Métodos de imputación basados en modelos: Métodos basados en la función de verosimilitud

Subdepartamento de Investigación Estadística

Departamento de Metodologías e Innovación Estadística

Instituto Nacional de Estadísticas

Miguel Alvarado

Mayo, 2023

Contents

1 Introducción

A partir de esta sección se presentan los métodos de imputación basados en modelos. Estos métodos se basan en dos conceptos fundamentales: la *función de verosimilitud* de los datos[1] y el *mecanismo* que genera los *datos faltantes*[2] (*missing data*). La presentación formal de ambos conceptos, así como sus implicancias, se desarrollan en las secciones que siguen a esta introducción. Sin embargo, para los propósitos de esta introducción es necesario hacer referencia al segundo concepto. De este modo, antes de presentar el marco teórico y los fundamentos inferenciales en que se basan los métodos de imputación basados en modelos, los que permiten sostener que este tipo de modelos son una solución satisfactoria a los inconvenientes que se derivan del problema de la falta de datos, es necesario discutir, al menos brevemente, el por qué los *métodos tradicionales*[3], presentados en la primera parte de este documento, son métodos cuyas soluciones son poco satisfactorias y cuestionables como alternativa de respuesta al mismo problema antes mencionado, pero además estos métodos podrían ser potencialmente problemáticos porque pueden introducir sesgos independientemente de mecanismo (Enders 2022, pag. 24).

Antes del trabajo de Donal B. Rubin (Rubin 1976), los análisis estadísticos con datos faltantes eran realizados a partir de suponer, implícita o explícitamente, que el mecanismo que genera los datos faltantes podía ser *ignorado*. Sin embargo, hasta ese entonces, la literatura estadística que estudiaba el problema de la falta de datos no había respondido a una pregunta anterior: *¿cuándo es apropiado ignorar el mecanismo que genera los datos faltantes?* (Rubin 1976, pag. 581). En este sentido, los métodos tradicionales simplemente asumen que el mecanismo que genera los datos faltantes puede ser ignorado, asumiendo que la falta de datos ocurre de manera *completamente aleatoria* en la muestra de datos, pero sin discutir sobre la validez de dicho supuesto. De manera más precisa, estos métodos asumen que el *mecanismo* que genera los datos faltantes es del tipo *MCAR*[4] (*Missing Completly At Random*), supuesto que, como se menciona a lo largo de la literatura, resulta sumamente restrictivo (Enders 2022, pag. 24) y, a menudo, *poco realista* (Van Buuren 2012, pag. 7). Entonces, dado que el supuesto *MCAR* es poco plausible, los métodos tradicionales y las inferencias que se desprenden de estos son cuestionables.

Como se describirá en los párrafos siguientes, los métodos tradicionales dependen de asumir supuestos poco verosímiles y, además, muchos de estos métodos simplemente carecen de algún tipo de fundamento inferencial. Por otro lado, aun cuando la aplicación práctica de los métodos tradicionales es sencilla, algunos de estos métodos pueden dificultar e incluso imposibilitar el cálculo de algunas estimaciones. Por último, en algunos de estos métodos se podrían requerir de la toma de decisiones que quedan al arbitrio de quienes implementan tales métodos. A continuación, de manera sucinta, se discute sobre las limitaciones e inconvenientes que presentan los métodos tradicionales.

[1] El término *datos* se entiende como un conjunto de arreglos rectangulares de datos o, más simple, una *matriz de datos*. Donde, las filas de la matriz de datos representan unidades, también llamados *casos*, *observaciones* o *elementos* según el contexto, y las columnas representan *características* o *variables* que son medidas para cada unidad. Las *entradas* o *celdas* en la matriz de datos son, casi siempre, números reales, ya sea que representen los valores de variables continuas, (i.e., ingresos), o que representen categorías de respuesta, que pueden estar ordenadas (i.e., nivel de educación) o no ordenadas (i.e., sexo). En este sentido, este documento trata sobre el análisis de dicha matriz de datos cuando *no se observan* algunas de las entradas de la matriz.

[2] Los términos *datos faltantes* y *falta de datos* se usan de manera indistinta a lo largo de este documento y hacen referencia a la ausencia del *valor* correspondiente que habrían sido recolectado de los casos (unidades, elementos) que conforman la muestra de datos.

[3] En algunos textos como (Enders 2022, secc. 1.7), estos métodos se denominan como *métodos antiguos*.

[4] En la siguiente sección se hace una presentación formal este y otros conceptos.

