Guía con lineamientos y orientaciones metodológicas para la imputación de datos estadísticos Instituto Nacional de Estadísticas

Departamento de Metodologías e Innovación Estadística
Subdepartamento de Investigación Estadística
Instituto Nacional de Estadística

Índice

1 Introducción

1.1 Definición

La guía con lineamientos y orientaciones metodológicas para la imputación de datos estadísticos es un documento que identifica y describe las herramientas y métodos de imputación, considerando aspectos de la imputación por donantes, imputación basada en modelos y entre ellas, tanto los métodos de imputación simple como los métodos de imputación múltiple. Se incluyen ejemplos de imputación en oficinas de estadísticas considerando encuestas conocidas por el público.

Adicionalmente, se realiza un diagnóstico y se revisan antecedentes bibliográficos de los métodos de imputación de datos aplicados institucionalmente. Esta información es clave para el desarrollo del documento, la cual permite conocer espacios de mejora y reconocimiento de métodos actualizados para la producción estadística.

Dentro del alcance de la guía se establecen los límites donde se inicia la imputación de datos faltantes, de esta forma facilita al lector la implementación de estos métodos y que resguarden el vínculo con los procesos de procesamiento de datos del subproceso de "imputar".

En resumen, la guía con lineamientos y orientaciones metodológicas para la imputación de datos estadísticos es un documento de referencia para las oficinas de estadísticas que les permite conocer y aplicar las mejores prácticas en materia de imputación de datos.

1.2 La Imputación

La imputación de datos es una técnica crucial en la recopilación y procesamiento de información tanto en encuestas como en registros administrativos, porque dichos constructos pueden presentar problemas de datos faltantes, lo que compromete la calidad y la integridad de la información a procesar. La imputación se utiliza para estimar y completar los valores faltantes utilizando diversas técnicas estadísticas y de modelamiento.

En el caso de las encuestas, los datos faltantes pueden surgir debido a respuestas omitidas por parte de los encuestados o por no conocer la respuesta a una pregunta en particular o porque se nieguen a contestar. La imputación desempeña un papel fundamental al proporcionar valores estimados para estos datos faltantes, permitiendo así obtener resultados más completos y precisos.

Por otro lado, los registros administrativos, como bases de datos administrativas, también pueden contener datos faltantes debido a errores de ingreso, registros incompletos o fallas en la recopilación de información. Por lo tanto, una contramedida es someter a controles de calidad estadística a los registros administrativos antes de que queden a disposición para ser usados. Esto en atención a que estos registros son valiosos para la producción de estadísticas, pero la presencia de datos faltantes, afecta su utilidad y confiabilidad. De manera que la imputación de datos en registros administrativos es esencial para llenar las brechas y garantizar que se disponga de datos completos y confiables para el análisis estadístico.

En ambos casos, la imputación busca minimizar los sesgos y preservar la integridad de los datos. Se utilizan diversos métodos y enfoques, como la imputación simple, la imputación múltiple y técnicas avanzadas basadas en modelos, para estimar los valores faltantes de manera precisa y robusta.

El el siguiente documento, buscamos realizar un primer diagnostico de como se aborda el problema de los datos faltantes (o perdidos) en nuestro INE. Los abordajes métodologicos los realizaremos tanto con notación matematica, basandonos en el lenguaje de programación R Core Team (2019) y diversos paquetes disponibles que iremos ilustrando en los diferentes apartados.

1.3 Evaluación de usos y constumbres vigentes en la institución

Siendo la realidad de nuestro INE la siguiente:

Tabla 1: Tabla resumen de métodos de imputación declarados

Método Imputación	n° productos	n° variables
regresión	xx	xx
razón	xx	xx
media	xx	xx
donantes hotdeck	xx	xx
donantes colddeck	xx	xx
multivariante	xx	xx
multiple	xx	xx
otros tipos de imputación	xx	xx

Nota:
Esta tabla busca ilustrar la realidad de uso de estos métodos y señalar la frecuencia con que son invocados, marcando la necesidad de documentación oficial, por una parte y por otra, consolidar dichos métodos a nivel institucional.

1.4 Mapa conceptual

Conociendo nuestra realidad quedamos en posición de poder orientarnos usando como directrices los alcances y logros presentes en el estado del arte de esta importante técnica, que nos permitímos exhibir mediante el siguiente diagrama conceptual, para dar el contexto necesario acerca de que hacemos y la preponderancia de los métodos usados en la institición.

2 El problema de los datos faltantes

El análisis de datos puede presentar dificultades cuando faltan valores en el conjunto de ellos. Cuando existen valores faltantes, los cálculos y las inferencias estadísticas pueden verse afectados. En el software R, los valores faltantes se indican con el símbolo `NA`.

En el análisis univariado, calcular la media de un conjunto de números es sencillo usando R. Sin embargo, cuando hay valores faltantes, el resultado de la media se vuelve indefinido. Afortunadamente, R ofrece la opción de eliminar los valores faltantes antes de calcular la media. Este parametro así calculado, puede difererir con la presencia de los valores reales.

En el análisis multivariado, el problema se vuelve más complicado. Por ejemplo, al ajustar un modelo de regresión lineal para predecir los niveles de una variable de interés que se

mide diariamente, si hay valores faltantes, R no puede continuar con el análisis. Una forma de solucionar esto es eliminar los registros incompletos antes de ajustar el modelo y nos exponemos a la misma conclusión que el parrafo anterior.

La eliminación de casos incompletos se conoce como eliminación de casos completos. Esto permite que los cálculos se realicen, pero puede introducir complejidades adicionales en la interpretación de los resultados, tal como hemos previsto en los parrafos anteriores.

Además de la eliminación de casos completos, existen otras formas de abordar los datos faltantes. Algunas de estas estrategias incluyen la imputación de valores faltantes, donde se estiman dichos valores en función de los datos existentes, y eventualmente se puede hacer uso de modelos estadísticos para resolver esta misma cuestión.

Es importante tener en cuenta que el enfoque elegido para tratar los datos faltantes puede afectar los resultados finales. No existe una solución única que funcione en todos los casos, por lo que es fundamental comprender las ventajas y limitaciones de cada enfoque. Ello demanda tener una comprensión global de los métodos disponibles y eventualmente proponer innovaciones metodológicas.

2.1 Introducción

Es tarea del Instituto Nacional de Estadística (INE) y otros organismos productores de estadística, proporcionar información estadística de alta calidad sobre muchos aspectos de la sociedad, lo más actualizada y precisa posible. Una de las dificultades para realizar esta tarea surge del hecho de que las fuentes de datos que se utilizan para la producción de resultados estadísticos, tanto las encuestas tradicionales como los datos provenientes de registros administrativos, contienen inevitablemente errores que pueden influir en las estimaciones de las cifras que se publicarán.

Atendiendo lo anterior, este CDC, esencialmente viene establecer la comprensión y aplicación de metodologías relacionadas con la imputación. El objetivo es que el INE pueda proporcionar información estadística de alta calidad sobre diversos aspectos de la sociedad, manteniéndola actualizada y precisa. Sin embargo, debemos tener en cuenta que el principal desafío radica en el hecho de que las fuentes de datos utilizadas para la producción de resultados estadísticos, tanto encuestas tradicionales como registros administrativos, siempre contienen valores faltantes. Por lo tanto, un esfuerzo atendible es intentar corregir estas ausencias, siempre y cuando existan las condiciones adecuadas, pero siempre teniendo presente que esto debe hacerse con precaución para evitar distorsiones en las estimaciones de las cifras a publicar.

Para evitar sesgos e inconsistencias sustanciales en las cifras de publicación, las oficinas de estadística llevan a cabo un extenso proceso de verificación de los datos recopilados.

Siendo así, que esta problematica es un obstaculo serio para el cumplimiento exitoso de la tarea de los INE y otros institutos de estadística y que caracterizaemos como la frecuente ausencia de datos. Esta falta puede considerarse como una forma simple de datos incorrectos, en el sentido de que los valores perdidos son fáciles de identificar. Sin embargo, estimar valores adecuados para estos datos faltantes es complicado.

Una vez que se han recopilado los datos, estos debe ser procesados, ya sea ingresandolos a un sistema, codificandolos e imputandolos. Los errores que surgen durante esta etapa se conocen como errores de procesamiento.

Además de los errores, también pueden surgir datos faltantes en la recopilación de datos. Estos datos faltantes pueden deberse a que un encuestado no conoce la respuesta a una pregunta o se niega a proporcionarla.

La falta de datos es un problema bien conocido que deben afrontar básicamente todos los institutos que recopilan datos sobre personas o empresas. En la literatura estadística, por lo tanto, se presta mucha atención a los datos faltantes. La solución más común para manejar los datos faltantes en conjuntos de datos es la imputación, donde los valores faltantes se estiman y se completan. Siendo relevante que la imputación preserve la distribución estadística del conjunto de datos, siendo por lo tanto, un problema complejo, especialmente para datos de alta dimensión.

2.1.1 El proceso estadístico

Los procesos para manejar datos faltantes, forman parte del quehacer de la producción estadistica y que observan en la práctica los INE. Este proceso de producción de información estadística consta de varias etapas, independientemente de la naturaleza de la encuesta (sociales, económicas, de género, seguridad, etc). Willeboordse (1998) distingue las siguientes fases en el proceso estadístico:

- Establecimiento de objetivos de la encuesta.
- Diseño de cuestionarios y diseño muestral.
- Recolección de datos y entrada de datos.
- Tratamiento y análisis de datos.
- Publicación y difusión de datos.

Siendo necesario detallar lo siguiente:

- **Establecimiento de los objetivos de la encuesta.** En la primera fase, se identifican los grupos de usuarios para la información estadística en consideración, se evalúan las necesidades de los usuarios, se exploran las fuentes de datos disponibles, se consulta a los encuestados potenciales sobre su disposición a cooperar, la encuesta se integra en el marco general de las encuestas, se especifican la población objetivo y las variables objetivo del producto previsto, y se diseña la tabla de resultados.
- **Diseño de cuestionarios y diseño muestral.** En la segunda fase se determina la utilidad potencial de los registros administrativos disponibles, se compara la población marco en el llamado Registro Estadístico con la población objetivo, se define el marco muestral, se selecciona el diseño muestral y el método de estimación, y el cuestionario diseñado. Existe un proceso de decisión sobre cómo recolectar los datos: cuestionarios en papel, entrevistas personales, entrevistas telefónicas o intercambio electrónico de datos.
- **Recolección e ingreso de datos.** En la tercera fase, se extrae la muestra, se recopilan datos de las unidades muestreadas y se ingresan en el sistema informático en la oficina de estadística. Durante esta fase, la oficina de estadística intenta minimizar la carga de respuesta y minimizar su ausencia.
- **Procesamiento y análisis de datos.** En esta fase es cuando se imputan los datos faltantes, se determinan las ponderaciones correspondientes, se estiman las cifras, se itengran los datos, para luego ser analizados (por ejemplo, si se trata de una encuesta

económica, se busca ajustar los efectos estacionales; si se tratase de una encuesta de migraciones, incorporar los flujos de entradas y salidas de localidades).

- **Publicación y difusión de datos.** Esta fase incluyeel establecimiento de políticas de publicación y difusión, la protección de los datos finales (tanto los datos tabulares como los microdatos, es decir, los datos de los encuestados individuales, mediante Control a Divulgación Estadística) contra la divulgación de información sensible y, por último, la publicación de los datos protegidos.

2.1.2 Tipos de Datos Perdidos: MCAR, MAR y MNAR

En el campo del análisis de datos, los datos faltantes son un problema común y demandan eventualmente ser abordados. Los datos perdidos pueden ocurrir por diversos motivos, y por lo tanto, comprender la naturaleza de la falta de datos es crucial para una imputación adecuada de los mismos. Donald B. Rubin (1976a) introdujo los conceptos de MCAR, MAR y MNAR para clasificar los problemas de datos perdidos según los mecanismos subyacentes.

- MCAR (Faltantes Completamente al Azar): Cuando la probabilidad de que los datos estén perdidos es la misma para todos los casos y no está relacionada con los datos en sí, se denomina MCAR. En otras palabras, la falta de datos ocurre al azar y no hay un patrón sistemático o relación con los datos observados. Aunque MCAR simplifica el problema, a menudo es irrealista para la mayoría de los conjuntos de datos. Por ejemplo, si una balanza se queda sin batería y esto resulta en mediciones de peso faltantes, se puede considerar un ejemplo de MCAR. Ignorar los datos faltantes en un escenario MCAR conduce a la pérdida de información, pero no introduce sesgo en el análisis.
- MAR (Faltantes al Azar): MAR es una clase más amplia que MCAR y asume que la probabilidad de falta de datos depende solo de los datos observados. En otras palabras, una vez que se tienen en cuenta los datos observados, la falta de datos es aleatoria. MAR reconoce que la falta de datos puede tener un patrón, pero ese patrón puede explicarse mediante variables observadas. Por ejemplo, si una balanza produce más valores faltantes cuando se coloca sobre una superficie blanda en comparación con una superficie dura, la falta de datos no es MCAR. Sin embargo, si conocemos el tipo de superficie y asumimos que la falta de datos no está relacionada con otros factores no observados dentro de cada tipo de superficie, entonces se puede considerar MAR. Los métodos modernos de manejo de datos perdidos a menudo parten del supuesto MAR.
- MNAR (Faltantes No al Azar): Cuando la falta de datos depende de factores no observados o desconocidos que no están relacionados ni con los datos observados ni con los datos faltantes en sí, se clasifica como MNAR. MNAR implica que la probabilidad de falta de datos varía por razones que no se entienden ni se miden. Es el tipo más complejo y desafiante de datos perdidos. Un ejemplo de MNAR es cuando el mecanismo de una balanza se desgasta con el tiempo, lo que resulta en más datos faltantes a medida que pasa el tiempo. Si se miden objetos más pesados más tarde, la distribución de las mediciones estará distorsionada. MNAR también incluye situaciones en las que ciertos grupos tienen menos probabilidades de responder, lo que conduce a resultados sesgados. El manejo de MNAR requiere información adicional sobre las causas de la falta de datos o la realización de análisis de sensibilidad para evaluar el impacto de diversos escenarios en los resultados.

Comprender la distinción entre MCAR, MAR y MNAR es fundamental porque diferentes métodos para imputar datos perdidos funcionan bajo diferentes supuestos. Soluciones simples, como la imputación de la media o el análisis de casos completos, generalmente asumen MCAR, lo cual puede no ser válido en conjuntos de datos del mundo real. Ignorar MAR o MNAR puede conducir a estimaciones sesgadas e inferencias estadísticas inválidas. Por lo tanto, es importante considerar el mecanismo de falta de datos y elegir estrategias adecuadas de imputación en consecuencia.

2.2 Defincione importantes acerca de datos incompletos

(https://stefvanbuuren.name/fimd/sec-idconcepts.html)

Existen diversas técnicas estadísticas que abordan problemas relacionados con datos incompletos. Supongamos que nuestro interés radica en conocer el ingreso promedio Q de una determinada población. Si tomamos una muestra de dicha población, es probable que las unidades no incluidas en la muestra presenten valores faltantes,porque no han sido medidos. De manera que se determine de forma inmediata el promedio de la población, que se vuelve imposible, dado que la existencia de uno o más valores faltantes hacen indefinido el cálculo del promedio. Así, tenemos qué el enfoque de datos incompletos se presenta como un marco conceptual necesario para analizar los datos como un problema de datos faltantes.

La estimación de un promedio en una población es un problema ampliamente conocido, el cual también puede resolverse sin hacer referencia a datos faltantes. No obstante, resulta útil reflexionar acerca de qué acciones tomaríamos si los datos estuvieran completos y cómo podríamos obtener datos completos. El enfoque de datos incompletos es de naturaleza general y abarca entre otros, el problema del muestreo, el modelo de contrastes orientado a la inferencia causal, el modelado estadístico de los datos faltantes y las técnicas de computación estadística (que contemplen las capacidades necesarias y suficientes para un debido procesamiento).

En el campo de la estadística, existen diversas fuentes de conocimiento que profundizan en la importancia y la amplitud del enfoque de datos perdidos. Los libros de A. Gelman et al. (2004), capítulo 7 y A. Gelman and Meng (2004) brindan discusiones detalladas sobre la generalidad y riqueza de esta perspectiva. Además, Roderick J. Little (2013) enumera diez ideas poderosas para el científico estadístico[1], y su consejo final es especialmente relevante:

My last simple idea is overarching: statistics is basically a missing data problem! Draw a picture of what's missing and find a good model to fill it in, along with a suitable (hopefully well calibrated) method to reflect uncertainty.

(Mi última idea simple es general: ¡las estadísticas son básicamente un problema de datos faltantes! Hacen un dibujo de lo que falta y encuentra un buen modelo para completarlo, junto con un método adecuado (con suerte bien calibrado) para reflejar la incertidumbre.)

[1]El artículo "En alabanza de la simplicidad, no de la matematización: Diez ideas simples y poderosas para el científico estadístico" de Roderick J. Little argumenta que la estadística debe ser vista como una herramienta para resolver problemas del mundo real, en lugar de como una rama de las matemáticas. Little identifica diez ideas simples y poderosas que han influido en su pensamiento sobre la estadística, en sus áreas de interés de investigación: datos faltantes, inferencia causal, muestreo de encuestas y modelado estadístico en general. El tema principal del artículo es que la estadística es un problema de datos faltantes, y el objetivo es predecir desconocidos con las medidas de incertidumbre adecuadas. Little concluye argumentando que la estadística es una herramienta poderosa que se puede utilizar para resolver problemas del mundo real. Sin embargo, advierte contra la tentación de ver la estadística como una rama de las matemáticas y de favorecer la complejidad matemática sobre la simplicidad. Él cree que la simplicidad es a menudo la clave para un análisis estadístico eficaz.

2.3 Circunstancias en la creación de datos perdidos

Existe una amplia distinción entre dos tipos de datos perdidos: datos perdidos intencionales y datos perdidos no intencionales. Los datos perdidos intencionales son planificados por el recolector de datos. Por ejemplo, los datos de una unidad pueden estar ausentes porque se excluyó a esa unidad de la muestra. Otra forma de datos perdidos intencionales es el uso de diferentes versiones del mismo instrumento para diferentes subgrupos, enfoque conocido como muestreo en matriz. Ahora para tener una idea cabal de esta situación, consulte a Gonzalez and Eltinge (2007) o a Graham (2012) para obtener una visión general al respecto. Además, los datos perdidos que ocurren debido al enrutamiento (sucesión programada de las respuestas) en un cuestionario son intencionales, al igual que los datos (por ejemplo, tiempos de supervivencia) que son datos censurados en algún momento porque el evento (por ejemplo, la muerte) aún no ha ocurrido. Un término relacionado en un contexto multinivel es el de datos sistemáticamente perdidos, que se refiere a variables que están ausentes para todos los individuos en un grupo porque la variable no se midió en ese grupo Resche-Rigon and White (2018).

Aunque a menudo se prevén, los datos perdidos no intencionales son no planificados y no están bajo el control del recolector de datos. Algunos ejemplos son: el encuestado omitió un ítem (negarse a responder el monto del ingreso), hubo un error en la transmisión de la información que provocó la pérdida de datos (intermitencia en la señal de internet), algunas unidades fueron excluidas antes de que pudiera completar el estudio (exclusión territorial por causas de seguridad o sanitarias), lo que resulta en datos parcialmente completos, o el encuestado fue seleccionado pero se negó a cooperar, tambien encontramos en un contexto multinivel la acepción de datos perdidos de forma esporádica, que dice relación para variables con valores faltantes para algunos pero no todos los individuos en un grupo.

Otra distinción importante es la *falta de respuesta de ítems* frente a la *falta de respuesta de unidades*. La falta de respuesta de ítems se refiere a la situación en la que el encuestado omitió uno o más ítems en la encuesta. La falta de respuesta de unidades ocurre si el encuestado se negó a participar, por lo que todos los datos de resultado están ausentes para este encuestado. Históricamente, los métodos para la falta de respuesta de ítems y la falta de respuesta de unidades han sido bastante diferentes, con la falta de respuesta de unidades abordada principalmente mediante métodos de ponderación, y la falta de respuesta de ítems abordada principalmente mediante técnicas de edición e imputación.

2.4 Mecanismo de Datos Faltantes (MA)

Previamente habiamos anunciado y descrito los tipos de datos perdidos, y ahora corresponde profundizar y contextualizar lo señalado en el parrafo 2.1.2. Cabe consignar un antes y un después del trabajo de Rubin (Donald B. Rubin 1976b), porque 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?* (Donald B. Rubin 1976b, 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 (Donald B. Rubin 1976b, pag. 582). Esto, dentro la literatura

Tabla 2: Ejemplos de motivos para la falta de datos en combinacion con datos faltantes intencionales/no intencionales con falta de respuesta de elementos/items.

	Intencional	No-intentional
Falta de respuesta de la unidad	Muestreo	Rechazo
		Autoselección
Falta de respuesta del Item	Matriz de muestreo	Salto de pregunta
	Derivación	Error de codificación

Nota:
La tabla muestra un cruce de ambas distinciones y proporciona algunos ejemplos típicos en cada una de las cuatro celdas. La distinción entre datos perdidos intencionales/no intencionales es la más importante. La distinción entre falta de respuesta de ítems/unidades indica cuánta información falta, mientras que la distinción entre datos perdidos intencionales y no intencionales indica por qué falta cierta información. Conocer las razones por las cuales los datos están incompletos es el primer paso hacia la solución.

que aborda el problema de la falta de datos, se conoce como situación 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*.

2.4.1 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 los campos vacios 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* (R. J. A. Little and Rubin 2020a, pag. 13).

Con el propósito de formalizar lo antes señalado, consideremos la siguiente notación[2]. 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*[3], $R = \{r_{ij}\}$, matriz de

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

[3]Como se menciona en (R. J. A. Little and Rubin 2020a, 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 (R. J. A. Little and Rubin 2020a, 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,

$n \times p$ 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 (R. J. A. Little and Rubin 2020a, 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, $Y = (Y_{obs}, Y_{miss})$[4]. 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 (R. J. A. Little and Rubin 2020a, pag. 13).

Tomando el trabajo de Rubin (Donald B. Rubin 1976b), Little y Rubin (R. J. A. Little and Rubin 2020a, 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.

2.4.2 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

and Hsu 2021a, pag. 7), entre otros.

[4]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).

para un mecanismo *MCAR* (He, Zhang, and Hsu 2021a, pag. 13), es[5]:

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

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 (R. J. A. Little and Rubin 2020a, pag. 78; Van Buuren 2012, pag. 7).

2.4.3 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 2021a, pag. 13), es[6]:

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

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*.

2.4.4 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

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

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

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

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

de R dado Y. Entonces, la distribución para un mecanismo *MNAR* (Van Buuren 2012, pag. 31), esˆ[Si se utiliza la *matriz indicadora de falta de respuesta*, M, equivalentemente, la distribución para un mecanismo *MNAR* (Enders 2022, pag. 11), es:

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

]:

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

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

2.5 Ignorabilidad (MA)

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 *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 (Donald B. Rubin 1976b) 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?* (Donald B. Rubin 1976b, 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 (Donald B. Rubin 1976b, 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 Métodos de Imputación Simple y Métodos tradicionales de Imputación

3.1 Introducción

Con frecuencia, faltan valores en las encuestas. En muchos casos, un encuestado no respondió una o más preguntas en una encuesta que se suponía que debía responder, mientras que sí respondió las otras preguntas. Esto se conoce como falta de respuesta de ítem (o algunas veces como falta de respuesta parcial). Hay varias razones para no responder una pregunta. El entrevistado puede no entender la pregunta, puede no saber la respuesta a la pregunta, puede olvidarse de responder la pregunta, puede negarse a responder la pregunta porque considera que la respuesta a la pregunta es información privada, puede negarse a responder la pregunta porque toma demasiado tiempo responder el cuestionario completo, etc. Además, en los registros, los valores perdidos ocurren con frecuencia para elementos que se supone que no faltan. Por último, es posible que los valores se hayan establecido como perdidos durante la fase de edición o que los datos simplemente se hayan perdido mientras se procesan en el instituto de estadística.

Es posible que algunas unidades de población a las que se les pidió que respondieran a un cuestionario no respondieran en absoluto. De manera similar, algunas unidades pertenecientes a la población sobre la que el instituto de estadística quisiera informar pueden faltar por completo en un registro. Estos casos, en los que faltan registros completos de posibles encuestados o unidades potenciales en un registro, se denominan no respuesta de unidad. A menos que, dicho de otra manera, siempre que hagamos referencia a valores perdidos en este documento, nos referiremos a valores perdidos debido a la falta de respuesta del elemento en lugar de la falta de respuesta de la unidad.

En el caso de una respuesta parcial, el investigador tiene que decidir si se ha dado un número suficiente de respuestas para considerar el registro como una respuesta y, por lo tanto, los elementos que faltan como una no respuesta del elemento, o si no se han dado suficientes respuestas y el registro debe ser considerado como falta de respuesta de la unidad. En el caso de la falta de respuesta de la unidad, ponderar la respuesta de la encuesta o las unidades disponibles en los registros es un enfoque válido para reducir el efecto de la falta de respuesta en las estimaciones de población (ver J. Bethlehem (2009)).

Otra forma de lidiar con los valores perdidos es *imputar*, es decir, estimar y completar, un valor factible para un valor faltante en el conjunto de datos. Esto se conoce como imputación. La imputación es parte del proceso de producción, es decir, el proceso que abarca toda la edición, la imputación y otras acciones realizadas para transformar los datos sin procesar en un conjunto de datos estadísticos listos para su análisis y tabulación. Sin embargo, la imputación no es un paso necesario en el proceso de producción: uno puede decidir dejar algunos valores faltantes y tratar de resolver el problema de estimación más tarde ponderando la encuesta o aplicando técnicas de análisis.

Hacemos una distinción entre imputación y *derivación*. Cuando se derivan variables, se crean nuevas variables. Estas variables pueden verse como funciones de las variables que ya están contenidas en el conjunto de datos. Al imputar valores perdidos, se crean valores en variables ya existentes.

Durante el proceso de edición estadística, los errores se detectan y corrigen. Cuando un valor se considera erróneo y se considera que el valor en sí no participa en el proceso de corrección, la sustitución de este valor por uno mejor, se considera imputación. En algunos casos, sin embargo, se considera que el valor original erróneo juega un papel importante en la estimación de un mejor valor. La modificación de tales valores no se llama imputación sino *corrección*.

Algunos valores perdidos faltan correctamente y deben reconocerse como tales para evitar que se imputen. Por ejemplo, los hombres no tienen que responder, ni pueden, responder a la pregunta cuando dieron a luz a su primer hijo, y las personas sin trabajo no responden ni pueden responder dónde están empleados. Respuestas como "no sé", "sin opinión" y "desconocido" también son valores válidos si se trata de respuestas a preguntas sobre el conocimiento o la opinión del encuestado.

Incluso cuando faltan valores injustificadamente, se puede decidir no imputar estos valores perdidos. Como ya mencionamos, en lugar de resolver el problema de los valores perdidos en el conjunto de datos por medio de la imputación, se puede intentar resolver este problema en una fase de estimación o análisis posterior. En el caso de los datos categóricos, también se tiene la opción de introducir una categoría especial para los datos faltantes: "desconocidos". Esta es una razón por la que la imputación se aplica con más frecuencia a las variables numéricas que a las categóricas y, por lo tanto, con mayor frecuencia a las variables de estadísticas económicas que para las estadísticas sociales.

Las razones importantes para imputar los datos faltantes en lugar de dejar los campos vacíos son obtener un conjunto de datos completo y mejorar la calidad de los datos. Un conjunto de datos completo, con registros completos, hace que sea más fácil agregar microdatos, construir tablas a partir de estos microdatos y garantizar la coherencia entre las tablas construidas. Por ejemplo, los valores perdidos en una variable Ocupación pueden resultar en una distribución diferente de una variable Edad en la tabla "Edad × Ocupación" que en una tabla "Edad × Sexo", a menos que los valores perdidos de Ocupación estén codificados como una categoría especial "desconocida". Cuando en una encuesta faltan valores para una variable numérica Ingreso, entonces solo se puede estimar un Ingreso para la subpoblación de personas que respondieron al cuestionario y no para la población en su conjunto. La imputación puede superar este problema, pero solo cuando los valores imputados sean de calidad suficientemente alta.

Cuando se quiere aplicar la imputación para mejorar la calidad de los datos, se debe tener claro qué aspecto de la calidad de los datos se quiere mejorar realmente. A menudo, el objetivo principal de un instituto de estadística es estimar medias y totales. En otros casos, el objetivo principal es estimar la distribución de una variable lo mejor posible, por ejemplo, la distribución de la renta entre varios grupos de la población. En otros casos más, se desea tener un conjunto de microdatos que los investigadores puedan utilizar para realizar muchos tipos diferentes de análisis estadísticos. Diferentes propósitos pueden llevar a diferentes imputaciones "óptimas". Sin embargo, es posible que los institutos nacionales de estadística (INE) generalmente prefieren (como máximo) un valor imputado por valor faltante para garantizar la coherencia entre las diversas tablas y otros resultados que publican. En general, el instituto de estadística recopila y procesa datos pueden determinar mejores estimaciones de los valores perdidos que otras organizaciones, ya que el instituto de estadística generalmente tiene mucha más información de respaldo que se puede utilizar como datos auxiliares en el proceso de estimación para producir las imputaciones.

A veces, el valor "verdadero" de un valor perdido se puede determinar con certeza a partir de las otras características de la unidad. En tal caso, se puede aplicar la imputación deductiva. Para la imputación deductiva, se pueden utilizar las restricciones de edición que también se utilizan en el proceso de edición. Si se puede utilizar la imputación deductiva, se prefiere este método a cualquier otro método de imputación. A veces, este método de imputación también se utiliza cuando el valor real no puede estimarse con certeza, sino solo con una probabilidad (muy) alta.

Cuando no se puede aplicar la imputación deductiva, a menudo todavía hay información adicional disponible (auxiliares o variables x) que permite predecir los valores perdidos en una variable y objetivo con bastante precisión. Si se puede construir un modelo que prediga bien la variable objetivo, se puede usar la *imputación basada en modelos* para mejorar la calidad del conjunto de datos o de las estimaciones de los parámetros (de población) de interés. Los valores predichos según el modelo seleccionado son las imputaciones o estimaciones de los valores perdidos. Los modelos de regresión, principalmente para variables numéricas, son los modelos de imputación más utilizados. La imputación basada en un modelo de regresión se denomina *imputación por regresión.*

Aparte de los modelos paramétricos (de regresión), los enfoques no paramétricos también se utilizan a menudo para obtener valores imputados. En particular, los métodos de donante hot-deck que copian el valor de otro registro para completar el valor faltante son alternativas populares que se pueden aplicar tanto a variables numéricas como categóricas. El objetivo de estos métodos es similar a la regresión, pero son algo más fáciles de aplicar cuando se deben imputar varios valores perdidos relacionados en un registro, y uno tiene como objetivo preservar las relaciones entre las variables. Cuando se aplica la imputación de donante, para cada no encuestado i se busca un registro de donante d que sea lo más similar posible al registro i con respecto a ciertas características de fondo que se (se consideran) relacionadas con la variable objetivo y.

La imputación de valores perdidos no implica necesariamente que los datos después de la imputación sean internamente coherentes, en el sentido de que se satisfacen todas las restricciones de edición. Se pueden agregar restricciones de edición como restricciones al proceso de imputación y, por lo tanto, garantizar que solo se imputen los valores permitidos y que no surjan inconsistencias después de la imputación. Un enfoque alternativo es imputar primero los valores faltantes sin tener en cuenta las restricciones, y luego ajustar los valores imputados para que se satisfagan todas las restricciones de edición.

3.2 Problemas generales en la aplicación de métodos de imputación

3.2.1 Modelos de imputación por subpoblación

Se puede construir un modelo de imputación para toda la población o para subpoblaciones, por ejemplo, definido por "rama de la industria" por "clase de tamaño" para las estadísticas comerciales, por separado. Distinguiendo ese tipo de clases de imputación (grupos de imputación) puede ser beneficioso cuando hay poca variación dentro de estas clases con respecto a los puntajes de la variable objetivo y, y los puntajes entre clases difieren fuertemente. Para la imputación por regresión, se puede ver la distinción de subpoblaciones como parte del proceso de modelado, ya que el análisis de regresión puede tener en cuenta variables x categóricas en el modelo de imputación. Esto se puede realizarse incorporando las variables

categóricas (y sus términos de interacción) correspondientes a estas subpoblaciones como variables *dummy* en el modelo de regresión. La imputación por donantes hot-deck está destinada a las variables x categóricas, es decir, a las subpoblaciones.

3.2.2 Ponderación

En la mayoría de los métodos de imputación, se tiene la opción de ponderar a los encuestados de los ítems, por ejemplo, estableciendo sus ponderaciones iguales a los recíprocos de las probabilidades de muestreo, o al aumento de ponderaciones que se obtienen después de corregir el muestreo, ponderaciones para la falta de respuesta selectiva de la unidad J. Bethlehem (2009). En el caso de la imputación por regresión lineal, esto implica que se utiliza la estimación por mínimos cuadrados ponderados en lugar de la estimación por mínimos cuadrados ordinarios para estimar los parámetros del modelo, y en el caso de la imputación aleatoria de donantes hot-desk implica que los donantes potenciales con un peso mayor tienen una mayor probabilidad de ser seleccionado como donante que los posibles donantes con un peso menor. El uso de ponderaciones no tiene ningún efecto sobre la imputación deductiva.

No hay un consejo claro sobre si usar poderación o no. Desde una perspectiva basada en modelos, cada registro se mide con la misma precisión, asumiendo residuos distribuidos de manera idéntica, independientemente de las probabilidades de muestreo o las probabilidades de respuesta. Desde esta perspectiva, si se cree en la validez del modelo de imputación, por lo tanto, no es necesario utilizar ponderaciones, y es incluso mejor no utilizar ponderaciones, porque la ponderación infla los errores estándar. Si se incluye la variable que contiene las ponderaciones (o las variables que se han utilizado para calcular estas ponderaciones) como variables explicativas en el modelo, la ponderación es innecesaria de todos modos. Al seleccionar las variables auxiliares para el modelo de imputación, se debe tener esto en cuenta.

Sin embargo, desde el punto de vista de la teoría del muestreo, las respuestas de una unidad de muestra son *"representativas"* para algunas unidades de población que no están incluidas en la muestra. Implícitamente, se asume más o menos que estas unidades de población habrían dado las mismas respuestas que la unidad muestral. Desde este punto de vista, y suponiendo que la unidad no responda selectivamente, la ponderación es necesaria para obtener resultados insesgados basados en el diseño de muestreo. La cuestión de la inferencia basada en modelos versus la inferencia basada en el diseño se analiza ampliamente en Skinner, Holt, and Smith (1989).

Andridge and Little (2009) muestran mediante simulación que cuando se utiliza la imputación de donantes por hot-deck para imputar valores faltantes de una variable numérica, con el objetivo de estimar la media poblacional de esta variable, el mejor enfoque es utilizar la ponderación muestral como una estratificación junto a variables auxiliares adicionales al formar clases de imputación.

3.2.3 Imputación Masiva

A veces, uno quiere imputar valores no solo para los ítems que no respondieron, pero para todas las unidades que no están en la muestra. Esto se conoce como imputación masiva, incluso si solo se va a imputar una variable objetivo. Después de la imputación masiva, es fácil calcular los totales y las medias de la población para la variable objetivo y, los totales se obtienen simplemente sumando todos los valores (observados o imputados) para y y las

medias dividiendo estos totales por el número de unidades de la población. Para la imputación Hot Deck ponderada, esto corresponde al uso del llamado estimador de postestratificación, y para la imputación por regresión con estimación por mínimos cuadrados ponderados; véase, por ejemplo, C.-E. Särndal and Lundstrom (2005), J. Bethlehem (2009).

En cuanto a la imputación de solo los ítems que no respondieron, para la imputación masiva nuevamente tenemos que la ponderación se vuelve menos importante si se incluyen más variables utilizadas para determinar la variable de ponderación como variables auxiliares en el modelo de imputación. Sin embargo, en ocasiones esto es imposible porque las variables que se utilizan para determinar la variable de peso solo están disponibles para las unidades de la muestra. En ese caso, la ponderación es una opción a considerar.

Las experiencias con la imputación masiva reportadas en la literatura conducen a diferentes conclusiones. Mientras que Kaufman and Scheuren (1997) informan que estaban decepcionados con el desempeño de la imputación masiva y que hasta el momento no ha cumplido con sus expectativas, Krotki and Creel (2005) concluyen que "la imputación masiva se está convirtiendo en una herramienta más en el conjunto de herramientas de los estadísticos de encuestas para los cuales existe una demanda cada vez mayor". Shlomo, De Waal, and Pannekoek (2009) informan buenos resultados de evaluación para la imputación masiva en un estudio de evaluación limitado. Nuestra conclusión general es que la imputación masiva es todavía un área relativamente inexplorada que ofrece muchas oportunidades para futuras investigaciones, pero lamentablemente aún no es una solución a todos los problemas para la falta de respuesta.

3.2.4 Selección de variables auxiliares

En este documento no tratamos en detalle cómo seleccionar variables e interacciones auxiliares adecuadas. Al igual que el análisis de regresión, la selección de variables auxiliares e interacciones adecuadas para el modelo de regresión es parte del análisis multivariante sobre el que existe una amplia literatura. La idea básica es buscar variables auxiliares o x que estén fuertemente correlacionadas con la variable objetivo y. Por lo general, es una cuestión de prueba y error en combinación con el análisis estadístico y el sentido común para seleccionar variables auxiliares e interacciones adecuadas para el modelo de regresión, pero también se pueden usar procedimientos de búsqueda hacia adelante o hacia atrás para agregar automáticamente variables auxiliares al modelo de regresión y eliminarlas del modelo de regresión, respectivamente. En general, estos procedimientos de búsqueda automática son de naturaleza no paramétrica.

En primer lugar, se seleccionan aquellas variables como variables auxiliares para el modelo de imputación para las que se puede esperar que también sean relevantes para el ítem no encuestados. En general, se utilizará el ítem encuestados para comprobar si las variables auxiliares son capaces de explicar bien la variable objetivo, ya que una prueba del modelo de imputación para los ítems no encuestados es imposible.

En segundo lugar, no incluir demasiadas variables auxiliares en un modelo de regresión. Las estimaciones de los parámetros de dicho modelo tendrían errores estándar grandes. Para obtener buenos valores imputados, es preferible un modelo con pocas variables auxiliares. Debemos considerar el principio de parsimonia.

En tercer lugar, cuando se aplica la imputación por donantes, no importa si se distinguen

muchas subpoblaciones. Incluso agregar variables que no tienen valor explicativo para la variable objetivo con el único propósito de obtener un donante único no es un problema. De hecho, esto puede verse como una alternativa para extraer un donante aleatorio de una clase de imputación. En este caso, hay que tener cuidado de no atraer al mismo donante con demasiada frecuencia.

Cuarto, al agregar variables secuencialmente al modelo de imputación, usar medidas de calidad como el aumento de R^2, la $prueba - F$, AIC y BIC (consulte, por ejemplo, Burnham and Anderson 2002 para obtener más información sobre AIC y BIC) para determinar los beneficios de agregar otra variable al modelo frente a no agregar esta variable. Cabe señalar que cuando se agregan secuencialmente variables al modelo de imputación, el orden de agregar variables al modelo es parte del proceso de selección del modelo.

En quinto lugar, a menudo es importante incluir variables de diseño, por ejemplo, variables que definen estratos de muestreo que tienen probabilidades de inclusión diferenciales.

3.2.5 Puntos atípicos

Si se producen valores atípicos en una variable objetivo numérica y entre los encuestados del ítem, se puede considerar limitar su influencia en el proceso de imputación. Por ejemplo, se puede realizar un análisis de regresión robusto, o dar a un donante potencial con un valor atípico en la variable objetivo y, es decir, atípico dados los valores de las variables auxiliares, una menor probabilidad de ser seleccionado como donante. Hay que tener en cuenta que los valores atípicos durante el proceso de imputación conducen a errores estándar más pequeños, pero puede dar lugar a resultados sesgados. Por lo tanto, hay que tener cuidado al tener en cuenta los valores atípicos durante el proceso de imputación y hay que considerar cuidadosamente qué estimaciones se quieren obtener. Por ejemplo, los métodos robustos suelen ser apropiados para pequeñas subpoblaciones porque, de lo contrario, los errores estándar pueden volverse demasiado grandes; para poblaciones grandes, los métodos robustos suelen ser menos apropiados o beneficiosos. Por ejemplo, cuando alguien gana un millón de dolares al año, se puede utilizar a esta persona como donante para obtener resultados de todo un país. Sin embargo, si esta persona vive en un barrio pobre de un pequeño pueblo, obviamente no es una buena idea utilizar a esta persona como donante para obtener resultados para este barrio. Para decidir cómo tratar los valores atípicos, se debe utilizar el conocimiento de la materia en combinación con el análisis estadístico.

3.2.6 Marcación de los datos imputados

Para la imputación, es importante que los valores faltantes estén claramente indicados en el conjunto de datos. Esto se puede hacer dando a los valores que faltan un código especial, como **-1**, **9** o **99**, si esto no genera confusión con los posibles valores correctos. Se considera una mala práctica codificar los valores faltantes por ceros cuando los ceros pueden ser respuestas correctas, o viceversa codificar ceros por valores faltantes. Ambas situaciones ocurren a veces en las encuestas económicas. Si se hace esto, no se puede hacer ninguna distinción entre los valores perdidos y los ceros verdaderos.

Cuando un instituto de estadística imputa valores faltantes para obtener un conjunto de datos completo que luego se entrega a investigadores externos, es importante documentar qué valores se han imputado y qué métodos se han aplicado para hacerlo, incluidas las variables

auxiliares incluidas en el modelo de imputación y los parámetros del modelo utilizado. Esto es necesario para que el investigador pueda determinar por sí mismo, si quiere usar los valores imputados por el instituto de estadística o si es mejor para sus objetivos de investigación imputar los valores que originalmente le faltaban. Además, para determinar los errores estándar correctos, el investigador puede necesitar saber qué valores se han imputado y qué método de imputación se ha aplicado.

3.3 Imputación de regresión

3.3.1 El modelo de imputación de regresión

Cuando se aplica la *imputación de regresión*, se utiliza un modelo de regresión adecuado, basado en cero, una o varias variables auxiliares x, para predecir un valor para un valor faltante y_i de la variable objetivo y en el registro i. Se obtiene una imputación del valor faltante, basada en la predicción del modelo. El modelo de regresión lineal está dado por:

$$y = \alpha + \beta_1 x_1 + ... + \beta_p x_p + \epsilon = \alpha + x^T \beta + \epsilon, \tag{4}$$

donde α, β,..., β_p denota los parámetros del modelo, y ϵ denota los residuos. Sustituyendo estimaciones de los parámetros produce una variable predicha:

$$\hat{y} = \hat{\alpha} + x^T \hat{\beta} \tag{5}$$

Esta variable predicha $\hat{y}$ se define tanto para los encuestados del ítem como para los no encuestados del ítem.

Para cada elemento que no responde en la variable objetivo y, se puede imputar la mejor predicción o se puede agregar un residuo aleatorio. Es decir, existen dos opciones básicas para determinar un valor imputado $\hat{y}$ para un ítem que no responde:

1. Sin residuos

$$\tilde{y_i} = \hat{y_i} = \hat{\alpha} + x_i^T \hat{\beta}. \tag{6}$$

2. Con residuos

$$\tilde{y_i} = \hat{y_i} + \epsilon_i = \hat{\alpha} + x_i^T \hat{\beta} + \epsilon_i. \tag{7}$$

Si en la fórmula (4) no se utilizan variables auxiliares x, esta fórmula se convierte en $y = \mu + \epsilon$ el valor esperado de y, y la fórmula (6) se reduce a $\tilde{y} = \tilde{\mu} = \overline{y}$. Esto es imputación media. Tratamos este método por separado debido a su popularidad en la práctica. La imputación sin utilizar información auxiliar sólo puede justificarse cuando existen pocos ítems que no respondan y las imputaciones apenas tienen efecto sobre los parámetros a estimar.

