.

El trabajo realizado deja abiertas nuevas líneas de investigación sobre temas no considerados en esta tesis. Entre ellos: evaluar la influencia de estrategias de imputación de datos cuando existen covariables dependientes del tiempo, comprobar el posible efecto de la proporción de datos censurados sobre la eficiencia de las estrategias de imputación y proponer modificaciones sobre las varianzas de los estimadores calculados a partir de matrices con datos imputados en forma exacta o mediante alguna aproximación.

Bibliografía

[1] Geert J. van der Heijden, A. R. Donders, Theo Stijnen, and Karel G. Moons. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: A clinical example. *Journal of Clinical Epidemiology*, 59(10):1102–1109, 2006.

[2] A. R. Donders, G. J. van der Heijden, T. Stijnen, and K. G. Moons. Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, 59(10):1087–91, 2006.

[3] Carme Borrell and Maica Rodríguez-Sanz. Aspectos metodológicos de las encuestas de salud por entrevista: Aportaciones de la Encuesta de Salud de Barcelona 2006. *Revista Brasileira de Epidemiologia*, 11(5):46–57, 2008.

[4] Cali Curley, Rachel M. Krause, Richard Feiock, and Christopher V. Hawkins. Dealing with missing data: A comparative exploration of approaches using the integrated city sustainability database. *Urban affairs review*, 55(2):591–615, 2019.

[5] Marcelle Manley. Psychiatric interview, history, and mental status examination. In *Kaplan & Sadock's Comprehensive Textbook of Psychiatry*, 2004.

[6] Calvin D. Croy and Douglas K. Novins. Methods for addressing missing data in Psychiatric and Developmental Research. *Journal of the American Academy of Child & Adolescent Psychiatry*, 44(12):1230–1240, 2005.

[7] Demián E. Rodante, Leandro N. Grendas, Soledad Puppo, Patricia Vidjen, Alicia Portela, Sasha M. Rojas, Luciana C. Chiapella, and Federico M. Daray. Predictors of short and long term recurrence of suicidal behavior in Borderline Personality Disorder. *Acta Psychiatrica Scandinavica*, 2019.

[8] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[9] James J. Heckman. Sample selection bias as a specification error. *Econometrica: Journal of the Econometric Society*, pages 153–161, 1979.

[10] Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[11] Roderick J. A. Little and Donald B. Rubin. *Statistical analysis with missing data. Second Edition.* Wiley Series in Probability and Statistics. John Wiley & Sons, 2014.

[12] Borja Seijo-Pardo, Amparo Alonso-Betanzos, Kristin P. Bennett, Verónica Bolón-Canedo, Julie Josse, Mehreen Saeed, and Isabelle Guyon. Biases in feature selection with missing data. *Neurocomputing*, 342:97–112, 2019.

[13] María P. Fernández-García, Guillermo Vallejo-Seco, Pablo Livácic-Rojas, and Ellian Tuero-Herrero. The (ir) responsibility of (under) estimating missing data. *Frontiers in Psychology*, 9:556, 2018.

[14] Roderick J. A. Little and Donald B. Rubin. *Statistical analysis with missing data. Third Edition.* Wiley Series in Probability and Statistics. John Wiley & Sons, 2019.

[15] Jae-On Kim and James Curry. The treatment of missing data in multivariate analysis. *Sociological Methods & Research*, 6(2):215–240, 1977.

[16] Rima Houari, Ahcène Bounceur, Abdelkamel Tari, and M. Tahar Kecha. Handling missing data problems with sampling methods. In *2014 International Conference on Advanced Networking Distributed Systems and Applications (INDS)*, pages 99–104. IEEE, 2014.

[17] A. M. G. Ali, S. J. Dawson, F. M. Blows, E. Provenzano, I. O. Ellis, L. Baglietto, D. Huntsman, C. Caldas, and P. D. Pharoah. Comparison of methods for handling missing data on immunohistochemical markers in survival analysis of breast cancer. *British Journal of Cancer*, 104(4):693–699, 2011.

[18] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[19] John C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27(4):857–871, 1971.

[20] Martijn Kagie, Michiel van Wezel, and Patrick J. F. Groenen. An empirical comparison of dissimilarity measures for recommender systems. *ERIM Report Series Research in Management*, 2009.

[21] Daniel J. Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

[22] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[23] Zhongheng Zhang. Multiple imputation with multivariate imputation by chained equation (MICE) package. *Annals of Translational Medicine*, 4(2), 2016.

[24] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–68, 2010.

[25] Alma B. Pedersen, Ellen M. Mikkelsen, Deirdre Cronin-Fenton, Nickolaj R. Kristensen, Tra My Pham, Lars Pedersen, and Irene Petersen. Missing data and multiple imputation in clinical epidemiological research. *Clinical Epidemiology*, 9:157, 2017.

[26] Yulia Sidi and Ofer Harel. The treatment of incomplete data: Reporting, analysis, reproducibility, and replicability. *Social Science & Medicine*, 209:169–173, 2018.

[27] Jonathan A.C. Sterne, Ian R. White, John B. Carlin, Michael Spratt, Patrick Royston, Michael G. Kenward, Angela M. Wood, and James R. Carpenter. Multiple imputation for missing data in Epidemiological and Clinical Research: Potential and pitfalls. *British Medical Journal*, 338:b2393, 2009.

[28] Ian R. White, Patrick Royston, and Angela M. Wood. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(4):377–399, 2011.

[29] Recai Yucel. Impact of the non-distinctness and non-ignorability on the inference by multiple imputation in multivariate multilevel data: A simulation assessment. *Journal of Statistical Computation and Simulation*, 87(9):1813–1826, 2017.

[30] Ian R. White and Patrick Royston. Imputing missing covariate values for the Cox model. *Statistics in Medicine*, 28(15):1982–1998, 2009.

[31] Stef van Buuren, Hendriek C. Boshuizen, and Dick L. Knook. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, 18(6):681–694, 1999.

[32] Chiu-Hsieh Hsu and Mandi Yu. Cox regression analysis with missing covariates via nonparametric multiple imputation. *Statistical Methods in Medical Research*, 28(6):1676–1688, 2018.

[33] Lihong Qi, Ying-Fang Wang, and Yulei He. A comparison of multiple imputation and fully augmented weighted estimators for Cox regression with missing covariates. *Statistics in Medicine*, 29(25):2592–2604, 2010.