El método de *análisis de casos completos*, también conocido como *eliminación por lista*, es la forma más simple de lidiar con los datos faltantes. En este método se eliminan todos los casos con uno o más datos faltantes cualquiera de las variables que conforman la muestra de datos. Si el mecanismo es *MCAR*, la eliminación por lista produce estimaciones insesgadas para las medias, las varianzas y los coeficientes de regresión; no obstante, los errores estándar y niveles de significancia *solo* son correctos para el conjunto reducido de casos completos, pero que a menudo son mayores en relación con todos los datos observados. Una clara desventaja de este método es que potencialmente se puede llegar a eliminar una parte considerable de los casos, especialmente si el número de variables con datos faltantes es grande (Van Buuren 2012, pag. 8). En efecto, como se señala en (Little and Rubin 2020, pag. 47), las desventajas que se derivan de la posible pérdida de información al descartar casos incompletos tiene dos aspectos: *pérdida de precisión* y *sesgo* cuando el mecanismo no es del tipo *MCAR*. El grado de sesgo y pérdida de precisión dependen no solo de la fracción de casos completos y del mecanismo de los datos faltantes, sino también de la medida en que las unidades completas e incompletas difieren y de las estimaciones de interés (Little and Rubin 2020, pag. 48).

El método de *eliminación por pares*, también conocido como *análisis de casos disponibles*, intenta remediar el problema de la pérdida de casos que se produce en el método de análisis de casos completos. En este método el cálculo de cualquier estimación de interés de alguna variable es realizado a partir de los casos disponibles en dicha variable. De este modo, las estimaciones de la variable Y se realizan a partir de los casos disponibles en la variable Y. De manera análoga, las estimaciones de la variable X se obtienen a partir de los casos disponibles en la variable X, así sucesivamente con el resto de las variables. El método es simple, puesto que usa toda la información disponible y produce estimaciones consistentes para las medias, correlaciones y covarianzas bajo el supuesto *MCAR* (Van Buuren 2012, pag. 10). Sin embargo, cuando estas estimaciones se toman en conjunto, aparecen inconvenientes considerables. Primero, las estimaciones pueden estar sesgadas si el mecanismo no es del tipo *MCAR* (Van Buuren 2012, pag. 10). Además, existen problemas al momento del cálculo computacional. Por ejemplo, la matriz de correlación puede no ser definida positiva, lo cual es un requisito para la mayoría de los procedimientos multivariantes. De igual modo, pueden ocurrir correlaciones que no están en el rango unitario $[-1, +1]$, un problema que proviene de utilizar diferentes subconjuntos de datos para el cálculo de las covarianzas y las varianzas. Otro problema es que no queda claro qué tamaño de muestra debe usarse para calcular los errores estándar (Van Buuren 2012, pag. 10).

El método de *imputación por la media (aritmética)*, también conocido como *sustitución por la media*, es un enfoque de *única* imputación que completa los datos faltantes para una variable con la media[5] de los datos observados. Este método no tiene justificación teórica y distorsiona las estimaciones de parámetros, independiente del *mecanismo* que genera la falta de datos (Enders 2022, pag. 25), puesto que este método distorsiona la distribución de los datos de varias maneras (Van Buuren 2012, pag. 11). La imputación por la media es una solución rápida y sencilla para abordar el problema de los datos faltantes. Sin embargo, este método subestima la varianza, altera las relaciones entre las variables, sesga casi cualquier estimación que no sea la media y sesga la estimación de la media cuando el mecanismo no es del tipo *MCAR*, por lo tanto su uso debe evitarse en general[6] (Van Buuren 2012, pag. 11).

[5] En el caso de variables categóricas, la imputación de los datos faltantes es realizada usando la *moda* de los datos observados (Van Buuren 2012, pag. 10).

[6] Teniendo en cuenta los sesgos que genera el método de imputación por la media, el uso de este método no suele ser recomendado. Un refinamiento sobre el método de la imputación por la media, es imputar a partir del uso de medias condicionales dados los valores observados (Little and Rubin 2020, pag. 70). Mayor detalle sobre este enfoque

El método de *imputación por regresión*, con el propósito de mejorar la imputación de los datos faltantes en la variable de interés, incorpora la información contenida en las otras variables que forman parte de la muestra de datos. El método parte por ajustar un modelo de regresión a partir de los datos observados en la muestra. Luego, el valor no observado en los datos es reemplazado por las *predicciones* bajo el modelo ajustado. De este modo, los valores imputados corresponden a los valores más *verosímiles* bajo el modelo ajustado (Van Buuren 2012, pag. 12). Sin embargo, al igual que en el método de imputación por la media, el conjunto de valores imputados presenta menor variabilidad que en los valores observados[7]. Si bien es posible que cada uno de los valores individuales imputados sean el mejor pronóstico bajo el modelo, resulta poco probable que los valores reales (pero no observados) de la variable imputada tengan tal distribución. La imputación de los datos faltantes a partir de este método también tiene un efecto sobre la correlación. Dado que la correlación de los datos imputados bajo el modelo ajustado es igual a 1 (Enders 2022, pag. 27), la correlación para el conjunto de los datos completos se ve necesariamente incrementada, en consecuencia las varianzas y correlaciones estimadas quedan sesgadas.