Si no se usa un término constante y solo una variable auxiliar numérica x, la fórmula (4) se convierte en $y = R_x + \epsilon$, y (6) se reduce a la imputación de razón.

Cuál de las dos opciones (6) o (7) debe elegirse depende del objetivo de la imputación. Con el fin de estimar medias y totales, no es necesario agregar un residual aleatorio al valor predicho,

e incluso puede dar lugar a resultados sesgados, a menos que la expectativa de los residuales sea igual a cero, pero si el objetivo también es estimar la variación de la variable objetivo y, la opción preferida es agregar un residuo aleatorio a las imputaciones.

Si se imputa la mejor predicción posible de acuerdo con el modelo de regresión para todos los valores faltantes, los datos imputados se suavizan mucho; es decir, todos los valores imputados se ajustan perfectamente a la regresión. Aparte de estimar totales y medias, los datos imputados serán bastante inútiles para otros tipos de análisis de los microdatos o, en algunos casos, incluso de datos tabulares. Un ejemplo simple es una estadística demográfica nacional de la población, donde para cada edad desconocida del esposo o la esposa se usa el modelo de imputación de que el esposo es 2 años mayor que la esposa. Tal modelo de imputación puede ser un buen modelo para la distribución de edades de esposos y esposas, pero si investigadores externos examinaran los microdatos más tarde, podrían descubrir un pico inesperado en la distribución de la diferencia de edad entre hombres y mujeres.

En general, la imputación de la mejor predicción posible según el modelo de regresión conduce a una subestimación de la variación de las puntuaciones ("regresión hacia la media"). Conduce a distribuciones con picos y colas que son demasiado delgadas, especialmente cuando la variable objetivo y tiene muchos valores faltantes y el modelo de regresión explica poco de la varianza de y. El efecto es más fuerte para la imputación media. Esto no genera ningún problema cuando solo se desea estimar medias y totales, pero sí lo es cuando se desea estimar distribuciones o medidas de dispersión.

Si también se desea estimar la distribución en lugar de solo los totales y las medias, se recomienda agregar un residuo aleatorio al mejor valor predicho posible. Para la imputación de regresión, el residuo ϵ en (7) se puede determinar de dos maneras:

(a) $\epsilon_i = \epsilon_d$ con ϵ_d el residual de un donante arbitrario o especialmente seleccionado.

(b) ϵ_i es construido desde una distribución estocastica. Por ejemplo, una distribución normal.

En el caso **(b)** la expectativa de la distribución normal generalmente es igual a cero, y la varianza a menudo se estima mediante el error residual del modelo de regresión (4).

La imputación de regresión se puede aplicar a grupos separados o clases de imputación, es decir, subpoblaciones. En ese caso, para cada grupo estimamos parámetros de modelo separados o incluso desarrollamos un modelo de imputación separado. Cada grupo se define aquí en términos de variables auxiliares.

Por lo general, se usa un modelo lineal (4), pero, en principio, también se pueden usar modelos no lineales. Un tipo específico de modelo no lineal es el modelo lineal generalizado (ver McCullagh and Nelder 1989), que tiene la forma:

$$y = f(x^T \beta). \tag{8}$$

El término residual ϵ se puede agregar explícitamente al modelo (8) o puede ser parte del modelo implícitamente.

Las variables auxiliares del modelo de regresión pueden ser variables continuas o variables dummy para representar las categorías de variables categóricas. Cuando solo se incluyen

variables auxiliares categóricas, el análisis de regresión lineal a veces también se denomina "análisis de varianza".

Cuando se imputa $\hat{y}_i$ mediante la fórmula (6), las imputaciones no tienen efecto en la estimación del total poblacional, si se utiliza el denominado estimador de regresión con el mismo modelo que el modelo de imputación; véase C.-E. Särndal and Lundstrom (2005)).

La imputación de regresión se aplica principalmente cuando y es una variable numérica. Cuando y es una variable categórica, también se puede usar un enfoque de regresión, pero luego se aplica una transformación de la variable objetivo, como en una regresión logística binaria o multinomial McCullagh and Nelder (1989). Para una variable binaria y con puntuaciones posibles de 0 y 1, el modelo de regresión logística es:

$$ln\frac{pr}{1-pr} = \alpha + \beta_1 x_1 + ... + \beta_p x_p \equiv \alpha + x^T\beta \tag{9}$$

con pr la probabilidad de que y asuma la puntuación 1, dadas las variables x y el modelo planteado. Si falta un valor de y, se puede estimar el parámetro β — por ejemplo, mediante el enfoque de máxima verosimilitud— y posteriormente estimar la probabilidad $\hat{pr}$ donde el valor de marca 1 es imputado por:

$$y = \frac{e^{\hat{\alpha}+x^T\hat{\beta}}}{1+e^{\hat{\alpha}+x^T\hat{\beta}}} = \frac{1}{1+e^{-\left(\hat{\alpha}+x^T\hat{\beta}\right)}} \tag{10}$$

Estas probabilidades se pueden calcular fácilmente mediante varios paquetes de software estadístico, como SPSS y R.

Más adelante señalamos la posibilidad de un análisis de regresión ponderado, si se quiere que el análisis de regresión refleje que algunos encuestados poseen un mayor peso que otros. La heterogeneidad de los términos residuales puede ser otra razón para realizar un análisis de regresión mediante estimación por mínimos cuadrados ponderados, en lugar de estimación por mínimos cuadrados ordinarios (no ponderados).

En este documento no discutimos la teoría del análisis de regresión en general. Hay muchos libros excelentes disponibles que tratan el análisis de regresión en detalle, (p. ej., Draper and Smith 1998). Con respecto a la selección del modelo, se hicieron algunos comentarios generales previamente.

3.3.1.1 Ejemplos de imputación de regresión

3.3.1.1.1 Ejemplo 1: Estadísticas de hogares holandeses Cada año, Oficina de Estadísticas de Holanda recibe una versión de la denominada Administración de base municipal. La Administración de Base Municipal contiene para cada domicilio los datos de las personas que viven en el domicilio, incluyendo las relaciones familiares. Sin embargo, falta información sobre cómo se componen exactamente los hogares que viven en la dirección. Para las Estadísticas de Hogares anuales es esencial saber qué personas viven en la misma dirección que constituyen un hogar de acuerdo con la definición aplicada en Estadísticas de los Países Bajos. A partir de 1999 se utiliza la Administración de Base Municipal para determinar las principales variables

Número de hogares y *Composición del hogar* a partir de la estructura de la familia o familias que habitan en el domicilio. Para más del 90% de las direcciones de la Administración de Base Municipal se puede construir la información de estas variables derivadas. Sin embargo, para las direcciones restantes, no se puede derivar ni el número de hogares ni la composición exacta del hogar. Para estas direcciones se utiliza la imputación, con modelos de imputación separados para diferentes situaciones.

En este ejemplo, discutimos el tipo más simple de direcciones con una composición del hogar desconocida: direcciones con dos personas no emparentadas, es decir, dos personas que no están casadas ni registradas como pareja entre sí y que no son familiares entre sí. Para estas direcciones se desconoce si las dos personas juntas constituyen un hogar o si son solteros y cada uno tiene su propio hogar.

En primer lugar, se aplica la imputación deductiva, mediante una regla deductiva: Cuando ambas personas comenzaron a vivir en el domicilio en la misma fecha según la Administración de Base Municipal, entonces se considera que constituyen un solo hogar. Esto conducirá a una ligera sobreestimación del verdadero número de hogares. Las direcciones restantes están vinculadas a la *Encuesta de Población Activa*. Para 1999 se obtuvieron así 1662 domicilios con dos personas. A partir de estos datos se construyó un modelo de imputación.

Mediante la información obtenida de los entrevistadores de la Encuesta de población activa y los datos reales recopilados por medio de la Encuesta de población activa, para cada una de las 1662 direcciones se determinó si la dirección contenía uno o dos hogares. A veces esto era bastante complicado por falta de respuesta o porque resultaba que había una diferencia entre la ocupación real y registrada de la dirección. La probabilidad de dos hogares resultó estar fuertemente correlacionada con la edad de ambas personas, en particular, con la diferencia de edad entre las dos personas, ya sea que las personas tuvieran o no el mismo género, independientemente del grado de urbanización e independientemente de el número de personas solteras en la dirección. Se desarrolló un modelo de regresión logística con estas variables como variables auxiliares.

A continuación, se utilizó la fórmula (10) para cada dirección con dos personas no emparentadas en la *Administración de Base Municipal* que no se vinculó a la *Encuesta de Fuerza Laboral* para estimar la probabilidad de que la dirección contenga dos hogares. La probabilidad estimada se usó para determinar "dos hogares" o "un hogar".

Este es un ejemplo de imputación de registro: se imputan todas las direcciones con un score desconocido en el *Número de hogares*. Las puntuaciones desconocidas son, además, muy selectivas. Es decir, *Número de viviendas* es una variable derivada que solo para grupos específicos no se puede determinar desde la *Administración de Base Municipal*. Solo al vincular el registro a una encuesta, se dispone de información estocástica sobre el número de hogares para esos grupos.

3.3.1.1.2 Ejemplo 2: Encuesta de bibliotecas públicas holandesas Para las variables continuas en las encuestas sociales, la imputación de regresión lineal se usaba, y se usa a menudo en la oficina de estadísticas de Holanda. Por otro lado, la imputación de regresión lineal, no ha sido utilizada, en estadísticas de los Países Bajos. Sin embargo, existen excepciones. Para la Encuesta de Bibliotecas Públicas se examinaron varios modelos de imputación de regresión, a saber, imputación de razón, imputación de regresión lineal e imputación de regresión no lineal

(ver Groot and Dekker 2001). El modelo de regresión lineal incluyó un término cuadrático, lo cual no es de uso común, lo que resultó en el modelo:

$$y = \alpha x^2 + \beta x + \epsilon,$$

donde como es habitual y es la variable objetivo, x la variable auxiliar y α y β los parámetros del modelo de regresión. El residual ϵ es un término de error estocástico con valor esperado cero y varianza σ^2.

También se examinó el siguiente modelo de regresión no lineal:

$$y = \beta x^{\alpha} + \epsilon,$$

donde y, x, α, β y ϵ se definen de la misma manera que antes. Para este conjunto de datos en particular, los modelos de imputación lineal parecían dar mejores resultados que el modelo no lineal.

Este ejemplo, así como varios otros en este documento, han sido encontrados en De Waal (2000). Aunque ha habido algunos cambios con respecto a los métodos de imputación aplicados en Oficina de Estadísticas de Holanda desde 1999, el panorama general no ha cambiado mucho. Las principales técnicas de imputación que se aplicaban en 1999 todavía se aplican en la actualidad. A menudo, las diferencias entre 1999 y ahora no son tanto con respecto a los métodos de imputación en sí mismos, sino más bien con respecto a las variables auxiliares utilizadas y la estimación de los parámetros del modelo.

3.3.2 Calidad de la imputación de la regresión

Es importante medir la calidad de las imputaciones. Como se mencionó anteriormente, un problema fundamental es que generalmente se desconocen los valores verdaderos. En muchos casos, las medias antes y después de la imputación difieren entre sí. Esto no es necesariamente motivo de alarma porque la falta de respuesta del elemento puede haber sido selectiva. Si hay superposición con otras encuestas, se pueden realizar validaciones externas para tener una idea de la calidad de las imputaciones. Sin embargo, a menudo existen diferencias en las definiciones de las variables y en las poblaciones entre las encuestas. Esto significa que las posibilidades de tales validaciones son limitadas en la práctica.

Dado que, por lo general, la calidad de las imputaciones no se puede probar, los indicadores de calidad para la imputación de regresión que se describen a continuación solo se basan en el ajuste del modelo en los ítems que responden.

Para el análisis de regresión lineal basado en la estimación de mínimos cuadrados, se puede usar la conocida medida R^2 para medir el ajuste del modelo en los encuestados del ítem. Esta medida de ajuste se puede utilizar para comparar diferentes modelos de imputación entre sí. Un requisito previo es que se pueda comparar una ganancia en la medida R^2 con un aumento en el número de grados de libertad. Esta medida de ajuste también se puede aplicar para la imputación de donantes, porque la imputación de donantes puede verse como una regresión en variables dummy. Para algunos modelos no lineales la probabilidad o una cantidad derivada de la probabilidad, como AIC y BIC (ver, p. ej., Burnham and Anderson 2002) o el R^2 de Nagelkerke (ver, p. ej., Nagelkerke et al. 1991), pueden usarse como un

indicador de el ajuste. Observamos que es teóricamente posible que aunque el modelo **A** tiene un mejor ajuste que otro modelo **B** para los ítems que respondieron, el modelo **A** tiene un peor ajuste que el modelo **B** para los ítems que no respondieron, es decir, que los residuos de las predicciones del modelo y el los valores verdaderos son en promedio mayores para el modelo **A** que para el modelo **B** para los que no respondieron.

Otra opción para tener una idea de la calidad de un modelo de imputación es realizar un experimento de simulación. Para un experimento de este tipo, se utiliza un conjunto de datos totalmente observados, donde se conocen los valores reales de las variables objetivo para todos los elementos de la población. Dicho conjunto de datos se obtiene considerando un conjunto de datos realistas previamente editado como "la población objetivo" o creando un conjunto de datos sintéticos a partir de un modelo estadístico. En el estudio de simulación, algunos de los valores verdaderos se eliminan temporalmente y se imputan nuevos valores para los valores eliminados. Si los valores imputados $\hat{y}_i$ están cerca de los valores originales y_i, es probable que la calidad del método de imputación sea alta. Al definir una métrica de distancia adecuada, se puede seleccionar el modelo de imputación o los parámetros del modelo más adecuados. Un ejemplo de una métrica de distancia de este tipo, es la desviación absoluta media entre los valores imputados y los verdaderos,

$$\frac{1}{I}\sum |\hat{y}_i - \hat{y}|$$

con I el número de registros imputados. A nivel agregado, se puede utilizar como métrica de distancia la media sobre los experimentos de simulación de la desviación absoluta media entre los valores agregados con y sin imputación,

$$\frac{1}{T}\sum_{t=1}^{T}\left|\tilde{Y}_t - Y\right|,$$

donde $\tilde{Y}_t$ es la suma de los valores de la variable y después de la imputación en el experimento $t-$*ésimo*, Y es la suma de los valores verdaderos de la variable y, y T es el número de experimentos de simulación realizados. Chambers (2004) presenta un gran número de medidas de evaluación para diferentes aspectos de la calidad de la imputación. Por otro lado, en Nordholt (1998) se describen otros ejemplos de experimentos de simulación.

En la oficina de estadísticas de Canada, se ha desarrollado un programa SAS llamado Generalized Simulation System (GENESIS) para realizar estudios de simulación; [Por ejemplo, Haziza (2003)]. En GENESIS, el usuario proporciona un archivo de población y elige un diseño de muestreo, un mecanismo de datos faltantes, una técnica de imputación y el número requerido de iteraciones. A continuación, el programa ejecuta la simulación solicitada. Se calculan varias métricas para evaluar la calidad de las imputaciones. Haziza (2006) brinda una excelente introducción al diseño y uso de estudios de simulación.

3.3.3 Conexión entre imputación y ponderación

Suponga que una población objetivo consta de elementos numerados $1, ..., N$. Estamos interesados en estimar, digamos, la población total de la variable objetivo y:

$$Y = \sum_{i=1}^{N} y_i \tag{11}$$

Se extrae una muestra s de la población objetivo y, por conveniencia notacional, suponemos que esta muestra consta de los elementos numerados $1, ..., n$. Usamos "obs" y "mis" para referirnos, respectivamente, a las partes de la muestra que responden y que no responden. Nuevamente, por conveniencia, asumimos que los elementos $1, ..., r$ son quienes contestan y que los elementos $r + 1, ..., n$ son quienes no contestan.

En esta situación, el modelo de regresión lineal (4) se puede utilizar (al menos) de tres maneras diferentes para obtener una estimación de (11):

I. Ponderar los datos de los encuestados aplicando el llamado estimador de regresión. II. Imputar las personas que no respondieron en la muestra utilizando la imputación de regresión, luego ponderar los datos de la muestra aplicando el estimador de regresión. III. Imputar tanto los elementos que no respondieron como los que no fueron muestreados utilizando la imputación de regresión.

Nos referiremos a estas estrategias como el *enfoque de ponderación*, el *enfoque combinado* y el *enfoque de imputación masiva*. Dado que se utiliza el mismo modelo de regresión en estos enfoques, uno podría esperar intuitivamente que todos produzcan la misma estimación. En esta subsección, mostramos que esto es de hecho cierto, bajo ciertas condiciones. C.-E. Särndal and Lundstrom (2005) identifican las estimaciones basadas en la ponderación y la imputación desde un punto de vista ligeramente diferente.

Comenzamos introduciendo algo más de notación. El vector $y = (y_1, ..., y_N)^T$ contiene los valores de y en la población, mientras que el vector más pequeño $y_{obs} = (y_1, ..., y_r)^T$ contiene los valores de y en la parte que responde solamente de la muestra. Además, un vector de información auxiliar, denotado por $x_i = (x_{i1}, ..., x_{ip})^T$, está disponible para todo $i = 1, ..., N$. Sea $X = (x_{ij})$ la matriz $N \times p$ con x_i^T como la i-ésima fila, para $i = 1, ..., N$. Las submatrices correspondientes para los elementos muestreados, los que respondieron y los que no respondieron se denotan, respectivamente, por X_s, X_{obs} y X_{mis}. Tenga en cuenta que "*mis*" se refiere al hecho de que y_i no se observa para estos elementos de población, mientras que x_i^T se observa completamente para todos los elementos de la población.

Por conveniencia, reescribimos ligeramente el modelo de regresión lineal como:

$$y = X\beta + \epsilon$$

siendo $\beta = (\beta_1, ..., \beta_p)^T$ y $\epsilon = (\epsilon_1, ...\epsilon_N)^T$ los vectores de coeficientes de regresión y residuos, respectivamente. En esta notación, el término constante α toma como uno de los coeficientes β, correspondiente a una columna de unos en X. Usando mínimos cuadrados ordinarios (ols), se encuentra que los coeficientes de regresión son

$$\beta = (X^T X)^{-1} X^T y.$$

Inicialmente, solo la parte observada de la muestra se puede utilizar para estimar los coeficientes de regresión. Teniendo en cuenta el diseño muestral con probabilidades de inclusión π_i, se obtiene una estimación asintóticamente insesgada de β mediante mínimos cuadrados ponderados (wls):

$$\hat{\beta}_{obs} = (X_{obs}^T \Pi_{obs}^{-1} X_{obs})^{-1} X_{obs} \Pi_{obs}^{-1} y_{obs}, \tag{12}$$

donde $\Pi_{obs}^{-1} = diag(1/\pi_1, ..., 1/\pi_r)$ es una matriz diagonal de pesos de diseño para los elementos que responden (cf. J. G. Bethlehem and Keller 1987; Knottnerus 2003 (págs. 118–123)). En el caso especial en que una muestra se obtenga mediante un muestreo aleatorio simple (con $\pi_1 = ... = \pi_N = n/N$), la matriz de pesos de diseño puede quedar fuera de esta expresión.

En el enfoque de ponderación, los coeficientes de regresión estimados se utilizan para calcular el llamado estimador de regresión:

$$\hat{Y}_W = \sum_{i=1}^{r} \frac{y_i}{\pi_i} + \left(\sum_{i=1}^{N} x_i^T - \sum_{i=1}^{r} \frac{x_i^T}{\pi_i} \right) \hat{\beta}_{obs} \tag{13}$$

(ver Knottnerus 2003). La lógica detrás del estimador de regresión es que la estimación directa para el total de y, con base en los pesos de diseño $(1/\pi_i)$, se ajusta agregando un término de corrección, que predice el error en la estimación directa, utilizando el hecho de que el verdadero valor de x_i se conoce para toda la población. Tenga en cuenta que el término entre paréntesis es solo la diferencia entre el vector de los totales reales de la población y el vector de estimaciones directas para las variables x.

El término de "ponderación" alude al hecho de que $\hat{Y}_W$ se puede escribir como una suma ponderada de los valores observados:

$$\hat{Y}_W = \sum_{i=1}^{r} w_i y_i,$$

con,

$$w_i = \frac{1}{\pi_i} \left[1 + \left(\sum_{k=1}^{N} \mathbf{x}_k^T - \sum_{k=1}^{r} \frac{\mathbf{x}_k^T}{\pi_k} \right) \left(\mathbf{x}_{obs}^T \Pi_{obs}^{-1} \mathbf{x}_{obs} \right)^{-1} \mathbf{x}_i \right]$$

(ver, por ejemplo, C.-E. Särndal, Swensson, and Wretman 2003). Un tratamiento general del enfoque de ponderación queda fuera del alcance de este documento.

En el enfoque combinado, la imputación de regresión se utiliza para imputar los datos que faltan para los elementos de la muestra que no responden. Por ahora, asumimos que no se agregan residuos aleatorios a las imputaciones:

$$\hat{y}_i = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{obs}, \qquad i = r+1, ..., n.$$

Luego aplicamos el enfoque de ponderación, usando los datos observados e imputados para estimar los coeficientes de regresión:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{s}} = \left(\mathbf{X_s}^T \boldsymbol{\Pi}_s^{-1} \mathbf{X_s} \right)^{-1} \mathbf{X_s}^T \boldsymbol{\Pi}_{\mathbf{s}}^{-\mathbf{1}} \tilde{\mathbf{y}}_{\mathbf{s}}, \tag{14}$$

con $\Pi_s^{-1} = diag(1/\pi_1, ..., 1/\pi_n)$ matriz diagonal de los pesos del diseño para los elementos muestreados, $\tilde{y}_s = y_1, ..., y_r, \tilde{y}_{r+1}, ...\tilde{y}_n$. Esto produce el siguiente estimador de regresión para (11):

$$\hat{Y}_{IW} = \sum_{i=1}^{r} \frac{y_i}{\pi_i} + \sum_{i=r+1}^{n} \frac{\tilde{y}_i}{\pi_i} + \left(\sum_{i=1}^{N} \mathbf{x_i^T} - \sum_{i=1}^{n} \frac{\mathbf{x_i^T}}{\pi_i} \right) \hat{\boldsymbol{\beta}}_{\boldsymbol{s}}.$$

Usando el hecho que $\tilde{y}_i = \mathbf{x_i^T} \hat{\boldsymbol{\beta}}_{\boldsymbol{obs}}$, esta expresión puede ser escrita de la siguiente forma:

$$\tilde{Y}_{IW} = \sum_{i=1}^{r} \frac{y_i}{\pi_i} + \left(\sum_{i=1}^{N} \mathbf{x_i^T} - \sum_{i=1}^{r} \frac{\mathbf{x_i^T}}{\pi_i} \right) \hat{\boldsymbol{\beta}}_{\boldsymbol{s}} + \sum_{i=r+1}^{n} \frac{\mathbf{x_i^T}}{\pi_i} \left(\hat{\boldsymbol{\beta}}_{\boldsymbol{obs}} - \hat{\boldsymbol{\beta}}_{\boldsymbol{s}} \right).$$

Comparando (13) y (15) observamos que la estimación ponderada y la estimación combinada son idénticas si se mantiene que $\hat{\boldsymbol{\beta}}_{\boldsymbol{s}} = \hat{\boldsymbol{\beta}}_{\boldsymbol{obs}}$.

Finalmente, en el enfoque de imputación masiva, el modelo de regresión se utiliza para imputar todos los valores no observados:

$$\hat{y}_i = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{obs}, \qquad i = r+1, ..., N.$$

Luego aplicamos el enfoque de ponderación, usando los datos observados e imputados para estimar los coeficientes de regresión:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{s}} = (\boldsymbol{X}_{\boldsymbol{s}}{}^T \boldsymbol{\Pi}_{\boldsymbol{s}}{}^{-1} \boldsymbol{X}_{\boldsymbol{s}})^{-1} \boldsymbol{X}_{\boldsymbol{s}} \boldsymbol{\Pi}_{\boldsymbol{s}}{}^{-1} \tilde{\boldsymbol{y}}_{\boldsymbol{s}}$$

con $\pi_s^{-1} = diag(1/\pi_1, ..., 1/\pi_n)$ una matriz diagonal de pesos del diseño para los elementos muestreados, y $\tilde{y}_s = (y_1, ..., y_r, \tilde{y}_1, ..., \tilde{y}_n)^T$. Esto produce el siguiente estimador de regresión para (11) :

$$\hat{Y}_{IW} = \sum_{i=1}^{r} \frac{y_i}{\pi_i} + \sum_{i=r+1}^{n} \frac{\tilde{y}_i}{\pi_i} + \left(\sum_{i=1}^{N} \mathbf{x_i^T} - \sum_{i=1}^{n} \frac{\mathbf{x_i^T}}{\pi_i} \right) \hat{\boldsymbol{\beta}}_{\boldsymbol{s}}.$$

Usando el hecho de que $\tilde{y}_i = \mathbf{x_i}^T \hat{\boldsymbol{\beta}}_{obs}$, esta expresión se puede escribir como:

$$\hat{Y}_{IW} = \sum_{i=1}^{r} \frac{y_i}{\pi_i} + \left(\sum_{i=1}^{N} \mathbf{x_i^T} - \sum_{i=1}^{r} \frac{\mathbf{x_i^T}}{\pi_i} \right) \hat{\boldsymbol{\beta}}_{\boldsymbol{s}} + \sum_{i=r+1}^{n} \frac{\mathbf{x_i^T}}{\pi_i} \left(\hat{\boldsymbol{\beta}}_{\boldsymbol{obs}} - \hat{\boldsymbol{\beta}}_{\boldsymbol{s}} \right). \tag{15}$$

Comparando (13) y (15), observamos que la estimación ponderada y la estimación combinada son idénticas si se cumple que $\hat{\boldsymbol{\beta}}_{\boldsymbol{s}} = \hat{\boldsymbol{\beta}}_{\boldsymbol{obs}}$.

Finalmente, en el enfoque de imputación masiva, el modelo de regresión se utiliza para imputar todos los valores no observados:

$$\tilde{y}_i = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{obs}, \qquad i = r+1, ..., N.$$

No es necesario ponderar y obtenemos la siguiente estimación para (11):

$$\hat{Y}_I = \sum_{i=1}^{r} y_i + \sum_{i=r+1}^{N} \mathbf{x_i^T} \hat{\boldsymbol{\beta}}_{obs}. \tag{16}$$

En general, esta estimación no necesita ser idéntica a $\hat{Y}_W$ o $\hat{Y}_{IW}$.

El teorema 3.1 a continuación establece condiciones generales tales que las tres estimaciones $\hat{Y}_W$,$\hat{Y}_{IW}$ y $\hat{Y}_i$ son idénticas. En la preparación de este resultado, observamos que si el vector N de unos está contenido en el espacio de columna de $\mathbf{X}$, entonces la estimación ponderada (13) se puede escribir de manera concisa como:

$$\hat{Y}_W = \sum_{i=1}^{N} \mathbf{x_i^T} \hat{\boldsymbol{\beta}}_{obs}, \tag{17}$$

Esto sigue del hecho bien conocido de que el vector de residuos observados $\hat{\boldsymbol{\varepsilon}}_{obs} = \boldsymbol{y}_{obs} - \mathbf{X}_{obs}\hat{\boldsymbol{\beta}}_{obs}$ es ortogonal a cada una de las columnas en $\mathbf{X}_{obs}$:

$$\mathbf{x}_{obs}^T \boldsymbol{\Pi}_{obs}^{-1} \hat{\boldsymbol{\varepsilon}}_{obs} = \mathbf{0}.$$

Tenga en cuenta que debido a que usamos wls (mínimos cuadrados ponderados) para obtener $\hat{\boldsymbol{\beta}}_{obs}$, la ortogonalidad no está definida con respecto a la métrica euclidiana habitual, sino con respecto a una métrica que involucra la multiplicación por $\boldsymbol{\Pi}_{obs}^{-1}$. Como hemos supuesto que el espacio columna de $\mathbf{X}$ contiene el vector de unos, se sigue en particular que:

$$\sum_{i=1}^{r} \frac{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{obs}}{\pi_i} = \sum_{i=1}^{r} \frac{\hat{\boldsymbol{\varepsilon}}_{obs,i}}{\pi_i} = (1, ..., 1)\boldsymbol{\Pi}_{obs}^{-1}\hat{\boldsymbol{\varepsilon}}_{obs} = \mathbf{0} \tag{18}$$

ver también (Knottnerus (2003), p. 121). De manera similar, si el vector constante de unos está contenido en el espacio columna de X, se cumple que

$$\hat{Y}_{IW} = \sum_{i=1}^{N} \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_s,$$

Al comparar (17) y (16), vemos inmediatamente que

$$\hat{Y}_I - \hat{Y}_W = \sum_{i=1}^{r} y_i - \sum_{i=1}^{r} \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{obs} = \sum_{i=1}^{r} \hat{\boldsymbol{\varepsilon}}_{obs,i}. \tag{19}$$

Para el muestreo aleatorio simple, esta expresión es igual a cero por (18), y concluimos que la estimación de ponderación y la estimación de imputación de masa son idénticas bajo el supuesto de que el espacio de columna de $\boldsymbol{X}$ contiene el vector de unos. Sin embargo, esto no se aplica a los diseños de muestreo generales, porque la suma no ponderada de los residuos observados no tiene por qué ser igual a cero bajo wls.

Presentamos ahora el resultado general.

Theorem 3.1. *Para todos los diseños de muestreo y todos los modelos de regresión lineal, siempre se cumple que $\hat{Y}_W = \hat{Y}_{IW}$. Además, si tanto el vector constante de unos como el vector de probabilidades de inclusión $\boldsymbol{\pi} = (\pi_1, ..., \pi_N)^T$ están contenidos en el espacio columna de la matriz auxiliar $\mathbf{X}$, entonces también se cumple que $\hat{Y}_W = \hat{Y}_{IW} = \hat{Y}_I$.*

Demostración. Para probar la primera afirmación, ya observamos que basta demostrar que $\hat{\boldsymbol{\beta}}_s = \hat{\boldsymbol{\beta}}_{obs}$. Intuitivamente, está claro que las estimaciones de los parámetros de un modelo de regresión no deberían cambiar si agregamos observaciones a las ecuaciones de regresión que se ajustan exactamente a un modelo previamente ajustado. Una prueba formal procede de la siguiente manera. Por definición, la estimación wls $\hat{\boldsymbol{\beta}}_s$ minimiza la función cuadrática $\left(\tilde{\boldsymbol{y}}_s - \mathbf{x}_s\hat{\boldsymbol{\beta}}_s\right)^T \boldsymbol{\Pi}_s^T \left(\tilde{\boldsymbol{y}}_s - \mathbf{x}_s\hat{\boldsymbol{\beta}}_s\right)$. Esta función se puede expresar como:

$$\left(\boldsymbol{y}_{obs} - \mathbf{x}_{obs}\hat{\boldsymbol{\beta}}_s\right) \boldsymbol{\Pi}_{obs}^{-1} \left(\boldsymbol{y}_{obs} - \mathbf{x}_{obs}\hat{\boldsymbol{\beta}}_s\right) + \left(\hat{\boldsymbol{\beta}}_{obs} - \hat{\boldsymbol{\beta}}_s\right)^T \boldsymbol{X}_{mis}^T \boldsymbol{\Pi}_{mis}^{-1} \boldsymbol{X}_{mis} \left(\hat{\boldsymbol{\beta}}_{obs} - \hat{\boldsymbol{\beta}}_s\right),$$

Con $\boldsymbol{\Pi}_{mis}^{-1} = diag(1/\pi_{r+1}, ..., 1/\pi_n)$. El primer término es una función cuadrática que se minimiza eligiendo $\hat{\boldsymbol{\beta}}_s = \hat{\boldsymbol{\beta}}_{obs}$. El segundo término es una función cuadrática que es igual a cero para esta elección de $\hat{\boldsymbol{\beta}}_s$. Dado que ambos términos no son negativos, concluimos que $\hat{\boldsymbol{\beta}}_s = \hat{\boldsymbol{\beta}}_{obs}$ es la estimación wls (por minimos cuadrados ponderados) con el enfoque combinado. Por lo tanto, la primera afirmación está probada.

Para probar la segunda afirmación, usamos la expresión (19). Bajo el supuesto de que el espacio columna de $\mathbf{X}$ contiene $\boldsymbol{\pi}$, la propiedad de ortogonalidad implica que

$$\sum_{i=1}^{r} \hat{\boldsymbol{\varepsilon}}_{obs,i} = (\pi_1, ..., \pi_r) \boldsymbol{\Pi}_{obs}^{-1} \hat{\boldsymbol{\varepsilon}}_{obs} = 0$$

y de ahí se concluye que $\hat{Y}_I = \hat{Y}_{IW}$.

Una manera obvia de satisfacer la condición de que el vector constante de unos está contenido en el espacio columna de $\mathbf{X}$ es incluir el término constante en las ecuaciones de regresión. De manera similar, la condición de que π esté contenido en el espacio de columnas de $\mathbf{X}$ se puede satisfacer agregando las probabilidades de inclusión al modelo de regresión. **Para el muestreo estratificado, basta con incluir todas las variables auxiliares que se utilizaron para construir los estratos en el modelo de regresión**. Tenga en cuenta que para el muestreo aleatorio simple, las dos condiciones son equivalentes.

3.3.3.0.1 Ejemplo 3 Considere un diseño de muestreo aleatorio simple (sin respuesta) y un modelo de regresión que involucre solo el término constante. De (12) se encuentra que el coeficiente de regresión estimado es

$$\hat{\beta}_{obs} = \frac{1}{r} \sum_{i=1}^{r} y_i,$$

es decir, la media observada de y. Usando este modelo en el enfoque de ponderación, la estimación de regresión es

$$\hat{Y}_W = \frac{N}{r} \sum_{i=1}^{r} y_i.$$

Por lo tanto, los pesos de diseño N/n se ajustan por falta de respuesta por un factor constante n/r.

En el enfoque combinado, a los que no respondieron se les imputa primero la media observada de y. En este caso, la parte de ponderación del enfoque combinado equivale a utilizar únicamente las ponderaciones de diseño N/n, porque el modelo de regresión no utiliza información externa a la muestra. Encontramos

$$\hat{Y}_{IW} = \frac{N}{r}\sum_{i=1}^{r} y_i + \frac{N}{r}\sum_{i=r+1}^{r} \tilde{y}_i = \frac{N}{n}\left(1 + \frac{n-r}{r}\right)\sum_{i=1}^{r} y_i = \frac{N}{r}\sum_{i=1}^{r} y_i$$

Finalmente, para el enfoque de imputación masiva, cada valor no observado de y se imputa con la media observada de y. Encontramos

$$\hat{Y}_I = \sum_{i=1}^{r} y_i + \sum_{i=r+1}^{N} \tilde{y}_i = \left(1 + \frac{N-r}{r}\right)\sum_{i=1}^{r} y_i = \frac{N}{r}\sum_{i=1}^{r} y_i.$$

Así, bajo muestreo aleatorio simple, el modelo de regresión que incluye solo el término constante produce la misma estimación para los tres enfoques.

3.3.3.0.2 Ejemplo 4 En este ejemplo, consideramos un diseño de muestreo general con una estimación basada en *post-estratificación*. La post-estratificación se usa ampliamente en la inferencia basada en encuestas por muestreo (ver, por ejemplo, J. Bethlehem 2009; o C. Särndal, n.d.). Esta técnica corresponde a un modelo de regresión que involucra, digamos, L variables dummy: $\mathbf{x}_i = (x_{i1}, ..., x_{iL})^T$ con $x_{il} = 1$ si el i-ésimo elemento de población cae en el l-ésimo post-estrato, y $x_{il} = 0$ en caso contrario. Asumiendo que los postestratos son mutuamente excluyentes, tenemos $\sum_l^L x_{il} = 1$ para todo i. Esto muestra que el vector constante de unos está contenido en el espacio columna de $\mathbf{X}$, aunque no está explícitamente incluido en el modelo.

De (12) el vector de coeficientes de regresión estimados es

$$\hat{\boldsymbol{\beta}}_{obs} = \left(\hat{\beta}_{obs,1}, ..., \hat{\beta}_{obs,L}\right)^T$$

con

$$\hat{\beta}_{obs,l} = \frac{\hat{Y}_{obs,l}}{\hat{N}_{obs,l}} = \frac{\sum_{i=1}^{r} x_{il} y_i/\pi}{\sum_{i=1}^{r} x_{il}/\pi}.$$

Usando el enfoque de ponderación con estratificación posterior, obtenemos la siguiente estimación de (11):

$$\hat{Y}_W = \sum_{l=1}^{L} N_l \frac{\hat{Y}_{obs,l}}{\hat{N}_{obs,l}}.$$

Bajo el enfoque combinado, los elementos que no responden de la muestra, $i = r+1, ..., n$, se imputan con $\tilde{y}_i = \sum_{l=1}^{L} x_{il}\hat{Y}_{obs,l}/\hat{N}_{obs,l}$.

A continuación, se calcula el estimador de regresión con posestratificación en función de la muestra imputada. El teorema 3.1 garantiza que el vector de coeficientes de regresión estimados $\hat{\boldsymbol{\beta}}_s = (\hat{\beta}_{s,1}, ..., \hat{\beta}_{s,L})^T$ es idéntico a $\hat{\beta}_{obs}$.

De hecho, esto es fácil de verificar directamente:

$$\begin{aligned}
\hat{\beta}_{s,l} &= \frac{\sum_{i=1}^{r} x_{il} y_i/\pi_i + \sum_{i=r+1}^{n} x_{il}\tilde{y}_i/\pi_i}{\sum_{i=l}^{n} x_{il}/\pi_i} \\
&= \frac{\hat{Y}_{obs,l} + \left(\hat{Y}_{obs,l}/\hat{N}_{obs,l}\right)\sum_{i=r+1}^{n} x_{il}/\pi_i}{\hat{N}_{obs,l} + \sum_{i=r+1}^{n} x_{il}/\pi_i} \\
&= \frac{\hat{Y}_{obs,l} + \left(\hat{Y}_{obs,l}/\hat{N}_{obs,l}\right)\sum_{i=r+1}^{n} x_{il}/\pi_i}{\hat{N}_{obs,l} + \sum_{i=r+1}^{n} x_{il}/\pi_i} \\
&= \frac{\hat{Y}_{obs,l}}{\hat{N}_{obs,l}} = \hat{\beta}_{obs,l}.
\end{aligned}$$

Resulta que

$$\hat{Y}_{IW} = \sum_{l=1}^{L} N_l = \frac{\hat{Y}_{obs,l}}{\hat{N}_{obs,l}},$$

por lo tanto, el enfoque de ponderación y el enfoque combinado producen la misma estimación posestratificada. Usando el enfoque de imputación masiva, cada elemento de población no observado, $i = r+1, ..., N$, se imputa con $\hat{y}_i = \sum_{l=1}^{L} x_{il}\hat{Y}_{obs,l}/\hat{N}_{obs,l}$.

Esto produce la siguiente estimación de (11):

$$\hat{Y}_i = \sum_{i=l}^{r} y_i + \sum_{l=1}^{L} (N_l - r_l)\,\frac{\hat{Y}_{obs,l}}{\hat{N}_{obs,l}},$$

donde r_l denota el número de encuestados en el postestrato $l - ésimo$, con $\sum_{l=1}^{L} r_l = r$.

En general, esta estimación no tiene por qué ser igual a las dos estimaciones anteriores. Supongamos, sin embargo, que estamos tratando con una muestra estratificada, donde los post-estratos son más pequeños que los estratos de muestra, y cada post-estrato cae exactamente en un estrato de muestra. En este caso común, todos los elementos en un estrato posterior tienen la misma probabilidad de inclusión, lo que significa que el vector de probabilidades de inclusión está contenido en el espacio de columna de $\mathbf{X}$, y el teorema 3.1 establece que la estimación de ponderación y la estimación de imputación de masa son idéntico. También podemos verificar esto directamente. Señalando que:

$$\frac{\hat{Y}_{obs,l}}{\hat{N}_{obs,l}} = \frac{1}{r_e}\sum_{i=1}^{r} x_{il} y_i.$$

en este caso, encontramos que

$$\hat{Y}_W = \hat{Y}_{IW} = \sum_{l=1}^{L} \frac{N_l}{r_l} \sum_{i=1}^{r} x_{il} y_i$$

y

$$\hat{Y}_I = \sum_{l=1}^{L} \left(\sum_{i=1}^{r} x_{il} y_i + \frac{N_l - rl}{rl} \sum_{i=1}^{r} x_{il} y_i \right) = \sum_{l=1}^{L} \frac{N_l}{rl} \sum_{i=1}^{r} x_{il} y_i.$$

Entonces, en este caso especial, se mantiene que $\hat{Y}_W = \hat{Y}_{IW} = \hat{Y}_I$ para las estimaciones posestratificadas.

Hasta ahora, solo hemos considerado la imputación sin residuos agregados. Si se utiliza la imputación estocástica, con un residuo aleatorio agregado a cada valor predicho, los tres enfoques ya no producen estimaciones idénticas. Suponiendo que los residuos se extraen de una distribución con media cero, aún se mantiene que las estimaciones del enfoque combinado y el enfoque de imputación masiva son iguales a $\hat{Y}_W$ en expectativa, bajo las condiciones del teorema 3.1.

3.4 Imputación de razón

3.4.1 El modelo de imputación de razón

En la práctica cotidiana de los institutos de estadística, un caso especial importante de imputación de regresión es la imputación de razón (o proporción). Cuando se aplica la imputación de razón para la variable objetivo y, se utiliza una sola variable auxiliar x para la cual la razón con la variable objetivo y es aproximadamente constante. Si R denota la relación entre y y x, entonces se imputa un valor faltante y_i mediante

$$\tilde{y}_i = R x_i \tag{20}$$

Generalmente, R no se conoce y se estima utilizando los registros para los que se conocen tanto x como y:

$$\hat{R} = \frac{\sum_{k \in obs} y_k}{\sum_{k \in obs} x_k} \tag{21}$$

donde, como antes, "obs" denota el conjunto de unidades observadas. La razón estimada $\hat{R}$, por lo tanto, es igual a la razón de las medias de la variable objetivo y y x para los encuestados del ítem para la variable y. Un ejemplo es cuando se estima una facturación desconocida (y) en función del número de empleados (x). Para R, entonces se usaría la rotación media por empleado. Para estimar la relación R, se puede decidir ponderar a los encuestados de los ítems con sus ponderaciones crecientes.

Sustituyendo (21) en (22) da

$$\tilde{y}_i = \hat{R} x_i = \frac{\sum_{k \in obs} y_k}{\sum_{k \in obs} x_k} x_i \tag{22}$$

La situación que ocurre con más frecuencia en la práctica es que x mide la misma característica que y, pero en un momento anterior. En este caso, escribimos las variables y y x como $y^{(t)}$, respectivamente $y^{(t-1)}$. La fórmula (20) se convierte entonces en

$$\tilde{y}_i^{(t)} = R y_i^{(t-1)} \tag{23}$$

donde R es el aumento (o disminución) relativo de la variable y desde el momento $t-1$ hasta t, y

$$\hat{R} = \frac{\sum_{k \in obs} y_k^{(t)}}{\sum_{k \in obs} y_k^{(t-1)}}$$

Se puede considerar la fórmula (20) como una ecuación de regresión sin intersección. Si un modelo con un intercepto conduce a un mejor ajuste, o si se desea agregar más variables auxiliares al modelo, el método de imputación de regresión más general puede ser más adecuado. En los INE, como La Oficina de Estadísticas de Holanda, generalmente no se agrega ningún residuo a (20), porque para muchas estadísticas en las que se aplica la imputación de razón, las medias y los totales son los productos más importantes.