[34] Andrea Marshall, Douglas G. Altman, Patrick Royston, and Roger L. Holder. Comparison of techniques for handling missing covariate data within prognostic modelling studies: A simulation study. *BMC Medical Research Methodology*, 10(1):7, 2010.

[35] Federica Barzi and Mark Woodward. Imputations of missing values in practice: Results from imputations of serum cholesterol in 28 cohort studies. *American Journal of Epidemiology*, 160(1):34–45, 2004.

[36] José M. Jerez, Ignacio Molina, Pedro J. García-Laencina, Emilio Alba, Nuria Ribelles, Miguel Martín, and Leonardo Franco. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine*, 50(2):105–115, 2010.

[37] Taane G. Clark and Douglas G. Altman. Developing a prognostic model in the presence of missing data: An ovarian cancer case study. *Journal of Clinical Epidemiology*, 56(1):28–37, 2003.

[38] David Collett. *Modelling survival data in Medical Research*. Chapman and Hall/CRC, 2015.

[39] Paul D. Allison. *Survival analysis using SAS: a practical guide*. Sas Institute, 2010.

[40] David R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[41] John D. Kalbfleisch and Ross L. Prentice. *The statistical analysis of failure time data*. John Wiley & Sons, 2011.

[42] Norman Breslow. Covariance analysis of censored survival data. *Biometrics*, 30(1):89–99, 1974.

[43] Bradley Efron. The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, 72(359):557–565, 1977.

[44] David R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

[45] Anastasios A. Tsiatis. A large sample study of Cox's regression model. *The Annals of Statistics*, 9(1):93–108, 1981.

[46] Tormod Næs. The asymptotic distribution of the estimator for the regression parameter in Cox's regression model. *Scandinavian Journal of Statistics*, 9(2):107–115, 1982.

[47] Kent R. Bailey. The asymptotic joint distribution of regression and survival parameter estimates in the Cox regression model. *The Annals of Statistics*, 11(1):39–48, 1983.

[48] Roderick J. A. Little and Donald B. Rubin. *Statistical analysis with missing data*. Wiley Series in Probability and Statistics. John Wiley & Sons, 1987.

[49] H.O. Hartley and R.R. Hocking. The analysis of incomplete data. *Biometrics*, 27(4):783–823, 1971.

[50] Carl Finkbeiner. Estimation for the multiple factor model when data are missing. *Psychometrika*, 44(4):409–420, 1979.

[51] Gustavo E.A.P.A. Batista and Maria C. Monard. A study of k-nearest neighbour as an imputation method. In *Hybrid Intelligent Systems*, pages 251–260. Front Artificial Intelligence Applications, 2002.

[52] Paul D. Allison. *Missing data*. Sage Publications, 2001.

[53] Rebecca R. Andridge and Roderick J.A. Little. A review of hot deck imputation for survey non-response. *International Statistical Review*, 78(1):40–64, 2010.

[54] Craig K. Enders. *Applied missing data analysis*. Guilford Press, 2010.

[55] Harvey Goldstein, James Carpenter, Michael G. Kenward, and Kate A. Levin. Multilevel models with multivariate mixed response types. *Statistical Modelling*, 9(3):173–197, 2009.

[56] Sander Greenland and William D. Finkle. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology*, 142(12):1255–1264, 1995.

[57] Donald B. Rubin. *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, 2004.

[58] Craig K. Enders. Multiple imputation as a flexible tool for missing data handling in Clinical Research. *Behaviour Research and Therapy*, 98:4–18, 2017.

[59] Michael Spratt, James Carpenter, Jonathan A.C. Sterne, John B. Carlin, Jon Heron, John Henderson, and Kate Tilling. Strategies for multiple imputation in longitudinal studies. *American journal of epidemiology*, 172(4):478–487, 2010.

[60] Robert E. Fay. *When Are Inferences from Multiple Imputation Valid?* US Census Bureau, 1992.

[61] Vincent Audigier, François Husson, and Julie Josse. Multiple imputation for continuous variables using a Bayesian principal component analysis. *Journal of Statistical Computation and Simulation*, 86(11):2140–2156, 2016.

[62] Shahab Jolani. Hierarchical imputation of systematically and sporadically missing data: An approximate Bayesian approach using chained equations. *Biometrical Journal*, 60(2):333–351, 2018.

[63] Trevor Hastie, Rahul Mazumder, Jason D. Lee, and Reza Zadeh. Matrix completion and low-rank SVD via fast alternating least squares. *The Journal of Machine Learning Research*, 16(1):3367–3402, 2015.

[64] Jeroen K. Vermunt, Joost R. Van Ginkel, L. Andries Van der Ark, and Klaas Sijtsma. Multiple imputation of incomplete categorical data using Latent Class Analysis. *Sociological Methodology*, 38(1):369–397, 2008.

[65] Julie Josse, Marie Chavent, Benot Liquet, and François Husson. Handling missing values with regularized iterative multiple correspondence analysis. *Journal of classification*, 29(1):91–116, 2012.

[66] Pier A. Ferrari, Paola Annoni, Alessandro Barbiero, and Giancarlo Manzi. An imputation method for categorical variables with application to nonlinear principal component analysis. *Computational Statistics & Data Analysis*, 55(7):2410–2420, 2011.

[67] Davide Vidotto, Jeroen K. Vermunt, and Katrijn van Deun. Bayesian multilevel latent class models for the multiple imputation of nested categorical data. *Journal of Educational and Behavioral Statistics*, 43(5):511–539, 2018.

[68] Samuel S. Wilks. Moments and distributions of estimates of population parameters from fragmentary samples. *The Annals of Mathematical Statistics*, 3(3):163–195, 1932.

[69] R.R. Sokal and C.D. Michener. A statistical method for evaluating systematic relationships. *Bull*, 38:1409–1438, 1958.

[70] Paul Jaccard. Nouvelles recherches sur la distribution florale. *Bulletin de la Société vaudoise des Sciences Naturelles*, 44:223–270, 1908.

[71] Shyam Boriah, Varun Chandola, and Vipin Kumar. Similarity measures for categorical data: A comparative evaluation. In *Proceedings of the 2008 SIAM International Conference on Data Mining*, pages 243–254. SIAM, 2008.