Bajo el supuesto que el mecanismo es del tipo *MCAR*, la imputación por regresión produce estimaciones insesgadas tanto para las medias (al igual que en el método de imputación por la media), como para los ponderadores del modelo de regresión ajustado para realizar la imputación de los datos faltantes, esto último si las variables explicativas en el modelo están completas. Por otro lado, como se ha mencionado, la variabilidad de los datos imputados queda subestimada de manera sistemática y el grado de subestimación depende de la varianza explicada y de la proporción de datos faltantes (Van Buuren 2012, pag. 12). La idea básica detrás del método de imputación por regresión es intuitivamente atractiva: las variables tienden a estar correlacionadas, por lo que se *reemplazan* los valores faltantes por predicciones que vienen de un modelo que toma prestada información importante de los datos observados. Aunque esta idea tiene sentido, como se ha mencionado, las imputaciones resultantes pueden introducir sesgos, cuya naturaleza y magnitud dependen del mecanismo de los datos faltantes y varían según las diferentes estimaciones (Enders 2022, pag. 27).

El método de *imputación por regresión estocástica* es un refinamiento del método de imputación por regresión, en el cual se agrega *variabilidad* a las predicciones del modelo ajustado (Van Buuren 2012, pag. 13). De este modo, este método también ajusta un modelo de regresión a partir de los datos observados, luego el valor no observado en los datos es reemplazado por las *predicciones* bajo el modelo ajustado, pero tomando el paso adicional de *agregar* a cada predicción un término de *ruido aleatorio* (*random noise*) desde una distribución normal. Al agregar estos residuos a las predicciones se reduce la correlación (Van Buuren 2012, pag. 13), se restaura la pérdida de variabilidad de los datos y se eliminan los sesgos asociados al método de imputación por regresión (Enders 2022, pag. 28).

El método de imputación por regresión estocástica no resuelve todos los problemas y hay muchas sutilezas que deben tenerse presentes[8]. No obstante, el método de imputación por regresión estocástica es el único método tradicional que, generalmente, es capaz de producir estimaciones insesgadas

se puede encontrar en (Little and Rubin 2020, secc. 4.2.2).

[7] La imputación por la media se puede considerar como un caso especial del método de imputación por regresión donde las variables explicativas (predictores) son variables indicadoras (*dummies*) para las celdas dentro de las cuales se imputa por la media (Little and Rubin 2020, pag. 68).

[8] Por ejemplo, al añadir un ruido aleatorio a las predicciones bajo el modelo ajustado es posible que para las predicciones localizadas en los extremos de la distribución, el valor a imputar quede fuera del rango factible de valores de la variable a imputar. Un ejemplo de esto puede encontrarse en (Van Buuren 2012, pag. 13), en cuyo ejemplo, una parte de las imputaciones son valores negativos en circunstancias que la variable a imputar solo puede tomar valores mayores o iguales a cero.

de los parámetros de interés cuando el mecanismo es del tipo *MAR*[9] (*Missing At Random*). Más importante aún, la idea central detrás del método de imputación por regresión estocástica (una imputación es igual a una predicción más un ruido aleatorio) constituye la base de técnicas de imputación más avanzadas y, como se verá más adelante, resurge con los métodos bayesianos e imputación múltiple (Enders 2022, pag. 29).

El método de *adelantar la última observación* (*Last Observation Carried Forward, LOCF*) es una técnica de datos faltantes para estudios longitudinales. Utilizar el método *LOCF* en estudios sociales y del comportamiento es bastante poco frecuente, siendo su uso más común en estudios médicos y ensayos clínicos. Como el nombre del método lo indica, la idea es tomar el último valor observado y *adelantarlo* (*trasladarlo*) en reemplazo de los datos faltantes de la actual muestra de datos. El método *LOCF* es conveniente en el sentido que genera un conjunto de datos completo. Sin embargo, este método asume que no existen cambios desde la última observación realizada y/o durante el período de tiempo en que se genera la nueva medición. La creencia popular indicaría que imputar los datos faltantes con datos *estables* en el tiempo, produciría una estimación conservadora de las diferencias entre los grupos bajo estudio. Sin embargo, la investigación empírica muestra que esto no es necesariamente cierto, ya que el método también puede exagerar las diferencias entre estos grupos. En efecto, la dirección y la magnitud del sesgo que se produce dependen de las características específicas de los datos, pero es probable que el método *LOCF* produzca estimaciones sesgadas de los parámetros de interés, incluso asumiendo que el mecanismo es del tipo *MCAR* (Enders 2022, pag. 31).