Sin embaro, incluso en los INE hay algunas excepciones. En el pasado, La Oficina de Estadísticas de Canada publicaba para algunas ramas de la industria una tabla con el número de empresas con una facturación más alta en comparación con el año anterior frente al número de empresas con una facturación más baja. Si se aplicara la imputación (23) y se estimara R en, digamos, 1.01, entonces para todos los items que no respondieron se supondría que su rotación habia crecido desde el momento $t-1$ hasta t lo cual es muy poco probable. Por lo tanto, para tales tablas es necesario un residuo a (23)

o imputación de razón, no se requiere software complejo. Las fórmulas (22) y (23) son fáciles de calcular una vez que se ha estimado R.

3.4.2 Ejemplo para la imputación de razón

3.4.2.1 Ejemplo 5 (Estadísticas Económica Estructural de Holanda) En La Oficina de Estadísticas de Holanda, para las estadísticas económicas estructurales se utiliza un procedimiento de imputación automatizado para las pequeñas y medianas empresas. Este procedimiento de imputación se basa principalmente en la imputación de razón. La disponibilidad de información auxiliar se examina en un orden fijo. Este orden, desde la información auxiliar más preferida hasta la información auxiliar menos preferida, viene dado por:

1. observación de la misma empresa en el año anterior $t-1$ (para todas las variables);
2. observación de la misma empresa en las Estadísticas Coyunturales del año t (solo si la Facturación es la variable objetivo y);
3. observaciones de unidades en el mismo estrato (definido por **"tamaño de la empresa"**× **"rama de la indistria"**) en el año t.

En otras palabras, si la falta de respuesta del ítem ocurre para una determinada empresa, primero se examina si la empresa tuvo una puntuación válida en la variable correspondiente en el año anterior. Si es así, se aplica la fórmula (23) con $y^{(t)}$ su valor de la variable

correspondiente en el año t, $y^{(t-1)}$ el valor en el año anterior y R una corrección de tendencia estimada. Para *Facturación*, la corrección de tendencia es igual al crecimiento (o reducción) de la facturación total.

Sin embargo, si $y_i^{(t-1)}$ es desconocido, por ejemplo, porque la empresa no se incluyó en la muestra del año anterior, se elige la segunda o la tercera opción, dependiendo de la variable objetivo. Estas opciones no son imputaciones de razón. En la segunda opción, el volumen de negocios total de la empresa bajo consideración en el año t como se observa en las *Estadísticas Coyunturales* se utiliza para imputar la observación faltante para el *Volumen de Negocios* en las *Estadísticas Estructurales de Empresas*. La opción 3 es un ejemplo de imputación por la media del grupo, con una combinación de "tamaño de la empresa" y "rama de la industria" como tipo de imputación.

3.5 Imputación media (Grupo)

3.5.1 El modelo de imputación media (Grupal)

Cuando se aplica la *imputación media*, un valor perdido se reemplaza por el valor medio de la variable correspondiente de las unidades que tienen un valor válido. Es decir, el valor imputado $\tilde{y}_i$ para un valor desconocido y_i viene dado por la media observada

$$\tilde{y}_i = \overline{y}_{obs} \equiv \frac{\sum_{k \in obs} y_k}{r}, \tag{24}$$

donde y_k es el valor observado del $k-$ésimo elemento que responde, "obs" es el conjunto de unidades observadas y r es el número de elementos que responden para la variable objetivo y.

Si se desea, se puede dar un peso diferente a los valores observados de la variable objetivo y, por ejemplo, porque estos datos se recopilan con un diseño de muestreo con diferentes pesos de inclusión. Sea w_i el peso (aumentado) del ítem-respondedor i. La imputación media resultante viene dada por

$$\tilde{y}_i = \overline{y}_{obs}^{(w)} \equiv \frac{\sum_{k \in obs} y_k}{\sum_{k \in obs} w_k}. \tag{25}$$

En general, este es un mejor estimador, es decir, menos sesgado, para la media de la población.

Cuando se aplica la imputación media, no se utiliza información auxiliar. El método solo se recomienda cuando se carece de información auxiliar, o cuando la información auxiliar disponible no está o apenas está relacionada con la variable objetivo y. Si la fracción de valores perdidos en una determinada variable es muy pequeña y las imputaciones difícilmente afectarán los parámetros (por ejemplo, el total de la población) a estimar, la imputación media puede ser un buen método. En muchos casos, sin embargo, el enfoque es demasiado simplista y conduce a datos de calidad inferior.

Como ya mencionamos, la imputación media conduce a una distribución pico.

Por lo tanto, el método solo se puede aplicar con éxito si solo se desea estimar las medias y los totales de la población. Sin embargo, la imputación media no es adecuada para estimar una distribución, por ejemplo, ingresos o para estimar una desviación estándar.

Cuando no se dispone de variables auxiliares adecuadas para una variable objetivo categórica y, se puede imputar el valor más común (la moda) o extraer de las categorías probabilidades proporcionales a las frecuencias observadas de estas categorías. Por lo general, no se recomienda imputar la moda porque esto suele conducir a estimaciones sesgadas de las frecuencias de la población. Extraer de las categorías con probabilidades proporcionales a las frecuencias observadas de estas categorías es lo mismo que usar la imputación aleatoria de donantes.

Cuando se aplica la *imputación de medias grupales*, un valor faltante se reemplaza por el valor medio de la variable correspondiente de las unidades que tienen un valor válido y además pertenecen al mismo grupo que el ítem que no responde. En el caso de la imputación de la media del grupo, primero se determinan las clases de imputación apropiadas, es decir, los grupos. El valor imputado para un valor faltante en el grupo h viene dado por

$$\tilde{y}_{hi} = \overline{y}_{h;obs} \equiv \frac{\sum_{k \in h \cap obs} y_k}{r_h}, \tag{26}$$

donde y_k es el valor observado para el $k - ésimo$ encuestado, y r_h es el número de ítems encuestados para la variable y en h.

Cuando se aplica la imputación de la media del grupo, la distribución tiene menos picos que para la imputación de la media, porque la variación entre grupos se tiene en cuenta al imputar los datos faltantes; sólo se desprecia la variación dentro de los grupos.

En otras palabras, mientras que la imputación de la media conduce a un gran pico, la imputación de la media de grupo conduce a varios picos más pequeños en la distribución de los datos imputados. Si la relación entre la varianza y la varianza interna es grande, el método también se puede usar para estimar medidas de dispersión con una precisión razonable, asumiendo la validez del modelo de imputación.

La imputación de la media de grupo conduce a los mismos totales generales y medias que el llamado estimador posterior a la estratificación, cuando los estratos del estimador posterior a la estratificación se utilizan como clases de imputación [ver, por ejemplo; J. Bethlehem (2009) y el Ejemplo 4 anterior].

En el caso de la imputación de medias grupales, se utiliza información auxiliar, a saber, una o más variables categóricas para construir los grupos. Cuanto más homogéneos sean los grupos —también denominados subpoblaciones, clases de imputación o estratos de imputación— con respecto a la variable a imputar, mejor será la calidad de las imputaciones. En la práctica, solo se puede probar la homogeneidad de los grupos para los que respondieron el ítem y, a menudo, se supone que los grupos no solo son homogéneos para los que respondieron el ítem, sino también para los que no respondieron. Distinguir diferentes grupos generalmente tiene más efecto para la imputación de medias (de grupo) que para la imputación de razón, porque las razones de los grupos son generalmente más homogéneas que las medias de los grupos.

Como de costumbre, la imputación media (de grupo) también se puede aplicar con un término de error estocástico. En cuanto a la imputación de razón, una de las principales ventajas de la imputación media (de grupo) es que no se necesita ningún software especial debido a la simplicidad del método. La imputación de la media (de grupo) se puede aplicar con casi cualquier paquete de software estadístico y con muchos otros paquetes de software.

3.5.1.1 Ejemplos de imputación media (grupal)

3.5.1.1.1 Ejemplo 5 (Estadísticas económicas estructurales holandesas) Como se explicó anteriormente, en las Estadísticas económicas estructurales holandesas, si falta información auxiliar sobre una empresa con una respuesta incompleta, se utiliza la imputación de la media del grupo. En ese caso, para un valor faltante, digamos, *Facturación*, se imputa la facturación media en la clase de imputación correspondiente. Las subpoblaciones se definen mediante combinaciones de *"tamaño de la empresa"* y *"rama de la industria"*. La fracción de muestreo es demasiado pequeña para distinguir todas las combinaciones de "tamaño de la empresa" y "rama de la industria". El procedimiento de imputación difiere algo para las empresas grandes y las menos grandes. La imputación de la media del grupo a menudo se utilizará para, por ejemplo, nuevas empresas para las que obviamente no se dispone de información auxiliar de un período anterior.

3.5.1.1.2 Ejemplo 6 (Encuesta de automatización holandesa) En la Encuesta de automatización holandesa se aplicó la imputación media (ver; De Waal (2000)). En este caso particular, los grupos de imputación no se basaron en los datos en sí, sino en la salida. Es decir, las celdas de publicación de las tablas que se liberaban definían los grupos de imputación. En algunos casos, se imputó la media general en lugar de la media del grupo, por ejemplo, cuando solo había unas pocas observaciones por celda de publicación.

3.6 Imputación de donantes Hot Deck

3.6.1 Introducción

Cuando se aplica la *imputación de donante Hot Deck*, para cada elemento que no responde i, se busca el registro de donante d en el conjunto de datos que tiene características que son lo más similares posible al elemento que no responde i, en la medida en que estas características están (se considera que están) correlacionadas con la variable objetivo y. Del donante seleccionado, la puntuación, y_d, se utiliza para imputar el valor faltante para el elemento que no responde i:

$$\tilde{y}_i = y_d \tag{27}$$

El elemento que no responde se denomina "receptor". La imputación del donante se puede aplicar tanto para una variable objetivo numérica como categórica y. Si faltan varios valores en un registro, en principio, se utiliza el mismo donante para imputar todos estos valores faltantes. Una excepción puede ser cuando los datos tienen que satisfacer restricciones.

Hay diferentes maneras de encontrar un donante. Estas formas se pueden subdividir en

1. métodos que utilizan clases de imputación;
2. métodos que buscan un donante minimizando una función de distancia (Hot Deck del vecino más cercano).

Ejemplos del primer tipo de métodos son *Hot Deck aleatorio* e *imputación Hot Deck secuencial*. Cuando se aplica la imputación aleatoria Hot Deck, las clases de imputación se forman en función de las variables auxiliares. De los donantes potenciales en la misma clase de imputación

que el receptor, es decir, con los mismos valores en las variables auxiliares que el receptor, se selecciona aleatoriamente un donante. Cuando se aplica la imputación Hot Deck secuencial, no se construyen clases de imputación explícitamente, sino que para cada ítem que no responde se imputa la puntuación de la variable objetivo en el primer registro posterior con las mismas puntuaciones en las variables auxiliares para el valor faltante.

La imputación aleatoria y secuencial de donantes Hot Deck se puede aplicar cuando las variables auxiliares son categóricas. Las variables numéricas se pueden utilizar como variables auxiliares categorizándolas temporalmente, es decir, clasificándolas en varias categorías. Para conjuntos de datos muy grandes en los que se desea aplicar la imputación Hot Deck, a veces se aplica Hot Deck secuencial simplemente para reducir el tiempo de cálculo y el uso de la memoria de la computadora.

Cuando se aplica la imputación del donante del vecino más cercano, no se forman clases de imputación y se permiten algunas diferencias entre las puntuaciones en las variables x del donante y el receptor. La imputación del vecino más cercano suele aplicarse cuando las variables auxiliares x son principalmente numéricas, y se perdería información si estas variables se categorizaran temporalmente para llevar a cabo el proceso de imputación. La imputación del vecino más cercano también puede tener en cuenta variables auxiliares cualitativas, siempre que se utilice una función de distancia adecuada. Dado que una función de distancia que mide la distancia entre un donante potencial y el receptor se minimiza cuando se aplica la imputación del donante del vecino más cercano, es esencial que la importancia de cada variable auxiliar x se cuantifique en forma de un peso apropiado.

Un caso especial de imputación del vecino más cercano es el coincidencia predictiva de medias, donde el donante del vecino más cercano se determina mediante un valor predictivo para la variable objetivo y, utilizando un modelo de regresión adecuado.

Además de la imputación *Hot Deck*, también existe la imputación *Cold Deck*. Aquí el valor imputado se toma de otro conjunto de datos, por ejemplo, el valor de la misma unidad en la misma variable en un momento anterior. Si el valor imputado del otro conjunto de datos es el valor correcto, podemos considerar esto como una imputación deductiva. Si el valor imputado es de la misma unidad en un momento anterior, simplemente usar este valor rara vez es una buena idea. En la mayoría de los casos, el valor imputado mejorará si se agrega un factor de tendencia al modelo. Esto llevaría a la imputación de razón.

La imputación de donantes también se utiliza cuando faltan varios valores por registro en variables (fuertemente) correlacionadas. Al elegir solo un donante para todos estos valores faltantes, se evita la inconsistencia entre los valores imputados. En tal caso, uno tiene que crear clases de imputación que sean homogéneas para varias variables objetivo simultáneamente. La imputación multivariante de donantes puede verse como una solución específica para el problema de la imputación multivariante.

3.7 Imputación Hot Deck aleatoria y secuencial

Cuando se aplica la imputación Hot Deck, se busca una unidad en el mismo conjunto de datos que tenga las mismas características que el destinatario; por ejemplo, una persona del mismo género, en la misma clase de edad, que vive en el mismo sector y trabaja en la misma rama de la industria. La idea es que si las características de dos individuos son las mismas,

los valores de la variable objetivo a imputar serán en muchos casos similares. Cuando se aplica la imputación Hot Deck aleatoria o secuencial, el donante y el receptor deben tener exactamente los mismos valores en las características de fondo, es decir, deben estar en la misma clase de imputación. Si en el ejemplo anterior no se puede encontrar ningún donante con las mismas cuatro características que el elemento que no responde, la clase de imputación es aparentemente demasiado pequeña. Para imputar un valor para este ítem-encuestado, se debe descartar al menos una de las cuatro características, o se deben combinar las clases de imputación. Si hay varios donantes potenciales en la clase de imputación correspondiente, se puede seleccionar un donante al azar. En lugar de seleccionar un donante al azar, también se puede agregar características adicionales de fondo, con la esperanza de obtener un único donante potencial solamente.

Es importante evitar que una unidad sea donante para muchos receptores. Es decir, el uso múltiple de una unidad como donante aumenta los errores estándar de las medias y los totales de la variable objetivo y, ya que los valores atípicos pueden "magnificarse". Se puede evitar esto, por ejemplo, en una cierta clase de imputación, permitiendo el uso múltiple de una unidad como donante solo cuando todas las unidades en esa clase de imputación hayan sido utilizadas como donante.

Previamente, ya hemos descrito que cuando se aplica la imputación Hot Deck secuencial, para cada ítem que no responde se imputa la puntuación en la variable objetivo y del primer ítem que responde en el conjunto de datos con las mismas características. Si en el conjunto de datos aparece una cantidad de elementos que no respondieron de la misma clase de imputación en rápida sucesión, es posible que obtengan su valor imputado del mismo donante. Para evitar esto, se puede ajustar el método secuencial de donante Hot Deck al no seleccionar el siguiente registro de donante potencial con las mismas características de antecedentes que el receptor, sino los primeros K registros de donantes potenciales con las mismas características de antecedentes que el receptor y luego seleccionar uno al azar de estos K registros como donante. Es decir, excluyendo al donante ya usado.

La imputación secuencial de Hot Deck se puede aplicar después de clasificar los registros en un orden aleatorio. De esta forma, el método a veces se denomina método de *imputación aleatoria secuencial Hot Deck*. La imputación Hot Deck secuencial también se puede aplicar sin clasificar aleatoriamente los registros. Luego, depende de cómo se construyan los datos si el método de imputación secuencial Hot Deck conducirá o no a medias y totales sesgados o no. En ambos casos los valores efectivamente imputados dependen del orden de los registros.

Los métodos secuenciales de imputación Hot Deck y cold deck son métodos de imputación deterministas. Después de clasificar los registros del conjunto de datos en orden aleatorio, la imputación secuencial de Hot Deck se convierte en un método estocástico. Como sugiere el nombre, el método aleatorio Hot Deck también es un método estocástico.

Para la imputación de donantes, G. Kalton (1983) ofrece una serie de métodos en los que la probabilidad de ser seleccionado como donante es proporcional al peso. Para evitar que una unidad con un peso pequeño sea seleccionada como donante para un receptor con un peso grande o viceversa, a menudo se asegura que los pesos del donante y del receptor no difieran mucho. Una forma de hacerlo es utilizar la variable de ponderación, o las variables que se utilizan para calcular la variable de ponderación, como variables auxiliares al seleccionar el donante.

3.8 Imputación vecino más cercano

Para aplicar la imputación Hot Deck del vecino más cercano, se debe definir una función de distancia $D(i,k)$ que mida la distancia entre dos unidades i y k, donde i es el ítem que no responde y k es un ítem arbitrario que responde. La función de distancia $D(i,k)$ se puede definir de muchas maneras diferentes. Una función de distancia general de uso frecuente es la llamada distancia de Minkowski

$$D(i,k) = \left(\sum_j |x_{ij} - x_{kj}|^z \right)^{1/z}, \tag{28}$$

donde las variables x son numéricas, y la suma se toma sobre todas las variables auxiliares; $x_{ij}(x_{kj})$ denota el valor de la variable x_j en el registro $i(k)$. Si se alcanza el valor más pequeño de $D_{(i,k)}$ para el ítem que responde $d[d = \arg\min_k D(i,k)]$, entonces se dice que el encuestado d es el vecino más cercano del ítem que no responde i y se convierte en su donante. Para $z = 2$, la distancia de Minkowski es la distancia euclidiana, y para $z = 1$ es la llamada distancia de manzana. Para z más grandes, las grandes diferencias entre x_{ij} y x_{kj} son "castigadas" con más fuerza.

Dejando a z tender a infinito, se obtiene la llamada *distancia minimax* (ver, por ejemplo Sande 1982; Roderick JA Little and Rubin 2002b)). La distancia minimax entre los registros i y k está definida por

$$D(i,k) = max_j \left| x_{ij} - x_{kj} \right|,$$

donde el máximo se toma sobre todas las variables auxiliares x_j. Esta medida de distancia se aplica a menudo en los INE. Era, por ejemplo, la distancia elegida para el módulo del vecino más cercano en el sistema de software de edición e imputación GEIS de Statistics Canada (Cotton (1991)).

Cuando se usa la distancia minimax, se elige un registro de donante tal que la diferencia absoluta máxima entre los valores de las variables auxiliares del donante y el receptor sea mínima. Esta forma de seleccionar un donante asegura que incluso el valor de la variable auxiliar más diferente del registro del donante se acerque al valor correspondiente del receptor. Por lo tanto, el método es robusto frente a la presencia de valores atípicos.

Una función de distancia aún más general que (28) es la función de distancia ponderada dada por

$$D_\gamma(i,k) = \left(\sum_j \gamma_j \left| x_{ij} - x_{kj} \right|^z \right)^{1/z} \tag{29}$$

El factor adicional γ_j es un peso que expresa la importancia de la variable x_j. Dado que solo es relevante el peso relativo, podemos suponer que $\sum_j \gamma_j = 1$. El peso de la variable x_j debe estar relacionado con la importancia, para una imputación precisa, de encontrar un donante con un valor similar en esta variable. En la práctica, las ponderaciones adecuadas suelen ser más fáciles de determinar cuando las variables x se normalizan por primera vez

para que tengan una varianza igual a 1. Esto evita una ponderación implícita cuando las variables se miden en unidades diferentes. También es posible tener en cuenta las covarianzas entre variables al definir $D(i,k)$, pero esto generalmente complica la determinación de pesos adecuados.

Una versión ponderada de la función de distancia minimax viene dada por

$$D_\gamma(i,k) = \mathrm{m}ax_j \gamma_j \left| x_{ij} - x_{kj} \right|$$

o, incluso más generalmente, por

$$D_\gamma(i,k) = \mathrm{m}ax \gamma_j d\left(x_{ij}, x_{kj}\right)$$

con $d(.,.)$ alguna medida de distancia univariada. Tal función de distancia es apropiada cuando se pretende encontrar un donante que no se desvíe mucho del receptor en ninguna de las variables x.

Un caso especial de imputación Hot Deck del vecino más cercano es el *método predictivo de emparejamiento de medias* [véase también Little (1988)]. Cuando se utiliza este método de imputación, primero se lleva a cabo una regresión lineal de la variable objetivo y sobre varias variables predictoras numéricas x, utilizando los registros sin ítems sin respuesta. A continuación, el modelo de regresión resultante se utiliza para predecir para cada registro un valor para la variable objetivo y por medio de la fórmula (5). El registro del donante para el elemento que no responde i viene dado por el elemento que responde d para el cual el valor predicho $\hat{y}_d$ es más cercano al valor predicho $\hat{y}_i$ para el elemento que no responde. Finalmente, se imputa el valor observado y_d del donante d; es decir, $\tilde{y}_i = y_d$ de acuerdo con la fórmula (27).

El coincidencia predictiva de medias es un caso especial de la imputación del vecino más cercano que puede ver observando que el coincidencia predictiva de medias que minimiza la función de distancia

$$D_{(i,k)} = \left| \tilde{y}(\mathbf{x}_i) - \hat{y}(\mathbf{x}_k) \right|, \tag{30}$$

siendo $\mathbf{x}_i$ el vector con variables predictoras para la unidad i que no responde y $\mathbf{x}_k$ el vector con las mismas variables para la unidad k. La distancia de la ecuación (30) se puede expresar como

$$D_{(i,k)} = \left| \mathbf{x}_i^T \boldsymbol{\beta} - \mathbf{x}_k^T \boldsymbol{\beta} \right| = \left| \sum_j \left(x_{ij} - x_{kj} \right) \beta_j \right| \tag{31}$$

donde $\hat{\beta}$ es el vector con coeficientes de regresión estimados $\hat{\beta}_j$. La expresión (31) muestra que el coincidencia predictiva de medias es un método del vecino más cercano con una función de distancia igual al valor absoluto de una suma ponderada de diferencias entre los valores de las variables predictoras. Tenga en cuenta que (31) implica que no es necesario calcular realmente los valores predichos para aplicar el coincidencia predictiva de medias.

Cuando se aplica la imputación del vecino más cercano, incluida la coincidencia de media predictiva, también se pueden seleccionar los K registros más cercanos y dibujar uno de esos K registros al azar, como describimos para la imputación secuencial de Hot Deck. Una ligera modificación es crear un donante potencial con una puntuación pequeña en la función de distancia con una probabilidad alta. Tener en cuenta el aumento de pesos, como en el método de imputación Hot Deck aleatorio ponderado, no tiene ningún efecto si uno se limita a seleccionar un vecino más cercano. Cuando se aplica el coincidencia predictiva de medias, es poco probable que el uso de ponderaciones elevadas tenga mucho efecto.

Se pueden combinar los métodos de imputación aleatoria y del vecino más cercano construyendo primero clases de imputación usando una o más características y luego aplicando el método del vecino más cercano dentro de estas clases. Esta es una forma de aplicar la imputación Hot Deck del vecino más cercano si uno tiene variables categóricas y numéricas. En este caso, las variables auxiliares categóricas se consideran más importantes que las numéricas. Más generalmente, se puede agregar una función de distancia para las variables categóricas a una función de distancia como (29) para variables numéricas y usar una suma ponderada de ambas funciones de distancia como la función de distancia combinada. A las diversas variables categóricas se les puede dar diferentes pesos.

En la Sección 3.1 hicimos una distinción entre "imputación" y el término más general "corrección". En el caso de la imputación, un valor faltante se reemplaza por un valor válido; la corrección de un valor erróneo por un valor válido solo se considera imputación si se considera que el valor erróneo original no juega ningún papel en el proceso de corrección. La imputación Hot Deck del vecino más cercano se puede extender fácilmente a la corrección donde el valor original juega un papel. Luego, la función de distancia se amplía mediante la adición de un término que expresa que el nuevo valor puede no desviarse mucho del valor original erróneo.

3.8.0.1 Ejemplos de imputación Hot Deck

3.8.0.1.1 Ejemplo 7 (Encuesta de Demanda de Vivienda Holandesa) En La Oficina de Estadísticas de Holanda, la imputación de donantes se ha aplicado con frecuencia en el pasado a la Encuesta de Demanda de Vivienda. Para la Encuesta de demanda de vivienda holandesa, las variables se dividieron en grupos, tales como "el hogar", "la vivienda actual", "la vivienda anterior", "la posición socioeconómica del encuestado" y otras. La imputación se realizó por grupo de variables. Dentro de un grupo de variables, las variables con las fracciones más bajas de valores perdidos se imputaron primero. Las variables discretas se imputaron principalmente mediante el método aleatorio Hot Deck. Las variables continuas se imputaron mediante el método Hot Deck aleatorio o mediante coincidencia predictiva de medias. Muchas variables relacionadas en diferentes grupos se imputaron mediante coincidencia de registros, o por medio de la regla del donante común, como también se denomina esta técnica (ver Nordholt 1998).

Es decir, solo se usó un registro de donante para imputar todos los valores faltantes en esas variables relacionadas por registro con falta de respuesta del ítem. Para otras variables, las variables que ya estaban imputadas se utilizaron como covariables para la imputación de valores faltantes que aún no habían sido imputados. En cuanto a la coincidencia de registros, una razón importante para este enfoque fue garantizar la consistencia interna

del registro imputado resultante. Nordholt (1998) da el ejemplo de que falta la Edad del encuestado y Edad de la pareja del encuestado. En tal caso, el valor imputado para la primera variable se utilizó como covariable durante la imputación de la segunda variable para evitar combinaciones de edad muy improbables. Para más detalles sobre la imputación de la Encuesta de demanda de vivienda holandesa, consulte Nordholt (1998); Schulte Nordholt and Hooft Van Huijsduijnen (1997).

3.8.0.1.2 Ejemplo 8 (Encuesta Holandesa sobre La Estructura de los Ingresos) La Encuesta sobre la estructura de los ingresos de Holanda se creó combinando tres fuentes de datos: la Encuesta sobre empleo y salarios, el sistema de registro del fondo de seguridad social y la Encuesta sobre la población activa. Se seleccionó un subconjunto de las variables disponibles en las tres fuentes para la Encuesta sobre la estructura de los ingresos. Estas variables se utilizaron en los procesos de pareamiento, imputación y ponderación. Solo se utilizaron coincidencias exactas entre las tres fuentes de datos. Después de hacer coincidir las tres fuentes de datos, surgió el problema de los valores faltantes en el conjunto de datos resultante. Para algunas variables este problema se resolvió por imputación, para otras variables se resolvió por ponderación. Para imputar los valores faltantes se aplicó el *método secuencial Hot Deck* por clase de imputación. Las variables relacionadas se imputaron simultáneamente usando el mismo modelo de imputación para evitar la introducción de inconsistencias dentro de un registro imputado. Se aplicó el *método de Hot Deck secuencial* y no el de *Hot Deck aleatorio*, porque se consideró que el número de registros era demasiado grande para utilizar este último método de manera eficiente. Se introdujo un componente aleatorio en el proceso de imputación al colocar los registros de donantes potenciales en orden aleatorio antes de que tuviera lugar la imputación real. Para obtener más información sobre la imputación de la Encuesta holandesa sobre la estructura de los ingresos, consulte Nordholt (1997); Nordholt (1998).

Los datos faltantes de la Encuesta holandesa sobre la estructura de los ingresos también se imputaron por medio de una red neuronal (ver Heerschap and Graaf (1999)). En particular, los valores perdidos de la variable Salario bruto anual se han imputado por medio de una red neuronal de perceptrón multicapa de alimentación hacia adelante (consulte, por ejemplo Fine 1999 para obtener más información sobre redes neuronales).

3.8.0.1.3 Ejemplo 9 (Estadísticas Holandesas de Mecanización en la Agricultura y Horticultura) La imputación de cold deck, en el sentido de que los datos de una empresa disponibles en un conjunto de datos se utilizan para imputar los datos faltantes de la misma empresa en otro conjunto de datos, se utiliza a menudo en Estadísticas de Holanda. Se aplicó, por ejemplo, para las Estadísticas de Mecanización en Agricultura y Horticultura, donde se utilizaron datos del llamado Censo Agropecuario para imputar valores faltantes. Por otro lado, los métodos de imputación Hot Deck rara vez se aplican en La Oficina de Estadísticas de Holanda para imputar datos económicos.

3.9 Un modelo de imputación general

En esta sección describimos nuevamente los diferentes métodos de imputación de las secciones anteriores, esta vez como casos especiales del mismo modelo lineal general. A su manera, queremos destacar tanto las diferencias como las similitudes entre estos métodos. Este enfoque

para integrar los diferentes métodos de imputación es similar al descrito en Graham Kalton (1986). El modelo lineal que consideramos es una extensión del modelo de regresión lineal tratado en la Sección 3.3:

$$y = \alpha + \beta_1 x_1 + ... + \beta_p x_p + \varepsilon^* = \alpha + \mathbf{x}^T \boldsymbol{\beta} + \varepsilon^* \tag{32}$$

donde α, β_1, β_p denota los parámetros del modelo y ε^* denota un residuo. La variable y es la variable objetivo para la que se encuentra un valor imputado. Las variables x se denominan variables auxiliares o predictoras y pueden ser variables continuas o variables dummy binarias $(0, 1)$ que representan las categorías de una variable categórica. Estas variables dummy se definen de la siguiente manera: Para una variable categórica Z con K categorías, K variables dummy $x_1^{(z)}, ..., x_k^{(z)}$ se crean, cada uno con valor $x_k^{(z)} = 1$ si el elemento pertenece a la categoría k de Z y $x_k^{(z)} = 0$ en caso contrario. Juntas, las variables ficticias K para las categorías de Z se denotarán por el vector z. Dado que se supone que las categorías de variables auxiliares categóricas son mutuamente excluyentes y exhaustivas, una de las dummies será 1 y las otras 0. Aparte de las variables categóricas como continuas, el modelo también puede contener interacciones entre estos dos tipos de variables. Estas interacciones se pueden incluir en el modelo generando, para una variable categórica de categoría K, K variables de interacción definidas como los productos $x_k^{(z)}.x_j$, para algunas variables continuas x_j y $k = 1, ..., K$. Juntas, las K variables para la interacción entre Z y x_j se denotarán como el producto $\mathbf{z}.x_j$.

En (32) se agrega un superíndice " $*$ " a ε para indicar que, en contraste con el modelo de regresión habitual, ε^* no es necesariamente la realización de una variable aleatoria con alguna distribución paramétrica, como la distribución normal. En cambio, ε^* también puede ser la realización de una distribución empírica no paramétrica definida por los propios datos y, en algunos casos, incluso puede calcularse de manera determinista. Al especificar si se usa un parámetro α, si se usan (y cuántos) predictores se usan, y cómo se determina ε^*, este modelo de imputación general cubre muchos métodos de imputación para una sola variable objetivo y.

Todos los casos especiales considerados aquí (excepto los métodos deductivos) usan estimaciones de los parámetros α y β, $\hat{\alpha}$ y $\hat{\beta}$, digamos. En general, estos parámetros se estiman mediante wls (minímos cuadrados poderados) utilizando los datos para los que se observan las variables y y x. Los estimadores wls de α y β se pueden escribir como

$$\left(\hat{\alpha}, \hat{\beta}^T\right)^T = \left(\sum_{i \in obs} \mathbf{x}_i \nu_i^{-1} \mathbf{x}_i^T\right)^{-1} \sum_{i \in obs} \left(\mathbf{x}_i \nu_i^{-1} y_i\right), \tag{33}$$

donde el subíndice i denota las unidades, "obs" denota los índices de las unidades utilizadas para estimar los parámetros y ν_i denota una estimación de la varianza de ε_i^*. Así como el valor esperado de y es una función de predictores o variables auxiliares (función media), ν_i también puede ser una función de variables auxiliares (función de varianza). Si se supone que vi es una constante, escribimos $\nu_i = c$ y la ecuación (33) se reduce al estimador de mínimos cuadrados. Se utilizarán otras tres suposiciones sobre ν_i.

I. La varianza es igual dentro de los grupos. En este caso tenemos $\boldsymbol{\nu}_i = \mathbf{z}_i^T \mathbf{c}$, con $\mathbf{c}$ un vector con parámetros de varianza, con longitud igual al número de categorías (o grupos) de la variable discreta Z .

II. La varianza es proporcional a una de las variables auxiliares, digamos x_j. Entonces tenemos $\nu_i = x_{ij}c$, siendo c la constante de proporcionalidad.

Los supuestos **I** y **II** se pueden combinar para obtener el tercer supuesto:

III. La varianza es proporcional a una variable auxiliar dentro de los grupos. En este caso $\nu_i = (\mathbf{z}_i \cdot x_{ij})^T \mathbf{c}$, siendo $\mathbf{c}$ un vector con constantes de proporcionalidad.

Utilizando el modelo de imputación general (32), se puede hacer una distinción importante entre imputación determinista y estocástica. Para los métodos de imputación estocástica, el residuo ε^* se basa en un proceso estocástico, como una extracción aleatoria de una distribución o una extracción aleatoria de residuos observados. Para la mayoría de los métodos de imputación deterministas, el residuo ε_i^* es igual a cero. Una excepción es la imputación del vecino más cercano (ver más abajo), donde el residuo es distinto de cero. Sin embargo, el residuo de la imputación del vecino más próximo se calcula de forma determinista. A continuación consideramos valores imputados generados por casos especiales del modelo (32). Esto conduce a una descripción concisa de una serie de métodos de imputación bien conocidos.

Los métodos de imputación determinista incluyen:

- *Imputación proxy* ($\tilde{y}_i = x_i$). El valor imputado es el valor de otra variable que se observa y, preferiblemente, cercana a la variable a imputar. No es necesario estimar parámetros. Esta es una forma de imputación deductiva. Algunos ejemplos son imputar, en una encuesta en curso, con un valor anterior, o imputar el valor faltante de la nacionalidad de un esposo con el valor de su esposa.
- *Imputación eductiva* usando ediciones de balance ($\tilde{y}_i = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}$). Supongamos que algunas variables en el registro tienen que satisfacer una relación de la forma

$$x_{i,p} = x_{i,1} + ... + x_{i,p-1}$$

y una de estas variables es el valor y_i faltante mientras que las demás se observan, entonces el valor imputado se puede calcular a partir de los valores observados en la edición. Las estimaciones de los parámetros son **+1** o **-1** y aunque podrían estimarse por mínimos cuadrados, pueden determinarse directamente sin más datos que el registro i.

- *Imputación media* ($\tilde{y}_i = \hat{\alpha}; \nu_i = c$). Para la imputación de la media no se utiliza la variable auxiliar x_j, y $\hat{\alpha}$ se estima por Mínimos Cuadrados Ordinarios (mco) porque $\nu_i = c$, lo que significa que $\hat{\alpha}$ será igual a la media de las unidades con respuestas en y, es decir, los ítems encuestados.
- *Imputación de la media del grupo* ($\tilde{y}_i = \mathbf{z}_i^T \hat{\boldsymbol{\beta}}$; $\nu_i = c$ ó $\nu_i = \mathbf{z}_i^T \boldsymbol{c}$). Para la media del grupo imputación, las medias y las varianzas son constantes dentro de cada grupo. Para la imputación determinista no importa si los parámetros se estiman por mco ($\nu_i = c$) o por (Mínimos Cuadrados Ponderados) wls con función de varianza $\nu_i = \mathbf{z}_i^T \boldsymbol{c}$; en ambos casos es $\tilde{y}_i$ igual al valor medio de los ítems encuestados en el grupo al que pertenece el

registro i. Para los modelos estocásticos considerados a continuación, las suposiciones sobre la estructura de la varianza sí marcan la diferencia.

- *Imputación de Razón* ($\tilde{y}_i = x_{ij}\hat{\beta}; \nu_i = x_{ij}c$) Para la imputación de razón, se utiliza una variable x auxiliar. El parámetro $\alpha = 0$ por definición. El estimador wls $\hat{\beta}_j$ es igual a la razón entre las medias de y y x_j entre las unidades con ambas variables observadas, como puede verificarse fácilmente tomando x_i en (33) como el puntaje x_{ij} en una sola variable x_j y sustituyendo $x_{ij}c$ para ν_i.
- *Imputación de razón de grupo* $\left(\tilde{y}_i = x_{ij}\hat{\beta}_j; \nu_i = x_{ij}c\right)$. Las varianzas son proporcionales a x_j pero con una constante de proporcionalidad c_k diferente para cada grupo k. Los componentes del estimador wls $\hat{\boldsymbol{\beta}}$ son los cocientes de las medias dentro del grupo de y y x_j. En general, si los valores auxiliares tanto en la función de media como en la función de varianza solo aparecen en interacción con la misma variable categórica, entonces el modelo se "desacopla" entre los grupos y es equivalente a aplicar el modelo a cada grupo por separado. Este es, por ejemplo, también el caso de la imputación de la media de grupo.
- *Regresión imputación (sin residuo)* $\left(\tilde{y}_i = \hat{\alpha} + \mathbf{x}_i\hat{\boldsymbol{\beta}}_j; \nu_i\right)$. La imputación de media, la imputación de razón y las variantes agrupadas de las mismas son casos especiales de imputación de regresión, cada uno de los cuales corresponde a funciones específicas de media y varianza. El caso general también cubre todas las demás especificaciones de la función de media y varianza.
- *Imputación del vecino más cercano* $\left(\tilde{y}_i = \left(A_{(\mathbf{x}_i)}\right)^T \hat{\boldsymbol{\beta}} + e_d\right)$. Para la imputación del vecino más próximo, se utiliza una aproximación $A(\mathbf{x}_i)$ a x_i. Esta aproximación es el valor x de otro registro (el donante d), con tanto y como x observados, que es el más cercano a x_i de acuerdo con alguna función de distancia especificada. El residual e_d es el residual realizado $y_d - \mathbf{x}_d^T\hat{\boldsymbol{\beta}}$ para el donante. Por lo tanto, para cualquier método de estimación de $\hat{\boldsymbol{\beta}}$, el valor imputado es igual al valor observado y_d del donante.
- *Coincidencia predictiva de medias* $\left(\tilde{y}_i = A(\mathbf{x}_i^T\hat{\boldsymbol{\beta}}) + e_d\right)$. Coincidencia predictiva de medias puede verse como un caso especial de imputación del vecino más cercano. En este caso, se busca un donante tal que el valor predicho de y, sin un término de error, sea el más cercano al valor predicho para el registro i. La imputación con este valor donante también se puede interpretar como una aproximación del valor predicho para el registro i por $A(\mathbf{x}_i^T\hat{\boldsymbol{\beta}}) = \mathbf{x}_d^T\hat{\boldsymbol{\beta}}$ y sumando el residual observado del donante.
- *Imputación secuencial Hot Deck* ($\tilde{y}_i = y_d$). Para la imputación Hot Deck secuencial, se considera que el donante es el primer registro en el conjunto de datos después del registro en consideración que tiene un valor válido para la variable objetivo y y los mismos valores de las variables predictoras categóricas. El valor imputado simplemente es igual al valor observado del donante. Por supuesto, esto también es cierto para los otros métodos de donantes.

Los métodos de imputación estocástica incluyen:

- *Imputación de regresión (con un residual de una distribución paramétrica)* $\left(\tilde{y}_i = \hat{\alpha} + \mathbf{x}_i^T\hat{\boldsymbol{\beta}} + \varepsilon_i^*; \varepsilon_i^* \sim (0, \nu_i)\right)$. Para la imputación de regresión con un resid-

ual de una distribución paramétrica, se usa un parámetro α y un vector de variables x numéricas o categóricas con el correspondiente vector de parámetros β y se agrega un residual. El residual se extrae de una distribución paramétrica, generalmente la distribución normal, con media cero y varianza ν_i. Al elegir funciones específicas de media y varianza, surgen versiones estocásticas de diversos submodelos del modelo de imputación de regresión.

- *Imputación de regresión (con un residuo observado).* La imputación de regresión con un residuo observado es similar a la imputación de regresión con un residuo de una distribución paramétrica. La única diferencia es que el residuo ahora se extrae del conjunto de residuos observados, digamos e_{obs}.
- *Imputación aleatoria de Hot Deck* ($\tilde{y}_i = \hat{\alpha} + \varepsilon_i^*; \varepsilon_i^* \sim (\mathbf{e}_{obs})$). Para la imputación aleatoria Hot Deck, $\hat{\alpha}$ se estima mediante mco y, por lo tanto, es igual a la media general de los encuestados. Se agrega un residuo $y_d - \hat{\alpha}$ de un donante elegido al azar. Este método puede verse como una variante de la imputación media estocástica con los residuos extraídos de la distribución empírica de residuos alrededor de la media.
- *Imputación aleatoria grupal de Hot Deck* ($\tilde{y}_i = \mathbf{z}_i^T \hat{\boldsymbol{\beta}} + \varepsilon_i^*; \varepsilon_i^* \sim)\mathbf{e}_{obs})$). Para la imputación aleatoria grupal Hot Deck, $\hat{\beta}$ consiste en las medias de los encuestados de las categorías de Z (los grupos). Se agrega un residuo $y_d - \mathbf{z}_i^T \hat{\boldsymbol{\beta}}$ de un donante seleccionado al azar con el mismo valor de $\mathbf{z}(\mathbf{z}_i = \mathbf{z}_d)$, es decir, un donante del mismo grupo que el receptor. Este método puede verse como una variante de la imputación estocástica de la media del grupo con los residuos extraídos de la distribución empírica de los residuos alrededor de las medias de su grupo.

Dado que los valores imputados se extraen aleatoriamente para la imputación estocástica, no pueden reproducirse en general. En el caso de imputación determinista, los valores imputados pueden reproducirse, dado el modelo de imputación seleccionado. La elección entre imputación estocástica y determinista en muchos casos depende de si se desea o no agregar un residuo al modelo. Sin embargo, la imputación del vecino más cercano, incluida la coincidencia de media predictiva, agrega implícitamente un residuo al modelo de imputación y es determinista, porque el donante seleccionado se encuentra de manera determinista mediante una función de distancia.

Otra distinción que se puede hacer es entre la imputación de regresión (ya sea determinista o estocástica), incluidos sus casos especiales de imputación de razón e imputación de la media (de grupo), y las técnicas de imputación de donantes. Para las técnicas de imputación de donantes, los residuos se extraen de alguna manera de los residuos observados y las variables auxiliares $x_1, ..., x_p$ solo se utilizan para construir clases de imputación o para basar una función de distancia. La elección entre la imputación de regresión y la imputación de donantes a menudo no es obvia. Algunas consideraciones al elegir entre estos dos enfoques pueden incluir lo siguiente:

- Dado que los valores de los donantes son valores realmente observados (y correctos), siempre son valores válidos para la variable de destino. Si la variable objetivo tiene un valor entero o no negativo, también lo serán los valores del donante. En general, los modelos de regresión no producirán valores predichos enteros y pueden generar predicciones negativas para variables no negativas.