[72] Brendan McCane and Michael Albert. Distance functions for categorical and mixed variables. *Pattern Recognition Letters*, 29(7):986–993, 2008.

[73] D. Randall Wilson and Tony R. Martinez. Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research*, 6:1–34, 1997.

[74] Lorenzo Beretta and Alessandro Santaniello. Nearest neighbor imputation algorithms: A critical evaluation. *BMC Medical Informatics and Decision Making*, 16(3):74, 2016.

[75] Michelle H. Cartwright, Martin J. Shepperd, and Qinbao Song. Dealing with missing software project data. In *Proceedings. 5th International Workshop on Enterprise Networking and Computing in Healthcare Industry (IEEE Cat. No. 03EX717)*, pages 154–165. IEEE, 2003.

[76] Ofer Harel and Xiao-Hua Zhou. Multiple imputation: Review of theory, implementation and software. *Statistics in Medicine*, 26(16):3057–3077, 2007.

[77] Joseph L. Schafer. *Analysis of incomplete multivariate data*. Chapman and Hall/CRC, 1997.

[78] Arthur B. Kennickell. Imputation of the 1989 Survey of Consumer Finances: Stochastic relaxation and multiple imputation. In *Proceedings of the Survey Research Methods Section of the American Statistical Association*, volume 1, page 41, 1991.

[79] Jaap J.P.L. Brand. *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets*. Thesis Erasmus University Rotterdam, 1999.

[80] Trivellore E. Raghunathan, James M. Lepkowski, John Van Hoewyk, Peter Solenberger, et al. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27(1):85–96, 2001.

[81] David Heckerman, David M. Chickering, Christopher Meek, Robert Rounthwaite, and Carl Kadie. Dependency networks for inference, collaborative filtering, and data visualization. *Journal of Machine Learning Research*, 1(Oct):49–75, 2000.

[82] Donald B. Rubin. Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica*, 57(1):3–18, 2003.

[83] Andrew Gelman. Parameterization and Bayesian modeling. *Journal of the American Statistical Association*, 99(466):537–545, 2004.

[84] Stef van Buuren. *Multivariate imputation by chained equations: MICE V1. 0 user's manual*. Leiden: TNO, 2000.

[85] Stef van Buuren. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16(3):219–242, 2007.

[86] Gerko Vink, Laurence E. Frank, Jeroen Pannekoek, and Stef van Buuren. Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica*, 68(1):61–90, 2014.

[87] Nathaniel Schenker and Jeremy M.G. Taylor. Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis*, 22(4):425–446, 1996.

[88] Tim P. Morris, Ian R. White, and Patrick Royston. Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology*, 14(1):75, 2014.

[89] Andrea Marshall, Douglas G. Altman, and Roger L. Holder. Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: A resampling study. *BMC medical research methodology*, 10(1):112, 2010.

[90] James M. Robins, Andrea Rotnitzky, and Lue P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

[91] Anoop D. Shah, Jonathan W. Bartlett, James Carpenter, Owen Nicholas, and Harry Hemingway. Comparison of random forest and parametric imputation models for imputing missing data using mice: a caliber study. *American journal of epidemiology*, 179(6):764–774, 2014.

[92] RStudio Team. *RStudio: Integrated Development Environment for R.* RStudio, Inc., Boston, MA, 2016.

[93] Jonathan Kropko and Jeffrey J. Harden. *coxed: Duration-Based Quantities of Interest for the Cox Proportional Hazards Model*, 2019. R package version 0.3.0.

[94] Jeffrey J. Harden and Jonathan Kropko. Simulating duration data for the Cox model. *Political Science Research and Methods*, 7(4):921–928, 2019.

[95] Matthias Templ, Andreas Alfons, Alexander Kowarik, and Bernd Prantner. VIM: visualization and imputation of missing values. *R package version*, 2(3), 2011.

[96] Daniel J. Stekhoven. Package 'missforest': Nonparametric Missing Value Imputation using Random Forest. *Swiss Federal Institute of Technology, Zürich, Switzerland*, 2013.

[97] Leandro N Grendas, Sasha M Rojas, Soledad Puppo, Patricia Vidjen, Alicia Portela, Luciana Chiapella, Demián E Rodante, and Federico M Daray. Interaction between prospective risk factors in the prediction of suicide risk. *Journal of affective disorders*, 258:144–150, 2019.

[98] M. Villar García, J. F. Pérez Prieto, M. Hernández Viadel, M. Renovell Farré, C. Leal Cercos, and M. Gómez Beneyto. Preparation of a SCID-II-based diagnostic tool for personality disorders. Spanish version. Translation and adaptation. *Actas luso-españolas de Neurología, Psiquiatría y ciencias afines*, 23(4):178–183, 1995.

[99] Viviana Montalván, Ruth Graver, María A. Oquendo, Enrique Baca-García, Miguel Morales, and John Mann. Spanish adaptation of the Buss-Durkee Hostility Inventory (BDHI). *European Journal of Psychiatry*, 15(2):101–112, 2001.

[100] Arnold H. Buss and Ann Durkee. An inventory for assessing different kinds of hostility. *Journal of Consulting Psychology*, 21(4):343, 1957.

[101] M. Bosc, A. Dubini, and V. Polin. Development and validation of a social functioning scale, the Social Adaptation Self-evaluation Scale. *European Neuropsychopharmacology*, 7(1):S57–S70, 1997.

[102] J. Bobes, M. P. Gonzalez, M. T. Bascaran, A. Corominas, A. Adan, J. Sánchez, and P. Such. Validation of the Spanish version of the Social Adaptation Scale in depressive patients. *Actas españolas de Psiquiatría*, 27(2):71–80, 1999.

[103] Joost R. van Ginkel, Marielle Linting, Ralph C.A. Rippe, and Anja van der Voort. Rebutting existing misconceptions about multiple imputation as a method for handling missing data. *Journal of Personality Assessment*, pages 1–12, 2019.

[104] Roderick J. Little, Ralph D'Agostino, Michael L. Cohen, Kay Dickersin, Scott S. Emerson, John T. Farrar, Constantine Frangakis, Joseph W. Hogan, Geert Molen-

berghs, Susan A. Murphy, et al. The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, 367(14):1355–1360, 2012.

[105] Erik Von Elm, Douglas G. Altman, Matthias Egger, Stuart J. Pocock, Peter C. Gøtzsche, Jan P. Vandenbroucke, für die STROBE Initiative, et al. Das strengthening the reporting of observational studies in epidemiology (STROBE-) statement. *Notfall+ Rettungsmedizin*, 11(4):260, 2008.