El método de imputación *Hot-Deck* imputa los valores faltantes utilizando los valores observados en casos "*similares*" en la muestra, estos últimos usualmente denominados como *donantes*[10]. Este método es común en la práctica de las encuestas y puede implicar esquemas muy elaborados para seleccionar los casos *donantes*[11]. La ventaja del método *Hot-Deck* es que, a diferencia del método de imputación por la media, la distribución de los valores muestreados de la variable a imputar no queda distorsionada por las imputaciones. Sin embargo, el incremento en la varianza que produce el método *Hot-Deck* puede ser no despreciable. Aun cuando se pueden lograr reducciones en la varianza adicional que se produce con el método *Hot-Deck*, por ejemplo mediante una selección más eficiente del esquema de muestreo, poniendo restricciones en el número de veces que un caso actúa como donante, usando los valores observados en la variable para formar estratos de muestreo para donantes o mediante el uso de un *Hot-Deck secuencial*; los *métodos de imputación múltiple* se deben preferir por sobre este método, puesto que los métodos de imputación múltiple no solo que pueden reducir el incremento de la varianza del muestreo a niveles insignificantes, sino que también proporcionan errores estándar válidos que tienen en cuenta la incertidumbre del proceso de imputación. Las estimaciones que se derivan del uso del método *Hot-Deck* son insesgadas solo bajo el supuesto que el mecanismo es del tipo *MCAR*; supuesto que, generalmente, es poco realista (Little and Rubin 2020, pag. 78).

El método de imputación *Cold-Deck* imputa los valores faltantes de una variable por un valor constante que proviene de una fuente externa, por ejemplo a partir del los datos de una encuesta anterior. La aplicación práctica de este método suele tratar los datos resultantes como una muestra completa, ignorando las consecuencias de la imputación. Una teoría satisfactoria para el análisis

[9] En la siguiente sección se hace una presentación formal este y otros conceptos.

[10] En este método, la imputación de los valores faltantes de un caso es realizada con los valores observados en algún otro caso *similar* al que se busca imputar. Sin embargo, cuando existen dos o más casos *similares*, pero con valores observados diferentes en las variables a imputar, la decisión sobre cuál caso tomar como *donante*, queda al arbitrio de quien realiza la imputación.

[11] En (Little and Rubin 2020, secc. 4.3.2) se puede encontrar mayor detalle sobre variantes del método *Hot-Deck*.

de datos obtenidos mediante el método de imputación *Cold-Deck* es inexistente (Little and Rubin 2020, pag. 69; Van Buuren 2012, pag. 7).

A manera de síntesis, se puede señalar que una limitación importante de los *métodos tradicionales* de imputación descritos en esta introducción es que los estimadores de la varianza de muestreo que son aplicados a los datos *completados* mediante estos métodos de imputación, al no tener en cuenta la incertidumbre asociada al proceso de imputación, a excepción del método de imputación por regresión estocástica, finalmente subestiman sistemáticamente la verdadera varianza de muestreo de las estimaciones. Por lo tanto, los errores estándar calculados a partir de los datos *completados* también se subestiman sistemáticamente, lo que implica que los *p-values* de las pruebas sean demasiado significativos y los intervalos de confianza sean demasiado estrechos (Little and Rubin 2020, pag. 81). Lo anterior ocurre incluso si el modelo utilizado para generar las imputaciones es el correcto, algo que, salvo para el caso antes mencionado, depende de asumir que el mecanismo que genera la falta de datos es del tipo $MCAR$, supuesto que, generalmente, es poco realista (Little and Rubin 2020, pag. 78).

Dado que los métodos tradicionales presentan limitaciones importantes que resultan insalvables y dado que estos métodos dependen de supuestos inverosímiles, lo que a continuación sigue en este documento es la presentación del marco teórico y los robustos fundamentos inferenciales en los que se basan los métodos de imputación basados en modelos; los cuales no presentan las limitaciones de los métodos tradicionales, ni dependen de supuestos inverosímiles.