- La regresión y la imputación del vecino más cercano naturalmente pueden hacer uso de variables auxiliares tanto categóricas como continuas. Los métodos aleatorios y secuenciales de Hot Deck se aplican dentro de grupos determinados por las combinaciones de valores de variables categóricas. Las variables continuas sólo pueden utilizarse como variables auxiliares si antes se discretizan, lo que implica una pérdida de información y también de poder predictivo si la relación entre las variables auxiliares y la variable objetivo es lineal.
- La cantidad de información auxiliar es más limitada para los métodos que utilizan una agrupación de datos (es decir, métodos aleatorios y secuenciales de Hot Deck) que para los métodos de regresión y del vecino más cercano. Debido a que los grupos están definidos por todas las combinaciones de valores posibles de las variables auxiliares, los grupos con solo unos pocos donantes o ninguno en absoluto se convertirán en un problema cada vez mayor a medida que aumenta el número de variables auxiliares (y categorías por variable). En regresión, el equivalente de la estructura de agrupación es incluir variables ficticias para todas las variables auxiliares categóricas y todas las interacciones entre estas variables. Esto también puede conducir a problemas de grupos vacíos y, en este caso, parámetros inestimables. Sin embargo, en el enfoque de regresión es fácil, e incluso habitual, reducir el modelo a uno más parsimonioso dejando de lado las interacciones entre variables pero conservando los efectos principales. Para los métodos basados en grupos, se usa una variable (con todas las interacciones) o no se usa en absoluto. Los métodos del vecino más cercano no presentan problemas técnicos cuando se utilizan muchas variables auxiliares; pero cuando aumenta el número de variables auxiliares, la discrepancia entre el donante y el receptor en cada variable tenderá a aumentar debido a la necesidad de emparejar lo mejor posible en más variables simultáneamente.
- La imputación de donantes se extiende más fácilmente que la imputación basada en modelos a problemas multivariados en los que se deben imputar varias variables en un registro, y la relación entre estas variables se debe preservar lo mejor posible.
- Para actividades donde no todos contribuyen, la distribución de los datos consiste en ceros para una parte de la población, los no participantes y valores positivos para los participantes. Algunos ejemplos son la cantidad de dinero gastado en vacaciones, la cantidad de kilómetros recorridos en automóvil y la cantidad de dinero invertida por una empresa en máquinas nuevas. Los métodos Hot Deck dan buenos resultados para este tipo de variables, en el sentido de que conservan la distribución. Sin embargo, si se aplica la imputación media, no se imputará ningún cero. La imputación de regresión tiene el mismo problema y, además, pueden ocurrir valores imputados negativos para dichas variables no negativas. Si el objetivo es estimar las medias de la población, esto no es un problema importante. Sin embargo, si el objetivo es estimar la variación de la variable objetivo o la fracción de participantes, no se pueden aplicar estas técnicas. Una opción para la imputación en tales casos, además de Hot Deck, es aplicar la imputación en dos pasos. Primero, se aplica una regresión logística o un método de imputación Hot Deck para "determinar", es decir, imputar, si un elemento que no responde es un participante o no, y luego se obtienen los valores imputados para los supuestos participantes por medio de un modelo de imputación de regresión lineal.

3.10 Imputación de Datos Longitudinales

Los datos longitudinales ocurren cuando las mismas variables de las mismas unidades se miden varias veces, en diferentes momentos. Por ejemplo, los datos de las encuestas de panel son datos longitudinales, al igual que los datos de un registro de población en diferentes momentos. En la Oficina Estadística de Holanda, se utiliza una encuesta de panel para la Encuesta de población activa. Un ejemplo de registro longitudinal en los Países Bajos es la Administración Base Municipal para la cual cada año se construye una nueva versión. La mayoría de los registros proporcionan información longitudinal cuando se vinculan varias versiones de diferentes momentos.

La imputación longitudinal se distingue de la mayoría de los demás métodos de imputación descritos en este capítulo porque para la imputación longitudinal se utilizan datos de las mismas unidades, en muchos casos sin utilizar información de otras unidades. Por lo tanto, para cada unidad, se tiene una serie de tiempo con uno o más valores faltantes que deben imputarse. Sin embargo, la imputación longitudinal está estrechamente relacionada con la imputación de razón.

Los valores faltantes en los datos longitudinales ocurren de dos maneras diferentes:

1. Valores faltantes que se distribuyen en el tiempo;
2. Valores faltantes debido a la caída del panel.

Cuando los valores faltantes se deben a abandonos, solo se puede usar información del pasado de las unidades correspondientes, y se pueden usar métodos de extrapolación o imputación de razón (consulte la Sección 3.4) para predecir los valores faltantes de los datos numéricos. Una técnica de uso frecuente es el último valor transferido donde simplemente se imputa el último valor observado de un período anterior. En el caso de los datos numéricos, el último valor transferido a menudo se aplica sin una corrección de tendencia, pero también se puede aplicar con una corrección de tendencia (consulte también la Sección 3.4 sobre imputación de razón).

Cuando los valores faltantes se distribuyen en el tiempo, la imputación a menudo se basará en la interpolación. Sin embargo, cuando el objetivo es proporcionar estimaciones lo antes posible después de que se disponga de una nueva oleada de datos longitudinales, se pueden volver a utilizar solo datos de momentos anteriores y, de nuevo, se deben aplicar técnicas de extrapolación o imputación de razón. La información en momentos posteriores solo se puede utilizar cuando se tiene tiempo suficiente para esperar olas posteriores de datos longitudinales o cuando se construyen imputaciones para varios momentos simultáneamente para obtener un conjunto de datos longitudinales lo mejor posible.

Para algunas encuestas, se comienza con un conjunto completo de datos, incluso antes de que se hayan observado datos para el período actual. Los valores imputados pueden, por ejemplo, imputarse en función de los valores de un período anterior $t-1$. Cada vez que se obtienen nuevos datos observados, estos valores reemplazan los valores imputados, después de lo cual los valores imputados restantes se actualizan, utilizando el modelo de imputación con parámetros del modelo actualizados. El proceso estadístico para tal enfoque difiere del enfoque estándar para la imputación de valores faltantes, y tiene la ventaja de que se puede obtener fácilmente una estimación de la población en cualquier momento, pero desde un punto de vista teórico no hay una gran diferencia entre los dos enfoques.

Las encuestas sociales típicamente involucran muchos datos categóricos. Para los datos categóricos, no se aplican las técnicas de imputación, extrapolación e interpolación de razón antes mencionadas. En su lugar, se pueden utilizar métodos de imputación hot deck donde las clases de imputación involucran variables que coinciden en un levantamiento diferente.

Los problemas causados por la falta de valores debido a la caída del panel generalmente también se pueden resolver mediante métodos de ponderación. Cuando se quiere estimar una estadística en un momento determinado, se puede considerar la deserción reciente como falta de respuesta unitaria y agregarla a la falta de respuesta de levantamientos anteriores del panel.

La deserción del panel en los registros a menudo está justificada y puede, por ejemplo, ser causada por la emigración y las muertes.

Para obtener más información sobre la imputación longitudinal, consulte Daniels and Hogan (2008)

Ya se ha dado una ilustración de la imputación longitudinal en el Ejemplo 3.4.2. A continuación damos ejemplos adicionales.

3.10.0.1 Ejemplo 10 (Estadísticas comerciales de Holanda) La imputación de datos longitudinales se aplica a menudo en Estadísticas de los Países Bajos para imputar los datos que faltan en las estadísticas económicas que se llevan a cabo de forma regular (mensual, trimestral o anual). Por ejemplo, los valores faltantes en las Estadísticas Empresariales Estructurales del Comercio Minorista se imputan por valores de un período anterior, ajustados por un factor de tendencia. El factor de tendencia generalmente se estima utilizando otros registros en el mismo grupo de imputación.

Para imputar los valores faltantes en las estadísticas anuales, a menudo se utilizan los datos de las estadísticas mensuales o trimestrales correspondientes, si están disponibles. Este enfoque se utilizó, por ejemplo, para las Estadísticas de la Industria del Petróleo.

3.10.0.2 Ejemplo 11 (Encuesta de la industria de la construcción en Holanda) Para la Encuesta de la Industria de la Construcción se utilizaron fuentes de datos internas y registros externos para guiar el proceso de imputación. La encuesta consistió en alrededor de un centenar de variables en su mayoría numéricas. En caso de que una empresa no respondiera, se determinaba el valor del volumen de negocios o el número de empleados de esa empresa. Esto se hizo volviendo a contactar a la empresa o usando datos que estaban disponibles de otra fuente. Se estimó la "estructura" de la empresa, es decir, las proporciones entre las diversas variables desconocidas y la variable observada. Esto se hizo usando datos del año anterior o usando la estructura promedio de las empresas que respondieron. Luego se imputaron los valores faltantes conservando esa estructura. A continuación ilustramos el procedimiento utilizado para la Encuesta de la Industria de la Construcción mediante un ejemplo sencillo.

3.10.0.3 Ejemplo 12 (Encuesta de Panel de Hogares de la Comunidad Europea) Nordholt (1998) describe la estrategia de imputación que se aplicó para imputar los valores faltantes de la Encuesta de Panel de Hogares de la Comunidad Europea (ECHP) (véase también Schulte Nordholt 1996). El Panel Socioeconómico Holandés fue parte de esa encuesta. El ECHP se

centró en los ingresos y el mercado laboral; otros temas incluyen salud, educación, vivienda y migración. La encuesta se llevó a cabo a nivel de hogar.

La ventaja del ECHP en comparación con las encuestas transversales fue que permitió el estudio de los patrones de movilidad de ingresos. Sin embargo, el hecho de que la ECHP fuera una encuesta de panel complicó el proceso de imputación.

La imputación de cada levantamiento por separado de los levantamientos anteriores podría conducir a extraños cambios en los ingresos. En lugar de esta estrategia de imputación ingenua, una mejor estrategia sería imputar todos los levantamientos disponibles simultáneamente. Sin embargo, debido a que esta estrategia de imputación conduciría a cambios en los resultados de todas los levantamientos anteriores después de la imputación de un nuevo levantamiento, se adoptó una estrategia más sencilla. La estrategia elegida fue tener en cuenta imputaciones de levantamientos anteriores al imputar un nuevo levantamiento, pero no se adaptaron las imputaciones realizadas durante encuestas anteriores.

Para imputar el primer levantamiento del ECHP, se utilizó la imputación aleatoria hot deck dentro de los grupos. Dado que las variables de ingresos son muy importantes en el ECHP, la imputación se concentró en estas variables. Además de variables como Género y Año de nacimiento, varias variables laborales importantes, como Situación laboral y Ocupación actual, se utilizaron como variables auxiliares para imputar valores faltantes de las variables de ingresos. Se tuvo cuidado de evitar inconsistencias entre las variables relacionadas, como el ingreso bruto y neto del trabajo por mes. Para imputar el Ingreso laboral bruto mensual y el Ingreso laboral neto mensual se utilizó el método de cotejo de registros. Se utilizó el ingreso bruto del trabajo por mes como variable principal para la imputación. Se utilizó el método hot deck dentro de grupos para imputar los valores faltantes de esta variable. Si el valor de la variable relacionada Ingresos netos del trabajo por mes también faltaba en un registro, esta variable se imputó utilizando el mismo registro del donante para evitar inconsistencias.

3.11 Métodos para la estimación de la varianza con datos imputados

3.11.1 El problema y enfoques para enfrentarlo

La mayoría de los métodos de imputación conducen a una subestimación de la varianza de la variable imputada si las imputaciones se tratan como si no fueran diferentes de las observaciones "reales". Como consecuencia, las fórmulas estándar para la varianza y el error estándar de los estimadores están sesgadas a la baja y los intervalos de confianza se vuelven demasiado estrechos. El siguiente ejemplo sencillo ilustra el problema.

Suponga que se dispone de una muestra aleatoria simple sin reemplazo (SRSWOR) de tamaño n de una población de tamaño N para estimar la media poblacional de y. Un estimador insesgado para la media poblacional es la media muestral $\tilde{y}$ con estimador de varianza dado por

$$\hat{V}(\overline{y}) = \left(\frac{1}{n} - \frac{1}{N}\right)\hat{\sigma^2}, \tag{34}$$

con $\sigma^{2\frown}$ un estimador de la varianza poblacional, σ^2, de y. Este estimador es en este caso $s^2 = \sum_{i=1}^{n}(y_i - \overline{y})^2/(n-1)$. Ahora suponga que solo respondieron los n_obs de las n unidades y el valor de y para las unidades que no respondieron se imputa con la media de las unidades que

respondieron, $\overline{y}_{obs}$. Sin más información se puede estimar la media poblacional con la media de los datos imputados, que es igual a $\overline{y}_{obs}$. Si se aplica la fórmula de varianza estándar (34) a los datos imputados, el estimador de la varianza de y sería $\sum_{i\in obs}(y_i - \overline{y}_{obs})^2/n - 1$,ya que los residuos $y_i - \overline{y}_{obs}$ se cancelan para los valores imputados. Este estimador de σ^2 estará sesgado a la baja. Se puede obtener un estimador correcto para σ^2 si asumimos que el mecanismo de falta de respuesta es uniforme, es decir, probabilidades de respuesta independientes e idénticas en todas las unidades de muestra, de modo que los residuos distintos de cero observados puedan verse como una submuestra aleatoria de los residuos para el total de la muestra. El estimador correcto para σ^2 es entonces

$$s^2_{obs} = \sum_{i\in obs}(y_i - \overline{y}_{obs})^2/(n_{obs} - 1). \tag{35}$$

Es importante señalar que, a diferencia de la imputación determinista, la imputación estocástica puede conducir a un estimador correcto de σ^2 mediante la fórmula estándar. Con la estimación σ^2_{obs} la estimación de la varianza (34) es aproximadamente igual a la estimación que se hubiera obtenido con respuesta completa. Sin embargo, esta sigue siendo una estimación insatisfactoria porque se espera que la pérdida de información debido a la falta de respuesta resulte en una disminución de la precisión en comparación con el caso de una respuesta completa.

Este ejemplo muy simple ya muestra algunos de los principales problemas en la estimación de la varianza después de la imputación. En primer lugar, las fórmulas estándar para las varianzas o los errores estándar de las estimaciones están sesgadas a la baja. Además, la estimación de la varianza poblacional σ^2 es demasiado baja, pero esto se puede compensar (bajo ciertas suposiciones) mediante una adaptación de la fórmula o agregando residuos aleatorios. Sin embargo, una fórmula estándar con una estimación correcta de σ^2 todavía subestima la varianza de las estimaciones, ya que no tiene en cuenta adecuadamente la pérdida de información debido a la falta de respuesta.

Para calcular los errores estándar válidos de las estimaciones, es necesario utilizar métodos desarrollados específicamente para tener en cuenta las imputaciones. Se han propuesto varios enfoques para este problema, y en esta sección distinguimos tres tipos de enfoques.

El enfoque analítico.

En este enfoque, las fórmulas habituales de varianza de muestreo repetido basadas en el diseño se amplían para tener en cuenta la falta de respuesta e imputación posterior. Tales extensiones de la teoría estándar han sido consideradas, entre otros, por C. Särndal (n.d.) y Deville and Särndal (1994). Una ruta que se toma a menudo para derivar fórmulas de varianza analítica es el enfoque de dos fases. Muestreo en dos fases significa que una muestra final se extrae en dos pasos: en la primera fase se extrae una muestra de la población y en la segunda fase se extrae una muestra de la muestra realizada en el primer paso. Cuando se aplica a la falta de respuesta, la muestra de la primera fase corresponde a las unidades de muestreo (respondedores y no encuestados) y la muestra de la segunda fase corresponde a la submuestra de las unidades que respondieron. Para derivar estimadores de varianza, se hace una suposición para el mecanismo de selección en la segunda fase (el mecanismo de falta de respuesta). Por lo general, se supone que, dentro de las clases, el mecanismo de falta de respuesta es uniforme. Una suposición alternativa, que también puede conducir a

estimadores de varianza válidos, es suponer que el mecanismo de falta de respuesta es MAR; depende de las variables auxiliares completamente observadas que se utilizan en el modelo de imputación pero no de la variable objetivo. El uso de esta última suposición también se denomina enfoque asistido por modelos.

El enfoque de remuestreo.

Los métodos de remuestreo para la estimación de la varianza basado en muestras complejas sin falta de respuesta ha sido una práctica común en la metodología de la encuesta (ver Wolter 1985). Su ventaja es que, mientras que las fórmulas analíticas de la varianza deben derivarse por separado para diferentes tipos de estimadores y pueden volverse bastante complejas, los métodos de remuestreo ofrecen un procedimiento computacional relativamente simple para obtener estimaciones de la varianza que es lo suficientemente general como para ser aplicable a muchos problemas de estimación. Esta ventaja es particularmente relevante para el caso de la estimación de la varianza con datos imputados, ya que las fórmulas tienden a ser más complicadas que sin la falta de respuesta. Por lo tanto, los métodos de remuestreo para la estimación de la varianza con datos imputados han recibido una atención considerable (J. N. Rao and Shao (1992); Shao and Sitter (1996)). Los métodos más comúnmente estudiados y aplicados son los esquemas de remuestreo jackknife y bootstrap. Además de estos métodos, también han aparecido algunos trabajos sobre replicación repetida balanceada (BRR) (Shao (2002)).

Imputación múltiple (IM).

En este enfoque, cada valor faltante se imputa varias veces, digamos M , y la variación entre las imputaciones M se utiliza para estimar el aumento de la varianza debido a la falta de respuesta y la imputación. La imputación múltiple fue originada por Donald B. Rubin (1978); D. Rubin (1987). A diferencia de la imputación única (rellenar un solo valor para cada uno que falta), la imputación múltiple se pensó desde el principio como un método para proporcionar no solo una solución al problema de los datos faltantes por imputación, sino también para reflejar la incertidumbre inherente a las imputaciones. El objetivo de la imputación múltiple es proporcionar conjuntos de datos de imputación múltiple que deberían permitir a los investigadores realizar diferentes tipos de análisis y obtener inferencias con errores estándar válidos, intervalos de confianza y pruebas estadísticas, de una manera sencilla. Se realizan análisis estándar en cada uno de los M conjuntos de datos y los resultados se combinan usando fórmulas relativamente simples para obtener inferencias válidas.

Los enfoques analítico y de remuestreo son similares en el sentido de que ambos se adhieren al marco de muestreo repetido basado en el diseño para la inferencia estadística. Los valores de las variables auxiliares (x) y las variables objetivo (y) se ven como fijos. La aleatoriedad se introduce mediante la selección aleatoria de unidades (de acuerdo con el diseño especificado) de la población, y la inferencia es con respecto a la distribución de muestreo asociada. Los INE aplican, casi sin excepción, la inferencia basada en el diseño, ya que está más en línea con su tarea principal de producir estadísticas descriptivas, a menudo simples, de poblaciones fijas. La imputación múltiple, por otro lado, se desarrolló como una técnica basada en un modelo bayesiano; las variables se ven como variables aleatorias con alguna distribución especificada por un modelo. Las observaciones se ven como sorteos aleatorios independientes generados por esta distribución específica. No obstante, (D. Rubin (1987) Capítulo 4) explica que las inferencias de **IM** también son válidas para el marco de muestreo repetido basado en el

diseño, siempre que las imputaciones sean "adecuadas", lo que implica, entre otras cosas, que las características del diseño de muestreo que influyen en la varianza del muestreo repetido se tienen en cuenta en el modelo de imputación. Para diseños complejos que involucran estratificación, probabilidades de selección desiguales y agrupamiento, esto puede conducir a modelos de imputación bastante complejos. En el número de junio de 1996 del Journal of the American Statistical Association (Donald B. Rubin (1996); Fay (1996); J. Rao (1996)) hay un debate sobre las ventajas y desventajas de los diferentes enfoques y su aplicabilidad para diferentes propósitos.

A continuación, damos algunos detalles de estos diferentes enfoques y referencias a la literatura relevante.

3.11.2 El enfoque analítico

Los estimadores analíticos de la varianza que se han derivado generalmente constan de dos componentes (C.-E. Särndal (1992)), lo que lleva a la siguiente descomposición de un estimador analítico de la varianza $\hat{V}_A$:

$$\hat{V}_A = V_{sam} + V_{imp} \tag{36}$$

El primer componente es la varianza del estimador que se habría obtenido si no hubiera falta de respuesta y el segundo componente es la varianza adicional por falta de respuesta. Más precisamente, este segundo componente es la varianza condicional por falta de respuesta dada la muestra realizada. El objetivo del enfoque analítico es obtener estimadores no sesgados para cada uno de estos componentes por separado y obtener un estimador para la varianza total como su suma.

Como ilustración, consideramos nuevamente la estimación de la media poblacional a partir de una muestra por Muestreo aleatorio simple sin reemplazo (SRSWOR) bajo imputación de la media y un mecanismo uniforme de falta de respuesta. La varianza muestral de un estimador sin falta de respuesta está dada en este caso por (34) y una estimación insesgada de σ^2 es s^2_{obs} dada por (35). Sustituyendo σ^2 es s^2_{obs} en (34) se obtiene un estimador para V_{sam}. Para obtener un estimador del segundo componente, observamos que bajo el mecanismo de respuesta uniforme, $\overline{y}_{obs}$ es un estimador de la media de la muestra completa y la varianza condicional, que esencialmente trata a la muestra como la población y trata a los encuestados como la muestra, por lo tanto, $V_{imp} = (1/n_{obs} - 1/n)s^2_{obs}$. Sustituyendo estas expresiones por V_{sam} y V_{imp} en (36), obtenemos

$$\hat{V}_A = \left(\frac{1}{n} - \frac{1}{N}\right) s^2_{obs} + \left(\frac{1}{n_{obs}} - \frac{1}{n}\right) s^2_{obs} = \left(\frac{1}{n_{obs}} - \frac{1}{N}\right) s^2_{obs} \tag{37}$$

lo cual es fácilmente aceptable porque es la varianza de extraer una muestra de tamaño de las N unidades de población directamente. Para métodos de imputación más avanzados y diseños de muestreo más complicados que en este ejemplo simple, la derivación de una fórmula de varianza puede volverse difícil y las fórmulas resultantes pueden ser bastante complejas. El caso de estimar la media de la población cuando se utiliza la imputación de razón también es un ejemplo en el que se puede derivar una fórmula de varianza simple bajo un mecanismo de falta de respuesta uniforme (J. N. Rao and Sitter (1995)). Este estimador es

$$\hat{V}_A = \left(\frac{1}{n} - \frac{1}{N}\right) s_{obs}^2 + \left(\frac{1}{n_{obs}} - \frac{1}{n}\right) s_e^2, \tag{38}$$

donde,

$$s_e^2 = \sum_{i \in obs} \left(y_i - R x_i\right)^2 / \left(n_{obs} - 1\right),$$

siendo x_i la variable auxiliar con valores observados para todas las unidades, y R el cociente de las medias de y y x para las unidades con y observaciones. Este estimador de varianza difiere del de imputación de medias solo en el segundo componente porque V_{sam} estima la varianza si no hubiera respuesta y, por lo tanto, no se ve afectado por el modelo de imputación. El segundo componente es más pequeño que el de la imputación media porque los residuos del modelo de razón son más pequeños que los del "modelo medio" Esto ilustra que la varianza de la imputación disminuye a medida que aumenta la precisión predictiva del modelo de imputación; si los residuos se vuelven cero, los datos reales se reproducen exactamente y la varianza de imputación desaparece.

C.-E. Särndal (1992); Deville and Särndal (1994); J. N. Rao and Sitter (1995), y C.-E. Särndal and Lundstrom (2005) han descrito, entre muchos otros, enfoques asistidos por modelos y de dos fases para diferentes diseños de muestreo y estimadores. Lee and Särndal (2002) proporcionan una buena descripción general y muchas referencias.

3.11.3 El enfoque de remuestreo

Un estimador de varianza jackknife para un parámetro θ para datos completos, suponiendo que el factor de corrección de población finita puede despreciarse, tiene la forma general

$$\hat{V}_j(\hat{\theta}) = \frac{n-1}{n} \sum_{j=1}^{n} \left(\hat{\theta}_{(j)} - \hat{\theta}\right)^2, \tag{39}$$

siendo $\hat{\theta}_{(j)}$ el estimador obtenido después de eliminar el elemento j de la muestra y $\hat{\theta}$ el estimador obtenido de la muestra completa.

Cuando este estimador jackknife se aplica a los datos imputados, J. N. Rao and Shao (1992) han demostrado que la varianza se subestima porque la varianza debida a la falta de respuesta y la imputación no se tiene en cuenta. También demostraron que se puede obtener un estimador de varianza correcto ajustando los valores imputados.

Este procedimiento de ajuste es equivalente a aplicar de nuevo el procedimiento de imputación cuando se elimina a un encuestado de la muestra. Por ejemplo, para la imputación de la media, esto significa que la media se vuelve a calcular cada vez que se elimina un encuestado. En general, esto significa que los parámetros para los modelos de imputación paramétrica varían entre las réplicas de jackknife, lo que captura la varianza de la estimación en estos parámetros. De manera similar, para los métodos basados en donantes, el grupo de donantes variará entre réplicas. Este método jackknife ajustado conduce a un estimador de varianza válido bajo el supuesto de falta de respuesta uniforme. También se demostró que es válido bajo **MAR**.

El método de remuestreo bootstrap para datos imputados ha sido descrito, entre otros, por Shao and Sitter (1996). Un esquema de remuestreo bootstrap extrae una gran cantidad (B, digamos) de muestras con reemplazo de la muestra original, con un tamaño igual al de la muestra original. Para cada una de estas B muestras bootstrap, se estima el parámetro de interés y la varianza del estimador se estima mediante la varianza de las B estimaciones bootstrap resultantes, de la siguiente manera:

$$\hat{V}_B(\hat{\theta}) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b - \bar{\hat{\theta}}\right)^2 \tag{40}$$

con $\hat{\theta}_b$ el estimador para la muestra bootstrap b y $\bar{\hat{\theta}} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b$.

Cuando se ha utilizado la imputación, la estimación del parámetro $\hat{\theta}_b$ se estima en la muestra de arranque imputada. Al igual que con jackknife, es esencial que el proceso de imputación se repita para cada muestra bootstrap por separado.

En la literatura se han tratado muchos artículos que muestran la aplicabilidad de los métodos jackknife y bootstrap a diferentes diseños de muestreo y métodos de imputación. Estos incluyen muestreo estratificado por conglomerados, regresión (estocástica) e imputación de razón, y métodos de imputación de donantes. El factor de corrección de población finita, que adquiere importancia si el muestreo se realiza dentro de estratos pequeños, como suele ser el caso en las estadísticas económicas, también puede incorporarse en estos métodos de remuestreo (ver J. Rao (1996)). Shao (2002) ofrece una descripción general de los métodos de replicación.

3.11.4 El enfoque de imputación múltiple

La imputación múltiple (**MI**) imputa cada valor faltante M veces; y para crear las imputaciones M se utiliza un modelo bayesiano. Un modelo bayesiano trata los parámetros como variables aleatorias y asigna a estos parámetros una distribución previa que refleja el conocimiento sobre los parámetros antes de que se observen los datos (se pueden elegir distribuciones previas que esencialmente no reflejan información previa sobre los parámetros). Cuando la información previa sobre los parámetros se combina con la información de los datos disponibles, se obtiene la distribución posterior de los parámetros que contiene toda la información disponible sobre los parámetros y es la base para la inferencia bayesiana.

Las imputaciones múltiples reflejan la incertidumbre en los parámetros mediante la creación de valores imputados utilizando un procedimiento de dos pasos. Primero, los valores de los parámetros se extraen de su distribución posterior (por ejemplo, β, σ_e^2 en el caso de imputación de regresión, con σ_e^2 la varianza residual). Luego, utilizando estos valores de parámetros, se realiza una imputación estocástica (por ejemplo, la predicción de regresión con un residuo aleatorio agregado). Esto completa un conjunto de datos imputados. Al creando M tiempos de la distribución posterior, se pueden generar M conjuntos de datos imputados, cada uno imputado por un modelo con diferentes valores para los parámetros. Para la mayoría de las inferencias, parece que un valor pequeño para M (a menudo se menciona cinco) es suficiente. La parte difícil de este procedimiento es construir a partir de la distribución posterior. Los algoritmos para realizar este paso se describen, por ejemplo, en Schafer (1997).

Las imputaciones múltiples se utilizan para obtener estimadores puntuales y de varianza para cualquier parámetro θ de interés. Esto se logra estimando primero este parámetro y su varianza, en cada uno de los conjuntos de datos imputados, y luego combinando las estimaciones M de acuerdo con las reglas definidas por Rubin (D. Rubin (1987)). Para el escalar θ estas reglas son las siguientes.

El estimador puntual MI de θ, $\hat{\theta}_{MI}$, es el promedio de los M estimadores obtenidos al aplicar el estimador $\hat{\theta}$ a cada uno de los M conjuntos de datos:

$$\hat{\theta}_{MI} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m. \tag{41}$$

con $\hat{\theta}_m$ el estimador puntual para el conjunto de datos m. El estimador de varianza MI consta de dos componentes: la varianza dentro de la imputación y la varianza entre imputaciones. La varianza dentro de la imputación, $\hat{V}_W$, se obtiene como el promedio de las estimaciones de la varianza de datos completas para cada uno de los M conjuntos de datos imputados:

$$V_W = \frac{1}{M} \sum_{m=1}^{M} \hat{V}_m, \tag{42}$$

con $\hat{V}_m$ la estimación de la varianza para el conjunto de datos m. La varianza entre imputaciones es la varianza de las M estimaciones puntuales de datos completos:

$$V_B = \frac{1}{M-1} \sum_{m=1}^{M} \left(\hat{\theta}_m - \hat{\theta}_{MI}\right)^2. \tag{43}$$

La estimación de la varianza **MI** de $\hat{\theta}_{MI}$ se obtiene como la suma de las varianzas intra y entre imputaciones con un factor de corrección para tener en cuenta el número finito de imputaciones múltiples:

$$\hat{V}_{MI} = V_W + \left(1 + \frac{1}{M}\right) V_B. \tag{44}$$

La varianza debida a la falta de respuesta se refleja en el componente de varianza entre imputaciones. En ausencia de falta de respuesta, todos los $\hat{\theta}_m$ son iguales, $V_B = 0$, y el estimador de varianza MI se reduce al estimador de varianza de datos completo.

Aunque la imputación múltiple se formuló originalmente para modelos paramétricos, se ha extendido a la imputación de donantes aleatorios. Para la imputación aleatoria de donantes, se selecciona un donante entre los encuestados con valores válidos en la(s) variable(s) objetivo. Estos encuestados se denominan grupo de donantes. Rubin y Schenker (1986) han ideado una versión MI de imputación de donantes aleatorios, llamada "Aproximate Bayesian Bootstrap (ABB)". Este procedimiento funciona de la siguiente manera. Primero, se extrae una muestra bootstrap del grupo de donantes, el grupo de donantes bootstrapeados. Luego, para cada no encuestado, se selecciona un donante, al azar y con reemplazo, del grupo de donantes bootstrapeados y se imputan los valores faltantes. Esto completa una imputación del conjunto

de datos. El procedimiento se repite M veces para crear las M imputaciones múltiples. La varianza entre imputaciones surge aquí porque el conjunto de donantes bootstrapeados cambia entre las imputaciones M.

La imputación múltiple se trata en D. Rubin (1987) ; Roderick JA Little and Rubin (2002b); y Schafer (1997). Donald B. Rubin (1996) revisa las muchas aplicaciones del método en los últimos 18+ años.

3.12 Imputación fraccionada.

Concluimos el tema de imputación con una breve discusión de la imputación fraccionaria. Tratamos esta técnica directamente después de nuestra discusión sobre la imputación múltiple porque la imputación fraccionada está estrechamente relacionada con la imputación múltiple, en el sentido de que tanto la imputación múltiple como la fraccionaria son ejemplos de *imputación repetida*. Es decir, tanto en la imputación múltiple como en la imputación fraccionaria, se imputan múltiples valores para un solo valor faltante. A continuación, primero explicamos la técnica de la imputación fraccionada y luego señalamos algunas diferencias entre la imputación múltiple y fraccionada.

La imputación fraccionada se puede aplicar a datos numéricos y generalmente se basa en la imputación hot deck [ver, por ejemplo, Kim y Fuller (2004)]. En la imputación fraccionada, las imputaciones en un registro de receptor i están determinadas por dos factores: los registros de donante k que se utilizan y los pesos de imputación $w_{ik}^* \geq 0$ que se asignan a estos registros de donantes. Los pesos de imputación w_{ik}^* satisfacen

$$\sum_{ik \in R} w_{ik}^* = 1,$$

donde R denota el conjunto de registros con un valor observado para la variable a imputar. Una ponderación w_{ik}^* expresa la fracción del valor del registro del donante k que se utiliza para el valor faltante del registro del receptor i. Para un registro de destinatario i con un valor faltante en la variable a imputar, el valor faltante se imputa por la media ponderada de los valores del donante para ese registro

$$y_i^* = \sum_{k \in R} w_{ik}^* y_k.$$

Dados los valores imputados para todos los registros de los destinatarios, un estimador lineal (ponderado) $\hat{\theta}$ se puede escribir como

$$\hat{\theta} = \sum_{i \in U} w_i y_i^* = \sum_{i \in U} w_i \left(\sum_{k \in R} w_{ik}^* y_k \right) = \sum_{k \in R} W_k^* y_k,$$

donde U denota todo el conjunto de datos, w_i las ponderaciones de diseño (por ejemplo, determinadas por el diseño de muestreo y las ponderaciones de corrección para la falta de respuesta de la unidad), y $W_k^* \equiv \sum_{i \in U} w_i w_{ik}^*$.

Una diferencia importante entre la imputación fraccionaria y la imputación múltiple es que la imputación fraccionaria se define en un marco frecuentista, mientras que la imputación múltiple se define en un marco bayesiano. De hecho, la imputación fraccionaria puede verse como una forma de imputación múltiple impropia. Otra diferencia importante es que el objetivo principal de la imputación fraccionada es mejorar la eficiencia del estimador puntual imputado al reducir la varianza que surge del proceso de imputación, mientras que el objetivo principal de la imputación múltiple es simplificar la estimación de la varianza de un estimador puntual. Una tercera diferencia es que la imputación fraccionaria se basa en la imputación hot deck, al menos en su forma habitual y tradicional (ver, por ejemplo, Qin, Rao, and Ren 2008 ; Kim J. K. and Fuller 2008 para extensiones recientes de la imputación de regresión) , mientras que la imputación múltiple se basa en extraer valores de la distribución predictiva posterior del valor faltante en un registro dados los valores observados.

Para obtener más información sobre la imputación fraccionaria, consulte Graham Kalton and Kish (1984); Fay (1996); J. K. Kim and Fuller (2004); Durrant et al. (2005) y Durrant and Skinner (2006).

4 Imputación Multivariante (JB)

4.1 Introducción

La imputación de valores perdidos que se trató previamente se enfocó solo sobre una base de variable por variable. En este enfoque, una imputación es la media condicional predicha de la variable objetivo, dados (algunos de los) valores observados, o una extracción de la distribución condicional de la variable objetivo. Cuando hay múltiples variables con valores faltantes para una unidad, este enfoque de imputación de variable única se puede aplicar a cada una de estas variables por separado. A menudo, sin embargo, una mejor alternativa en tales casos es modelar la distribución simultánea de las variables que faltan e imputar con la media multivariada condicional de las variables que faltan o un sorteo de esta distribución. Este enfoque simultáneo o multivariante es el tema de este apartado.

En muchas aplicaciones en las que hay una serie de variables con valores faltantes que requieren imputación, algunas variables con valores faltantes se utilizan como predictores en los modelos de imputación para otras. En tales situaciones, los métodos de imputación de una sola variable tienen algunos inconvenientes, como se discutirá más adelante, que pueden paliarse mediante el uso de métodos de imputación de múltiples variables.

La estructura de un conjunto de datos de cuatro variables con valores faltantes se ilustra en la tabla 1. Las columnas de esta tabla representan las cuatro variables, indicadas por z, x_1, x_2 y x_3. Las filas representan observaciones con valores faltantes en las mismas variables, y los valores faltantes se indican con M y los valores observados con O. El patrón de datos faltantes y observados en cada una de estas filas se denomina patrón de datos faltantes. En este ejemplo, se observa una variable, z, para todas las unidades. Las variables para las que es seguro que siempre se observan son variables obtenidas del marco muestral y otras variables de datos administrativos que pueden vincularse al marco muestral. Las otras tres variables en la tabla 1 tienen valores faltantes para algunas unidades, y los ocho patrones diferentes de datos faltantes son posibles configuraciones de datos faltantes que existen en esta situación.

Tabla 3: Patrones de datos perdidos.

z	x_1	x_2	x_3	Patrón
O	M	M	M	A
O	M	M	O	B
O	M	O	M	C
O	O	M	M	D
O	M	O	O	E
O	O	M	O	F
O	O	O	M	E
O	O	O	O	H

Un método de imputación de una sola variable, como los modelos de imputación de regresión o regresión logística discutidos previamente, utiliza un modelo para la variable faltante en una unidad con las variables observadas en esa unidad como predictores. Por ejemplo, para imputar x_1 en unidades con el patrón de datos A, un modelo con solo z como predictor puede imputar x_1. En el patrón de datos B, x_1 se puede imputar usando z y x_3 como predictores; y en el patrón de datos E, x_1 se puede imputar usando las otras tres variables como predictores. La información para estimar estos modelos se vuelve más limitada a medida que aumenta el número de predictores con valores perdidos. Por ejemplo, el modelo con solo z como predictor se puede estimar usando todas las unidades con x_1 observadas (patrones D, F, G y H), pero el modelo con z, x_2 y x_3 como predictores solo se puede estimar usando el unidades con patrón de datos faltantes H. Esto puede conducir fácilmente a una falta de observaciones para estimar el modelo de manera confiable, en cuyo caso se debe elegir un modelo más simple que excluya algunos de los predictores disponibles. Los métodos de imputación multivariable discutidos en este documento también usan modelos que predicen las perdidas dados los valores observados en una unidad, pero para estimar estos modelos, estos métodos usan efectivamente los datos observados disponibles para todas las unidades en lugar de solo los datos con todos los predictores y la variable observada objetivo.

Una segunda ventaja importante de los métodos de imputación multivariable es que estos métodos pueden reproducir correlaciones entre variables mucho mejor que los métodos de imputación de una sola variable. Los métodos de imputación de una sola variable pueden, hasta cierto punto, preservar la relación entre la variable objetivo y las variables predictoras ya que esa relación es parte del modelo, pero la relación entre una variable objetivo y otra variable que falta y, por lo tanto, no está incluida en el el modelo no se conservará. Los métodos de imputación multivariada generan imputaciones a partir de una distribución simultánea estimada de las variables con valores faltantes que reflejarán las correlaciones entre estas variables y, por lo tanto, tienen por objeto producir imputaciones que preserven estas correlaciones.

Una última ventaja que debe mencionarse es que, a diferencia de los métodos de imputación de una sola variable, los métodos de imputación multivariante adecuados pueden tener en cuenta automáticamente algunas restricciones de edición. Por ejemplo, los métodos multivariantes para datos continuos basados en la distribución normal multivariante respetan las restricciones de igualdad lineal, como las ediciones de balance, siempre que los datos observados se ajusten

a estas ediciones. Este caso particular y una serie de otros modelos multivariantes que tienen en cuenta las restricciones de edición se tratarán más adelante, especialmente cuando hablemos de la imputación bajo restricciones de edición. Para datos categóricos, las ediciones que prescriben que ciertas combinaciones de valores no pueden ocurrir, se cumplirán si las variables involucradas se imputan simultáneamente y la interacción entre estas variables se tiene en cuenta en el modelo. Por ejemplo, si las personas con categoría de edad menor de 14 años no pueden casarse, la distribución de probabilidad estimada sobre las combinaciones de categorías de edad y estado civil asignará probabilidad cero a la combinación de valores (edad = *menos de 14 años*; estado civil = *casado*) siempre que no existan datos observados con esta combinación de valores. En consecuencia, ningún valor imputado caerá en esta combinación de categorías.

La imputación multivariante no se limita a la imputación basada en modelos explícitos: los enfoques hot-deck se extienden de manera casi trivial a situaciones multivariantes. Con múltiples valores faltantes en la unidad que se va a imputar (el receptor), un método hot-deck multivariado simplemente imputa todos los valores faltantes del mismo donante. Esto puede verse como un sorteo de la distribución multivariada empírica de los valores perdidos. Las reglas de edición que involucran solo variables imputadas se cumplirán, asumiendo que el donante cumple con las reglas de edición, pero las reglas de edición que contienen tanto variables imputadas como observadas pueden no cumplirse porque los valores de estas variables provienen de unidades diferentes.

Esta sección proporciona una introducción a la imputación basada en modelos multivariados para datos continuos y categóricos y algunos antecedentes sobre la estimación de dichos modelos en presencia de datos faltantes. Otras explicaciones más amplias de esta metodología se presentan, por ejemplo, en Roderick JA Little and Rubin (2002b) y Schafer (1997). Una alternativa para encontrar un modelo multivariado apropiado es modelar la distribución de cada variable por separado, condicionada a las otras variables (apropiadamente imputadas). Este llamado enfoque de imputación secuencial tiene ventajas para datos con variables de diferentes tipos (categóricas, continuas, semicontinuas) y datos sujetos a restricciones.

El resto de esta sección comienza analizando los dos modelos multivariados que se aplican con mayor frecuencia con fines de imputación, el modelo normal multivariado para datos continuos y el modelo multinomial para datos categóricos. Ambos modelos y las imputaciones derivadas de ellos dependen de parámetros desconocidos que deben estimarse a partir de los datos incompletos. El enfoque de máxima verosimilitud para este problema de estimación es el tema de la sección @ref(secEMV_datcompletos), que incluye una discusión del famoso algoritmo de maximización de expectativas (algoritmo EM) que fue presentado en su forma general por Arthur P. Dempster, Laird, and Rubin (1977b) y es ampliamente aplicado a problemas de estimación con datos incompletos. En la sección final de este apartado se analiza la aplicación de un enfoque de imputación multivariable a los datos de una encuesta entre las bibliotecas públicas.

4.2 Modelos de imputación multivariante

4.2.1 Imputación de datos multivariantes continuos

La imputación de regresión, como se describe en la sección 3.3, reemplaza un valor faltante de la unidad i en la variable objetivo x_t con una predicción de ese valor basada en el modelo.

$$x_{i,t} = \mathbf{x}_{i,p}^T \boldsymbol{\beta} + \varepsilon_{i,t}, \tag{45}$$

siendo $x_{i,p}$ el vector con los valores de las variables predictoras para la unidad i, β el vector con los parámetros de regresión y $\varepsilon_{i,t}$ un residuo aleatorio con esperanza cero. Aquí usamos una notación que es ligeramente diferente de la notación usada en la sección 3.1 (donde se usó y para representar la variable objetivo) pero más conveniente para la generalización al caso multivariado.

Al restar las medias de ambos lados de (45), este modelo también se puede expresar como

$$x_{i,t} - \mu_t = \left(\mathbf{x}_{i,p} - \boldsymbol{\mu}_p\right)^T \boldsymbol{\beta} + \varepsilon_{i,t} \tag{46}$$

con μ_t la media de $x_{i,t}$ y μ_p el vector medio de las variables predictoras. Para los datos completos, el estimador de mínimos cuadrados de β ahora se puede escribir como

$$\beta = \left[\sum_i \left(\mathbf{x}_{i,p} - \boldsymbol{\mu}_p\right)\left(\mathbf{x}_{i,p} - \boldsymbol{\mu_p}\right)^T\right]^{-1} \sum_i \left(\mathbf{x}_{i,p} - \mu_p\right)\left(x_{i,t} - \mu_t\right) = \sum\nolimits_{p,p}^{-1} \sum_{p,t}, \tag{47}$$

con $\sum_{p,p}$ la matriz de covarianza muestral de las variables predictoras y $\sum_{p,t}$ el vector con las covarianzas muestrales entre las variables predictoras y la variable objetivo.

El modelo (46) se puede extender al caso donde hay un vector de variables objetivo y luego se escribe como

$$\mathbf{x}_{i,t} = \boldsymbol{\mu}_t + \boldsymbol{B}_{t,p}\left(\mathbf{x}_{i,p} - \boldsymbol{\mu}_p\right) + \boldsymbol{\varepsilon}_{i,t}, \tag{48}$$

siendo $\mathbf{x}_{i,t}$ el vector de valores de las variables objetivo para la unidad i, $\boldsymbol{\mu}_t$ el vector medio de las variables objetivo, $\boldsymbol{B}_{t,p}$ la matriz con coeficientes de regresión con su número de filas igual al número de variables objetivo y su número de columnas igual a el número de variables predictoras, y $\boldsymbol{\varepsilon}_{i,t}$, vector de perturbaciones aleatorias con esperanza cero. Esta notación ignora el hecho de que la matriz de coeficientes depende de i en el sentido de que las variables predictoras y las variables a predecir dependen del patrón de datos faltantes; pero en el caso que esto pueda llevar a confusión, haremos explícita la dependencia del patrón de datos perdidos. La presentación relativamente compacta de este modelo es posible porque aquí se supone que para todas las variables objetivo se utilizan los mismos predictores. Los modelos en los que este no sea el caso, se denominan modelos de regresión aparentemente no relacionada (Seemingly Unrelated Regression,SUR) y tienen una estructura más complicada (véase, por ejemplo, T. Amemiya 1985).