[106] Cattram D. Nguyen, John B. Carlin, and Katherine J. Lee. Model checking in multiple imputation: an overview and case study. *Emerging themes in Epidemiology*, 14(1):8, 2017.

[107] N. Solaro, A. Barbiero, G. Manzi, and P.A. Ferrari. A simulation comparison of imputation methods for quantitative data in the presence of multiple data patterns. *Journal of Statistical Computation and Simulation*, 88(18):1–32, 2018.

[108] Stef Van Buuren. *Flexible imputation of missing data*. Chapman and Hall/CRC, 2018.

Anexo I: Pseudo-códigos correspondientes a los métodos de imputación

Algoritmo 1 k-vecinos cercanos (adaptación de [21])

Entrada: Matriz de datos $\boldsymbol{X}$ de tamaño $n \times r$, distancia a calcular entre unidades, número de vecinos k, medida de agregación según tipo de variable a imputar.

1: **para** $j = 1$ **to** r **hacer**
2: **para todo** x_{ij} con $i \in (1, \cdots, n)$, tal que x_{ij} es un dato faltante **hacer**
3: **para todo** $i' \in (1, \cdots, i-1, i+1, \cdots, n)$ **hacer**
4: Calcular la distancia $d_{ii'}$ entre $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{i(j-1)}, x_{i(j+1)}, \cdots, x_{ir})$ y $\boldsymbol{x}_{i'} = (x_{i'1}, \cdots, x_{i'(j-1)}, x_{i'(j+1)}, \cdots, x_{i'r})$.
5: **fin para**
6: Seleccionar las k unidades con menor distancia a $\boldsymbol{x}_i$.
7: Calcular sobre las k unidades seleccionadas la medida de agregación adecuada respecto a la variable X_j. **Salida:** x^*_{ij}.
8: Imputar x_{ij} mediante x^*_{ij}.
9: **fin para**
10: **fin para**

Algoritmo 2 missForest (adaptación de [21, 107])

Entrada: Matriz de datos $\boldsymbol{X}$, de tamaño $n \times r$, criterio de convergencia γ, $k = 0$.

1: Conformar el vector $\boldsymbol{s}$ de subíndices de columnas de $\boldsymbol{X}$ ordenados de acuerdo con la cantidad de datos faltantes, comenzando por el índice de la columna con menor cantidad de datos faltantes.
2: Imputar mediante algún algoritmo simple los valores faltantes de $\boldsymbol{X}$, obteniendo $\boldsymbol{X}^{*(0)}$.
3: **mientras** no se cumpla γ **hacer**
4: Asignar a $\boldsymbol{X}$ la matriz $\boldsymbol{X}^{*(k)}$.
5: $k = k + 1$
6: **para todo** $j \in \boldsymbol{s}$ **hacer**
7: Ajustar Random Forest considerando como respuesta $\boldsymbol{x}_j^{obs}$ y como predictores $\boldsymbol{X}_{-j}^{obs}$ para obtener el Random Forest entrenado.
8: Imputar los valores $\boldsymbol{x}_j^{miss}$ aplicando el modelo Random Forest entrenado a $\boldsymbol{X}_{-j}^{miss}$. **Salida:** vector con valor imputados $\boldsymbol{x}_j^{*(k)}$.
9: **fin para**
10: Conformar la matriz $\boldsymbol{X}^{*(k)} = \left(\boldsymbol{x}_1^{*(k)}, \cdots, \boldsymbol{x}_r^{*(k)}\right)$ con los valores predichos.
11: Actualizar γ.
12: **fin mientras**

Salida: Matriz de datos completa $\boldsymbol{X}^{*(k)}$.

Algoritmo 3 Imputación múltiple por ecuaciones encadenadas (adaptación de [108])

Entrada: Matriz de datos $\boldsymbol{X}$, de tamaño $n \times r$, s: número de iteraciones, m número de imputaciones, modelo predictivo para cada tipo de variable a imputar.

Imputar los datos faltantes de $\boldsymbol{X}$ mediante algún algoritmo simple. Se obtiene $\boldsymbol{X}^{(0)}$.
para $q = 1$ **to** m **hacer**
  para $t = 1$ **to** s **hacer**
    para $j = 1$ **to** r **hacer**
      Definir $\boldsymbol{X}_{-j}^{obs} = \left(\boldsymbol{x}_1^{(t)}, \cdots, \boldsymbol{x}_{j-1}^{(t)}, \boldsymbol{x}_{j+1}^{(t-1)}, \cdots, \boldsymbol{x}_r^{(t-1)}\right)$.
      Estimar $\boldsymbol{\beta}_j$ considerando $\boldsymbol{x}_j^{obs}$ vector respuesta y $\boldsymbol{X}_{-j}^{obs}$, matriz de variables explicativas, según el modelo adecuado para $\boldsymbol{X}_j$. **Salida:** $\hat{\boldsymbol{\beta}}_j$, $\hat{\boldsymbol{V}}_j$ (matriz de covarianzas de $\hat{\boldsymbol{\beta}}_j$) y $\hat{\sigma}_j$ (raiz cuadrada del error cuadrático medio).
      Hacer $\sigma^* = \hat{\sigma}\sqrt{\frac{(n^{obs}-r)}{g}}$, donde n^{obs} es el número de datos observados para la variable X_j, r es la cantidad de parámetros del modelo estimado y g es un valor aleatorio obtenido a partir de una distribución χ^2 de $n^{obs} - r$ grados de libertad.
      Hacer $\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}} + \frac{\sigma^*}{\hat{\sigma}}\boldsymbol{u}_1\hat{\boldsymbol{V}}^{\frac{1}{2}}$, donde $\boldsymbol{u}_1$ es un vector fila de r valores independientes seleccionados al azar a partir de una distribución Normal estándar y $\hat{\boldsymbol{V}}^{\frac{1}{2}}$ es la descomposición de Cholesky de $\hat{\boldsymbol{V}}$.
      para todo i tal que x_{ij} es faltante **hacer**
        $x_{ij}^* = \boldsymbol{\beta}^*\boldsymbol{y}_i + u_{2i}\sigma^*$, donde u_{2i} es un valor obtenido al azar a partir de una distribución Normal estándar y $\boldsymbol{y}_i$ es la fila de $\boldsymbol{X}_{-j}^{obs}$ correspondiente a la i-ésima unidad.
        Imputar x_{ij}^{miss} con x_{ij}^*.
      fin para
      Hacer $\boldsymbol{x}_j^{(t)} = \left(\boldsymbol{x}_j^{*(t)}, \boldsymbol{x}_j^{obs}\right)$.
    fin para
  fin para
  Hacer $\boldsymbol{X}^{*(q)} = \left(\boldsymbol{X}^{*(s)}, \boldsymbol{X}^{obs}\right)$.
  Estimar los parámetros de interés del vector $\boldsymbol{Q}$ mediante $\boldsymbol{X}^{*(q)}$. **Salida:** $\hat{\boldsymbol{Q}}^{(q)}$.
fin para
Amalgamar las estimaciones: $\bar{\boldsymbol{Q}} = \frac{1}{m}\sum_{k=1}^{m}\hat{\boldsymbol{Q}}^{(k)}$
Calcular varianza intra-imputación: $\boldsymbol{W} = \frac{1}{m}\sum_{k=1}^{m}\boldsymbol{W}^{(k)}$
Calcular varianza entre-imputaciones: $\boldsymbol{B} = \frac{1}{(m-1)}\sum_{k=1}^{m}(\hat{\boldsymbol{Q}}^{(k)} - \bar{\boldsymbol{Q}})^2$.
Calcular la varianza de las estimaciones: $var(\bar{\boldsymbol{Q}}) = \boldsymbol{W} + \left(1 + \frac{1}{m}\right)\boldsymbol{B}$.