2 Mecanismo de Datos Faltantes

Como se ha mencionado, hasta antes del trabajo de Rubin (Rubin 1976), los análisis estadísticos con datos faltantes eran realizados a partir de suponer, implícita o explícitamente, que el mecanismo que genera los datos faltantes podía ser *ignorado*, pero sin dar respuesta a la importante pregunta sobre: *¿cuándo es apropiado ignorar el mecanismo que genera los datos faltantes?* (Rubin 1976, pag. 581). Rubin, sin embargo, logró establecer las *condiciones necesarias* (*weakest conditions*) sobre el mecanismo que genera los datos faltantes, tal que *siempre* es apropiado *ignorar* dicho mecanismo al momento de realizar inferencia sobre la distribución de los datos (Rubin 1976, pag. 582). Esto, dentro la literatura que aborda el problema de la falta de datos, se conoce como el con el de *ignorabilidad* del mecanismo. En este sentido, en esta sección, junto con introducir uno de los dos conceptos fundamentales en que se basan los métodos basados en modelos, se presentan las *condiciones necesarias* sobre el *mecanismo* que genera los datos faltantes y que dan paso al importante concepto de *ignorabilidad*.

2.1 Patrón y Mecanismo de Datos Faltantes

Dentro de la literatura, dos conceptos que suelen prestarse para la confusión son los de: *patrón de datos faltantes* y *mecanismo de datos faltantes*. El *patrón de datos faltantes* se refiere a la configuración o disposición de los datos observados y los no observados (*missing data*) dentro un conjunto de datos. En tanto, el *mecanismo de datos faltantes* describe las posibles relaciones entre los datos y la *propensión* que tienen de ser no observados. En términos generales, el *patrón de datos faltantes* describe *dónde* están las celdas vacias en los datos, mientras que el *mecanismo de datos faltantes* describe *cómo* se generan los datos faltantes (Enders 2022, pag. 2), y en particular, si la falta de datos está relacionada con los valores subyacentes de las variables en el conjunto de datos. Hacer esta clara diferencia entre ambos conceptos es importante, pero entender el concepto de *mecanismo* resulta crucial, puesto que las propiedades de los métodos de imputación basados en modelos dependen en gran medida de la naturaleza de las *dependencias* al interior del *mecanismo* (Little and Rubin 2020, pag. 13).

Con el propósito de formalizar lo antes señalado, consideremos la siguiente notación[12]. Denotemos por $Y = \{y_{ij}\}$, la matriz de $n \times p$ que contiene los valores de los *datos completos*. Los *datos faltantes* ocurren cuando el valor de algunos datos, y_{ij}, son *no observados* (*missing*). Para describir lo anterior, se define la *matriz indicadora de respuesta*[13], $R = \{r_{ij}\}$, matriz de $n \times p$ de ceros y unos que define el *patrón* de los datos faltantes. De este modo, los elementos de Y y R se denotan por y_{ij} y r_{ij}, respectivamente, donde $i = 1, \ldots, n$ y $j = 1, \ldots, p$. Si el valor de y_{ij} es observado, entonces $r_{ij} = 1$ y si y_{ij} es no observado, $r_{ij} = 0$. Luego, Y denota los *datos completos*, esto es, los *datos observados*, Y_{obs}, y los *datos no observados*, Y_{miss}. Entonces, $Y = (Y_{obs}, Y_{miss})$.

Siguiendo el trabajo de Rubin, el papel crucial que toma el *mecanismo* dentro del análisis estadístico con datos faltantes puede ser formalizado a través de tomar la *matriz indicadora de respuesta*, R, como una variable aleatoria y asignarle una distribución de probabilidades (Little and Rubin 2020,

[12] Esta misma notación se encuentra en (Van Buuren 2012, pag. 30; He, Zhang, and Hsu 2021, pag. 7) y, con algunas diferencias, en (Enders 2022, pag. 4-5; Little and Rubin 2020, pag. 8-9).

[13] Como se menciona en (Little and Rubin 2020, pag. 9), alternativamente, la matriz indicadora de respuesta, R puede ser denotada por la *matriz indicadora de falta de respuesta*, $M = \{m_{ij}\}$, donde $m_{ij} = 0$, si el valor de y_{ij} es observado y $m_{ij} = 1$, si y_{ij} es no observado. El uso de M se encuentra, entre otros textos, en (Little and Rubin 2020, pag. 9; Enders 2022, pag. 5); en tanto, el uso de R, y que sigue este documento, se encuentra en (Van Buuren 2012, pag. 30; He, Zhang, and Hsu 2021, pag. 7), entre otros.

pag. 9). De esta forma, para todo dato se establece una cierta probabilidad de ser un dato *no observado*; es decir, un *dato faltante* (*missing*). Luego, el proceso aleatorio que gobierna las probabilidades de (no) observar el valor de un dato se denomina *mecanismo de respuesta* (o *mecanismo de falta de datos*) (Van Buuren 2012, pag. 6). Luego, el *mecanismo* puede ser formulado como un modelo estadístico para la matriz indicadora de respuesta, R, dado los datos, $Y = (Y_{obs}, Y_{miss})$[14]. Sin pérdida de generalidad, el *mecanismo* que genera los datos faltantes es caracterizado por la distribución condicional de R dado Y, es decir, $f(R|Y_{obs}, Y_{miss}, \phi)$; donde ϕ denota los parámetros (desconocidos) del modelo formulado para R y la función $f(\cdot|\cdot)$ denota una distribución de probabilidades (Little and Rubin 2020, pag. 13).