El estimador de mínimos cuadrados para $\boldsymbol{B}_{t,p}$ se construye a partir de los mismos componentes que en el caso univariado (47):

$$\hat{\boldsymbol{B}}_{t,p} = \sum\nolimits_{t,p} \sum\nolimits_{p,p}^{-1} \tag{49}$$

pero, en este caso, $\sum_{t,p}$ es una matriz con las covarianzas de las variables predictoras con las variables en objetivo de un vector.

Para usar el modelo (48)) con fines de imputación, necesitamos las estadísticas $\boldsymbol{\mu}_t$, $\sum_{p,p}$ y $\sum_{t,p}$, o estimaciones de las mismas. Como se señaló antes, las variables en el vector medio y las matrices de covarianza variarán entre los patrones de datos faltantes, pero en cualquier caso, estos parámetros se pueden obtener como particiones del vector medio $\boldsymbol{\mu}$, por ejemplo y la matriz de covarianza, por ejemplo, de todas las variables. En presencia de datos faltantes, podríamos basar las estimaciones de estos parámetros en los registros completamente observados, pero el número de unidades sin ningún valor faltante puede ser bastante pequeño, especialmente en casos con muchas variables. Especialmente en estos casos que la cantidad de parámetros en la matriz de covarianza es grande y la necesidad de una gran cantidad de observaciones para estimar esta matriz es mayor. Por el momento asumimos que de alguna manera hemos obtenido estimaciones satisfactorias de estos parámetros. Usando estas estimaciones, las imputaciones de regresión para las variables faltantes en un registro i sin un residuo se pueden obtener por la media predicha

$$\hat{\mathbf{x}}_{i,t} = \hat{\boldsymbol{\mu}}_{\boldsymbol{t}} + \hat{\boldsymbol{B}}_{t,p} \left(\hat{\mathbf{x}}_{i,p} - \hat{\boldsymbol{\mu}}_{\boldsymbol{p}} \right), \tag{50}$$

y se puede obtener una imputación con un vector residual aleatorio obtenido por

$$\hat{\mathbf{x}}_{i,t} = \hat{\boldsymbol{\mu}}_{\boldsymbol{t}} + \hat{\boldsymbol{B}}_{t,p} \left(\hat{\mathbf{x}}_{i,p} - \hat{\boldsymbol{\mu}}_{\boldsymbol{p}} \right) + \mathbf{e}_{i,t} \tag{51}$$

donde $\mathbf{e}_{i,t}$ es un valor aleatorio de la distribución de los residuos de la regresión de $\mathbf{x}_t$ sobre $\mathbf{x}_p$. Si se asume que la distribución conjunta de todas las variables es normal multivariante, entonces la distribución condicional de $\mathbf{x}_t$ dado $\mathbf{x}_p$ también es normal multivariante con vector medio dado por (8.6) y matriz de covarianza $\sum_{t,t.p'}$, por ejemplo. Los residuos entonces se distribuyen de forma normal multivariante con vector de media cero y matriz de covarianza igual a la matriz de covarianza de la distribución condicional. Por las propiedades de la distribución normal multivariante, una estimación de esta matriz de covarianza condicional o residual se puede obtener a partir de una estimación de la matriz de covarianza $\hat{\sum}$ de todas las variables según

$$\hat{\sum_{t,t.p'}} = \hat{\sum_{t,t}} - \hat{\sum_{t,p}}^{-1} \hat{\sum_{p,t}}, \tag{52}$$

donde las matrices en el lado derecho se pueden extraer todas desde $\hat{\sum}$. Añadir tales residuos debe ayudar no solo a preservar las varianzas de las variables imputadas sino también a preservar las correlaciones entre las variables imputadas. La estimación de $\hat{\sum}$ en presencia de datos faltantes se discutirá en la Sección 8.3.2.

4.2.2 Imputación de datos categóricos multivariados (AA)

Imputación multivariante de datos categóricos. Una variable binaria tiene solo dos valores posibles y estos valores se pueden codificar, arbitrariamente, como 1 y 0. Las observaciones sobre una variable multicategoría x con K categorías se pueden codificar con K variables dummys x_k, cada una de las cuales toma los valores 0 o 1 y $\sum_k (x_k) = 1$. La probabilidad de una observación en la categoría k de x es la probabilidad de que $x_k = 1$ y se denotará por π_k. Cuando hay múltiples variables categóricas, los resultados se pueden

representar mediante conjuntos de variables ficticias que representan las categorías de cada una de ellas. Por ejemplo, si tenemos tres variables $x^{(1)}$, $x^{(2)}$, y $x^{(3)}$, con categorías indicadas por $k = 1, \ldots, K$, $l = 1, \ldots, L$ y $m = 1, \ldots, M$ respectivamente, se crearán variables ficticias K, L y M. Una observación en la categoría k de $x^{(1)}$, l de $x^{(2)}$ y m de $x^{(3)}$ corresponderá entonces al evento $x^{(1)} = x^{(2)} = x^{(3)} = 1$. La probabilidad de este evento será denotada por π_{klm}. Las probabilidades π_{klm} para todos los k, l, m juntos representan la distribución de probabilidad sobre todos los posibles resultados conjuntos de las tres variables.

Las observaciones sobre múltiples variables categóricas también se pueden representar en una clasificación cruzada multidimensional. La notación para el caso completamente general es algo engorrosa y, por lo tanto, aquí usamos un ejemplo tridimensional para transmitir la idea. La generalización a más dimensiones (variables) será sencilla. Para tres variables con números de categorías K, L y M, los posibles resultados se pueden organizar en una clasificación cruzada tridimensional con K filas, L columnas y M capas. Cada una de las celdas $K \times L \times M$ en esta matriz tridimensional corresponde a uno de los posibles resultados conjuntos de las tres variables, y cada unidad se puede clasificar en una y solo una de las celdas. Si denotamos la celda correspondiente a $x^{(1)} = k$, $x^{(2)} = l$, y $x^{(3)} = m$ por klm, la probabilidad de que una unidad caiga en una celda klm es la probabilidad de celda π_{klm}. Las probabilidades bivariadas marginales correspondientes a las combinaciones de categorías de $x^{(1)}$ y $x^{(2)}$ se pueden obtener como $\pi_{kl} = \sum_m \pi_{klm}$ para todo k y l, y las otras probabilidades marginales bivariadas se pueden obtener de manera similar obtenido por sumatoria sobre las categorías de la tercera variable. Las probabilidades marginales univariadas se denotan por π_k, π_l y π_m y pueden expresarse como sumas de dos variables de las probabilidades trivariadas o sumas de una variable de las probabilidades bivariadas.

Para una muestra de n unidades, se puede contar el número de unidades que caen en cada celda de la clasificación cruzada. La clasificación cruzada con recuentos en las celdas se denomina tabla de contingencia. Cuando las probabilidades de las celdas son iguales para cada unidad y los puntajes de las variables son independientes entre las unidades, los conteos en la clasificación cruzada se distribuyen multinomialmente. Para el ejemplo de tres variables, la distribución multinomial da la probabilidad de observar el vector realizado de recuentos de celdas $n = (n_{111}, n_{112}, ..., n_{KLM})$. Para las tablas generales de varios factores, es más conveniente representar los recuentos de celdas (y las probabilidades) mediante un solo índice como $n = (n_1, ..., n_j, ..., n_C)$, siendo C el número de celdas. En esta notación, la distribución multinomial se puede escribir como

$$P(n|\pi) = \frac{n!}{\prod_{j=1}^{C} n_j!} \prod_{j=1}^{C} \pi_j^{n_j}, \tag{53}$$

con $\pi = (\pi_1, ..., \pi_C)$ el vector con probabilidades de las celdas. Una propiedad importante de la distribución multinomial, que utilizan los métodos de imputación basados en esta distribución, es que la distribución condicional sobre un subconjunto de celdas también es una distribución multinomial. Supongamos que se da que n_S unidades caen en un subconjunto S de las celdas y que π_S es la parte del vector π correspondiente a estas celdas, entonces la distribución de los conteos en las celdas C_S pertenecientes a S es multinomial con parámetro $\pi_S^* = \pi_S / \sum_{j \in S} \pi_j$, con el subíndice s denotando la parte del vector de parámetros correspondiente a las celdas en S.

Representaciones de datos faltantes en variables categóricas. Una unidad para la que faltan algunas de las variables no puede clasificarse en una de las combinaciones de categorías o celdas definidas por todas las variables consideradas, pero puede clasificarse parcialmente según las variables observadas. Por ejemplo, si falta la tercera variable en el ejemplo de tres variables, la unidad se puede clasificar en la tabla marginal bivariante $K \times L$, pero no en la tabla tridimensional completa. Más generalmente, todas las unidades con las mismas variables observadas pueden clasificarse en una tabla de contingencia correspondiente a esas variables. Tales tablas son diferentes para cada patrón de datos faltantes y son todas tablas marginales de la tabla de contingencia de datos completa. Por ejemplo, para el caso de tres variables, las unidades a las que les falta una variable pueden clasificarse en una de las tablas marginales bivariadas indicadas como $x^{(1)}$ x $x^{(2)}$, $x^{(1)}$ x $x^{(3)}$, y $x^{(2)}$ x $x^{(3)}$, y las unidades a las que les faltan dos variables pueden clasificarse en una de las tablas marginales univariadas correspondientes a $x^{(1)}$, $x^{(2)}$, y $x^{(3)}$. Los datos que se resumen en este conjunto de tablas de contingencia se denominan tabla de contingencia con márgenes suplementarios Roderick JA Little and Rubin (2002c).

Para mejor ilustrar exhimos un ejemplo simple con los datos de la Tabla 8.2. Los datos son ficticios pero similares a los datos de una encuesta sobre el comportamiento del movimiento. El ejemplo se refiere a datos sobre dos variables: Posesión de automóvil (Propietario de automóvil) y Posesión de licencia de conducir (Licencia), obtenidos de 90 personas mayores de 18 años. Para 60 de estas 90 personas se obtienen respuestas sobre ambas variables, para 20 personas solo se obtiene la respuesta sobre Dueño de Auto y 10 personas solo responden sobre la variable Licencia. Para este ejemplo, se supone que se conocen las probabilidades de las celdas bivariadas dadas en el último panel de la Tabla 8.2. En general, por supuesto que este no es el caso, pero se pueden obtener estimaciones de estas probabilidades a partir de los datos incompletos. Este problema de estimación se analiza en la Sección 8.3.

Imputación para Variables Multinomiales. Supongamos que queremos imputar el valor faltante de Propietario de automóvil para los 5 encuestados con puntaje Sí en Licencia. Para este problema consideramos la distribución condicional de la variable faltante, dada la observada. A partir de las probabilidades de celda, vemos que la distribución condicional sobre las categorías de Propietario de automóvil, dada Licencia = Sí, es $\{0.588/0.743, 0.155/0.743\} = \{0.79, 0.21\}$. Más formalmente, si indexamos las filas de la tabla con k y las columnas de la tabla con l, esta distribución condicional se puede escribir como $\pi_{k|l=1} = \pi_{kl} / \sum_k \pi_{kl}$. Usando la notación de índice único para las cuatro celdas $(j = 1, ..., 4)$ correspondientes a $k, l = (1, 1)(1, 2)(2, 1)(2, 2)$, la distribución condicional para las unidades con Licencia = Sí es la distribución sobre las dos celdas correspondientes a $j = 1$ y $j = 3$. Los valores esperados para las variables ficticias correspondientes a las cuatro celdas, dado que Licencia = Sí, son iguales a las probabilidades condicionales 0.79 y 0.21 para x_1 y x_3, respectivamente, y cero para las demás celdas.

Estos valores esperados condicionales pueden usarse como imputaciones, al igual que los valores esperados condicionales se usan como imputaciones para variables continuas. Pero contrariamente a las expectativas condicionales para variables continuas, las expectativas condicionales para variables ficticias son obviamente diferentes de las observaciones reales; no son iguales a cero o uno y hay múltiples dummies con valores distintos de cero. Sin embargo, para el propósito de estimar los totales de las celdas en una tabla de contingencia, esto no tiene por qué ser un problema. Una alternativa es, al igual que con las variables continuas, extraer un valor imputado de la distribución condicional. En este caso, esto implica extraer

de la distribución multinomial sobre las dos celdas con probabilidad condicional distinta de cero y con $n = 1$. Esto dará como resultado una cuenta de 1 en una de estas dos celdas, junto con un valor de 1 para la correspondiente variable ficticia.

En el caso general, consideramos, para cada celda en un margen suplementario, la distribución condicional sobre las celdas que contribuyen al total en esa celda marginal. La imputación de la información faltante se puede realizar entonces estableciendo las dummies correspondientes a las celdas contribuyentes iguales a sus valores esperados o extrayendo de la distribución multinomial condicional sobre estas celdas.

Los métodos de imputación descritos anteriormente darán como resultado un conjunto de datos completo a nivel de unidad. En este sentido, estos enfoques son similares a los enfoques para datos continuos. Para datos categóricos, sin embargo, existe la alternativa de imputar a nivel agregado. En el enfoque agregado, no se imputan valores para las variables ficticias, pero se crea una tabla de contingencia de dimensión completa en la que se clasifican todas las unidades. Esto se logra distribuyendo el total de cada celda en cada margen suplementario entre las celdas de la tabla completa que contribuyen a esa celda. En el ejemplo anterior, las cinco unidades con puntaje Sí en la licencia se pueden distribuir en las celdas con $j = 1$ y $j = 3$ usando un sorteo de una distribución multinomial con $n = 5$ y las mismas probabilidades que antes.

4.3 La imputación según el método E-M (AA)

4.3.1 Introducción

Para este CDC, resulta fundamental lograr la comprensión y aplicación de las metodologías que dicen relación con la Imputación Múltiple, en estricta atención a que las bases de datos contienen muchas variables de interés y es natural que que entre ellas existan relaciones de interes y que podamos usarlas para resolver la problemática de datos faltantes en toda la base que sea motivo de análisis, en aras de que el INE pueda proporcionar información estadística de alta calidad sobre muchos aspectos de la sociedad, lo más actualizada y precisa posible. Ya sabemos que el principal óbice para nuestro cometido surge del hecho de que las fuentes de datos que se utilizan para la producción de resultados estadísticos, tanto las encuestas tradicionales como los datos administrativos, contienen datos faltantes, por lo tanto esta ausencia debe intentar ser corregida, si existen las condiciones para tal cometido, aunque siempre debemos tener en cuenta que este es un paso "ad extremis", siendo los más atentos en recomendar "no imputar", para evitar distorsiones en las estimaciones de las cifras de publicación.

4.3.1.1 Estimación de Máxima Verosimilitud en presencia de datos perdidos

Desde (De Waal, Pannekoek, and Scholtus (2011) pp. 285 – 293) Handbook in Survey Methodology

El algoritmo E-M es un procedimiento computacional iterativo general, diseñado para obtener estimaciones de máxima verosimilitud de parámetros a partir de datos incompletos. En esta sección se hace una breve introducción de este algoritmo y a su aplicación en dos enfoques de imputación estándar, tanto para variables continuas y para variables discretas. Dado que el algoritmo está diseñado principalmente como una forma de maximizar la función de probabilidad en presencia de datos faltantes, primero describimos algunos conceptos básicos de la teoría de la probabilidad. Se puede encontrar más información sobre la teoría de

la probabilidad incluidos, en muchos textos que son estándar sobre estadística matemática, como Cox and Hinkley (1979), Wilks (1964) y Stuart and Ord (1991).

4.3.1.2 Estimación de Máxima Verosimilitud con datos completos La estimación de máxima verosimilitud supone que se ha especificado un modelo estadístico que describe el proceso estocástico que generá los datos. En el caso multivariado, los datos consisten en n observaciones x_i $(i = 1, \ldots, n)$ en J variables cada una. Y, como posible modelo estadístico para dichos datos, se puede suponer que estos datos pueden describirse como n realizaciones independientes de un vector aleatorio x que tiene una distribución normal multivariada con un vector de medias μ y una matriz de covarianzas Σ. En general, un modelo estadístico especifica la distribución o función de densidad de probabilidad conjunta de las n observaciones, hasta un vector desconocido de parámetros θ. Si se supone que las n observaciones son idénticamente distribuidas de forma independiente (iid) cuya función de densidad es $f(x|\theta)$, entonces la densidad conjunta viene dada por

$$f\left(X|\theta\right) = \prod_{i=1}^{n} f\left(x_i|\theta\right), \tag{54}$$

La función de verosimilitud es similar a $f(X|\theta)$ pero con el papel de X y θ invertidos; mientras que $f(X|\theta)$ describe la densidad de probabilidad de valores variables de X o un valor dado de θ, la función de probabilidad describe la densidad de probabilidad de un valor de X, el realmente observado, para valores variables de θ. Para expresar esto, la función de verosimilitud se define como $l(\theta|X) = f(X|\theta)$. El estimador de máxima verosimilitud $\widehat{\theta}$ de θ se define como el valor de θ que maximiza la función de verosimilitud sobre el espacio de parámetros Ω de θ. Maximizar la función de verosimilitud es equivalente a maximizar el logaritmo de la función de verosimilitud, que a menudo es más conveniente, por lo que podemos escribir

$$\widehat{\theta} = arg \max_{\theta \in \Omega} L\left(\theta|X\right) = arg \max_{\theta \in \Omega} \sum_{i=1}^{n} L\left(\theta|x_i\right) \ , \tag{55}$$

Donde

$$L\left(\theta|x_i\right) = ln\left[l(\theta|x_i)\right]$$

Generalmente, esta maximización se lleva a cabo igualando a cero las derivadas de primer orden con respecto a los parámetros. Las ecuaciones resultantes son denominadas las ecuaciones de verosimilitud. Si dichas ecuaciones no pueden resolverse en forma cerrada, se utiliza un procedimiento iterativo para este propósito. Por lo general, los algoritmos de búsqueda de raíces o ceros, según Newton-Raphson o Fisher se aplican para maximizar las funciones logarítmicas de verosimilitud.

4.3.1.3 Estimación de Máxima Verosimilitud con datos incompletos Cuando el conjunto de datos X está¡ incompleto, los valores faltantes pueden especificarse mediante una matriz indicadora de valores faltantes M con las mismas dimensiones que X y con elementos $m_{ij} = 1$

si x_{ij} se observa y $m_{ij} = 0$ si falta x_{ij}. El valor de una fila de M, digamos m_i, identifica el patrón de valores observados y faltantes para todas las variables y se denomina patrón de respuesta.

Una observación x_i puede, después de una permutación de sus elementos, dividirse como $x_i = \left(x_{i,obs}^T,\ x_{i,mis}^T\right)^T$ con $x_{i,obs}$ el vector con valores observados y $x_{i,mis}$ el vector con valores faltantes.

En presencia de datos faltantes, los datos que se observan son $x_{i,obs}$ y m_i. El proceso que generá los datos faltantes se denomina mecanismo de datos faltantes. En general, se supone que los m_i observados son realizaciones de una variable aleatoria con una distribución denominada distribución del mecanismo de datos faltantes (ver Donald B. Rubin (1976a) y Roderick JA Little and Rubin (2002c)).

Para continuar con la inferencia de probabilidad, ahora tenemos que especificar un modelo estadístico que especifique la distribución conjunta de los datos observados hasta algunos parámetros desconocidos. Para construir dicho modelo, primero escribimos la densidad conjunta de x_i y m_i y factorizamos esta densidad usando reglas estándar para probabilidades condicionales de la siguiente manera:

$$f\left(x_{i,obs}, x_{i,mis}, m_i|\theta, \phi\right) = f\left(x_{i,obs}, x_{i,mis}, m_i|\theta\right) f\left(m_i|x_{i,obs}, x_{i,mis}, \phi\right), \tag{56}$$

El primer factor del lado derecho especifica la densidad de x_i en ausencia de datos faltantes, y el segundo factor especifica la distribución del indicador de datos faltantes. La distribución de los datos faltantes depende de un vector de parámetros desconocido ϕ y, en general, también de los valores x observados y faltantes. La densidad de las observaciones $(x_{i,obs}, m_i)$ ahora se puede obtener integrando (8.12) sobre la distribución de los valores faltantes, llegando así a

$$f\left(x_{i,obs}, m_i|\theta, \phi\right) = \int f\left(x_{i,obs}, x_{i,mis}, m_i|\theta\right) f\left(m_i|x_{i,obs}, x_{i,mis}, \phi\right) dx_{i,mis}, \tag{57}$$

Esta expresión aún depende de los valores perdidos desconocidos, y la inferencia de probabilidad no puede proceder sin suposiciones adicionales. Por lo general, se hacen suposiciones sobre la dependencia del mecanismo de datos faltantes en los valores de x. Con respecto a esta dependencia, se distinguen los siguientes tres casos (ver también la Sección 1.3):

1. El mecanismo de datos faltantes no depende en absoluto de los valores de x; este mecanismo se llama PÉRDIDA COMPLETAMENTE ALEATORIA Missing Completely at Random (MCAR).
2. El mecanismo de datos faltantes depende de los valores x observados pero no de los valores x faltantes; esto se llama PÉRDIDA ALEATORIA Missing at Random (MAR).
3. El mecanismo de datos faltantes depende tanto de los valores de x faltantes como de los observados; esto se llama PÉRDIDA NO ALEATORIA Not Missing at Random (NMAR).

Claramente, el supuesto 1 es más fuerte que el supesto 2, que a su vez es más fuerte que el supuesto 3. De hecho Roderick JA Little and Rubin (2002c) y Donald B. Rubin (1976a), lo destacan y fueron quienes también acuñaron la terminología MCAR, MAR, NMAR. El

supuesto más importante y ampliamente utilizado es MAR, ya que da como resultado métodos prácticos para la inferencia sin tener que suponer demasiado. El supuesto MAR nos permite reescribir (8.13) como

$$\begin{aligned} f\left(x_{i,obs}, m_i|\theta,\phi\right) &= \int f\left(x_{i,obs}, x_{i,mis}|\theta\right) f\left(m_i|x_{i,obs},\phi\right) dx_{i,mis} \\ &= f\left(m_i|x_{i,obs},\phi\right) \int f\left(x_{i,obs}, x_{i,mis}|\theta\right) dx_{i,mis}, \\ &= f\left(m_i|x_{i,obs},\phi\right) f\left(x_{i,obs}|\theta\right) \end{aligned} \tag{58}$$

La distribución conjunta de las observaciones $(x_{i,obs}, m_i)$ ahora es el producto de dos factores, el primero dependiendo de ϕ y el segundo dependiendo de θ. Si estos parámetros son distintos en el sentido de que no están funcionalmente relacionados, se dice que el mecanismo de datos faltantes es ignorable, y Donald B. Rubin (1976a) muestra que la inferencia sobre θ puede basarse únicamente en el segundo factor, es decir, a través de los datos observados de máxima verosimilitud o logaritmo de máxima verosimilitud.

$$L\left(\theta|X_{obs}\right) = \sum_i \ln\left\{f\left(x_{i,obs}|\theta\right)\right\}$$

La importancia de este resultado es que, sin necesidad de especificar el mecanismo de datos faltantes, la probabilidad de los datos observados puede proporcionar una inferencia válida sobre θ no solo si el mecanismo faltante es puramente aleatorio (MCAR), sino también si este mecanismo depende de la x siempre que esta dependencia se limite a los valores x observados (MAR).

4.3.1.4 El algoritmo EM Incluso después de la simplificación causada por asumir MAR y a pesar de su apariencia concisa, la probabilidad de los datos observados $L\left(\theta|X_{obs}\right)$ sigue siendo una función complicada, en general. Esto se debe al hecho de que $x_{i,obs}$ son vectores de longitud variable que contienen observaciones sobre diferentes conjuntos de variables observadas x_j, lo que hace que la verosimilitud logarítmica de los datos observados sea mucho más compleja que la verosimilitud logarítmica de los datos completos correspondientes. En consecuencia, las derivadas de la verosimilitud o logverosimilitud necesarias para maximizar esta función también son funciones complicadas. El algoritmo Expectation-Maximization (EM) es un algoritmo iterativo que elude estas dificultades al completar los valores esperados para las funciones de los valores faltantes que aparecen en la probabilidad en un paso (el paso de expectativa o paso E) y maximizando el en este verosimilitud completada en el otro paso (el paso de maximización o paso M) e iterando estos dos pasos.

Para describir el algoritmo EM, primero expresamos la densidad del vector de datos como

$$\begin{aligned} f\left(x_i|\theta\right) &= f\left(x_{i,obs}, x_{i,mis}|\theta\right) \\ &= f\left(x_{i,obs}|\theta\right) f\left(x_{i,mis}|x_{i,obs},\theta\right)' \end{aligned} \tag{59}$$

La contribución de la unidad i al logaritmo de verosimilitud se puede escribir como

$$L\left(\theta|x_i\right) = ln\left[f\left(x_{i,obs}|\theta\right)\right] + ln\left[f\left(x_{i,mis}|x_{i,obs},\theta\right)\right], \tag{60}$$

y, al definir

$$\ln\left\{f\left(X_{mis}|X_{obs},\theta\right)\right\} = \sum_i \ln\left\{f\left(x_{i,mis}|x_{i,obs},\theta\right)\right\}$$

el logaritmo de verosimilitud, se puede escribir como

$$L\left(\theta|X\right) = L\left(\theta|X_{obs}\right) + ln\left[f\left(X_{mis}|X_{obs},\theta\right)\right], \tag{61}$$

Este logaritmo de verosimilitud de datos completos no puede evaluarse ya que $L\left(\theta|X\right)$ y $\ln\left[f\left(X_{mis}|X_{obs},\theta\right)\right]$ dependen de los datos no observados. Sin embargo, la expectativa sobre los datos faltantes de este logaritmo de verosimilitud es una función que se puede maximizar, y el algoritmo EM utiliza esta maximización como un dispositivo para maximizar el logaritmo de verosimilitud de los datos observados. Al tomar expectativas de los términos que involucran los datos faltantes, obtenemos

$$\int f\left(X_{mis}|X_{obs},\theta\right) L\left(\theta|X\right) dX_{mis} = L\left(\theta|X\right) + \int f\left(X_{mis}|X_{obs},\theta\right)\left\{\ln\left[f\left(X_{mis}|X_{obs},\theta\right)\right]\right\} dX_{mis}$$

decimos,

$$Q(\theta) = L\left(\theta|X_{obs}\right) + H(\theta), \tag{62}$$

Las expectativas en (8.18) se toman con respecto a la densidad de los datos faltantes dado X_{obs} y θ. En las iteraciones EM, tenemos que $\theta^{(t)}$, la estimación actual de $\hat{\text{I}}$¸ se toma como el valor de este θ dado y, siguiendo a Roderick JA Little and Rubin (2002c), esto se puede expresar escribiendo

$$Q\left(\theta|\theta^{(t)}\right) = L\left(\theta|X_{obs}\right) + H\left(\theta|\theta^{(t)}\right) \qquad (paso\ E), \tag{63}$$

El cálculo de la expectativa del logaritmo de la verosimilitud de los datos completos $L\left(\theta|X\right)$ es el paso E del algoritmo EM. El siguiente paso, el paso M, es la maximización del logaritmo de la verosimilitud esperada, es decir,

$$\theta^{(t+1)} = arg\ \max_{\theta} Q\left(\theta|\theta^{(t)}\right) \qquad (\ paso\ M), \tag{64}$$

Un resultado clave relacionado con la convergencia del algoritmo EM es que una secuencia de pasos E y M tiene la propiedad de que $L\left(\theta^{(t+1)}|X_{obs}\right) \geq L\left(\theta^{(t)}|X_{obs}\right)$; es decir, la probabilidad logarítmica de los datos observados aumenta en cada iteración. Este resultado fue probado por Arthur P. Dempster, Laird, and Rubin (1977b). Para ver esto, escribimos la diferencia entre la verosimilitud logarítmica de los datos observados en dos iteraciones consecutivas como

$$L\left(\theta^{(t+1)}|X_{obs}\right) - L\left(\theta^{(t)}|X_{obs}\right) = Q\left(\theta^{(t+1)}|\theta^{(t)}\right) - Q\left(\theta^{(t)}|\theta^{(t)}\right) + H\left(\theta^{(t)}|\theta^{(t)}\right) - H\left(\theta^{(t+1)}|\theta^{(t)}\right), \tag{65}$$

La primera diferencia del lado derecho de @ref(8.21) no es negativa debido al paso de maximización. Se puede demostrar que la segunda diferencia tampoco es negativa usando la siguiente forma de la desigualdad de Jensen (Rudin (2012)): para una variable aleatoria y con densidad de probabilidad $p(y)$ y una función convexa φ, se cumple que

$$\varphi\left(\int p(y)g(y)dy\right) \leq \int p(y)\varphi\left(g(y)\right)dy,$$

con el signo de desigualdad invertido para φ cóncavo. Para aplicar esta desigualdad, primero desarrollamos la segunda diferencia en (8.21), usando (8.18), como

$$H\left(\theta^{(t)}|\theta^{(t)}\right) - H\left(\theta^{(t+1)}|\theta^{(t)}\right) = \int -f\left(X_{mis}|X_{obs},\theta^{(t)}\right)\ln\left\{\frac{f\left(X_{mis}|X_{obs},\theta^{(t+1)}\right)}{f\left(X_{mis}|X_{obs},\theta^{(t)}\right)}\right\}dX_{mis}$$

que es de la forma $\int p(y)\varphi\left(g(y)\right)dy$, con $\varphi = -ln$ y por lo tanto mayor o igual que

$$-ln\int f\left(X_{mis}|X_{obs},\theta^{(t+1)}\right)dX_{mis} = -ln(1) = 0$$

Así, hemos demostrado que ambas diferencias en el lado derecho de (8.21) son mayores o iguales a cero. La igualdad se mantiene solo si $Q\left(\theta^{(t+1)}|\theta^{(t)}\right) = Q\left(\theta^{(t)}|\theta^{(t)}\right)$, lo que significa que el paso de maximización no puede aumentar más la probabilidad de los datos observados.

4.3.1.4.1 El Algoritmo EM para Familias Exponenciales. Muchos modelos estadísticos se basan en distribuciones de probabilidad o funciones de densidades que son miembros de la familia exponencial regular. Esta familia incluye una amplia variedad de distribuciones, incluidas las distribuciones normal, de Bernoulli, binomial, binomial negativa, exponencial, gamma y multinomial (ver, por ejemplo, McCullagh (2019) y Cox and Hinkley (1979)). La familia exponencial es de interés para las estadísticas teóricas porque las propiedades derivadas de esta familia se trasladan directamente a los muchos miembros de esta distribución que se utilizan en aplicaciones prácticas. Una de esas propiedades que es relevante para el algoritmo EM es que el logaritmo de la verosimilitud se puede escribir de una manera que hace que sea particularmente fácil tomar la expectativa requerida para el paso E (cf. Roderick JA Little and Rubin (2002c) y Schafer (1997)). Para el caso de n observaciones $x_i \sim iid$, el logaritmo de verosimilitud de la familia exponencial está dado por:

$$L\left(\theta|X\right) = \eta(\theta)^T T(X) + ng(\theta) + c, \tag{66}$$

Donde $\eta(\ )$ es una función que transforma el vector de parámetros θ en lo que se denomina el parámetro canónico. Esta es solo una reparametrización que no afecta la cantidad de

parámetros. La función $T(\ )$ extrae de los datos $T(X)$ los estadísticos suficientes para estimar $\eta(\theta)$ o, equivalentemente, θ y da como resultado un vector con el mismo número de elementos que θ. La constante c no contiene los parámetros y se puede ignorar para maximizar la verosimilitud logarítmica. El logaritmo de verosimilitud sin esta constante se denomina núcleo de la función de logaritmo de verosimilitud.

Dado que el núcleo del logaritmo de verosimilitud de datos completo @ref(8.22) depende de los datos solo a través de los estadísticos suficientes y es una función lineal de estos estadísticos, la expectativa sobre los datos faltantes del núcleo logarítmico de verosimilitud se obtiene reemplazando los estadísticos suficientes por sus expectativas. Por lo tanto, el paso E se puede realizar reemplazando los componentes $T_k(X)$ de $T(X)$, con $k = 1, \cdots, K$ y K el número de parámetros, en el logaritmo de verosimilitud de datos completo @ref(8.22) por sus expectativas $E_{X_{mis}} T_k(X)$. Para distribuciones familiares exponenciales regulares, las ecuaciones de verosimilitud toman una forma simple particular; equiparan las estadísticas suficientes a sus valores esperados bajo el modelo $f\left(X|\theta\right)$. El estimador de máxima verosimilitud se obtiene entonces como la solución de θ de estas ecuaciones de verosimilitud, es decir, como la solución de

$$E\left(T(X)\right) = T(X), \tag{67}$$

donde E denota la expectativa sobre X bajo el modelo. En el paso M del algoritmo EM, la maximización se aplica a la probabilidad logarítmica con las estadísticas suficientes reemplazadas por sus valores esperados.

Esto resulta en la resolución de las ecuaciones

$$E\left(T(X)\right) = T_{X_{mis}}(X), \tag{68}$$

En resumen, para aplicar el algoritmo EM a una verosimilitud logarítmica familiar exponencial, primero identificamos las estadísticas suficientes para estimar los parámetros en el caso de datos completos. Luego, se iteran los siguientes dos pasos:

Paso E: Evaluar la expectativa sobre los datos faltantes de las estadísticas suficientes.

Paso M: Resuelva las ecuaciones de probabilidad de datos completas con las estadísticas suficientes reemplazadas por sus valores esperados.

4.3.1.5 EM para datos normales multivariados. Ahora aplicaremos el algoritmo EM para estimar el vector de medias y la matriz de covarianza de una distribución normal multivariada. La verosimilitud de la densidad normal multivariada, o más precisamente el núcleo de la verosimilitud ya que ignoramos las constantes, se puede escribir como

$$l\left(\theta|X\right) = \left|\Sigma\right|^{\frac{-n}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left(x_i - \mu\right)^T \Sigma^{-1}\left(x_i - \mu\right)\right)$$

y para el log verosimilitud que obtenemos (ver, por ejemplo, Schafer (1997), Capítulo 5)

$$
\begin{aligned}
L\left(\theta|X\right) &= -\frac{n}{2}\ln|\Sigma| + \sum_{i=1}^{n}\mu^T\Sigma^{-1}x_i - \frac{1}{2}\sum_{i=1}^{n}x_i^T\Sigma^{-1}x_i - \frac{n}{2}\mu^T\Sigma^{-1}\mu \\
&= \mu^T\Sigma^{-1}x_i - \frac{1}{2}\sum_{i=1}^{n}tr\left(\Sigma^{-1}x_ix_i^T\right) - \frac{n}{2}\left(\ln|\Sigma| - \mu^T\Sigma^{-1}\mu\right)
\end{aligned}, \tag{69}
$$

donde hemos utilizado que el escalar $x_i^T\Sigma^{-1}x_i$ es igual a su traza y que $tr(AB)$ es $tr(BA)$ si A y B son matrices de dimensiones tales que ambos productos están definidos. La última expresión es de la forma @ref(8.22) con estadísticas suficientes

$$
\begin{aligned}
T_1 &= \sum_{i=1}^{n}x_i \\
T_2 &= \sum_{i=1}^{n}x_ix_i^T
\end{aligned}, \tag{70}
$$

y los correspondientes valores esperados

$$
\begin{aligned}
ET_1 &= n\mu \\
ET_2 &= n\left(\Sigma + \mu\mu^T\right)
\end{aligned}, \tag{71}
$$

Los estimadores de máxima verosimilitud para μ y Σ se obtienen equiparando los estadísticos suficientes a sus expectativas, con el conocido resultado de que μ y Σ son estimados por el vector de media muestral y la matriz de covarianza muestral.

Para aplicar el paso E del algoritmo EM, debemos tomar la expectativa sobre los datos faltantes de las estadísticas suficientes. Dado que las estadísticas suficientes son sumas de valores x *de* (T_1) o sumas de productos de valores x *de* (T_2), el cálculo del valor esperado de las estadísticas suficientes requiere el cálculo de la expectativa de valores x faltantes y la expectativa de productos de x valores que implican valores perdidos. Para la expectativa condicional de los valores faltantes en $x_i, x_{i,mis}$, dada $x_{i,obs}$ y las estimaciones actuales de μ y Σ tenemos de @ref(8.5) y @ref(8.6)

$$
\begin{aligned}
E\left(x_i, x_{i,mis}|x_{i,obs}, \widehat{\mu}, \widehat{\Sigma}\right) &= \widehat{x}_{mis} \\
E\left(x_i, x_{i,mis}|x_{i,obs}, \widehat{\mu}, \widehat{\Sigma}\right) &= \widehat{\mu}_{mis} + \widehat{B}_{mis,obs}\left(x_{i,obs} - \widehat{\mu}_{obs}\right) \\
E\left(x_i, x_{i,mis}|x_{i,obs}, \widehat{\mu}, \widehat{\Sigma}\right) &= \widehat{\mu}_{mis} + \widehat{\Sigma}_{mis,obs}\widehat{\Sigma}_{obs,obs}^{-1}\left(x_{i,obs} - \widehat{\mu}_{obs}\right),
\end{aligned}
$$

que son las predicciones de regresión para las variables que faltan en el registro i con las variables observadas en ese registro como predictores. La expectativa condicional sobre los datos que faltan de T_1, digamos T_1^*, ahora se puede escribir como

$$T_1^* = E_{X_{mis}} T_1 = \sum_{i=1}^{n} \begin{pmatrix} \widehat{x}_{i,mis} \\ x_{i,obs} \end{pmatrix} = \sum_{i=1}^{n} x_i^*, \tag{72}$$

Para evaluar la expectativa condicional de los productos de las variables x, primero consideramos la descomposición

$$x_{i,j} = \widehat{x}_{i,j} + \widehat{e}_{i,j}$$

$\widehat{e}_{i,j}$ es el residuo $x_{i,j} - \widehat{x}_{i,j}$. Ahora el producto de dos valores de x se puede expresar como

$$x_{i,j}x_{i,k} = \widehat{x}_{i,j}\widehat{x}_{i,k} + \widehat{x}_{i,j}\widehat{e}_{i,k} + \widehat{x}_{i,k}\widehat{e}_{i,j} + \widehat{e}_{i,j}\widehat{e}_{i,k}$$

Dado que los residuales tienen una expectativa cero, los términos con un residual desaparecen al tomar la expectativa. La expectativa del producto de los dos residuales es igual a la covarianza residual. Sin embargo, esta covarianza residual o condicional es cero si se observan uno o ambos valores $x_{i,j}$ y $x_{i,k}$ porque el condicionamiento de los valores observados implica que $\widehat{x}_{i,j} = x_{i,j}$ y $\widehat{e}_{i,j} = 0$ si se observa $x_{i,j}$. Por lo tanto, la expectativa condicional sobre los datos faltantes de T_2 y T_2^*, digamos, se puede expresar como

$$T_2^* = E_{X_{mis}} T_2 = \sum_{i=1}^{n} x_i^* \left(x_i^*\right)^T + V_i, \tag{73}$$

con V_i la matriz con elementos $(V_i)_{jk}$ que son iguales a los elementos correspondientes de la matriz de covarianza residual dada por @ref(8.8) si faltan tanto $x_{i,j}$ como $x_{i,k}$ y cero en caso contrario.

El algoritmo EM ahora procede de la siguiente manera:

Paso E: dadas las estimaciones actuales $\mu^{(t)}$ y $\Sigma^{(t)}$ > de μ y Σ, calcule las expectativas condicionales de los > estadísticos suficientes $T_1^{*(t)}$ y $T_2^{*(t)}$.

Paso M: al usarlas de esta manera, completa suficientes estadísticas, actualice los parámetros mediante las ecuaciones de probabilidad de datos completas:

$$mu^{(t+1)} = \frac{1}{n} T_1^{*(t)}, \tag{74}$$

$$\Sigma^{(t+1)} = \frac{1}{n} T_2^{*(t)} - \mu^{(t+1)} \left(\mu^{(t+1)}\right)^T, \tag{75}$$

4.3.2 EM para datos multinomiales

El Algoritmo EM para datos multinomiales es mucho más simple que el de datos normales multivariados y sigue de cerca la imputación agregada de datos categóricos discutida en la Sección 8.2.2. A partir de la distribución multinomial @ref(8.9), vemos que el núcleo del logaritmo de verosimilitud multinomial se puede escribir como

$$L\left(\pi|n\right) = \sum_{j=1}^{C} n_j \ln\left(\pi_j\right), \tag{76}$$

La distribución multinomial es un miembro de la familia exponencial, las estadísticas suficientes son los conteos de campos n_j, y el logaritmo de verosimilitud multinomial es claramente lineal en las estadísticas suficientes. Por lo tanto, los estimadores de máxima verosimilitud para las probabilidades de los campos se obtienen igualando los recuentos de campos con sus expectativas y, por lo tanto, están dados por

$$\widehat{\pi}_j = \frac{n_j}{n}, \tag{77}$$

Ahora, considere un conjunto de campos S que se suman a un campo en un cálculo suplementario como se describe en la **Sección 8.2.2.XXX** Sea $n_{j,obs}$ $(j \in S)$ el número de conteos de las unidades completamente observadas en estos campos. Además, deje que el recuento en el cálculo marginal suplementario sea $n_{s,mis}$; entonces el valor esperado del conteo de campos n_j, dado $n_{j,obs}$ y $n_{s,mis}$, está dado por

$$L\left(n_j|\widehat{\pi}_j, n_{j,obs}, n_{s,mis}\right) = n_{j,obs} + n_{s,mis} \cdot \widehat{\pi}_j^S \qquad\qquad (j \in S), \tag{78}$$

con el $\widehat{\pi}_j^S$ las probabilidades condicionales estimadas $\widehat{\pi}_j / \sum_{j \in S} \widehat{\pi}_j$. Por lo tanto, distribuimos el recuento total en el campo del margen suplementario sobre los campos que contribuyen a él de acuerdo con la estimación actual de la distribución condicional sobre estos campos. Este proceso se repite para cada campo en un margen suplementario; por lo tanto, todas las observaciones, con valores perdidos o no, se clasifican en la tabla de contingencia de dimensión completa. Esto completa el paso E.

Luego, el paso M simplemente calcula nuevas estimaciones de los parámetros π_j usando (77) para los datos completos obtenidos en el paso E.

El resultado del algoritmo EM es una tabla de contingencia completa. En este sentido, los valores faltantes se imputan a nivel agregado: se han imputado los recuentos de campos faltantes, pero no los valores faltantes en los registros subyacentes (consulte también la **Sección 8.2.2**), Johansen (1988)

5 Métodos de imputación basados en modelos: Introducción (MA)

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[7] y el *mecanismo* que genera los *datos faltantes*[8] (*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*[9], 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 (Donald B. Rubin 1976b), 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?* (Donald B. Rubin 1976b, 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*[10] (*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

[7] 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 *posiciones* 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.

[8] 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.

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

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

que presentan los métodos tradicionales.

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 (R. J. A. Little and Rubin 2020a, 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 (R. J. A. Little and Rubin 2020a, 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[11] 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

[11] 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).

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[12] (Van Buuren 2012, pag. 11).

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[13]. 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).

[12]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 (R. J. A. Little and Rubin 2020a, pag. 70). Mayor detalle sobre este enfoque se puede encontrar en (R. J. A. Little and Rubin 2020a, secc. 4.2.2).