Salida: Estimación de los parámetros $\bar{\boldsymbol{Q}}$ y su varianza $var(\bar{\boldsymbol{Q}})$.

Anexo II: Código utilizado para simulaciones

Este programa se estructura en 5 partes, para cuya ejecución se selecciona la parte 3 de acuerdo al mecanismo de pérdida que se desea evaluar. Las partes 2, 3 y 4 se ejecutan dentro de un proceso iterativo en el que varía el índice s desde 1 hasta el número de simulaciones a realizar (*sim*, 5000 en este caso).

Parte 1: Configuración

```
#Carga los paquetes a utilizar
paquetes <- c("readxl", "dplyr", "survival", "hydroGOF", "broom", "VIM",
              "missForest", "mice", "coxed")
lapply(paquetes, require, character.only = TRUE)

n <- 100 #Tamano de la base #Cambiar a 200, 500 y 1000
sim <- 5000 #Numero de simulaciones
prop <- 0.1 #Proporcion de datos faltantes #Cambiar a 0.2, 0.3 y 0.4
set.seed(276) #Semilla de arranque. # =276 si p=0.1, =123 si p=0.2
# = 127 si p=0.3 e =777 si p=0.4

cuali <- c(2,5) #Vector de indices de variables cualitativas
cuanti <- c(1,3,4) #Vector de indices de variables cuantitativas

#Prepara matrices para guardar resultados de las simulaciones
estim <- array(0,c(sim,6,12))
InfCI <- array(0,c(sim,6,12))
SupCI <- array(0,c(sim,6,12))
nrmse <- array(0,c(sim,length(cuanti),12))
MICE_nrmse <- array(0,c(5,length(cuanti),1))
meannrmse <- array(0,c(12,length(cuanti)))
ecm <- array(0,c(12,6))
cobertura <- array(0,c(12,6))
MICE_FCE <- array(0,c(5,length(cuali),1))
FCE <- array(0,c(sim,length(cuali),12))
meanFCE <- array(0,c(12,length(cuali)))
times <- array(NA,c(sim,12))
meantimes <- array(NA,c(1,12))

par <- c("B1","B2","B3","B4","B5","B6")
metodo <- c("CCA","RVI","K3-","k3+","K5-","k5+","k9-","k9+",
            "MF-","MF+","MICE-","MICE+")

dimnames(estim) <- list(c(1:sim),par,metodo)
dimnames(InfCI) <- list(c(1:sim),par,metodo)
dimnames(SupCI) <- list(c(1:sim),par,metodo)
dimnames(nrmse) <- list(c(1:sim),cuanti,metodo)
dimnames(MICE_nrmse) <- list(c(1:5),cuanti)
dimnames(meannrmse) <- list(metodo,cuanti)
dimnames(ecm) <- list(metodo,par)
```

```
dimnames(cobertura) <- list(metodo,par)
dimnames(MICE_FCE) <- list(c(1:5),cuali)
dimnames(FCE) <- list(c(1:sim),cuali, metodo)
dimnames(meanFCE) <- list(metodo, cuali)

colnames(times) <- metodo
colnames(meantimes) <- metodo
```

Parte 2: Generación de matrices de datos simulados

```
X1 <- rnorm(n = n, mean = 0, sd = 1)
X2 <- rbinom(n=n, size=1, prob=0.3)
X3 <-  runif(n,0,10)
X4 <- rexp(n, rate = 1)
y <- c(1,2,3)
X5 <- colSums(rmultinom(prob=c(0.3,0.2,0.5),size=1, n=n)*y)
Xa <- ifelse(X5==2,1,0)
Xb <- ifelse(X5==3,1,0)
simdata <- cbind(X1,X2,X3,X4,Xa,Xb)

my.hazard <- function(t){
  dnorm((log(t) - log(50))/log(10)) /
    (log(10)*t*(1 - pnorm((log(t) - log(50))/log(10))))
}

simdata2 <- sim.survdata(T=1000, X=simdata, num.data.frames = 1, censor=0.2,
                         beta=c(-0.10,0.25,-0.13,0.50,0.35,-0.15), hazard.fun=my.hazard)
v <- simdata2$data

coef <- c(-0.10,0.25,-0.13,0.50,0.35,-0.15)
coefcorr <- matrix(t(coef),nrow=sim,ncol=6,byrow=T)

for (j in 1:n){
  ifelse(simdata2$data$Xa[j]==1, simdata2$data$X5[j] <- 2,
         ifelse (simdata2$data$Xb[j]==1, simdata2$data$X5[j] <- 3, simdata2$data$X5[j]
    <- 1))
}
cox <- coxph(Surv(y, failed) ~ X1+X2+X3+X4+Xa+Xb, data = simdata2$data, method="breslow")

data <- as.data.frame(cbind(simdata2$data[,c(1:4,9,7:8)]))
data$X2 <- as.factor(data$X2)
data$X5 <- as.factor(data$X5)

data2 <- data #Database con missing
miss <- prop*n #Nro. de datos a eliminar por variable
miss <- round(miss,0)
```