Tomando el trabajo de Rubin (Rubin 1976), Little y Rubin (Little and Rubin 2020, secc. 1.3) introdujeron un sistema de clasificación para el problema de datos faltantes que es virtualmente universal en la literatura. Este trabajo describe tres tipos de *mecanismos* o *procesos aleatorios* que describen diferentes maneras en las que la probabilidad de los *datos faltantes* se relaciona con los datos: *falta de datos completamente aleatoria* (*Missing Completly At Random, MCAR*), *falta de datos aleatoria* (*Missing At Random, MAR*) y *falta de datos no aleatoria* (*Missing Not At Random, MNAR*). Desde una perspectiva práctica, estos diferentes tipos de mecanismo son de vital importancia, porque funcionan como supuestos estadísticos para un análisis de datos faltantes (Enders 2022, pag. 3-4); lo que hace importante un análisis formal de cada uno de estos.

2.1.1 Missing Completly At Random (MCAR)

Si la probabilidad de no observar el valor de un dato; es decir, ser un dato faltante ($R = 0$), es la misma para todas las observaciones, se dice que la *falta de datos es completamente aleatoria*, esto es, el mecanismo es del tipo *MCAR* (Van Buuren 2012, pag. 7). El mecanismo del tipo *MCAR* establece que la probabilidad de ser un dato faltante *no* esta relacionada con los *datos completos* (i.e., datos observados y no observados) (Enders 2022, pag. 6). La definición formal involucra la distribución condicional de R dado Y. Entonces, la distribución para un mecanismo *MCAR* (He, Zhang, and Hsu 2021, pag. 13), es[15]:

$$f(R = 0|Y_{obs}, Y_{miss}, \phi) \quad = \quad f(R = 0|\phi) \tag{1}$$

Esto es, la probabilidad de los datos faltantes *no* está relacionada con los datos y solo depende de los parámetros ϕ. En palabras simples, el lado derecho de la ecuación dice que todos los casos o elementos tienen la misma probabilidad de ser un dato faltante, y los parámetros ϕ (Enders 2022, pag. 6). Una consecuencia muy importante de un proceso de este tipo es que se pueden ignorar muchas de las complejidades que surgen debido a la falta de datos, a parte de la pérdida obvia de información. No obstante, como ya se ha mencionado, aún cuando esta situación resulta sumamente conveniente, el mecanismo *MCAR* es una situación poco realista (Little and Rubin 2020, pag. 78; Van Buuren 2012, pag. 7).

[14] Es decir, un modelo que establece la relación entre R e Y, donde una parte de Y, son *datos observados*, Y_{obs} y, otra parte, son *datos no observados*, Y_{miss}. En tanto, R es completamente observado. Un otro concepto que suele mencionarse dentro la literatura es el de *modelo de respuesta* o *modelo de falta de datos* y se refiere al modelo particular del *mecanismo* (Van Buuren 2012, pag. 6).

[15] Si se utiliza la *matriz indicadora de falta de respuesta*, M, equivalentemente, la distribución para un mecanismo *MCAR* (Enders 2022, pag. 6), es:

$$f(M = 1|Y_{obs}, Y_{miss}, \phi) \quad = \quad f(M = 1|\phi)$$

2.1.2 Missing At Random (MAR)

Si la probabilidad de no observar el valor de un dato; es decir, ser un dato faltante, es la misma solo dentro de grupos definidos por los datos *observados*, se dice que la *falta de datos es aleatoria*, esto es, el mecanismo es del tipo *MAR* (Van Buuren 2012, pag. 7). El mecanismo del tipo *MAR* establece que la probabilidad de ser un dato faltante esta relacionada con los *datos observados*, pero *no* con los *datos no observados* (Enders 2022, pag. 8). La definición formal involucra la distribución condicional de R dado Y. Entonces, la distribución para un mecanismo *MAR* (He, Zhang, and Hsu 2021, pag. 13), es[16]:

$$f\left(R=0|Y_{obs},Y_{miss},\phi\right) \quad = \quad f\left(R=0|Y_{obs},\phi\right) \tag{2}$$

Esto es, la probabilidad de los datos faltantes está relacionada *solo* con las parte observada de los datos y de los parámetros ϕ. En palabras simples, el lado derecho de la ecuación dice que los valores que se hubieran observado en Y_{miss}, no contiene información adicional sobre los datos faltantes, distinta a la aportada por los datos observados Y_{obs} (Enders 2022, pag. 8). Este mecanismo es más general que el primero y resulta un supuesto más realista que suponer un mecanismo del primer tipo. Como veremos, los métodos modernos de imputación, generalmente, suponen que la falta de datos es generado por un mecanismo del tipo *MAR*.