[13]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 los campos dentro de los cuales se imputa por la media (R. J. A. Little and Rubin 2020a, pag. 68).

El método de imputación por regresión estocástica no resuelve todos los problemas y hay muchas sutilezas que deben tenerse presentes[14]. No obstante, el método de imputación por regresión estocástica es el único método tradicional que, generalmente, es capaz de producir estimaciones insesgadas de los parámetros de interés cuando el mecanismo es del tipo *MAR*[15] (*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*[16]. Este método es común en la práctica de las encuestas y puede implicar esquemas muy elaborados para seleccionar los casos *donantes*[17]. 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

[14]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.

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

[16]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.

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

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 (R. J. A. Little and Rubin 2020a, 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 de datos obtenidos mediante el método de imputación *Cold-Deck* es inexistente (R. J. A. Little and Rubin 2020a, 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 (R. J. A. Little and Rubin 2020a, 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 (R. J. A. Little and Rubin 2020a, 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.

6 Estimación por Máxima Verosimilitud con Datos Incompletos

Breve introducción.

6.1 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.

$$L(\theta|Y) \propto f(Y|\theta) \tag{79}$$

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

$$\hat{\theta}_{ML} = \arg\max_{\theta} L(\theta|Y) \tag{80}$$

6.2 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:

$$L(\theta, \phi|Y_{obs}, R) \propto f(Y_{obs}, R|\theta, \phi) \tag{81}$$

$$\begin{aligned}
L(\theta, \phi|Y_{obs}, R) &\propto f(Y_{obs}, R|\theta, \phi) \\
&= \int f(Y, R|\theta, \phi)\, dY_{miss} \\
&= \int f(Y|\theta)\, f(R|Y, \phi)\, dY_{miss} \\
&= \int f(Y_{obs}, Y_{miss}|\theta)\, f(R|Y_{obs}, Y_{miss}, \phi)\, dY_{miss} \\
&= \int f(Y_{obs}, Y_{miss}|\theta)\, f(R|Y_{obs}, \phi)\, dY_{miss} \\
&= f(R|Y_{obs}, \phi) \int f(Y_{obs}, Y_{miss}|\theta)\, dY_{miss} \\
&= f(R|Y_{obs}, \phi)\, f(Y_{obs}|\theta)
\end{aligned} \tag{82}$$

- 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:

$$L\left(\theta, \phi|Y_{obs}, R\right) \quad \propto \quad f\left(Y_{obs}|\theta\right) \tag{83}$$

6.3 El Algoritmo de Esperanza y Maximización (EM)

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

6.4 Modelos para Datos Categóricos

6.4.1 Datos Categoricos Binarios

6.4.2 Datos Categoricos no Binarios

6.5 Modelos para Datos Continuos

6.6 Modelos para Mezclas de Distribuciones

7 Métodos de Pseudo-Verosimilitud para Datos Incompletos

8 Estimación Bayesiana con Datos Incompletos

8.1 Estimación Bayesiana: Conceptos básicos

8.2 Métodos Bayesianos con Datos Incompletos: Marco teórico general

9 Imputación Múltiple (FM)

9.1 Los fundamentos del enfoque de la imputación múltiple

La imputación múltiple (MI, por sus siglas en inglés), introducida por Donald B. Rubin (1988), es un enfoque para manejar datos faltantes en estudios estadísticos. El enfoque de Donald B. Rubin para la imputación múltiple, tal como se describe en Donald B. Rubin (2004), es un método para tratar los datos faltantes en los análisis estadísticos donde asume que los datos son "Missing At Random" (MAR), lo que significa que la probabilidad de que un valor sea faltante puede depender de los datos observados, pero no de los datos faltantes en sí.

Esta técnica permite generar valores razonables para datos que faltan, basándose en la distribución de los datos observados. El principio básico es que la imputación debería reflejar la incertidumbre acerca de los valores faltantes, generando varias versiones imputadas diferentes, lo que lleva a la "multiplicidad" en la imputación (Van Buuren 2018). Un manejo inadecuado de los datos faltantes en un análisis estadístico puede conducir a estimaciones sesgadas y/o ineficientes de parámetros como las medias o los coeficientes de regresión, y errores estándar sesgados que resultan en intervalos de confianza y pruebas de significancia incorrectas. En todos los análisis estadísticos, se hacen algunas suposiciones sobre los datos faltantes.

El marco de trabajo de Roderick JA Little and Rubin (2002a) se utiliza a menudo para clasificar los datos faltantes como: (i) faltantes completamente al azar (MCAR, por sus siglas en inglés - la probabilidad de que los datos falten no depende de los datos observados o no observados), (ii) faltantes al azar (MAR - la probabilidad de que los datos falten no depende de los datos no observados, condicionados a los datos observados) o (iii) faltantes no al azar (MNAR - la probabilidad de que los datos falten sí depende de los datos no observados, condicionados a los datos observados). Por ejemplo, en una encueste de hogares, los datos acerca del ingreso son MAR si es más probable que las personas con mayores años de estudio declaren en dicha variable (y los años de estudio se incluye en el análisis), pero son MNAR si las personas con ingresos altos son más propensas a no declarar sus ingresos en la encuesta que otras personas con iguales años de escolaridad. No es posible distinguir entre MAR y MNAR solo a partir de los datos observados, aunque la suposición de MAR puede hacerse más plausible recolectando más variables explicativas e incluyéndolas en el análisis.

Bajo el paradigma de imputación múltiple, la idea es generar múltiples conjuntos de datos donde cada valor faltante para un conjunto de datos Y_{mis} es reemplazado con un conjunto de valores plausibles, creando así múltiples versiones completas del conjunto de datos. Supongamos se generan M conjuntos de datos posibles, los resultados de estos M análisis se combinan en una única estimación y una única medida de incertidumbre. Este enfoque tiene la ventaja de reflejar adecuadamente la incertidumbre sobre los valores faltantes en las estimaciones finales, lo que puede dar lugar a inferencias más precisas y confiables en presencia de datos faltantes. En este método, la incertidumbre de la imputación se tiene en cuenta mediante la creación de estos múltiples conjuntos de datos. El proceso de imputación múltiple puede dividirse en tres fases:

- Imputación: Durante la fase de imputación, se generan M conjuntos de datos completos, donde M es el número de imputaciones. Cada conjunto de datos se crea reemplazando

los valores faltantes con estimaciones basadas en un modelo de imputación. Este modelo se ajusta a los datos observados y también incorpora la variabilidad aleatoria, lo que significa que las imputaciones son diferentes en cada uno de los M conjuntos de datos. Es decir, para un conjunto de datos con valores faltantes, se generan M imputaciones para cada valor faltante. Por lo tanto, a partir de un conjunto de datos original con datos faltantes, generamos M conjuntos de datos completos. Si denotamos la m-ésima imputación para el i-ésimo valor faltante como $y_{i,m}$, entonces, para cada i, generamos $y_{i,1}, y_{i,2}, \ldots, y_{i,M}$.

- Análisis: En la fase de análisis, se lleva a cabo el análisis estadístico de interés en cada uno de los M conjuntos de datos completos como si fueran datos completos sin faltantes. Cada uno de estos M conjuntos de datos se analiza por separado utilizando el análisis estadístico completo de los datos. Si denotamos el estimador de interés como θ, entonces para cada conjunto de datos completado obtenemos un estimado $\hat{\theta}_m$ para $m = 1, 2, \ldots, M$. Esto resulta en M conjuntos de estimaciones y estadísticas de prueba.
- Combinación: En la fase de combinación, las M estimaciones y estadísticas de prueba de los conjuntos de datos imputados se combinan para producir una única estimación y estadística de prueba. La combinación tiene en cuenta tanto la variabilidad dentro de cada conjunto de datos imputados (debido a la variabilidad de muestreo) como la variabilidad entre los conjuntos de datos imputados (debido a la incertidumbre en el proceso de imputación). La estimación final de θ se calcula como el promedio de las M estimaciones, es decir, $\hat{\theta} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_j$.

9.2 Implementación general de los métodos de imputación múltiple

Supongamos θ es una cantidad de interés a calcular de una población estadística, ya sea una media, total poblacional, coeficiente de regresión, etc. Note que θ es una característica de la población estadística y no depende de características de un determinado diseño. Dado que esta cantidad θ solo es posible calcularla con la población completa, se suele calcular un estimador $\hat{\theta}$ del parámetro poblacional.

El objetivo es encontrar un estimador insesgado de θ tal que la esperanza de $\hat{\theta}$ sobre todas las muestras posibles de los datos completos Y sea igual al parámetro poblacional deseado, es decir, se busca que $E(\hat{\theta}|Y) = \theta$. Note que la incertidumbre acerca de la estimación $\hat{\theta}$ depende acerca del conocimiento que se tiene acerca del vector Y_{mis}. En ese sentido, si fuese posible generar valores para Y_{mis} de manera exacta, entonces la incertidumbre acerca de la estimación $\hat{\theta}$ se reduciría o bien no existiría incertidumbre acerca de la estimación generada para el parámetro poblacional.

Sea $P(\theta|Y_{\text{obs}})$ la distribución a posteriori de θ, esta distribución puede ser descompuesta integrando sobre la distribución conjunta del vector $(Y_{\text{obs}}, Y_{\text{mis}})$, es decir:

$$P(\theta|Y_{\text{obs}}) = \int P(\theta, Y_{\text{mis}}|Y_{\text{obs}})dY_{\text{mis}} \tag{84}$$

$$= \int P(\theta|Y_{\text{obs}}, Y_{\text{mis}})P(Y_{\text{mis}}|Y_{\text{obs}})dY_{\text{mis}} \tag{85}$$

$$= \lim_{M\to\infty} \frac{1}{M}\sum_{m=1}^{M} P(\theta|Y_{\text{obs}}, Y_{\text{mis}}^{(m)}) \tag{86}$$

$$\approx \frac{1}{M}\sum_{m=1}^{M} P(\theta|Y_{\text{obs}}, Y_{\text{mis}}^{(m)}) \tag{87}$$

Donde $M > 1$ y $Y_{\text{miss}}^{(m)}$ es obtenida como una realización de la distribución de $P(Y_{\text{mis}}|Y_{\text{obs}})$ para $m = 1, ..., M$.

Dado que se desea hacer inferencia sobre el parámetro θ es de interés conocer la distribución de $P(\theta|Y_{\text{obs}})$ pues utiliza la información que se tiene, por otra parte $P(\theta|Y_{\text{obs}}, Y_{\text{mis}})$ es la distribución hipotética del parámetro sobre los datos completos y $P(Y_{\text{mis}}|Y_{\text{obs}})$ es la distribución de los valores perdidos dados los valores observados.

De la ecuación (85), seria posible obtener M imputaciones $\dot{Y}_{\text{mis}}$ a partir de la distribución $P(Y_{\text{mis}}|Y_{\text{obs}})$, con ello, se podria calcular la cantidad θ a partir de la distribución de $P(\theta|Y_{\text{obs}}, \dot{Y}_{\text{mis}})$. Van Buuren (2018) muestran que la media posteriori de $P(\theta|Y_{\text{obs}})$ es igual a

$$E(\theta|Y_{\text{obs}}) = E(E[\theta|Y_{\text{obs}}, Y_{\text{mis}}]|Y_{\text{obs}}) \tag{88}$$

En otras palabras, la media posteriori de θ bajo repetidas imputaciones de los datos.

Suponga que $\hat{\theta}_m$ es la estimación usando la m-ésima imputación, la estimación de las M estimaciones combinadas es igual a

$$\bar{\theta} = \frac{1}{M}\sum_{m=1}^{M} \hat{\theta}_m \tag{89}$$

En un caso multivariado, es posible que $\bar{\theta}_m$ contenga k parámetros y por tanto sea un vector de dimensión $k \times 1$. La varianza de la distribución a posteriori $P(\theta|Y_{\text{obs}})$ se puede escribir como la suma de dos componentes, esto es:

$$V(\theta|Y_{\text{obs}}) = E(V(\theta|Y_{\text{obs}}, Y_{\text{mis}})|Y_{\text{obs}}) + V(E(\theta|Y_{\text{obs}}, Y_{\text{mis}})|Y_{\text{obs}}) \tag{90}$$

La primera componente de (90) puede interpretarse como la media de las repetidas imputaciones a posteriori de la varianza de θ (La cuál será denominada como intra-varianza) mientras que la segunda componente es la varianza entre las medias de θ estimadas con la distribución a posteriori (la cuál será llamada entre-varianza).

Si denotamos $\bar{U}_\infty$ y B_∞ como la intra y entre varianzas cuando $M \to \infty$ entonces se tiene que $T_\infty = \bar{U}_\infty + B_\infty$ corresponde a la varianza posteriori de θ. Cuando M es finito, podemos calcular la media de las varianzas de las imputaciones como

$$\bar{U} = \frac{1}{M} \sum_{m=1}^{M} \bar{U}_m \tag{91}$$

donde $\bar{U}_m$ corresponde a la matriz de varianzas covarianzas de $\hat{\theta}_m$ obtenida de la m-ésima imputación. La estimación insesgada de las varianzas entre las M estimaciones realizadas está dadada por

$$B = \frac{1}{M-1} \sum_{m=1}^{M} (\hat{\theta}_m - \bar{\theta})(\hat{\theta}_m - \bar{\theta})' \tag{92}$$

Para calcular la varianza total T cuando M es finito, es necesario incorporar el hecho de que $\bar{\theta}$ es estimado usando un número de imputaciones finita. Donald B. Rubin (2004) muestra que dicho factor corresponde a $\frac{B}{M}$. Por tanto, la varianza total T de la estimación $\bar{\theta}$ a travéz de las M imputaciones puede ser escrita como

$$\begin{aligned} T &= \bar{U} + B + \frac{B}{M} \\ &= \bar{U} + \left(1 + \frac{1}{M}\right) B \end{aligned}$$

Steele, Wang, and Raftery (2010) investigaron alternativas para obtener estimaciones de T utilizando mezclas de distribuciones normales. En este escenario, cuando existe normalidad multivariante y M no es grande, estos métodos producen estimaciones ligeramente más eficientes de T.

9.2.1 Consideraciones técnicas en la aplicación de los métodos de imputación

Considerando que es necesario realizar inferencia sobre la estimación puntual $\bar{\theta}$ y la varianza estimada definida como T, diversos autores (Donald B. Rubin 1988; Van Buuren 2018; Geert Molenberghs et al. 2014) proponen utilizar la distribución t.

Van Buuren (2018) menciona que la inferencia de un solo parámetro se aplica cuando $k = 1$, o bien si $k > 1$ pero además la prueba se repite para cada uno de los k componentes en el parámetro. Dado que la varianza total T es desconocida, $\bar{\theta}$ sigue una distribución t en lugar de la normal. Las pruebas univariadas para la imputación se basan en la aproximación:

$$\frac{\theta - \bar{\theta}}{\sqrt{T}} \sim t_\nu \tag{93}$$

donde t_ν es una distribución t-student con ν grados de libertad. Con lo anterior podemos por tanto construir un intervalo de $(1-\alpha)100\%$ para $\bar{\theta}$ definido en la siguiente ecuación

$$\bar{\theta} \pm t_{\nu,1-\alpha/2}\sqrt{T} \tag{94}$$

donde $t_{\nu,1-\alpha/2}$ corresponde al cuantil de probabilidad $1-\alpha/2$ de t_ν. Supongamos se desea testear la hipotesis nula $\theta = \theta_0$ para un valor en específico de θ_0. El valor-p del test se puede calcular como

$$P_s = \Pr\left[F_{1,\nu} > \frac{(\theta_0 - \bar{\theta})^2}{T}\right] \tag{95}$$

donde $F_{1,\nu}$ es una distribución (F) Fisher-Snedecor con 1 y ν grados de libertad.

Utilizando una aproximación estándar de tipo Satterthwaite, Donald B. Rubin (1988) calculó los grados de libertad de la distribución de $\bar{\theta}$ dado los M conjunto de datos imputados como:

$$\nu = (M-1)\left[1 + \frac{\bar{U}}{(1+\frac{1}{M})B}\right]^2 \tag{96}$$

La ecuacion anterior puede ser reescrita como

$$\nu = (M-1)\left[1 + \frac{1}{r_M}\right]^2 \tag{97}$$

donde $r_M = \frac{(1+m^{-1})B}{\bar{U}}$ es conocida como el incremento de varianza relativa (RVI por sus siglas en inglés) debido a los valores faltantes, considerando que $\bar{U}$ representa la varianza de la estimación $\bar{\theta}$ cuando no existe variación entre los valores estimados $\hat{\theta}_m$, en cuyo caso $B = 0$.

Por otra parte, para θ podemos definir el ratio

$$\lambda_M = \frac{(1+m^{-1})B}{T} \tag{98}$$

el cuál puede ser interpretado como la proporción de varianza que se puede atribuir a la información perdida.

Si $\lambda_M = 0$, la información perdida no añade variación extra a la variación del muestreo, lo cuál ocurre excepcionalmente solo si se recrea de manera perfecta dicha información perdida. Por contraparte si $\lambda_M = 1$ toda la variabilidad es causada por la información faltante. Si $\lambda_M > 0.5$ la influencia del modelo de imputación en el resultado final es mayor que el modelo considerando los datos completos (Van Buuren 2018). Notar que $r_M = \lambda_M/(1-\lambda_M)$.

Una cantidad estrechamente relacionada con λ_M se denomina "fracción de información faltante" (FMI, por sus siglas en inglés), puede ser calculada comparando la "información" en la densidad posteriori (t) aproximada, definida como el negativo de la segunda derivada de la densidad log-posterior, con la de la densidad posteriori hipotética de los datos completos, dando como resultado (Donald B. Rubin 1988):

$$\gamma_M = \frac{r_M + \frac{2}{\nu+3}}{1+r_M} \tag{99}$$

Es fácil ver que $\gamma_M \to r_M/(1+r_M) = \lambda_M$ cuando $M \to \infty$. Esto permite observar que el efecto de los datos faltantes es una combinación de la actual cantidad de información perdida y el grado con el cuál aquella información de los datos incompletos contribuye a la estimación de interes mediante el modelo de imputación.

Barnard and Rubin (1999) muestran que la ecuación (97) puede producir valores en los grados de libertad que son mayores al tamaño muestral en los datos completos cuando la muestra es pequeña. Debido a esto, desarrollaron una adaptación para tamaños de muestras pequeñas teniendo en cuenta dicho problema. Se define ν_{old} como los grados de libertad de la ecuación (97) y ν_{com} los grados de libertad de $\bar{\theta}$ cuando se tiene los datos completos sin valores perdidos. En este caso, si se tienen k parámetros para un tamaño muestral de n, entonces $\nu_{com} = n - k$. Los grados de libertad de los datos observados que tienen en cuenta la información faltante es

$$\nu_{obs} = \frac{\nu_{com}+1}{\nu_{com}+3}\nu_{com}(1-\lambda) \tag{100}$$

Los grados de libertad ajustados que se utilizarán para las pruebas en imputación múltiplese puede escribir de manera concisa como

$$\nu = \frac{\nu_{old}\nu_{obs}}{\nu_{old}+\nu_{obs}} \tag{101}$$

Van Buuren (2018) señala que para la cantidad de la ecuación (101) siempre se tiene que $\nu \leq \nu_{com}$. Si $\nu_{com} = \infty$ entonces (101) se reduce a (97).

9.2.2 Número de imputaciones a realizar

La imputación múltiple es una técnica de simulación por lo que $\bar{\theta}$ y su varianza total estimada T están sujetas a errores de simulación. En ese sentido, la fórmula dada por

$$T_m = \left(1 + \frac{\gamma_0}{m}\right) T_\infty \tag{102}$$

es la relación entre la varianza del parámetro estimado en un escenario con un número finito de imputaciones (T_m) y la varianza del parámetro estimado en un escenario con un número infinito de imputaciones (T_∞).

Aquí, m representa el número de imputaciones múltiples y γ_0 es la fracción de información perdida. Esta cantidad es equivalente a la proporción esperada de observaciones que faltan en el caso de que Y sea una variable que no tenga covariab-les asociadas. Sin embargo, esta proporción suele ser menor si existen covariables que pueden predecir el valor de Y. Cuando m tiende a infinito, la varianza del estimador tiende a T_∞, es decir, se reduce la varianza debido al error de simulación. Sin embargo, en la práctica, rara vez se alcanza el límite de $m = \infty$ y se usa un número finito de imputaciones.

La cercanía de T_m a T_∞ es una medida de qué tan bien se ha estimado la varianza del parámetro. En teoría, cuanto mayor sea m, más cercano será T_m a T_∞, lo que significa que la varianza estimada es más precisa. Sin embargo, aumentar el número de imputaciones

también aumenta la carga computacional, por lo que se debe encontrar un equilibrio. Según Bodner (2008), en la mayoría de los escenarios prácticos, se pueden obtener buenos resultados con solo 20-40 imputaciones múltiples.

El intervalo de confianza para la estimación depende tanto de ν como de m. Royston (2004) sugiere un criterio para determinar m basado en el coeficiente de confianza $t_\nu\sqrt{T}$, y propone que el coeficiente de variación de $\log(t_\nu\sqrt{T})$ debería ser inferior a 0.05. Este criterio tiene el efecto de reducir el intervalo de confianza en un 10%, lo que implica que se necesitarían al menos $m > 20$ imputaciones.

En su estudio, Bodner (2008) examinó la variabilidad de tres medidas específicas con diferentes números de imputaciones múltiples (m): el ancho del intervalo de confianza del 95%, el valor p y γ_0 (la proporción real de información perdida). En este contexto, Bodner estableció un criterio para selecci## Algoritmo Data Augmentation (DA)

El algoritmo DA considera la distribución posteriori conjunta de θ y Y_{mis} condicional en Y_{obs}, $P(\theta, Y_{\text{mis}}|Y_{\text{obs}})$, y obtiene muestras de forma iterativa (Tanner y Wong, 1987). Note que

$$P(\theta, Y_{\text{mis}}|Y_{\text{obs}}) = P(\theta|Y_{\text{mis}}, Y_{\text{obs}})P(Y_{\text{mis}}|Y_{\text{obs}}) = P(Y_{\text{mis}}|\theta, Y_{\text{obs}})P(\theta|Y_{\text{obs}}) \tag{103}$$

En segundo lugar si consideramos Y_{mis} también como un "parámetro", podemos extraer $P(\theta, Y_{\text{mis}}|Y_{\text{obs}})$ iterando entre el muestreo de $P(\theta|Y_{\text{mis}}, Y_{\text{obs}})$ y $P(Y_{\text{mis}}|Y_{\text{obs}}, \theta)$ mediante gibss sampling. Estas muestras constituyen iteraciones de cadenas de Markov Monte Carlo (MCMC). Cuando las muestras de ambos $P(\theta|Y_{\text{mis}}, Y_{\text{obs}})$ y $P(Y_{\text{mis}}|Y_{\text{obs}}, \theta)$ convergen, también obtenemos muestras de Y_{mis} que convergen a la distribución predictiva posterior, $P(Y_{\text{mis}}|Y_{\text{obs}})$, creando así múltiples imputaciones (He, Zhang, and Hsu 2021b).

La estrategia DA se puede esbozar de forma algorítmica de la siguiente manera:

1. Derivar la distribución posterior de datos completos, $P(\theta|Y) = P(\theta|Y_{\text{obs}}, Y_{\text{mis}})$, bajo un prior $\pi(\theta)$.
2. Comenzar con una estimación o suposición del parámetro θ, digamos $\theta^*(t)$ (donde $t = 0$ en la primera iteración).
3. Extraer valores de datos faltantes de la distribución condicional: $Y^*(t)_{\text{mis}} \sim P(Y_{\text{mis}}|Y_{\text{obs}}, \theta^*(t))$ (el paso I, para "Imputación").
4. Extraer un nuevo valor del parámetro de su distribución posterior de datos completos, "insertando" el nuevo valor extraído de Y_{mis}: $\theta^*(t+1) \sim P(\theta|Y_{\text{obs}}, Y^*(t)_{\text{mis}})$ (el paso P, para "Posterior").
5. Repetir los Pasos 3 y 4 (el paso I y el paso P) hasta que se alcance la convergencia para $(\theta^*(t), Y^*(t)_{\text{mis}})$, digamos en $t = T$. Las muestras de $Y^*(T)_{\text{mis}}$ constituyen el m-ésimo conjunto de imputaciones, $Y_{\text{mis}}^{(m)}$.
6. Repetir los Pasos 2 al 5 de forma independiente M veces para crear múltiples conjuntos de imputaciones.

Una ventaja del algoritmo DA sobre el algoritmo DB es que trabajar en $P(\theta|Y_{\text{mis}}, Y_{\text{obs}})$ suele ser más fácil que $P(\theta|Y_{\text{obs}})$ porque el primero es condicional en datos completos y no está limitado por el patrón de datos faltantes. Tenga en cuenta que, a diferencia del algoritmo DB,

el algoritmo DA requiere iteraciones entre el paso I y el paso P. Cuando θ contiene múltiples componentes, el paso P del algoritmo DA puede constar de múltiples pasos, extrayendo cada parámetro condicional en otros parámetros y valores faltantes como en el Gibss sampling (He, Zhang, and Hsu 2021b).

9.2.3 Imputación basada en Regresión lineal con Data Augmentation

La regresión lineal es frecuentemente el modelo preferido para la imputación de variables continuas que siguen una distribución normal.

$$Y_{\text{obs}}|X;\boldsymbol{\beta} \sim \text{N}\left(X\boldsymbol{\beta}, \sigma^2\right) \tag{104}$$

Donde, $\hat{\beta}$ es el estimador del parámetro (o un vector de tamaño k) del modelo que se ajusta a las observaciones de datos Y_{obs}. Asimismo, $\mathbf{V}$ representa la matriz de varianzas-covarianzas de $\hat{\beta}$, y $\hat{\sigma}$ es la estimación de la varianza del modelo ajustado.

Para poder realizar la imputación, es necesario obtener los parámetros de imputación σ^* y β^* de la distribución a posteriori de σ y β. Estos parámetros de imputación se utilizan para generar valores imputados que respeten la incertidumbre en las estimaciones de los parámetros de la regresión.

En la imputación múltiple, estos valores imputados se generan para cada conjunto de datos, y luego los resultados de cada conjunto de datos imputado se combinan para generar una estimación final que tiene en cuenta tanto la variabilidad dentro de los conjuntos de datos imputados como la variabilidad entre ellos.

Esto asegura que la estimación final refleje tanto la incertidumbre en la estimación de los parámetros de la regresión como la incertidumbre debido a los valores faltantes.

1) Se genera σ^* como

$$\sigma^* = \hat{\sigma}\sqrt{(n_{\text{obs}} - k)/\text{g}} \tag{105}$$

donde g es una realización aleatoria de una distribución $\chi^2_{n_{\text{obs}}-k}$.

2) se genera $\boldsymbol{\beta}^*$ como

$$\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}} + \frac{\sigma^*}{\hat{\sigma}}\mathbf{u_1}V^{\frac{1}{2}} \tag{106}$$

donde $\mathbf{u_1}$ es una fila de k realizaciones independientes de una distribución normal estandar y $V^{\frac{1}{2}}$ es la descomposición de cholesky de $\mathbf{V}$

El valor imputado y_i^* para cada observación faltante y_i se obtiene como

$$y_i^* = \boldsymbol{\beta}\mathbf{x}_i + u_{2i}\sigma^* \tag{107}$$

donde u_{2i} es una realización aleatoria proveniente de una distribución normal estándar.

9.2.4 Imputación multivariada para datos continuos

9.2.4.1 Modelos multivariados utilizando la distribución normal Si se asume que $Y = (Y_1, Y_2)$ sigue una distribución normal bivariada. La distribución condicional de Y_i dado Y_j (donde $i, j \in \{1, 2\}$ y $j \neq i$) es un modelo de regresión lineal normal univariado:

$$f(Y_1|Y_2) = \mathcal{N}(\beta_{01} + \beta_{11}Y_2, \tau_1^2), \tag{108}$$

$$f(Y_2|Y_1) = \mathcal{N}(\beta_{02} + \beta_{12}Y_1, \tau_2^2). \tag{109}$$

Suponga que $\mu = (\mu_1, \mu_2)^t$, $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$, y $\theta = (\mu, \Sigma)$. En los modelos de regresión lineal (108) y (109), los parámetros (β's y τ^2's) son reparametrizaciones de θ:

$$\beta_{01} = \mu_1 - \beta_{11}\mu_2, \quad \beta_{11} = \frac{\sigma_{12}}{\sigma_2^2}, \quad \tau_1^2 = \sigma_1^2 - \frac{\sigma_{12}^2}{\sigma_2^2},$$

$$\beta_{02} = \mu_2 - \beta_{12}\mu_1, \quad \beta_{12} = \frac{\sigma_{12}}{\sigma_1^2}, \quad \tau_2^2 = \sigma_2^2 - \frac{\sigma_{12}^2}{\sigma_1^2}.$$

Después de imponer una distribución a priori plana para θ tal como $\pi(\mu, \Sigma) \propto |\Sigma|^{-1}$, los componentes principales del algoritmo de imputación pueden esbozarse de la siguiente manera:

1. Comenzar con una estimación o suposición del parámetro θ, digamos $\theta^*(t)$ (donde $t = 0$ en la 1ra iteración). Reparametrizar $\theta^*(t)$ a $\beta^*(t)$'s y $\tau^{2*}(t)$'s usando la fórmula anterior.
2. Paso-I: para los casos faltantes en Y_1 donde sus valores se observan en Y_2, imputarlos usando el Modelo
   eqrefeq:fmnmv1 con los parámetros extraídos (es decir, $\beta^*(t)_{01}, \beta^*(t)_{11}, \tau^{2*}(t)_1$) del Paso 1; imputar los casos faltantes en Y_2 de manera similar usando el Modelo
   eqrefeq:fmnmv2; si ambas variables faltan, pueden generarse a partir de $\mathcal{N}_2(\mu^*(t), \Sigma^*(t))$.
3. Paso-P: una vez que Y_1 y Y_2 se completan a partir del Paso 2, extraer $\mu^*(t+1)$ y $\Sigma^*(t+1)$ como: $\frac{\Sigma^*(t+1)}{(n-1)} \sim$ Inverse-Wishart$(S(t+1), n-1)$, y $\mu^*(t+1) \sim \mathcal{N}_2(Y^{(t+1)}, \frac{\Sigma^*(t+1)}{n})$, donde $Y^{(t+1)}$ y $S(t+1)$ son la media muestral y la matriz de covarianza a partir de los datos completados $Y^{(t+1)}$, respectivamente; transformar $\theta^*(t+1)$ a $\beta^*(t+1)$'s y $\tau^{2*}(t+1)$'s.
4. Repetir los Pasos 2 y 3 hasta que la convergencia para $(\theta^*(t), Y^*(t)_{\text{mis}})$ se satisfaga, digamos en $t = T$. Las extracciones de $Y^*(T)_{\text{mis}}$ constituyen el primer ($m = 1$) conjunto de imputaciones, $Y_{\text{mis}}^{(m)}$.
5. Repetir los Pasos 1 a 4 de forma independiente M veces.

bajo un modelo normal bivariado se puede generalizar lo anterior a un modelo normal multivariado con p-dimensiones. Primero note que si dividimos Y de p-variado en dos partes, Y_1 y Y_2, con dimensiones p_1 y p_2 donde $p = p_1 + p_2$, entonces $Y = (Y_1, Y_2)^t \sim \mathcal{N}((\mu_1, \mu_2)^t, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix})$, donde μ_1 y μ_2 son vectores de $p_1 \times 1$ y $p_2 \times 1$, respectivamente, y Σ_{ij} son matrices de covarianza con dimensiones $p_i \times p_j$ para $i, j = 1, 2$. La distribución condicional de Y_2 dado Y_1 es una distribución normal p_2-variada: $f(Y_2|Y_1) \sim \mathcal{N}(\mu_{2|1}, \Sigma_{2|1})$,

donde $\mu_{2|1} = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(Y_1 - \mu_1)$ es la media predicha a partir de la regresión lineal p_2-variada de Y_2 en Y_1, y $\Sigma_{2|1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ es la matriz de covarianza residual de esa regresión.

Para llevar a cabo JM para más de dos variables, necesitamos dividir a los sujetos en grupos según qué variables tienen valores faltantes. Luego, para el grupo con Y_2-variables faltantes y Y_1-variables observadas, sus imputaciones pueden generarse usando el modelo de regresión lineal multivariado caracterizado por $f(Y_2|Y_1) $ como se describió anteriormente. El Paso-I del algoritmo DA pasa por todos estos grupos y luego es seguido por el Paso-P.

9.2.4.2 Modelos multivariados para datos no normales y mezcla de distribuciones Para variables continuas multivariadas que muestran una fuerte asimetría y/o colas pesadas, una estrategia JM conveniente es aplicar algunas transformaciones para hacer que las suposiciones de normalidad sean más plausibles y luego llevar a cabo la imputación del modelo normal multivariado en la escala transformada.

Otra estrategia efectiva para datos continuos no normales es utilizar modelos de mezcla. Los modelos de mezcla permiten un modelado conjunto flexible, ya que pueden reflejar automáticamente estructuras distribucionales y de dependencia complejas. Las mezclas finitas de distribuciones normales [mclachlan2000finite] son una herramienta poderosa para el modelado estadístico en una amplia variedad de situaciones. Fraley and Raftery (2002) y Marron and Wand (1992) mostraron que muchas distribuciones de probabilidad pueden ser bien aproximadas por modelos de mezcla finita. Por otra parte, Priebe (1994) mostró que con 10,000 observaciones, una densidad lognormal puede ser bien aproximada por una mezcla de 30 componentes normales.

Para esbozar la idea, sea $Y = (Y_1, \ldots, Y_n)$ que comprende n observaciones completas, donde cada Y_i es un vector p-dimensional. Supongamos que cada individuo pertenece exactamente a uno de los K componentes de mezcla latentes (grupos o clases). Para $i = 1, \ldots, n$, sea $Z_i \in \{1, \ldots, K\}$ que indica el componente del individuo i, y sea $\pi_k = \Pr(Z_i = k)$. Supongamos que $\pi = (\pi_1, \ldots, \pi_K)$ es el mismo para todos los individuos. Dentro de cualquier componente k, supongamos que las p variables siguen una distribución normal multivariada específica del componente con media μ_k y varianza Σ_k. Sea $\theta = (\mu, \Sigma, \pi)$, donde $\mu = (\mu_1, \ldots, \mu_K)$ y $\Sigma = (\Sigma_1, \ldots, \Sigma_K)$. El modelo de mezcla finita se puede expresar como

$$Y_i|Z_i, \mu, \Sigma \sim \mathcal{N}(\mu_{Z_i}, \Sigma_{Z_i}), \tag{110}$$

$$Z_i|\pi \sim \text{Multinomial}(\pi_1, \ldots, \pi_K). \tag{111}$$

Al marginalizar sobre Z_i's, este modelo de mezcla es equivalente a

$$f(Y_i|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \Sigma_k). \tag{112}$$

Para completar una especificación bayesiana del modelo, se pueden imponer distribuciones previas comunes para θ. Por ejemplo, podemos especificar $\pi(\mu_k|\Sigma_k) \sim \mathcal{N}(\mu_0, \tau^{-1}\Sigma_k)$ y $\pi(\Sigma_k) \sim$ Inverse-Wishart(m, Λ) $(k = 1, \ldots K)$ para una media vectorial previa μ_0, un parámetro de precisión previo escalar τ, y grados de libertad previos m para la matriz de

covarianza previa Λ (A. E. Gelman et al. 2013). Consideremos además, $\pi(\pi_1, \ldots, \pi_K) \sim$ Dirichlet$(1, \ldots, 1)$, la distribución de Dirichlet.

Como una mezcla de distribuciones normales multivariadas, el modelo es lo suficientemente flexible como para capturar características distribucionales como asimetría y relaciones no lineales que una sola distribución normal multivariada no lograría codificar. Por ejemplo, cuando $K = 1$, el modelo de mezcla normal es el modelo normal típico. Al establecer $K = 2$ y $\mu_1 = \mu_2$ en los modelos (110) y (111), los datos con valores atípicos pueden considerarse como surgidos de dos clases: una contiene la mayoría de los valores normales y la otra comprende valores extremos/atípicos que tienen varianzas infladas. Esto también se conoce como el modelo normal contaminado en (R. J. A. Little and Rubin 2020b, capitulo 12). Además, aunque K a menudo se fija y se establece previamente, puede tratarse como desconocidos y obtenerse de manera impulsada por los datos mediante métodos bayesianos avanzados. Esta opción amplía aún más la flexibilidad y utilidad de los modelos de mezcla normal, especialmente para datos de alta dimensión (He, Zhang, and Hsu 2021b).

Es importante resaltar que los indicadores de agrupación Z_i's no se observan y son probabilísticos. Esta característica hace que el ajuste del modelo y la imputación sean bastante complicados. A menudo se requieren rutinas de código o computacionales específicas. Las aplicaciones de modelos de mezcla normal para imputaciones de datos faltantes se pueden encontrar, por ejemplo, en Elliott and Stettler (2007), Böhning et al. (2007) y H.-J. Kim et al. (2014).

En el mismo espíritu que los modelos de mezcla normal, otra estrategia JM para datos no normales es imponer modelos que puedan acomodar características no normales de los datos utilizando parámetros adicionales (es decir, además de μ y Σ) e incluir modelos normales como casos especiales. Por ejemplo, Liu (1995) desarrolló modelos de imputación que asumen una familia t-multivariada. Es bien sabido que en la familia t, los grados de libertad ν controlan el comportamiento de la cola de la distribución; a medida que $\nu \to \infty$, la distribución t converge a una distribución normal. He and Raghunathan (2012) consideraron una extensión multivariada de la familia gh Tukey (1977), que es una transformación de una variable normal estándar para acomodar diferentes asimetrías y alargamientos de la distribución de variables no normales, y la transformación está controlada por varios parámetros desconocidos. Además de los modelos normales multivariados contaminados, (R. J. A. Little and Rubin 2020b, capitulo 12) proporcionaron algunos ejemplos sobre la familia de modelos normales multivariados ponderados. También similar a los modelos de mezcla normal, un desafío técnico de usar estos modelos para la imputación es que las distribuciones posteriores de los parámetros pueden ser bastante complejas. El algoritmo DA correspondiente podría necesitar código específico o necesitar ser ejecutado con la ayuda de paquetes de software bayesianos.

9.2.5 Imputación de variables binarias

En el caso de variables binarias, es posible utilizar un modelo logistico de la forma

$$\text{logit}P(Y_{\text{obs}} = 1|\mathbf{x}\beta) = \boldsymbol{\beta x} \tag{113}$$

Sea $\hat{\beta}$ la estimación del parametro del modelo ajustado a los individuos con la información observada Y_{obs} y su matriz de varianzas covarianzas $\mathbf{V}$. Sea $\boldsymbol{\beta}^*$ una realización de la

distribución aposteriori de $\boldsymbol{\beta}$ aproximada por MVN$(\hat{\boldsymbol{\beta}}, \mathbf{V})$

Para cada observación perdida y_{miss} tomamos $p^* = [1 + \exp(-\boldsymbol{\beta}^* \mathbf{x}_i)]^{-1}$ y generamos la imputación y_i^* como

$$y_i^* = \begin{cases} 1 & \text{si } \ u_i < p_i^* \\ \\ 0 & \text{En otro caso.} \end{cases} \tag{114}$$

Donde u_i es una realización aleatoria de una distribución uniforme en $(0, 1)$

9.2.6 Imputación de variables categóricas no ordenadas

En el caso de variables categóricas no ordenadas con $L > 2$ categorías pueden ser modeladas usando una regresión logística multinomial, donde cada categoría tiene una regresión logistica que se compara con otra categoría determinada (digamos $l = 1$)

$$P(y_{\text{obs}} = l|\mathbf{x}, \boldsymbol{\beta}) = \left[\sum_{l'=1}^{L} \exp(\beta_{l'} \mathbf{x})\right]^{-1} \exp(\boldsymbol{\beta}_l \mathbf{x}) \tag{115}$$

donde β_l es un vector de dimension $k = \dim(\mathbf{x})$ y $\boldsymbol{\beta_1} = 0$.

Sea $\boldsymbol{\beta}^*$ una realización proveniente de una distribución normal aproximada a la distribución aposteriori de $\boldsymbol{\beta} = (\boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_L)$ vector de $k(L-1)$. Para cada observación perdida y_{miss}, sea $p_{il}^* = P(y_i = l|\mathbf{x_i}, \boldsymbol{\beta^*})$ la probabilidad de estar en cada categoría y $c_{il} = \sum_{l'=1}^{l} p_{il'}^*$. Se define cada valor imputado y_i^* como

$$y_i^* = 1 + \sum_{l'=1}^{L} I(u_i > c_{il}) \tag{116}$$

donde u_i es una realización aleatoria proveniente de una distribución uniforme en $(0, 1)$ y $I(u_i > c_{il}) = 1$ si $u_i > c_{il}$. 0 en otro caso.

9.2.7 Imputación multivariada para datos categóricos

9.2.8 Imputación mixta de variables categóricas y continuas

9.3 Analisis de sensibilidad

9.4 Consideraciones de imputar en encuestas complejas

Según Medina and Galván (2007), es importante recordar que los métodos de imputación presuponen ciertas características en la distribución de los datos faltantes, pero no abordan explícitamente el mecanismo que condujo a la selección de las unidades de observación. De forma incorrecta, estos métodos suelen asumir que los datos provienen de una muestra aleatoria y que todas las unidades tienen igual probabilidad de ser seleccionadas.

Las encuestas de hogares se realizan bajo diseños de muestreo complejos. Algunos autores, como Binder (1996) y Binder and Sun (1996), han planteado dudas sobre la validez de los métodos de imputación múltiple en tales contextos.

Se reconoce que la ausencia de datos es un problema inherente en todas las encuestas, y es habitual buscar procedimientos para completar la información. A pesar de ello, los procedimientos existentes se enfocan principalmente en analizar el patrón de datos faltantes, sin considerar que las unidades de observación pueden tener diferentes probabilidades de selección.

Por otra parte, es importante considerar el hecho de que los métodos de imputación múltiple asumen que los datos observados y/o completos siguen cierta distribución, pero bajo el paradigma de las encuestas, Kish (1965) es conocido por enfatizar que en el muestreo de poblaciones no son las observaciones individuales las que siguen una distribución, sino más bien las probabilidades de selección asignadas a cada elemento en la muestra. Esto es un principio fundamental en el diseño de muestreo y análisis de encuestas, que ayuda a garantizar que la muestra sea representativa de la población en su conjunto.

Incluso en casos donde la falta de respuesta es baja, Medina and Galván (2007) aconsejan analizar los ponderadores asociados a los datos faltantes. Puede suceder que un pequeño número de hogares en la muestra representen una porción importante de la población total, y un criterio de imputación inadecuado podría introducir sesgos difíciles de identificar y evaluar.

Los autores enfatizan que en los diseños de muestreo complejos, la selección de observaciones depende del método de estratificación y conglomeración del marco de muestreo, así como del vector de ponderaciones asociado a las diferentes unidades en la muestra.

Además, Medina and Galván (2007) señalan que la estratificación, la conglomeración y las ponderaciones deben tenerse en cuenta a la hora de imputar datos. En el contexto de una encuesta de hogares, la imputación no solo debe considerar el patrón de datos faltantes, sino también las probabilidades de selección de las unidades de observación.

Este enfoque aborda algunas de las limitaciones de los métodos de imputación tradicionales. Por ejemplo, al considerar las probabilidades de selección, se puede mitigar el riesgo de introducir sesgos en las estimaciones debido a la sobre o subrepresentación de ciertos grupos en la muestra.

Asimismo, al tener en cuenta la estratificación y la conglomeración, se pueden preservar las correlaciones entre las unidades dentro de cada estrato o conglomerado, que a menudo se pierden en los métodos de imputación que tratan cada unidad de observación de forma independiente. Sin embargo, también se debe tener en cuenta que, independientemente del método de imputación utilizado, siempre habrá cierta incertidumbre asociada con la imputación de datos faltantes. Por lo tanto, es importante manejar con cuidado los datos imputados y tener en cuenta esta incertidumbre al hacer inferencias a partir de los datos.