Parte 3: Generación de datos perdidos

a. Mecanismo de pérdida: MCAR

```
for (k in 1:5){ #Para cada variable
  i <- 1
  while (i < miss+1) { #Repite hasta alcanzar el nro. de missing en c/var.
    j <- sample(1:n, 1) #Elige los ID al azar
    if (is.na(data2[j,k]) == 'FALSE') {data2[j,k] <- NA #Controla que no sea NA
    i <- i+1}  #Lleva el conteo de los NA generados
  }}
```

b. Mecanismo de pérdida: MAR

```
NoX2 <- subset(data,data$X2==0)
  dim <- dim(NoX2)[1]
  SiX2 <- subset(data,data$X2==1)
  i <- 1
  while (i < miss+1) { #Repite hasta alcanzar el nro. de missing en c/variable
    j <- sample(1:dim, 1) #Elige los ID al azar
    if (is.na(NoX2$X1[j]) == 'FALSE') {NoX2$X1[j] <- NA #Controla que no sea NA
    i <- i+1}  #Lleva el conteo de los NA generados
  }
  data2 <- rbind(NoX2,SiX2)
```

```
  medX3<- median(data$X3,na.rm=TRUE)
  MasX3 <- subset(data2,data2$X3>medX3)
  dim <- dim(MasX3)[1]
  MenosX3 <- subset(data2,data2$X3<=medX3 )

  r <- 1
  while (r < miss+1) { #Repite hasta alcanzar el nro. de missing en c/variable
    j <- sample(1:dim, 1) #Elige los ID al azar
    if (is.na(MasX3$X5[j]) == 'FALSE') {MasX3$X5[j] <- NA #Controla que no sea NA
    r <- r+1}  #Lleva el conteo de los NA generados
  }
  data2 <- rbind(MasX3,MenosX3)
  data2 <- arrange(data2, Id)
  data2 <- data2[,-8]
```

c. Mecanismo de pérdida: MNAR

```
X2 <- subset(data,data$X2==1)
dim <- dim(X2)[1]
NX2 <- subset(data,data$X2==0)
i <- 1
while (i < miss+1) { #Repite hasta alcanzar el nro. de missing en c/variable
  j <- sample(1:dim, 1) #Elige los ID al azar
  if (is.na(NX2$X2[j]) == 'FALSE') {NX2$X2[j] <- NA #Controla que no sea NA
  i <- i+1}  #Lleva el conteo de los NA generados
}
data2 <- rbind(X2,NX2)

medX4<- median(data$X4,na.rm=TRUE)
MasX4 <- subset(data2,data2$X4>medX4)
MenosX4 <- subset(data2,data2$X4<=medX4)
dim <- dim(MenosX4)[1]

r <- 1
while (r < miss+1) { #Repite hasta alcanzar el nro. de missing en c/variable
  j <- sample(1:dim, 1) #Elige los ID al azar
  if (is.na(MenosX4$X4[j]) == 'FALSE') {MenosX4$X4[j] <- NA #Controla que no sea NA
  r <- r+1}  #Lleva el conteo de los NA generados
}
data2 <- rbind(MasX4,MenosX4)
data2 <- arrange(data2, Id)
data2 <- data2[,-8]
```