2.1.3 Missing Not At Random (MNAR)

Si la probabilidad de no observar el valor de un dato; es decir, ser un dato faltante, *no* es la misma para todas las observaciones, se dice que la *falta de datos es no aleatoria*, esto es, el mecanismo es del tipo *MNAR*. El mecanismo del tipo *MNAR* establece que la probabilidad de ser un dato faltante esta relacionada con los *datos observados* y, también, con los *datos no observados* (Enders 2022, pag. 11). La definición formal involucra la distribución condicional de R dado Y. Entonces, la distribución para un mecanismo *MNAR* (Van Buuren 2012, pag. 31), es[17]:

$$f\left(R=0|Y_{obs},Y_{miss},\phi\right) \tag{3}$$

A diferencia de los mecanismos anteriores, la distribución condicional de R dado Y no se simplifica.

2.1.4 Ignorabilidad

Hasta aquí, poco se ha dicho sobre los parámetros ϕ del modelo formulado para R. La razón es bastante simple, esos parámetros no tienen algún valor en si mismos y, generalmente, son desconocidos. En tal sentido, el análisis de los datos faltantes se simplificaría si se pudiera simplemente

[16]Si se utiliza la *matriz indicadora de falta de respuesta*, M, equivalentemente, la distribución para un mecanismo *MAR* (Enders 2022, pag. 8), es:

$$f\left(M=1|Y_{obs},Y_{miss},\phi\right) \quad = \quad f\left(M=1|Y_{obs},\phi\right)$$

[17]Si se utiliza la *matriz indicadora de falta de respuesta*, M, equivalentemente, la distribución para un mecanismo *MNAR* (Enders 2022, pag. 11), es:

$$f\left(M=1|Y_{obs},Y_{miss},\phi\right)$$

ignorar estos parámetros. Por su parte, la importancia práctica de haber realizado una distinción clara entre los diferentes tipos de mecanismo (*MCAR*, *MAR* y *MNAR*) es que esto permite clarificar las condiciones bajo las cuales es posible estimar con precisión los parámetros que si son de nuestro interés, sin necesidad de conocer los parámetros ϕ.

En el trabajo desarrollado por Rubin (Rubin 1976) se presentan dos modelos: el modelo que es el foco del análisis y un modelo que describe el mecanismo de datos faltantes. Sin pérdida de generalidad, supongamos que estos modelos tienen parámetros θ y ϕ, respectivamente. Los parámetros en ϕ, cualesquiera que sean, son esencialmente una *molestia*, porque no están relacionados con los objetivos que motivaron la investigación de las unidades que conforman la muestra de datos. Entonces, cabe preguntarse *¿en qué situaciones podemos estimar simplemente θ a partir de los datos observados sin preocuparnos de estimar el modelo para los datos faltantes o los parámetros en ϕ?*. Esto es la esencia del concepto de *ignorabilidad* del mecanismo; es decir, volvemos a la importante pregunta planteada por Rubin: *¿cuándo es apropiado ignorar el mecanismo que genera los datos faltantes?* (Rubin 1976, pag. 581).

El trabajo de Rubin logró establecer las *condiciones necesarias* (*weakest conditions*) sobre el *mecanismo* que genera los datos faltantes, tal que *siempre* es apropiado *ignorar* dicho mecanismo al momento de realizar inferencia sobre la distribución de los datos (Rubin 1976, pag. 582). De este modo, se dice que el *mecanismo* puede ser ignorado si: 1. Los datos faltantes siguen un mecanismo del tipo *MAR*, y 2. Los parámetros ϕ no contienen información sobre los parámetros de interés θ (es decir, ϕ y θ son distintos).

Como se verá en las secciones siguientes, el concepto de *ignorabilidad* tiene implicancias muy importantes en cuanto a la aplicación de los métodos de imputación basados en modelos. En este sentido, las condiciones que dan paso al concepto de ignorabilidad, son igual de importantes.

3 Estimación por Máxima Verosimilitud con Datos Incompletos

Breve introducción.