A pesar de que Binder and Sun (1996) demuestran que bajo el supuesto de un diseño de muestreo aleatorio simple y sin remplazo se pueden generar estimaciones precisas para medias y totales, siempre que se utilicen métodos bayesianos (bootstrap), Binder (1996) conjetura que la imputación múltiple no es adecuada en diseños complejos en los que existen más de una etapa de selección, conglomeración y probabilidades de selección desiguales puesto que las expresiones que se deben aplicar para la estimación de la variaza se complejizan. A pesar

de ello, para los autores que proponen la imputación múltiple no resulta una preocupación cómo dichas imputaciones afectan las estimaciones finales ante un muestreo multietápico.

Medina and Galván (2007) también enfatizan que los procedimientos de imputación no deben ser considerados como una solución definitiva para la falta de datos, sino como una herramienta que permite el manejo de los datos faltantes de una manera más rigurosa y estructurada. Los investigadores deben ser conscientes de las suposiciones subyacentes en cada método de imputación y su posible impacto en las conclusiones derivadas de los datos imputados. El autor señala además que aún "persiste el desafío de desarrollar algoritmos de imputaciones robustos que tengan en cuenta el diseño de la muestra y las probabilidades de selección de las observaciones".

La aplicación de métodos de imputación en diseños de muestreo complejos requiere un cuidado adicional. Es importante recordar que estos métodos deben adaptarse al diseño de muestreo particular y a la estructura de los datos faltantes. No todos los métodos de imputación son adecuados para todos los tipos de datos o diseños de muestreo.

Finalmente, los investigadores deben ser conscientes de que incluso los métodos de imputación más sofisticados no pueden reemplazar completamente los datos faltantes. A pesar de las técnicas de imputación, siempre existe el riesgo de sesgo debido a la falta de datos. Por lo tanto, es fundamental minimizar la cantidad de datos faltantes en la etapa de diseño y recolección de datos, y tratar los datos faltantes de manera adecuada en la etapa de análisis.

10 Anexo: Definiciones y aplicaciones en R

10.1 Deiniciones

En el lenguaje de programación R, utilizamos herramientas como `tidyverse` y el paquete R de `naniar` para enseñar a manejar y analizar datos faltantes de manera efectiva. El paquete `naniar` es una herramienta muy útil para explorar, visualizar y manejar valores faltantes en R.

```
library(tidyverse)
library(naniar)
```

La estadística Gertrude Mary Cox (Monroe and McVay (1980)) dijo una vez: "Lo mejor que se puede hacer con los datos faltantes es no tener ninguno". Si bien esto es cierto, no es el mundo en el que vivimos. Trabajar con datos del mundo real significa trabajar con datos faltantes. Para ser un gran analista, necesitamos saber cómo lidiar con los valores faltantes. Comprender cómo funcionan los datos faltantes es importante, ya que pueden tener efectos inesperados en tu análisis. Por ejemplo, ajustar un modelo lineal en datos con valores faltantes elimina fragmentos de datos. Esto significa que tus decisiones no se basan en la evidencia correcta. Reemplazar los valores faltantes, lo que se llama imputación, debe hacerse con mucho cuidado, ya que insertar solo la media puede llevar a estimaciones y decisiones deficientes.

En este apartado aprenderemos sobre qué son los valores faltantes, cómo encontrar datos faltantes, cómo manipular y limpiar datos faltantes, por qué faltan datos y cómo imputar valores faltantes. Por lo tanto, asumiremos que tenemoss experiencia básica a intermedia con

R, experiencia en la creación de gráficos utilizando `ggplot2`, experiencia en el uso de `dplyr` para manipular datos y experiencia en ajustar modelos lineales en R.

¿Qué son los valores faltantes?

Antes de comenzar, debemos definir los valores faltantes. Los valores faltantes son valores que deberían haberse registrado, pero no lo fueron. Pensemos en esto de esta manera: puedes no haber registrado accidentalmente que viste un pájaro, esto es un valor faltante. Esto es diferente a registrar que no se observaron pájaros. `R` almacena los valores faltantes como `NA`, que significa no disponible.

¿Cómo puedo verificar si tengo valores faltantes?

Los valores faltantes no saltan y gritan "¡Estoy aquí!". Por lo general, están ocultos, como una aguja en un pajar. Para detectar valores faltantes, usaresmo `any_na`, que devuelve `TRUE` si hay valores faltantes y `FALSE` si no los hay. `are_na` pregunta "¿son estos `NA`?" y devuelve `TRUE/FALSE` para cada valor. are_na nos muestra 3 valores `TRUE`, que corresponden a 3 valores faltantes. Para evitar contar cada `TRUE` manualmente, `n_miss` cuenta el número de valores faltantes. Y `prop_miss` proporciona la proporción de valores faltantes, lo que proporciona un contexto importante: Por ejemplo, el 50% de los datos está faltando.

¿Qué sucede cuando mezclamos valores faltantes con nuestros cálculos? Necesitamos saber qué sucede para poder estar preparados para encontrar estos casos. La regla general es: Los cálculos con `NA` devuelven `NA`. Digamos que tienes la altura de tres amigos: Sophie, Dan y Fred. La suma de sus alturas devuelve `NA`, esto se debe a que no conocemos la suma de un número y `NA`.

Hay algunas "trampas" que debes tener en cuenta al trabajar con datos faltantes: Por ejemplo, `NaN` significa "Not a Number" (No es un número) y se obtiene de operaciones como la raíz cuadrada de -1. **R** interpreta `NaN` como un valor faltante. `NULL` es un valor vacío pero no es faltante. Esto es sutilmente diferente de los valores faltantes: un cubo vacío no tiene agua faltante. `Inf` es un valor infinito, y se obtiene de ecuaciones como 10 dividido por 0 y no es faltante.

Por último, debeos tener cuidado con las declaraciones condicionales con valores faltantes. Por ejemplo, `NA` o `TRUE` es `TRUE`. `NA` o `FALSE` es `NA`. `NA + NaN` es `NA`. `NaN + NA` es `NaN`.

10.2 Usando y encontrando valores faltantes

Al trabajar con datos faltantes, hay algunos comandos con los que deberías estar familiarizado - en primer lugar, debes poder identificar si hay valores faltantes y dónde se encuentran.

Usando las herramientas `any_na()` y `are_na()`, identifica qué valores faltan.

```
# creamos x, un vector, con valores NA, NaN, Inf, ".", y "missing"
x <- c(NA, NaN,Inf, ".", "missing")

# Usamos any_na() y are_na() para explorar los valores missings
any_na(x)
```

```
## [1] TRUE
```

```
are_na(x)
```

```
## [1]  TRUE FALSE FALSE FALSE FALSE
```

¿Cuántos valores faltantes hay?

Una de las primeras cosas que desearás comprobar en un nuevo conjunto de datos es si existen valores faltantes y cuántos hay.

Podrías usar `are_na()` y contar los valores faltantes, pero la forma más eficiente de contarlos es usar la función `n_miss()`. Esto te dirá el número total de valores faltantes en los datos.

Luego puedes encontrar el porcentaje de valores faltantes en los datos con la función `pct_miss`. Esto te dirá el porcentaje de valores faltantes en los datos.

También puedes encontrar el complemento de estos valores, cuántos valores completos hay, usando `n_complete` y `pct_complete`.

```
# Usando el dataframe de ejemplo de estatuta (heights) y pesos (weights) dat_hw
dat_hw <- read.table("dealing/dat_hw.txt", h=T,  dec=".")
```

```
# Usa n_miss() para contar el numero total de valores missing en dat_hw
naniar::n_miss(dat_hw)
```

```
## [1] 30
```

```
# Usa n_miss() en dat_hw$weight para contar el numero total de valores missing
naniar::n_miss(dat_hw$weight)
```

```
## [1] 15
```

```
# Usa n_complete() en dat_hw para contar el numero total de valores completos
n_complete(dat_hw)
```

```
## [1] 170
```

```
# Usa n_complete() en dat_hw$weight para contar el numero total de valores completos
n_complete(dat_hw$weight)
```

```
## [1] 85
```

```
# Usamos prop_miss() y prop_complete() en dat_hw para contar el numero total de valore
prop_miss(dat_hw)
```

```
## [1] 0.15
```

```
prop_complete(dat_hw)
```

```
## [1] 0.85
```

Resumiendo la ausencia de datos

Ahora que comprendes el comportamiento de los valores faltantes en R y cómo contarlos, escalaremos nuestros resúmenes para casos (filas) y variables, utilizando `miss_var_summary()`

y `miss_case_summary()`, y también exploraremos cómo se pueden aplicar a grupos en un dataframe utilizando la función `group_by` de `dplyr`.

```
# Summarise missingness in each variable of the `airquality` dataset
miss_var_summary(airquality)
```

```
## # A tibble: 6 x 3
##   variable n_miss pct_miss
##   <chr>     <int>    <dbl>
## 1 Ozone        37    24.2
## 2 Solar.R       7     4.58
## 3 Wind          0     0
## 4 Temp          0     0
## 5 Month         0     0
## 6 Day           0     0
```

```
# Summarise missingness in each case of the `airquality` dataset
miss_case_summary(airquality)
```

```
## # A tibble: 153 x 3
##     case n_miss pct_miss
##    <int>  <int>    <dbl>
##  1     5      2     33.3
##  2    27      2     33.3
##  3     6      1     16.7
##  4    10      1     16.7
##  5    11      1     16.7
##  6    25      1     16.7
##  7    26      1     16.7
##  8    32      1     16.7
##  9    33      1     16.7
## 10    34      1     16.7
## # i 143 more rows
```

```
# Return the summary of missingness in each variable, grouped by Month, in the `airqua
airquality %>% group_by(Month) %>% miss_var_summary()
```

```
## # A tibble: 25 x 4
## # Groups:   Month [5]
##    Month variable n_miss pct_miss
##    <int> <chr>     <int>    <dbl>
##  1     5 Ozone         5     16.1
##  2     5 Solar.R       4     12.9
##  3     5 Wind          0      0
##  4     5 Temp          0      0
##  5     5 Day           0      0
##  6     6 Ozone        21     70
##  7     6 Solar.R       0      0
##  8     6 Wind          0      0
```

```
##  9     6 Temp          0      0
## 10     6 Day           0      0
## # i 15 more rows
```

```
# Return the summary of missingness in each case, grouped by Month, in the `airquality
airquality %>% group_by(Month) %>% miss_case_summary()
```

```
## # A tibble: 153 x 4
## # Groups:   Month [5]
##    Month  case n_miss pct_miss
##    <int> <int>  <int>    <dbl>
##  1     5     5      2       40
##  2     5    27      2       40
##  3     5     6      1       20
##  4     5    10      1       20
##  5     5    11      1       20
##  6     5    25      1       20
##  7     5    26      1       20
##  8     5     1      0        0
##  9     5     2      0        0
## 10     5     3      0        0
## # i 143 more rows
```

10.3 Evaluación de la no respuesta

10.3.1 Actividad

Considerando el siguien set de datos.

```
library(dplyr)
library(ggplot2)
biopics <- read.csv("curso_imputacion/biopics.csv")
```

Muestra las primeras 10 observaciones de los datos `biopics` y familiarízate con las variables.

```
# Muestra las primeras 10 observaciones
head(biopics, 10)
```

```
##    country year earnings sub_num sub_type sub_race non_white sub_sex
## 1       UK 1971       NA       1 Criminal     <NA>         0    Male
## 2    US/UK 2013   56.700       1    Other  African         1    Male
## 3    US/UK 2010   18.300       1  Athlete     <NA>         0    Male
## 4   Canada 2014       NA       1    Other    White         0    Male
## 5       US 1998    0.537       1    Other     <NA>         0    Male
## 6       US 2008   81.200       1    Other    other         1    Male
## 7       UK 2002    1.130       1 Musician    White         0    Male
## 8       US 2013   95.000       1  Athlete  African         1    Male
## 9       US 1994   19.600       1  Athlete     <NA>         0    Male
## 10   US/UK 1987    1.080       2   Author     <NA>         0    Male
```

```
# Obtiene el numero de valores perdidos por variable
biopics %>%
    is.na() %>%
    colSums()
```

```
##    country       year   earnings    sub_num   sub_type  sub_race non_white    sub_sex
##          0          0        324          0          0        197          0          0
```

10.3.2 Reconociendo los mecanismos de datos faltantes

En este ejercicio, se presentarán seis escenarios diferentes en los que faltan algunos datos. Intenta asignar a cada uno de ellos el mecanismo de datos faltantes más probable. Como recordatorio, aquí hay algunas pautas generales:

Si la razón de la falta de datos es puramente aleatoria, es MCAR. Si la razón de la falta de datos puede explicarse por otra variable, es MAR. Si la razón de la falta de datos depende del valor faltante en sí mismo, es MNAR.

Pérdida Completamente Aleatoria (MCAR)	Pérdida Aleatoria (MAR)	Pérdida no Aleatoria (NMAR)
Mientras se etiquetaba manualmente los datos, el etiquetador accidentalmente dejó algunas entradas sin etiquetar.	En una encuesta de salud, se observan datos faltantes en el peso. Se sospecha que los valores de la variable de peso faltan más para un género que para otro.	Es común que los simpatizantes de la extrema derecha no lo admitan en las encuestas electorales.
En un conjunto de datos que contiene resultados de exámenes escolares, algunos niños no tienen el resultado porque estaban enfermos y no asistieron al examen.	Está realizando el seguimiento de las ubicaciones de los visitantes de un sitio web. Si están utilizando una VPN (lo cual se sabe), el seguimiento no es confiable y a menudo se registran valores faltantes.	En las encuestas, las personas ricas tienen más probabilidades de no revelar sus ingresos.

Figura 1: alt text

10.3.2.0.1 Prueba t para perdida MAR Preparación de los datos

De los tres, MAR es posiblemente el más importante de detectar, ya que muchos métodos de imputación asumen que los datos son MAR. En este ejercicio práctico con `R`, buscaremos identificar si el patrón de pérdida es MAR.

Trabajaremos con los datos de la base `biopics`. El objetivo es probar si el número de valores faltantes en `earnings` difiere por género del sujeto. En este ejercicio, solo se preparan los datos para aplicar una *prueba t*. Primero, se crea una variable ficticia que indica la falta de datos en `earnings`. Luego, se divide por género filtrando los datos para mantener uno de los géneros y luego sacando la variable ficticia. Para filtrar, puede ser útil imprimir el `head()` de `biopics` en la consola y examinar la variable de género.

```
# Crea una variable dummy para la perdida en el gasto
biopics <- biopics %>%
  mutate(missing_earnings = is.na(earnings))

# Obtiene la perdida del gasto para hombres
```

```
missing_earnings_males <- biopics %>%
  filter(sub_sex == "Male") %>%
  pull(missing_earnings)

# Obtiene la perdida del gasto para mujeres
missing_earnings_females <- biopics %>%
  filter(sub_sex == "Female") %>%
  pull(missing_earnings)
```

Interpretación

En el ejercicio anterior, hemos preparado dos vectores con los valores faltantes de ingresos para cada sexo: `perdidos_gastos_hombres` y `perdidos_gastos_mujeres`. Ambos están disponibles en tu espacio de trabajo. Ahora es posible realizar la **prueba t** para verificar si sus medias difieren significativamente entre sí con el siguiente script

```
# Ejecuta el t-test
t.test(missing_earnings_males, missing_earnings_females)
```

```
##
##  Welch Two Sample t-test
##
## data:  missing_earnings_males and missing_earnings_females
## t = 1.1116, df = 294.39, p-value = 0.2672
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.03606549  0.12969214
## sample estimates:
## mean of x mean of y
## 0.4366438 0.3898305
```

El resultado muestra que no existe diferencia estadísticamente significativa ($\alpha > 0.05$) entre ambos grupos. Por lo tanto, se concluye que la perdida es MAR.

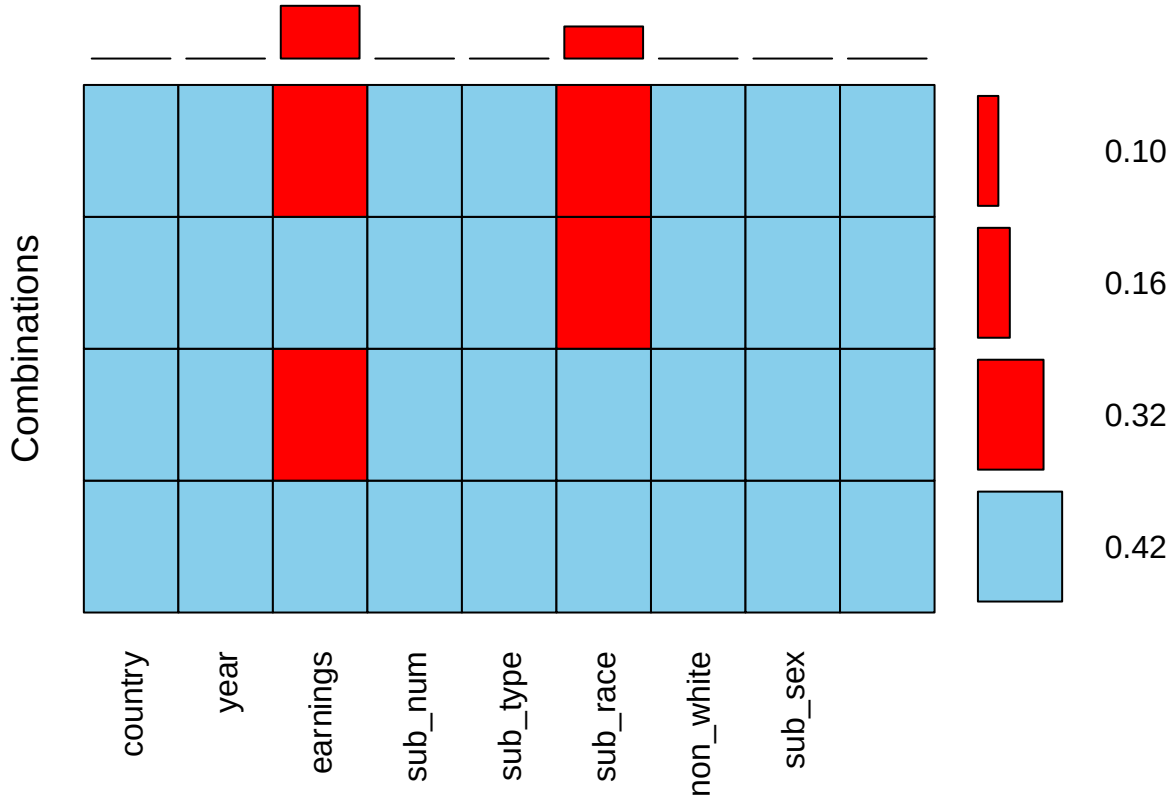
10.3.3 Aggregation plot

El gráfico de agregación proporciona la respuesta a la pregunta básica que uno puede hacer sobre un conjunto de datos incompleto: ¿en qué combinaciones de variables faltan datos y con qué frecuencia? Es muy útil para obtener una visión general de alto nivel de los patrones de ausencia de datos. Por ejemplo, hace visible inmediatamente si hay alguna combinación de variables que faltan juntas con frecuencia, lo que podría sugerir alguna relación entre ellas.

En este ejercicio, primero aplicaremos el gráfico de agregación para los datos de `biopics` y luego practicarás sacando conclusiones basadas en él. ¡Vamos a hacer algunos gráficos!

```
# Load the VIM package
library(VIM)

# Draw an aggregation plot of biopics
```

```
biopics %>%
    aggr(combined = TRUE, numbers = TRUE)
```



10.3.4 Cuestiones aclaratorias

Basado en el gráfico de agregación que acaba de crear, ¿cuál de las siguientes afirmaciones es falsa?

Posibles respuestas:

a. El 10% de las observaciones tienen valores faltantes tanto en `earnings` como en `sub_race`.

b. Hay más valores faltantes en `sub_race` que en `earnings`.

c. El 42% de las observaciones no tiene entradas faltantes.

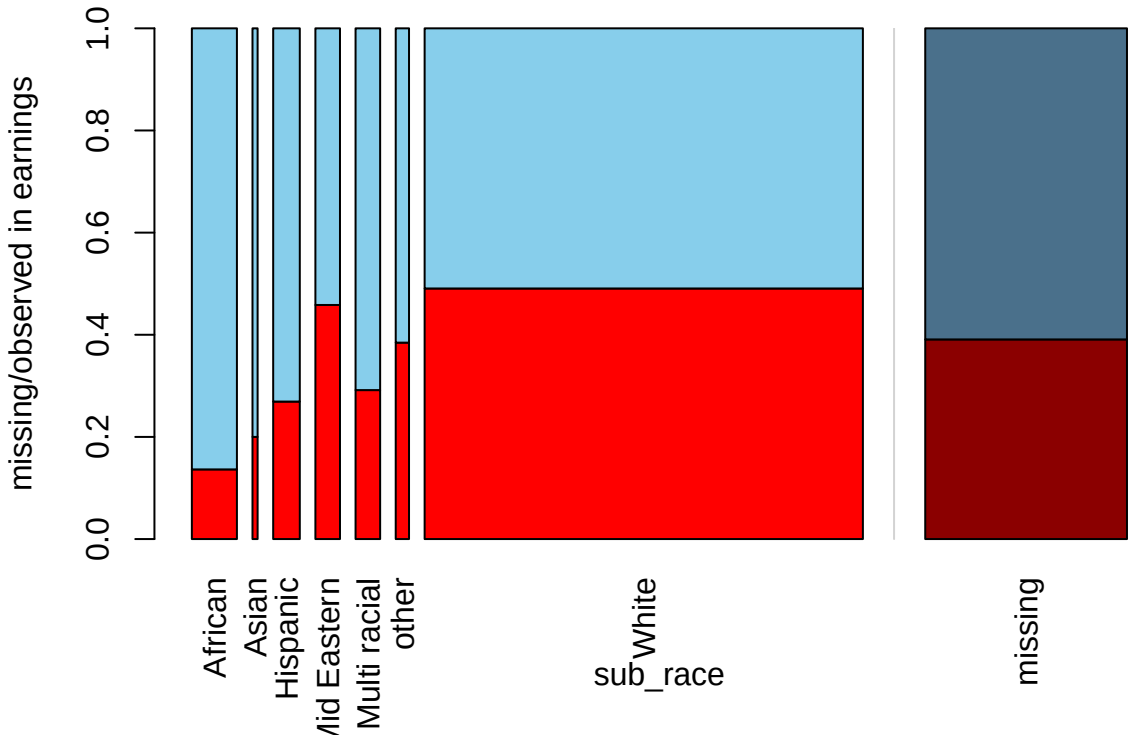
d. Exactamente dos variables en los datos `biopics` tienen valores faltantes.

10.3.5 gráfico de Mosaico

El gráfico de agregación que has dibujado en el ejercicio anterior te dio una visión general de alto nivel de los datos faltantes. Si estás interesado en la interacción entre variables específicas, un gráfico de Mosaico es el camino a seguir. Te permite estudiar el porcentaje de valores faltantes en una variable para diferentes valores de la otra, lo cual es conceptualmente muy similar a los test t que has estado realizando en la lección anterior.

En este ejercicio, dibujarás un gráfico de Mosaico para investigar el porcentaje de datos faltantes en `earnings` para diferentes categorías de `sub_race`. ¿Hay más datos faltantes en `earnings` para algunas razas específicas del personaje principal de la película? ¡Vamos a descubrirlo! El paquete `VIM` ya ha sido cargado para ti.

```
# Draw a spine plot to analyse missing values in earnings by sub_race

biopics %>%
    dplyr::select(sub_race,earnings) %>%
    spineMiss()
```



Cuestiones aclaratorias

Basándose en la gráfica de Mosaico que acabas de crear, ¿cuál de las siguientes afirmaciones es falsa?

Opciones de respuesta:

a. En la gran mayoría de las películas, el personaje principal es blanco.

b. Cuando el sujeto principal es africano, es más probable que tengamos información completa sobre las ganancias.

c. En lo que respecta a las ganancias y la subraza, los datos parecen ser MAR.

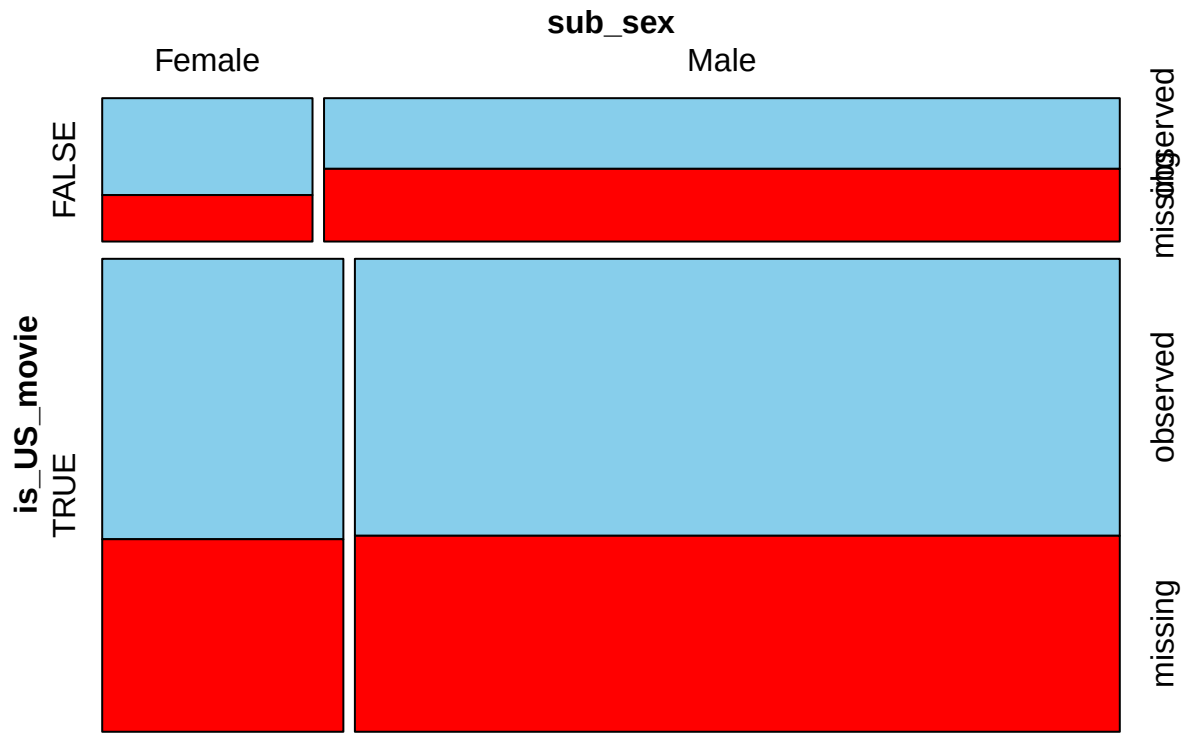
d. La raza que aparece con menos frecuencia en los datos tiene alrededor del 40% de las ganancias faltantes. (incorrecta)

10.3.6 Mosaic plot

La gráfica de Mosaico que hemos creado en el ejercicio anterior permite estudiar los patrones de datos faltantes entre dos variables a la vez. Esta idea se generaliza a más variables en forma de un gráfico de mosaico.

En este ejercicio, comenzarás por crear una variable ficticia que indique si Estados Unidos participó en la producción de cada película. Para hacer esto, utilizarás la función `grepl()`, que verifica si la cadena pasada como su primer argumento está presente en el objeto pasado como su segundo argumento. Luego, crearemos un gráfico de mosaico para ver si el género del sujeto se correlaciona con la cantidad de datos faltantes en `earnings` tanto para películas estadounidenses como no estadounidenses.

```
# Prepare data for plotting and draw a mosaic plot
biopics %>%
    # Create a dummy variable for US-produced movies
    mutate(is_US_movie = grepl("US", country)) %>%
    # Draw mosaic plot
    mosaicMiss(highlight = "earnings",
               plotvars = c("is_US_movie", "sub_sex"))
```



10.3.7 Olfateando el peligro de la imputación por la media

Uno de los métodos de imputación más populares es la imputación por media, en la cual los valores faltantes en una variable se reemplazan con la media de los valores observados en esa variable. Sin embargo, en muchos casos, este enfoque simple es una mala elección. A veces, una mirada rápida a los datos puede alertarnos sobre los peligros de la imputación por la media.

En esta etapa, trabajaremos con una submuestra de los datos del proyecto de *Atmósfera Tropical Oceánica* (`tao`). El conjunto de datos consiste en mediciones atmosféricas tomadas en dos períodos de tiempo diferentes en cinco ubicaciones distintas. Los datos vienen con el paquete `VIM`.

En este ejercicio, nos familiarizaremos con los datos y realizaremos un análisis simple que indicará cuáles podrían ser las consecuencias de la imputación por la media.

```
data(tao, package = "VIM")
names(tao)<-tolower(names(tao))
names(tao)<-sub("[.]", "_", names(tao))
names(tao)<-sub("[.]", "_", names(tao))

# Imprime las primeras 10 observaciones
head(tao, 10)
```

```
##    year latitude longitude sea_surface_temp air_temp humidity uwind vwind
## 1  1997        0      -110            27.59    27.15     79.6  -6.4   5.4
## 2  1997        0      -110            27.55    27.02     75.8  -5.3   5.3
## 3  1997        0      -110            27.57    27.00     76.5  -5.1   4.5
## 4  1997        0      -110            27.62    26.93     76.2  -4.9   2.5
## 5  1997        0      -110            27.65    26.84     76.4  -3.5   4.1
## 6  1997        0      -110            27.83    26.94     76.7  -4.4   1.6
## 7  1997        0      -110            28.01    27.04     76.5  -2.0   3.5
## 8  1997        0      -110            28.04    27.11     78.3  -3.7   4.5
## 9  1997        0      -110            28.02    27.21     78.6  -4.2   5.0
## 10 1997        0      -110            28.05    27.25     76.9  -3.6   3.5
```

```
# Obtiene el numero de valores perdidos por columna
tao %>%
  is.na() %>%
  colSums()
```

```
##             year         latitude        longitude sea_surface_temp
##                0                0                0                3
##         air_temp         humidity            uwind            vwind
##               81               93                0                0
```

```
# Calculate the number of missing values in air_temp per year
tao %>%
  group_by(year) %>%
  summarize(num_miss = sum(is.na(air_temp)))
```

```
## # A tibble: 2 x 2
##    year num_miss
##   <int>    <int>
## 1  1993        4
## 2  1997       77
```

10.3.8 Imputación de la media en temperatura

Imputar la media en la temperatura puede ser arriesgado. Si la variable que se está imputando está correlacionada con otras variables, esta correlación podría ser destruida por los valores imputados. Lo viste en el ejercicio anterior cuando analizaste la variable `air_temp`.

Para averiguar si estas preocupaciones son válidas, en este ejercicio realizarás una imputación de la media en `air_temp`, creando también un indicador binario para mostrar dónde se imputan los valores. Será útil en el siguiente ejercicio, cuando evaluarás el desempeño de tu imputación. ¡Vamos a completar esos valores faltantes!

```
tao_imp <- tao %>%
  # Create a binary indicator for missing values in air_temp
  mutate(air_temp_imp = ifelse(is.na(air_temp), TRUE, FALSE)) %>%
  # Impute air_temp with its mean
  mutate(air_temp = ifelse(is.na(air_temp), mean(air_temp, na.rm = TRUE), air_temp))

# Print the first 10 rows of tao_imp
head(tao_imp, 10)
```

```
##    year latitude longitude sea_surface_temp air_temp humidity uwind vwind
## 1  1997        0      -110            27.59    27.15     79.6  -6.4   5.4
## 2  1997        0      -110            27.55    27.02     75.8  -5.3   5.3
## 3  1997        0      -110            27.57    27.00     76.5  -5.1   4.5
## 4  1997        0      -110            27.62    26.93     76.2  -4.9   2.5
## 5  1997        0      -110            27.65    26.84     76.4  -3.5   4.1
## 6  1997        0      -110            27.83    26.94     76.7  -4.4   1.6
## 7  1997        0      -110            28.01    27.04     76.5  -2.0   3.5
## 8  1997        0      -110            28.04    27.11     78.3  -3.7   4.5
## 9  1997        0      -110            28.02    27.21     78.6  -4.2   5.0
## 10 1997        0      -110            28.05    27.25     76.9  -3.6   3.5
##    air_temp_imp
## 1         FALSE
## 2         FALSE
## 3         FALSE
## 4         FALSE
## 5         FALSE
## 6         FALSE
## 7         FALSE
## 8         FALSE
## 9         FALSE
## 10        FALSE
```

```
head(filter(tao_imp, air_temp_imp==TRUE))
```

```
##   year latitude longitude sea_surface_temp air_temp humidity uwind vwind
## 1 1997        0       -95             27.69 25.02925     79.8  -0.6   4.2
## 2 1997        0       -95             27.63 25.02925     74.5  -3.9   5.8
## 3 1997        0       -95             27.51 25.02925     76.3  -2.8   5.3
## 4 1997        0       -95             27.54 25.02925     81.6  -2.3   5.6
## 5 1997        0       -95             27.47 25.02925     81.9  -4.6   6.1
## 6 1997        0       -95             27.44 25.02925     74.2  -4.4   6.5
##   air_temp_imp
## 1         TRUE
## 2         TRUE
## 3         TRUE
## 4         TRUE
## 5         TRUE
## 6         TRUE
```

Nos damos cuenta que no tiene mucho sentido imputar por la media, ya que puede agregar inconsistencias entre las variables correlacionadas.

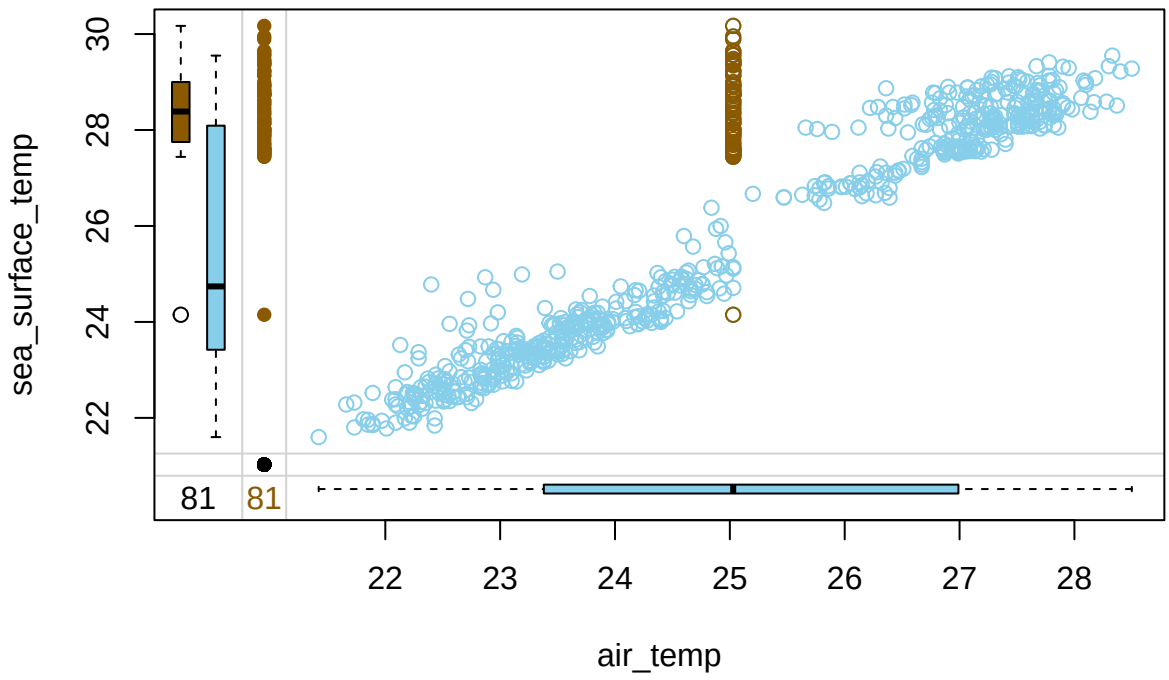
10.3.9 Evaluar la calidad de la imputación con un marginplot

En el último ejercicio, hemos imputado la media de `air_temp` y hemos agregado una variable indicadora para denotar cuáles valores fueron imputados, llamada `air_temp_imp`. Ahora es momento de ver qué tan bien funciona esto.

Al examinar los datos de `tao`, podríamos haber notado que también contiene una variable llamada `sea_surface_temp`, que razonablemente se esperaría que esté positivamente correlacionada con `air_temp`. Si ese es el caso, esperaríamos que estas dos temperaturas sean altas o bajas al mismo tiempo. Imputar la temperatura media del aire cuando la temperatura del mar es alta o baja rompería esta relación.

Para averiguarlo, en este ejercicio seleccionaremos las dos variables de temperatura y la variable indicadora y las usaremos para crear un `marginplot`.

```
# Creamos un marginplot de air_temp vs sea_surface_temp
tao_imp %>%
  select(air_temp, sea_surface_temp, air_temp_imp) %>%
  marginplot(delimiter = "imp")
```



10.3.10 Imputación por hot-deck

La imputación por hot-deck es un método simple que reemplaza cada valor faltante en una variable por el último valor observado en esa variable. Es muy rápido, ya que solo se necesita una revisión por los datos, pero en su forma más simple, hot-deck a veces puede romper las relaciones entre las variables.

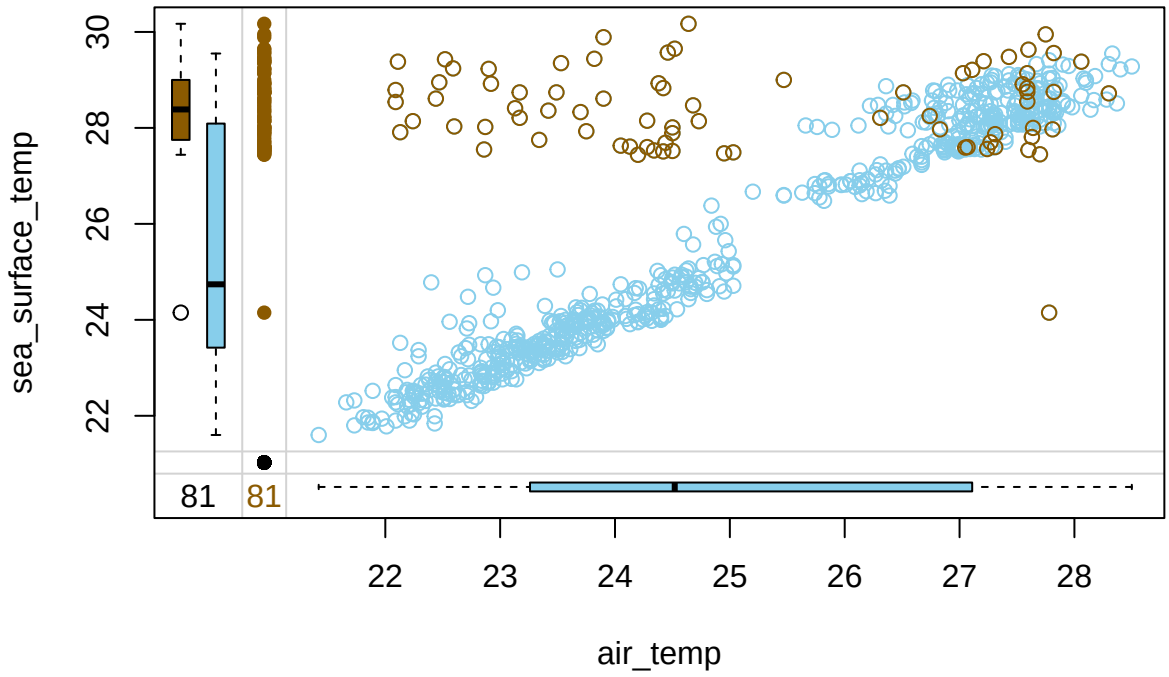
En este ejemplo, lo probaremos en el conjunto de datos `tao`. Imputaremos los valores faltantes en la columna de temperatura del aire `air_temp` por hot-deck y luego visualizaremos un gráfico de margen (`marginplot`) para analizar la relación entre los valores imputados y la columna de temperatura de la superficie del mar `sea_surface_temp`.

```
# Load VIM package
library(VIM)

# Impute air_temp in tao with hot-deck imputation
tao_imp <- hotdeck(tao, variable = "air_temp")

# Check the number of missing values in each variable
tao_imp %>%
    is.na() %>%
    colSums()
```

```
##             year         latitude        longitude sea_surface_temp
```

```
##                0                0                0                3
##        air_temp         humidity            uwind            vwind
##                0               93                0                0
##     air_temp_imp
##                0
```

```
# Draw a margin plot of air_temp vs sea_surface_temp
tao_imp %>%
    select(air_temp, sea_surface_temp, air_temp_imp) %>%
    marginplot(delimiter = "imp")
```



¿Se ve bien la imputación? Observa las observaciones en la parte superior izquierda del gráfico con los datos de `air_temp` imputados y los valores altos en `sea_surface_temp`. Estas observaciones deben haber sido precedidas por observaciones con bajos valores de `air_temp` en el `data frame`, y por lo tanto, después de la imputación hot-deck, terminaron siendo valores atípicos con `air_temp` bajos y `sea_surface_temp` altos.

10.3.11 Hot-deck trucos y consejos I: imputando dentro de dominios

Un truco que puede ayudar cuando la imputación por hot-deck rompe las relaciones entre las variables es imputar dentro de los dominios. Esto significa que si la variable a imputar está correlacionada con otra variable categórica, se puede ejecutar hot-deck por separado para cada una de sus categorías.