Parte 4: Imputación de datos y estimación del modelo de regresión de Cox

```
#Estimacion bajo CCA
cox <- coxph(Surv(y, failed) ~ X1+as.factor(X2)+X3+X4+as.factor(X5), data = data2,
    method="breslow")

resul<- tidy(cox)
estim[s,,1] <- t(resul[,2])
InfCI[s,,1] <- t(resul[,6])
SupCI[s,,1] <- t(resul[,7])

#Imputacion por valores "representativos"

impu <- data2
t <- proc.time()

#Imputacion por la mediana
impu$X1[(is.na(impu$X1))] <- median(impu$X1[!is.na(impu$X1)])
impu$X3[(is.na(impu$X3))]  <- median(impu$X1[!is.na(impu$X3)])
impu$X4[(is.na(impu$X4))]  <- median(impu$X1[!is.na(impu$X4)])

#Imputacion por el modo
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}

v <- subset(impu$X2, impu$X2!='NA')
modo <- getmode(v)
```

```
impu$X2[(is.na(impu$X2))] <- modo

v <- subset(impu$X5, impu$X5!='NA')
modo <- getmode(v)
impu$X5[(is.na(impu$X5))] <- modo

tiempo <- proc.time() - t
times[s,2] <- tiempo[3]

cox <- coxph(Surv(y, failed) ~ X1+as.factor(X2)+X3+X4+as.factor(X5), data = impu,
    method="breslow")

resul <- tidy(cox)
estim[s,,2] <- t(resul[,2])
InfCI[s,,2] <- t(resul[,6])
SupCI[s,,2] <- t(resul[,7])

nrmse[s,,2] <- hydroGOF::nrmse(impu[,cuanti], data[,cuanti])

r <- 1
for (i in cuali){
  for (j in 1:n){
    FCE[s,r,2] <- ifelse(impu[j,i]== data[j,i],FCE[s,r,2],FCE[s,r,2]+1)
  }
  r <- r+1}

#Imputacion valores faltantes mediante el metodo K-nearest neighbor

# Opcion (-): solo las covariables son consideradas en la distancia
# Opcion (+): se consideran tambien el tiempo de censura o evento y el Status
# en el calculo de la distancia

#Para k=3

#Opcion (-)
t <- proc.time()
k3a <- kNN(data2, dist_var=c("X1","X2","X3","X4","X5"), k=3)
tiempo <- proc.time() - t
times[s,3] <- tiempo[3]

cox <- coxph(Surv(y, failed) ~ X1+as.factor(X2)+X3+X4+as.factor(X5), data = k3a,
    method="breslow")

resul <- tidy(cox)

estim[s,,3] <- t(resul[,2])
InfCI[s,,3] <- t(resul[,6])
SupCI[s,,3] <- t(resul[,7])

nrmse[s,,3] <- hydroGOF::nrmse(k3a[,cuanti], data[,cuanti])
k3a <- k3a[,c(1:7)]

r <- 1
for (i in cuali){
  for (j in 1:n){
    FCE[s,r,3] <- ifelse(k3a[j,i]== data[j,i],FCE[s,r,3],FCE[s,r,3]+1)
  }
  r <- r+1}
print (c(s,"K3-␣ok"))

#Opcion (+)
t <- proc.time()
k3b <- kNN(data2, k=3)
tiempo <- proc.time() - t
times[s,4] <- tiempo[3]

cox <- coxph(Surv(y, failed) ~ X1+as.factor(X2)+X3+X4+as.factor(X5), data = k3b,
    method="breslow")

resul <- tidy(cox)
estim[s,,4] <- t(resul[,2])
InfCI[s,,4] <- t(resul[,6])
SupCI[s,,4] <- t(resul[,7])

nrmse[s,,4] <- hydroGOF::nrmse(k3b[,cuanti], data[,cuanti])
```

```
k3b <- k3b[,c(1:7)]
r <- 1
for (i in cuali){
  for (j in 1:n){
    FCE[s,r,4] <- ifelse(k3b[j,i]== data[j,i],FCE[s,r,4],FCE[s,r,4]+1)
  }
  r <- r+1}

#Para k=5

#Opcion (-)
t <- proc.time()
k5a <- kNN(data2, dist_var=c("X1","X2","X3","X4","X5"), k=5)
tiempo <- proc.time() - t
times[s,5] <- tiempo[3]

cox <- coxph(Surv(y, failed) ~ X1+as.factor(X2)+X3+X4+as.factor(X5), data = k5a,
    method="breslow")

resul <- tidy(cox)

estim[s,,5] <- t(resul[,2])
InfCI[s,,5] <- t(resul[,6])
SupCI[s,,5] <- t(resul[,7])

nrmse[s,,5] <- hydroGOF::nrmse(k5a[,cuanti], data[,cuanti])
k5a <- k5a[,c(1:7)]

r <- 1
for (i in cuali){
  for (j in 1:n){
    FCE[s,r,5] <- ifelse(k5a[j,i]== data[j,i],FCE[s,r,5],FCE[s,r,5]+1)
  }
  r <- r+1}

#Opcion (+)
t <- proc.time()
k5b <- kNN(data2, k=5)
tiempo <- proc.time() - t
times[s,6] <- tiempo[3]

cox <- coxph(Surv(y, failed) ~ X1+as.factor(X2)+X3+X4+as.factor(X5), data = k5b,
    method="breslow")

resul <- tidy(cox)
estim[s,,6] <- t(resul[,2])
InfCI[s,,6] <- t(resul[,6])
SupCI[s,,6] <- t(resul[,7])

nrmse[s,,6] <- hydroGOF::nrmse(k5b[,cuanti], data[,cuanti])

k5b <- k5b[,c(1:7)]
r <- 1
for (i in cuali){
  for (j in 1:n){
    FCE[s,r,6] <- ifelse(k5b[j,i]== data[j,i],FCE[s,r,6],FCE[s,r,6]+1)
  }
  r <- r+1}

#Para k=9

#Opcion (-)
t <- proc.time()
k9a <- kNN(data2, dist_var=c("X1","X2","X3","X4","X5"), k=9)
tiempo <- proc.time() - t
times[s,7] <- tiempo[3]

cox <- coxph(Surv(y, failed) ~ X1+as.factor(X2)+X3+X4+as.factor(X5), data = k9a,
    method="breslow")

resul <- tidy(cox)

estim[s,,7] <- t(resul[,2])
InfCI[s,,7] <- t(resul[,6])
```

```
SupCI[s,,7] <- t(resul[,7])

nrmse[s,,7] <- hydroGOF::nrmse(k3a[,cuanti], data[,cuanti])
k9a <- k9a[,c(1:7)]

r <- 1
for (i in cuali){
  for (j in 1:n){
    FCE[s,r,7] <- ifelse(k9a[j,i]== data[j,i],FCE[s,r,7],FCE[s,r,7]+1)
  }
  r <- r+1}

#Opcion (+)
t <- proc.time()
k9b <- kNN(data2, k=9)
tiempo <- proc.time() - t
times[s,8] <- tiempo[3]

cox <- coxph(Surv(y, failed) ~ X1+as.factor(X2)+X3+X4+as.factor(X5), data = k9b,
    method="breslow")

resul <- tidy(cox)
estim[s,,8] <- t(resul[,2])
InfCI[s,,8] <- t(resul[,6])
SupCI[s,,8] <- t(resul[,7])

nrmse[s,,8] <- hydroGOF::nrmse(k9b[,cuanti], data[,cuanti])

k9b <- k9b[,c(1:7)]
r <- 1
for (i in cuali){
  for (j in 1:n){
    FCE[s,r,8] <- ifelse(k9b[j,i]== data[j,i],FCE[s,r,8],FCE[s,r,8]+1)
  }
  r <- r+1}

#MissForest
#Opcion (-)
t <- proc.time()
mf <- missForest(data2[,c(1:5)]) #Para que impute sin considerar status y dias
tiempo <- proc.time() - t
times[s,9] <- tiempo[3]

#Unir los datos imputados con status y dias
mf <- cbind.data.frame(mf$ximp,data2[,c(6,7)])

cox <- coxph(Surv(y, failed) ~ X1+as.factor(X2)+X3+X4+as.factor(X5), data = mf,
    method="breslow")

resul <- tidy(cox)
estim[s,,9] <- t(resul[,2])
InfCI[s,,9] <- t(resul[,6])
SupCI[s,,9] <- t(resul[,7])

nrmse[s,,9] <- hydroGOF::nrmse(mf[,cuanti], data[,cuanti])

r <- 1
for (i in cuali){
  for (j in 1:n){
    FCE[s,r,9] <- ifelse(mf[j,i]== data[j,i],FCE[s,r,9],FCE[s,r,9]+1)
  }
  r <- r+1}

#Opcion (+)
t <- proc.time()
mfb <- missForest(data2)
tiempo <- proc.time() - t
times[s,10] <- tiempo[3]

mfb <- mfb$ximp #Se queda con los datos imputados y descarta otros resultados

cox <- coxph(Surv(y, failed) ~ X1+as.factor(X2)+X3+X4+as.factor(X5), data = mfb,
    method="breslow")

resul <- tidy(cox)
```

```
estim[s,,10] <- t(resul[,2])
InfCI[s,,10] <- t(resul[,6])
SupCI[s,,10] <- t(resul[,7])

nrmse[s,,10] <- hydroGOF::nrmse(mfb[,cuanti], data[,cuanti])

r <- 1
for (i in cuali){
  for (j in 1:n){
    FCE[s,r,10] <- ifelse(mfb[j,i]== data[j,i],FCE[s,r,10],FCE[s,r,10]+1)
  }
  r <- r+1}

#MICE
#Opcion (-)
t <- proc.time()
impua <- mice(data2[,c(1:5)]) #Para que impute sin considerar status y dias
tiempo <- proc.time() - t
times[s,11] <- tiempo[3]

analyses <- as.list(1:impua$m)
for (nim in 1:impua$m){
  compl <- cbind.data.frame(complete(impua,nim),data2[,c(6,7)])

  #Analisis de regresion de Cox para cada conjunto imputado
  analyses[[nim]] <- coxph(Surv(y, failed) ~ X1+as.factor(X2)+X3+X4+as.factor(X5), data
    = compl, method="breslow")
  r <- 1
  for (i in cuali){
    for (j in 1:n){
      MICE_FCE[nim,r,] <- ifelse(compl[j,i]==
    data[j,i],MICE_FCE[nim,r,],MICE_FCE[nim,r,]+1)
    }
    r <- r+1}
  MICE_nrmse[nim,,] <- hydroGOF::nrmse(compl[,cuanti], data[,cuanti])
}
FCE[s,,11] <- colMeans(MICE_FCE)
MICE_FCE <- array(0,c(5,length(cuali),1))
nrmse[s,,11] <- colMeans(MICE_nrmse)
MICE_nrmse <- array(0,c(5,length(cuanti),1))
object <- list(call=call, call1=impua$call, nmis=impua$nmis, analyses=analyses)
oldClass(object) <- c("mira", "coxph")
coxmicea <- summary(pool(object))

estim[s,,11] <- t(coxmicea[,1])
InfCI[s,,10] <- t(resul[,6])
SupCI[s,,10] <- t(resul[,7])

#Opcion (+)
t <- proc.time()
impub <- mice(data2) #Para que impute considerando status y dias
tiempo <- proc.time() - t
times[s,12] <- tiempo[3]

analyses <- as.list(1:impub$m)
for (nim in 1:impub$m){
  compl <- cbind.data.frame(complete(impub,nim))

  #Analisis de regresion de Cox para cada conjunto imputado
  analyses[[nim]] <- coxph(Surv(y, failed) ~ X1+as.factor(X2)+X3+X4+as.factor(X5), data
    = compl, method="breslow")
  r <- 1
  for (i in cuali){
    for (j in 1:n){
      MICE_FCE[nim,r,] <- ifelse(compl[j,i]==
    data[j,i],MICE_FCE[nim,r,],MICE_FCE[nim,r,]+1)
    }
    r <- r+1}
  MICE_nrmse[nim,,] <- hydroGOF::nrmse(compl[,cuanti], data[,cuanti])
}
FCE[s,,12] <- colMeans(MICE_FCE)
MICE_FCE <- array(0,c(5,length(cuali),1))
nrmse[s,,12] <- colMeans(MICE_nrmse)
MICE_nrmse <- array(0,c(5,length(cuanti),1))
```

```
object <- list(call=call, call1=impub$call, nmis=impub$nmis, analyses=analyses)
oldClass(object) <- c("mira", "coxph")
coxmiceb <- summary(pool(object))

estim[s,,12] <- t(coxmiceb[,1])
InfCI[s,,10] <- t(resul[,6])
SupCI[s,,10] <- t(resul[,7])
```