3.1 Estimación por Máxima Verosimilitud: Conceptos básicos

- Suponga que Y denota los datos completos y θ es el parámetro de interés.
- La función de densidad conjunta de los datos es: $f(Y|\theta)$.
- Se define la *función de verosimilitud* $L(\theta|Y)$ como alguna función de θ, proporcional a la densidad conjunta de los datos f.

$$L(\theta|Y) \quad \propto \quad f(Y|\theta) \tag{4}$$

- Luego, el estimador por máxima verosimilitud de θ, $\hat{\theta}_{ML}$ es tal que:

$$\hat{\theta}_{ML} \quad = \quad \arg\max_{\theta} L(\theta|Y) \tag{5}$$

3.2 Métodos de Máxima Verosimilitud con Datos Incompletos: Marco teórico general

- Suponga que $Y = (Y_{obs}, Y_{miss})$ denota los datos completos y θ es el parámetro de interés.
- La función de densidad conjunta de los datos *completos* es: $f(Y|\theta)$.
- R es la matriz de respuesta y ϕ es el parámetro del modelo de respuesta.
- $f(R|Y_{obs}, Y_{miss}, \phi)$ es el mecanismo de datos faltantes.
- Dado que la información observada en los datos incluye Y_{obs} y R, la función de verosimilitud de los datos observados se puede expresar como:

$$L(\theta, \phi|Y_{obs}, R) \quad \propto \quad f(Y_{obs}, R|\theta, \phi) \tag{6}$$

$$\begin{aligned}
L(\theta, \phi|Y_{obs}, R) &\propto f(Y_{obs}, R|\theta, \phi) \\
&= \int f(Y, R|\theta, \phi)\, dY_{miss} \\
&= \int f(Y|\theta)\, f(R|Y, \phi)\, dY_{miss} \\
&= \int f(Y_{obs}, Y_{miss}|\theta)\, f(R|Y_{obs}, Y_{miss}, \phi)\, dY_{miss} \\
&= \int f(Y_{obs}, Y_{miss}|\theta)\, f(R|Y_{obs}, \phi)\, dY_{miss} \\
&= f(R|Y_{obs}, \phi) \int f(Y_{obs}, Y_{miss}|\theta)\, dY_{miss} \\
&= f(R|Y_{obs}, \phi)\, f(Y_{obs}|\theta)
\end{aligned} \tag{7}$$

- Un mecanismo de datos faltantes se denomina *ignorable* si esta es MAR y los parámetros θ y ϕ son *distinguibles*.
- La ecuación anterior muestra que, bajo un *mecanismo de datos faltantes ignorable*, para realizar inferencias sobre θ, solo necesitamos trabajar con $f\left(Y_{obs}|\theta\right)$ en lugar de $f\left(Y_{obs}, R|\theta, \phi\right)$.
- De este modo, es suficiente trabajar con la función de verosimilitud de los datos observados, ignorando el mecanismo de datos faltantes, pues:

$$L\left(\theta, \phi|Y_{obs}, R\right) \quad \propto \quad f\left(Y_{obs}|\theta\right) \tag{8}$$

3.3 El Algoritmo de Esperanza y Maximización (EM)

Presentar brevemente el paper seminal de (Dempster, Laird, and Rubin 1977).

3.4 Modelos para Datos Categóricos

3.4.1 Datos Categoricos Binarios

3.4.2 Datos Categoricos no Binarios

3.5 Modelos para Datos Continuos

3.6 Modelos para Mezclas de Distribuciones

4 Métodos de Pseudo-Verosimilitud para Datos Incompletos

5 Estimación Bayesiana con Datos Incompletos

5.1 Estimación Bayesiana: Conceptos básicos

5.2 Métodos Bayesianos con Datos Incompletos: Marco teórico general

Bibliografía

Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society: Series B (Methodological)* 39 (1): 1–22.

Enders, Craig K. 2022. *Applied Missing Data Analysis*. Guilford Publications.

He, Yulei, Guangyu Zhang, and Chiu-Hsieh Hsu. 2021. *Multiple Imputation of Missing Data in Practice: Basic Theory and Analysis Strategies*. CRC Press.

Little, Roderick J. A., and Donald B. Rubin. 2020. *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. John Wiley & Sons.

Molenberghs, G., G. Fitzmaurice, M. G. Kenward, A. Tsiatis, and G. Verbeke. 2015. *Handbook of Missing Data Methodology*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press.

Pigott, Therese D. 2001. "A Review of Methods for Missing Data." *Educational Research and Evaluation* 7: 353–83.

Rubin, Donald B. 1976. "Inference and Missing Data." *Biometrika* 63 (3): 581–92.

Tan, Ming T., Guo L. Tian, and Kai W. Ng. 2010. *Bayesian Missing Data Problems: EM, Data Augmentation and Noniterative Computation*. Chapman & Hall/CRC Biostatistics Series. CRC Press.

Van Buuren, Stef. 2012. *Flexible Imputation of Missing Data*. Chapman & Hall/CRC Interdisciplinary Statistics Series. CRC Press.