Parte 5: Resultados

```
for (j in 1:6){
  for (i in 1:12){
    ecm[i,j] <- hydroGOF::mse(estim[,j,i], coefcorr[,j])
  }}
ECM <- round(ecm,5)
#Matriz de promedios de los NRMSE por variable cuantitativa
#y metodo de imputacion
for (i in 1:12){
  meannrmse[i,] <- colMeans (nrmse[,,i])
}

#Matriz con porcentaje de IC que cubren la estimacion correcta,
#por parametro y metodo
for (s in 1:sim){
  for (j in 1:6){
    for (i in 1:12){
      cobertura[i,j] <- ifelse(coefcorr[1,j]>InfCI[s,j,i] & coefcorr[1,j]<SupCI[s,j,i],
                               cobertura[i,j]+1,cobertura[i,j])
    }
  }
}

meanFCE <- colMeans(FCE)/(prop*n)*100
meantimes <- colMeans(times)

for (i in 1:12){
  a <- colMeans (estim[,,i])
  print(a)
}
```

Anexo III: Análisis descriptivo de las covariables del caso-estudio

Cuadro 6.1: Análisis descriptivo de las variables correspondientes al estudio sobre el tiempo hasta el reintento de suicidio en pacientes con trastorno límite de la personalidad.

Variable	Medida	Valor
Hostilidad	Media	46.93
	Desvío estándar	9.46
	Mediana	48.00
	Datos faltantes [n (%)]	16 (10.96%)
Funcionamiento psicosocial	Media	31.12
	Desvío estándar	6.92
	Mediana	32.00
	Datos faltantes [n (%)]	12 (8.22%)
Intentos previos		
Ninguno	n (%)	29 (19.86%)
1 o 2	n (%)	55 (37.67%)
3 o más	n (%)	62 (42.47%)
Abuso sexual infantil		
Sí	n (%)	70 (47.95%)
No	n (%)	68 (46.58%)
Datos faltantes	n (%)	8 (5.48%)
Edad	Media	33.18
	Desvío estándar	9.62
	Mediana	33.00
Edad del primer intento de suicidio	Media	24.47
	Desvío estándar	10.81
	Mediana	24.00
	Datos faltantes [n (%)]	13 (8.90%)
Días de seguimiento (tiempo hasta el evento o censura)	Media	411.55
	Desvío estándar	248.94
	Mediana	390.00
Estado	Reintento de suicidio	48 (32.88%)
	Censura por derecha	98 (67.12%)

Figura 6.1: Distribución de las variables explicativas del tiempo hasta el reintento de suicidio.