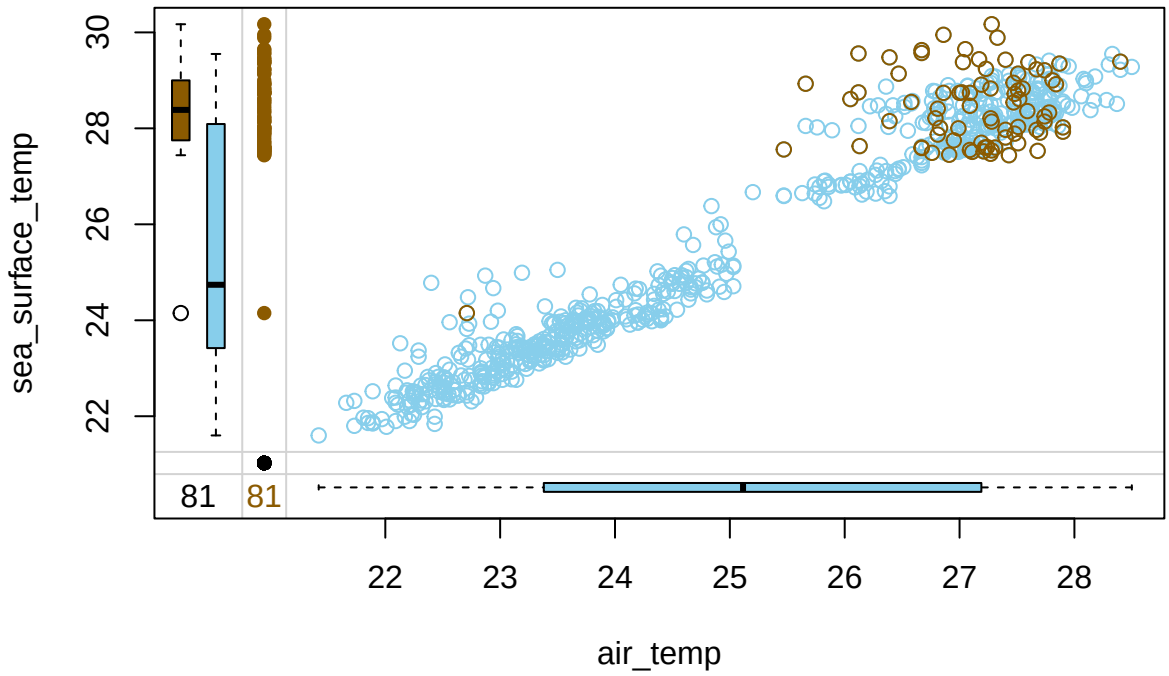
Por ejemplo, se podría esperar que la temperatura del aire dependa del tiempo, ya que

estamos viendo que las temperaturas promedio aumentan debido al calentamiento global. El indicador de tiempo que tenemos disponible en los datos de `tao` es una variable categórica, `year`. Primero, comprobaremos si la temperatura media del aire es diferente en cada uno de los dos años estudiados y luego ejecutaremos hot-deck dentro de los dominios de los años. Finalmente, volveremos a crear el `marginplot` para evaluar el rendimiento de la imputación.

```
# Calculate mean air_temp per year
tao %>%
    group_by(year) %>%
    summarize(average_air_temp = mean(air_temp, na.rm = TRUE))
```

```
## # A tibble: 2 x 2
##    year average_air_temp
##   <int>            <dbl>
## 1  1993             23.4
## 2  1997             27.1
```

```
# Hot-deck-impute air_temp in tao by year domain
tao_imp <- hotdeck(tao, variable = "air_temp", domain_var = "year")

# Draw a margin plot of air_temp vs sea_surface_temp
tao_imp %>%
    select(air_temp, sea_surface_temp, air_temp_imp) %>%
    marginplot(delimiter = "imp")
```



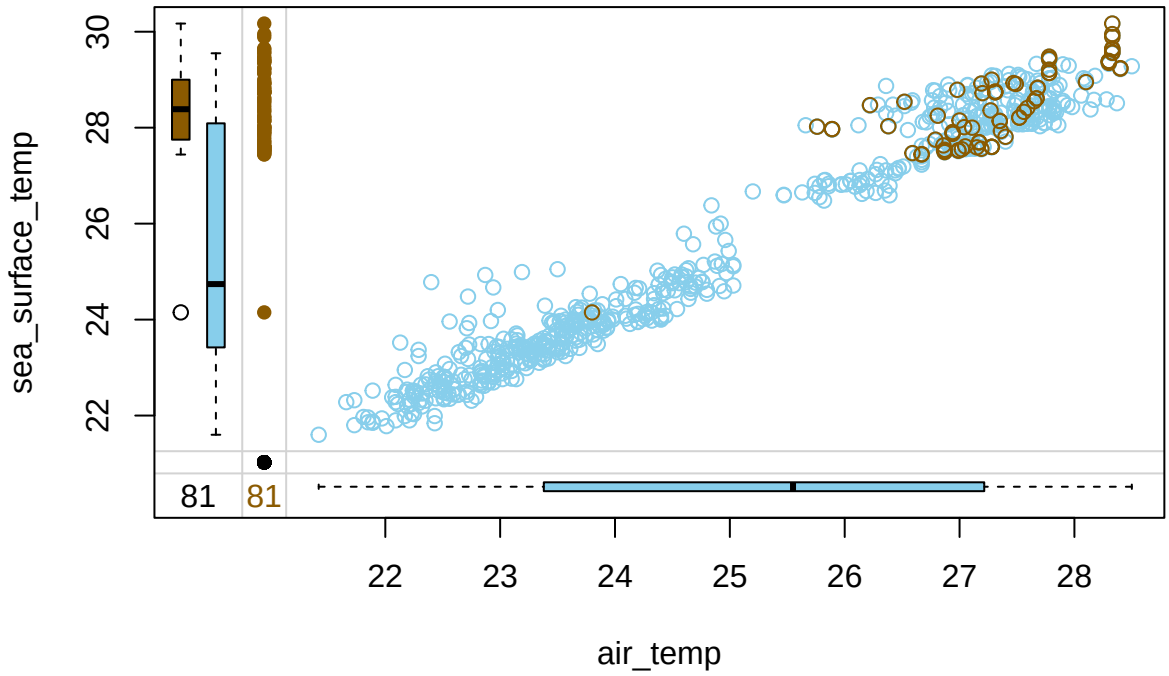
Los resultados se ven mucho mejor esta vez. Sin embargo, si observas la esquina superior derecha del gráfico, verás que la varianza en los valores imputados (naranja) es algo mayor que entre los valores observados (azul). Veamos si podemos mejorar aún más en el próximo ejercicio.

10.3.12 Hot-deck trucos y consejos II: ordenando por variables correlacionadas

Otro truco que puede mejorar el rendimiento de la imputación hot-deck es ordenar los datos por variables correlacionadas con la que queremos imputar.

Por ejemplo, en todos los `marginplot` que hemos estado usando recientemente, se ha visto que la temperatura del aire está fuertemente correlacionada con la temperatura de la superficie del mar, lo cual tiene mucho sentido. Podemos aprovechar este conocimiento para mejorar la imputación hot-deck. Si primero ordenamos los datos por `sea_surface_temp`, entonces cada valor imputado de `air_temp` vendrá de un donante con una `sea_surface_temp` similar.

```
# Hot-deck-impute air_temp in tao ordering by sea_surface_temp
tao_imp <- hotdeck(tao, variable = "air_temp", ord_var = "sea_surface_temp")

# Draw a margin plot of air_temp vs sea_surface_temp
tao_imp %>%
    select(air_temp, sea_surface_temp, air_temp_imp) %>%
    marginplot(delimiter = "imp")
```



Esta vez la imputación parece no afectar la relación entre las temperaturas del aire y la superficie del mar: si no fuera por los colores, probablemente no sabriamos cuáles son los valores imputados. La imputación hot-deck, posiblemente mejorada con la imputación por dominios o el ordenamiento, es un método rápido y sencillo que puede funcionar bien en muchas situaciones. Sin embargo, a veces puede ser necesario un enfoque más complejo.

10.3.13 Elegir el número de vecinos

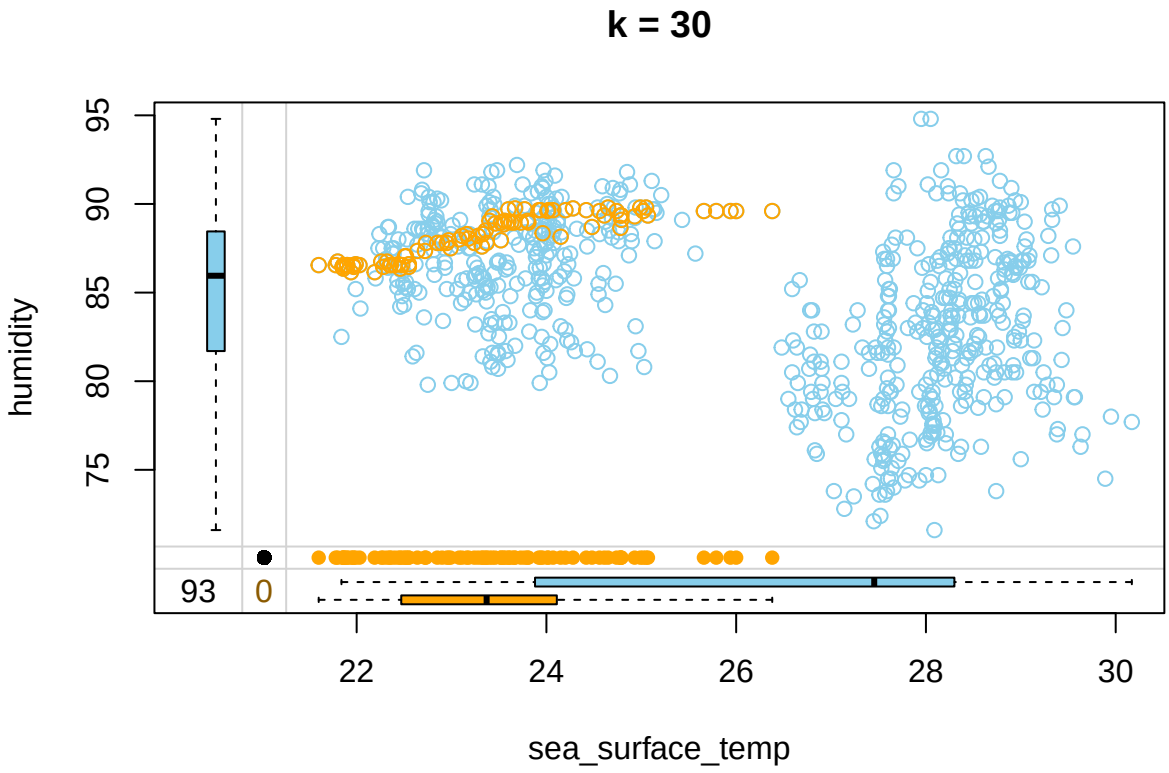
La imputación de k-Nearest-Neighbors (o kNN) imputa los valores faltantes en una observación en función de los valores que provienen de las k otras observaciones más similares a ella. El número de estas observaciones similares, llamadas vecinos, se consideran que es un parámetro que debe elegirse de antemano.

¿Cómo elegir k? Una forma es probar diferentes valores y ver cómo afectan las relaciones entre los datos imputados y observados.

Intentemos imputar `humidity` en los datos de `tao` utilizando tres valores diferentes de k y ver cómo se ajustan los valores imputados a la relación entre `humidity` y `sea_surface_temp`.

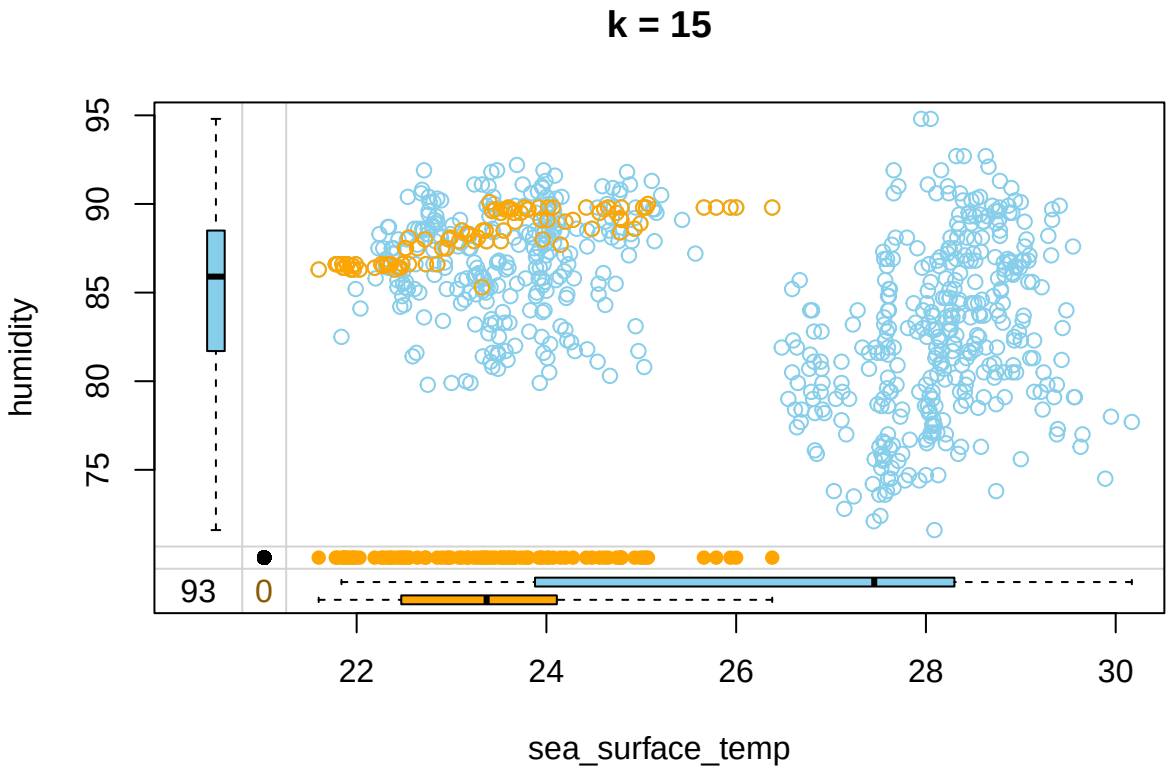
Imputamos `humidity` con la imputación de kNN usando 30 vecinos y visualizadolo mediante un `marginplot()` de `sea_surface_temp` vs `humidity`.

```
# Impute humidity using 30 neighbors
tao_imp <- kNN(tao, k = 30, variable = "humidity")

# Draw a margin plot of sea_surface_temp vs humidity
tao_imp %>%
    select(sea_surface_temp, humidity, humidity_imp) %>%
    marginplot(delimiter = "imp", main = "k = 30")
```



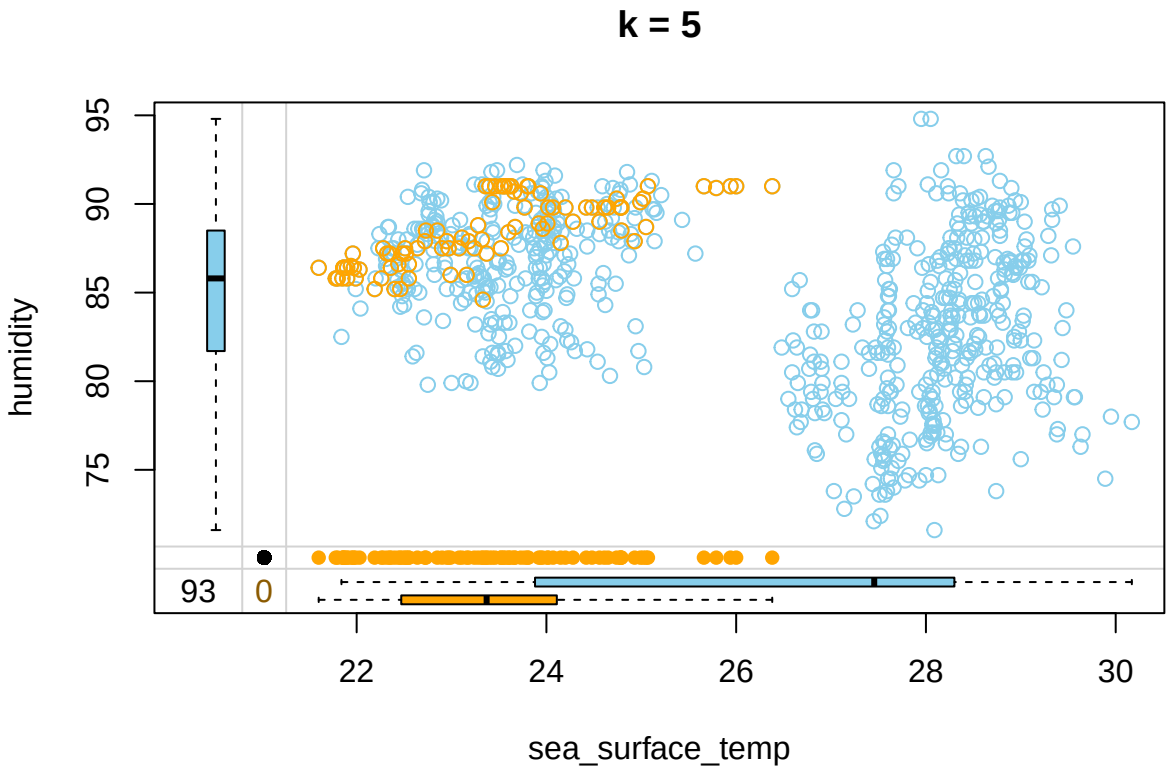
Ahora, imputamos `humidity` con imputación kNN usando 15 vecinos y vemos mediante el `marginplot` de `sea_surface_temp` vs `humidity`.

```
# Impute humidity using 15 neighbors
tao_imp <- kNN(tao, k = 15, variable = "humidity")

# Draw a margin plot of sea_surface_temp vs humidity
tao_imp %>%
    select(sea_surface_temp, humidity, humidity_imp) %>%
    marginplot(delimiter = "imp", main = "k = 15")
```



Finalmente, imputamos `humidity` con imputación kNN usando 5 vecinos y visualizando los resulados mediante `marginplot` de `sea_surface_temp` vs `humidity`.

```
# Impute humidity using 5 neighbors
tao_imp <- kNN(tao, k = 5, variable = "humidity")

# Draw a margin plot of sea_surface_temp vs humidity
tao_imp %>%
    select(sea_surface_temp, humidity, humidity_imp) %>%
    marginplot(delimiter = "imp", main = "k = 5")
```



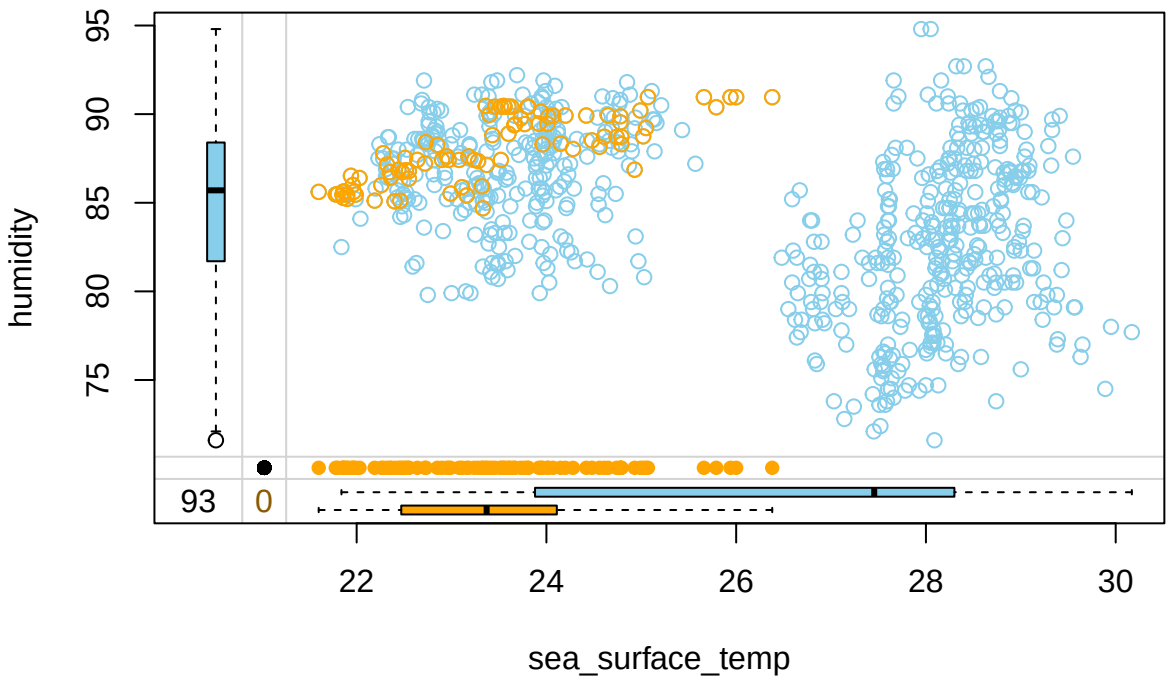
10.3.14 kNN trucos y consejos I: ponderando los donantes

Una variación de la imputación `kNN` que se aplica con frecuencia utiliza la llamada agregación ponderada por distancia. Lo que esto significa es que cuando agregamos los valores de los vecinos para obtener un reemplazo para un valor faltante, lo hacemos usando la media ponderada y las ponderaciones son las distancias invertidas de cada vecino. Como resultado, los vecinos más cercanos tienen más impacto en el valor imputado.

En este ejercicio, aplicamos la agregación ponderada por distancia mientras imputamos los datos de `tao`. Esto solo requerirá dar dos argumentos adicionales a la función `kNN()`.

```
# Load the VIM package
library(VIM)

# Impute humidity with kNN using distance-weighted mean
tao_imp <- kNN(tao,
               k = 5,
               variable = "humidity",
               numFun = weighted.mean,
               weightDist = TRUE)

tao_imp %>%
    select(sea_surface_temp, humidity, humidity_imp) %>%
```

```
    marginplot(delimiter = "imp")
```



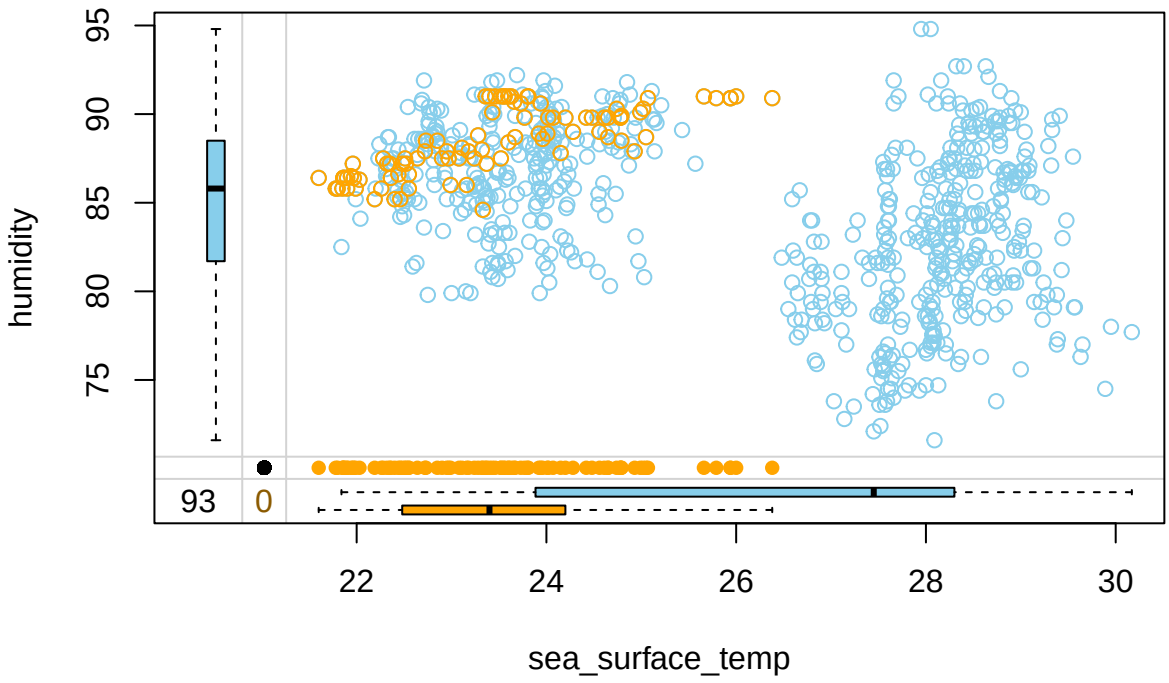
Trucos y consejos de kNN II: ordenar variables

Mientras el algoritmo de k-Nearest Neighbors recorre las variables en los datos para imputarlos, calcula las distancias entre observaciones utilizando otras variables, algunas de las cuales ya han sido imputadas en los pasos anteriores. Esto significa que si las variables ubicadas al principio de los datos tienen muchas valores faltantes, entonces el cálculo de la distancia posterior se basa en muchos valores imputados. Esto introduce ruido en el cálculo de la distancia.

Por esta razón, una buena práctica es ordenar las variables por el número de valores faltantes antes de realizar la imputación kNN. De esta manera, cada cálculo de distancia se basa en tantos datos observados y tan pocos datos imputados como sea posible.

```
# Get tao variable names sorted by number of NAs
vars_by_NAs <- tao %>%
  is.na() %>%
  colSums() %>%
  sort(decreasing = FALSE) %>%
  names()

# Sort tao variables and feed it to kNN imputation
tao_imp <- tao %>%
```

```
  select(vars_by_NAs) %>%
  kNN(k= 5)

tao_imp %>%
    select(sea_surface_temp, humidity, humidity_imp) %>%
    marginplot(delimiter = "imp")
```



El kNN que acabamos de programar debería ser más preciso y resistente a imputaciones defectuosas, así que recordemos ordenar las variables primero antes de realizar la imputación con kNN.

10.3.15 Imputación con regresión lineal

A veces, se puede utilizar el conocimiento del dominio, la investigación previa o simplemente el sentido común para describir las relaciones entre las variables en sus datos. En tales casos, la imputación basada en modelos es una gran solución, ya que permite imputar cada variable de acuerdo con un modelo estadístico que puede especificar uno mismo, teniendo en cuenta cualquier suposición que pueda tener sobre cómo las variables impactan entre sí.

Para variables continuas, una elección de modelo popular es la regresión lineal. Siempre puede incluir un cuadrado o un logaritmo de una variable en los predictores. En este caso, trabajaremos con el paquete `simputation` para ejecutar una sola imputación de regresión lineal en los datos `tao` y analizar los resultados.

```
# Lee la libreria simputation
library(simputation)

# Imputa air_temp y humidity con una regresion lineal
formula <- air_temp + humidity ~ year + latitude + sea_surface_temp
tao_imp <- impute_lm(tao, formula)
```

```
# Obtenemos el numero de valores missing por columna
tao_imp %>%
  is.na() %>%
  colSums()
```

```
##             year         latitude        longitude sea_surface_temp
##                0                0                0                3
##         air_temp         humidity            uwind            vwind
##                3                2                0                0
```

```
# Imprime las celdas de tao_imp en donde air_temp o humidity siguen missing
tao_imp %>%
  filter(is.na(air_temp) | is.na(humidity))
```

```
##   year latitude longitude sea_surface_temp air_temp humidity uwind vwind
## 1 1993        0       -95               NA       NA       NA  -5.6   3.1
## 2 1993        0       -95               NA       NA       NA  -6.3   0.5
## 3 1993       -2       -95               NA       NA     89.9  -3.4   2.4
```

La regresión lineal falla cuando al menos uno de los predictores está ausente. En este caso, fue `sea_surface_temp`. En el próximo ejercicio, lo solucionaremos inicializando los valores faltantes antes de ejecutar `impute_lm()`.

10.3.16 Inicialización de valores perdidos e iteración sobre variables

Como acabamos de ver, la ejecución de `impute_lm()` podría no llenar todos los valores perdidos. Para asegurarte de imputar todos ellos, debemos inicializar los valores perdidos con un método simple, como la imputación de hot-deck que de la sección anterior, que simplemente retroalimenta el último valor observado.

Además, una sola imputación generalmente no es suficiente. Se basa en los valores iniciales básicos y podría estar sesgada. Un enfoque adecuado es iterar sobre las variables, imputándolas una a la vez en las ubicaciones donde originalmente faltan.

En este ejercicio, primero inicializaremos los valores perdidos con la imputación de hot-deck y luego iteraremos cinco veces sobre `air_temp` y `humidity` de los datos `tao` para imputarlos con la regresión lineal.

```
# Inicializa los valores missing con hot-deck
tao_imp <- hotdeck(tao)

# Crea un indicador booleano desde donde air_temp y humidity son missing
missing_air_temp <- tao_imp$air_temp_imp
```

```
missing_humidity <- tao_imp$humidity_imp

for (i in 1:5) {
  # Define air_temp como NA en los lugares donde faltaban originalmente y re-imputa
  tao_imp$air_temp[missing_air_temp] <- NA
  tao_imp <- impute_lm(tao_imp, air_temp ~ year + latitude + sea_surface_temp + humidity
  # Define humidity como NA en los lugares donde faltan originalmente y  re-imputa
  tao_imp$humidity[missing_humidity] <- NA
  tao_imp <- impute_lm(tao_imp, humidity ~ year + latitude + sea_surface_temp + air_temp
}
```

Esa es una aproximación apropiada a la imputación basada en modelos que acabamos de codificar, pero, ¿cómo sabemos que 5 es el número adecuado de iteraciones para ejecutar?.

10.3.17 Detectando convergencia

¿Cuántas iteraciones son necesarias? Cuando los valores imputados no cambian con la nueva iteración, podemos detenernos.

Ahora extenderás nuestro código para calcular las diferencias entre las variables imputadas en las iteraciones subsiguientes. Para hacer esto, usaremos la función de cambio porcentual promedio absoluto (`mapc`), definida de la siguiente manera:

```
mapc <- function(a, b) {   mean(abs(b - a) / a, na.rm = TRUE) }
```

`mapc()` es una función que devuelve un solo número que te dice cuánto difiere b de a. La usaremos para verificar cuánto cambian las variables imputadas en las iteraciones siguientes. En base a esto, decidiremos cuántas iteraciones son necesarias.

Los indicadores booleanos `missing_air_temp` y `missing_humidity` son usados aquí, al igual que los datos de `tao_imp` inicializados con hot-deck.

```
mapc<- function(a, b) {
  mean(abs(b - a) / a, na.rm = TRUE)
}
```

```
diff_air_temp <- c()
diff_humidity <- c()

for (i in 1:5) {
  # Asigna el resultado de la iteración anterior (o inicialización) a prev_iter
  prev_iter <- tao_imp
  # Imputa air_temp y humidity en las ubicaciones que originalmente faltaban
  tao_imp$air_temp[missing_air_temp] <- NA
  tao_imp <- impute_lm(tao_imp, air_temp ~ year + latitude + sea_surface_temp + humidity
  tao_imp$humidity[missing_humidity] <- NA
  tao_imp <- impute_lm(tao_imp, humidity ~ year + latitude + sea_surface_temp + air_temp
  # Calcula MAPC para air_temp y humidity y los incluye a la iteración anterior de MAP
  diff_air_temp <- c(diff_air_temp, mapc(prev_iter$air_temp, tao_imp$air_temp))
```

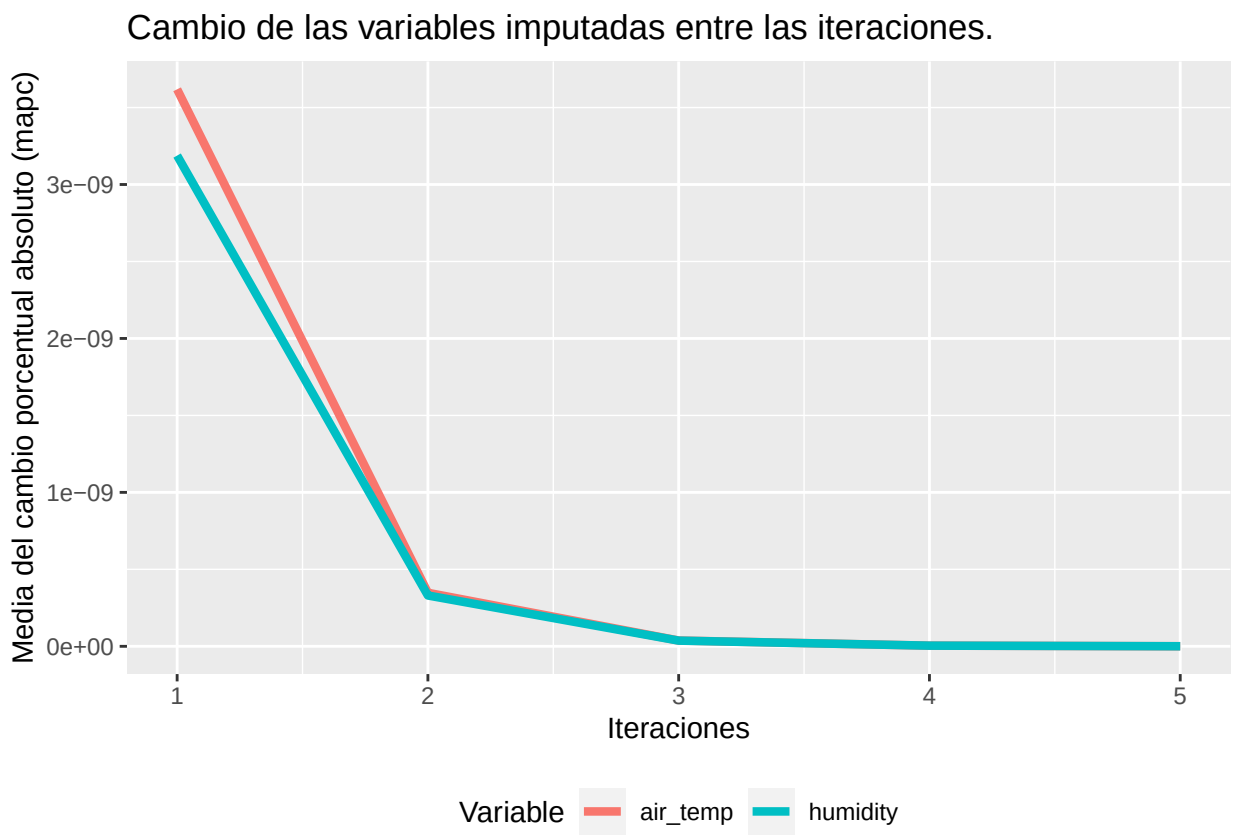
```
  diff_humidity <- c(diff_humidity, mapc(prev_iter$humidity, tao_imp$humidity))
}
```

¿Cuál es un número suficiente de iteraciones para ejecutar, según las diferencias almacenadas en `diff_air_temp` y `diff_humidity`?

Para responder a esta pregunta, puedemos imprimir los dos vectores en la consola y analizar los números, o trazarlos usando la función proporcionada: simplemente ejecutamos `plot_diffs(diff_air_temp, diff_humidity)` en la consola.

```
plot_diffs <- function(a, b) {
  data.frame("mapc" = c(a, b),
             "Variable" = c(rep("air_temp", length(a)),
                            rep("humidity", length(b))),
             "Iteraciones" = c(1:length(a), 1:length(b))) %>%
    ggplot(aes(Iteraciones, mapc, color = Variable)) +
    geom_line(size = 1.5) +
    ylab("Media del cambio porcentual absoluto (mapc)") +
    ggtitle("Cambio de las variables imputadas entre las iteraciones.") +
    theme(legend.position = "bottom")
}
```

```
plot_diffs(diff_air_temp, diff_humidity)
```



10.3.18 Imputación por regresión logística

Una opción popular para imputar variables binarias es la regresión logística. Desafortunadamente, no hay una función similar a `impute_lm()` que lo haga. Por eso crearemos una función para ello.

Llamemos a la función `impute_logreg()`. Su primer argumento será un data frame `df`, cuyos valores faltantes se han inicializado y solo contiene valores faltantes en la columna a imputar. El segundo argumento será una fórmula para el modelo de regresión logística.

La función hará lo siguiente:

1. Mantendrá las ubicaciones de los valores faltantes.
2. Construirá el modelo.
3. Realizará predicciones.
4. Reemplazará los valores faltantes con las predicciones.

No te preocupes por la línea que crea `imp_var` - esto es solo una forma de extraer el nombre de la columna a imputar de la fórmula.

```
impute_logreg <- function(df, formula) {
  # Extrae el nombre de la variable respuesta
  imp_var <- as.character(formula[2])
  # Guarda los lugares donde la respuesta es missing
  missing_imp_var <- is.na(df[imp_var])
  # Ajusta una regresion del modo logistica
  logreg_model <- glm(formula, data = df, family = binomial)
  # Predice la respuesta y la convierte 0s y 1s
  preds <- predict(logreg_model, type = "response")
  preds <- ifelse(preds >= 0.5, 1, 0)
  # Imputa los valores missing con las predicciones
  df[missing_imp_var, imp_var] <- preds[missing_imp_var]
  return(df)
}
```

La función está completamente operativa y se puede enchufar en el bucle sobre las variables que viste en la sección previa, al igual que `impute_lm()` del paquete `simputation`. Pronto, combinaremos estos dos para imputar tanto variables continuas como binarias. Pero antes, mejoraremos `impute_logreg()` para que reproduzca mejor la variabilidad en los datos imputados.

10.3.19 Crear una distribución condicional

Simplemente llamar a `predict()` en un modelo siempre devolverá el mismo valor para los mismos valores de los predictores. Esto da como resultado una pequeña variabilidad en los datos imputados. Para aumentarla y que la imputación replique la variabilidad de los datos originales, que podemos extraer de la distribución condicional. Esto significa que en lugar de siempre predecir 1 cuando el modelo devuelve una probabilidad mayor que `0.5` , podemos extraer la predicción de una distribución binomial descrita por la probabilidad devuelta por el modelo.

Trabajaremos en el código del ejercicio anterior. La siguiente línea fue eliminada:

```
preds <- ifelse(preds >= 0.5, 1, 0)
```

Nuestra tarea es llenar su lugar con la creación de una distribución binomial.

```
impute_logreg <- function(df, formula) {
 # Extrae el nombre de la variable respuesta
 imp_var <- as.character(formula[2])
 # Guarda las posiciones donde la respuesta es missing
 missing_imp_var <- is.na(df[imp_var])
 # Ajusta una regresion del modo logistico
 logreg_model <- glm(formula, data = df, family = binomial)
 # Predice la respuesta
 preds <- predict(logreg_model, type = "response")
 # Toma una muestra de las predicciones de la distribución binomial
 # preds <- ifelse(preds >= 0.5, 1, 0)
 preds <- rbinom(length(preds), size = 1, prob = preds)
 # Imputa los valores missing con las predicciones
 df[missing_imp_var, imp_var] <- preds[missing_imp_var]
 return(df)
}
```

Crear la distribución condicional hará que la variabilidad de los datos imputados sea mucho parecida a la del conjunto de datos observados originales. Con esta potente función en nuestras manos, ahora podemos diseñar un flujo de imputación basado en modelos que se encargue tanto de variables continuas como binarias.

10.3.20 Imputación basada en modelos con varios tipos de variables

En este ejercicio, combinaremos lo que hemos aplicado hasta ahora sobre imputación basada en modelos para imputar diferentes tipos de variables en los datos de `tao`.

Nuestra tarea es iterar sobre las variables como lo hemos hecho previamente e imputar dos variables:

`is_hot`, una nueva variable binaria que se creó a partir de `air_temp`, que es 1 si `air_temp` está a 26 grados o más y 0 de lo contrario; `humidity`, una variable continua con la que ya estamos familiarizados.

Tendremos que utilizar la función de regresión lineal que aprendimos antes, así como la función para la regresión logística.

```
tao$is_hot<-ifelse(tao$air_temp>= 26, 1,0)
```

```
# Inicializamos los valores missing con hot-deck
tao_imp <- hotdeck(tao)

# Creamos el indicador booleano desde donde is_hot y humidity son missing
missing_is_hot <- tao_imp$is_hot_imp
missing_humidity <- tao_imp$humidity_imp
```

```
for (i in 1:3) {
  # Define is_hot como NA en los lugares donde fue originalmente missing y re-imputa
  tao_imp$is_hot[missing_is_hot] <- NA
  tao_imp <- impute_logreg(tao_imp, is_hot ~ sea_surface_temp)
  # Define humidity como NA en los lugares donde fue originalmente missing y re-imputa
  tao_imp$humidity[missing_humidity] <- NA
  tao_imp <- impute_lm(tao_imp, humidity ~ sea_surface_temp + air_temp)
}
```

10.3.21 Imputación con bosques aleatorios

Un enfoque de aprendizaje automático para la imputación puede ser más preciso y más fácil de implementar en comparación con modelos estadísticos tradicionales. Primero, no requiere que especifiques relaciones entre variables. Además, los modelos de aprendizaje automático como los *random forest* son capaces de descubrir relaciones altamente complejas y no lineales y explotarlas para predecir valores faltantes.

En este ejercicio, usaremos el paquete `missForest`, que construye un bosque aleatorio separado para predecir valores faltantes para cada variable, uno por uno. Llamaremos a la función de imputación sobre los datos de películas, `biopics`, con los que hemos trabajado anteriormente y luego extraeremos los datos completos, así como los errores de imputación estimados.

```
# leemos nuevamente los datos
biopics <-read.csv("curso_imputacion/biopics.csv")
# cargamos la libreria
library(missForest)

# trasformación de character a factor
biopics <- type.convert(biopics, as.is=FALSE)


# imputa los datos de biopics usando missForest
imp_res <- missForest(biopics)

# Extrae los datos imputados y revisa por valores missing
imp_data <- imp_res$ximp
print(sum(is.na(imp_data)))
```

```
## [1] 0
```

```
# Extrae e imprime los errores de imputacion
imp_err <- imp_res$OOBerror
print(imp_err)
```

```
##      NRMSE        PFC
## 0.02057198 0.04698582
```

En el ejercicio anterior hemos extraído los errores de imputación estimados a partir de la salida de `missForest`. Esto te dio dos números:

el error cuadrático medio raíz normalizado (NRMSE) para todas las variables continuas; la proporción de entradas falsamente clasificadas (PFC) para todas las variables categóricas.

Sin embargo, podría darse el caso de que el modelo de imputación funcione muy bien para una variable continua y muy mal para otra. Para diagnosticar tales casos, basta con decirle a `missForest` que produzca estimaciones de error por variable. Esto se hace estableciendo el argumento `variablewise` en `TRUE`.

```
# Imputa los datos de biopics con missForest calculando los errores por variable
imp_res <- missForest(biopics, variablewise = TRUE)

# Extrae e imprime los errores de imputacion
per_variable_errors <- imp_res$OOBerror
print(per_variable_errors)
```

```
##          PFC          MSE          MSE          MSE          PFC          PFC
##    0.0000000    0.0000000 1310.6849382    0.0000000    0.0000000    0.1950355
##          MSE          PFC
##    0.0000000    0.0000000
```

```
# Renombra las columnas para incluir el nombre de las variables
names(per_variable_errors) <- paste(names(biopics),
                                    names(per_variable_errors),
                                    sep = "_")

# Imprime los errores renombrados
print(per_variable_errors)
```

```
##   country_PFC      year_MSE  earnings_MSE   sub_num_MSE  sub_type_PFC
##     0.0000000     0.0000000  1310.6849382     0.0000000     0.0000000
##  sub_race_PFC non_white_MSE   sub_sex_PFC
##     0.1950355     0.0000000     0.0000000
```

Observa cómo produjimos una serie de medidas de error en lugar de las dos por defecto que habiamos visto antes. Ahora podemos evaluar la calidad de imputación para cada variable por separado. Esto es útil cuando necesitamos saber cómo se desempeña el modelo para una variable en particular que deseamos modelar o analizar más a fondo.

10.3.22 Trade-off velocidad-precisión

En este sentido existen dos paramétros que podemos ajustar para influir en el rendimiento de los bosques aleatorios (*random forest*):

. Número de árboles de decisión en cada bosque. . Número de variables utilizadas para la división dentro de los árboles de decisión.

Aumentar cada uno de ellos puede mejorar la precisión del modelo de imputación, pero también requerirá más tiempo para ejecutarse. En este ejercicio, exploraremos estas ideas

ajustando `missForest()` a los datos de `biopics` dos veces con diferentes configuraciones. Mientras seguimos estos pasos, pongamos atención a los errores que imprimiremos y al tiempo que tomará la ejecución del código.

```
# Determina el tiempo inicial del primer enfoque
 t <- proc.time()

# Define el numero de arboles a 5 y el numero de variables usadas para dividir en 2
imp_res <- missForest(biopics, mtry = 2, ntree = 5)
tiempo1<-proc.time() - t
# Imprime los resultados de los errores de la imputacion
print(imp_res$OOBerror)
```

```
##      NRMSE        PFC
## 0.02471563 0.08284884
```

```
# Determina el tiempo inicial del segundo enfoque
 t <- proc.time()
# Define el numero de arboles a 50 y el numero de variables usadas para dividir en 6
imp_res <- missForest(biopics, mtry = 6, ntree = 50)
tiempo2<-proc.time() - t
# Imprime los errores resultantes de la imputacion
print(imp_res$OOBerror)
```

```
##      NRMSE        PFC
## 0.02153514 0.04476950
```

Compara los errores y los tiempos de ejecución de los dos modelos de imputación. ¿Puedes ver una relación? Como dicen, "no hay nada gratuito". Para obtener una imputación más precisa, tuvimos que invertir más tiempo de computación.

```
tiempo1
```

```
##    user  system elapsed
##    0.06    0.00    0.07
```

```
tiempo2
```

```
##    user  system elapsed
##    3.63    0.00    3.66
```

10.3.23 La imputación y el modelado en una función

Siempre que realice cualquier análisis o modelado en datos imputados, debe tener en cuenta la incertidumbre de la imputación. Ejecutar un modelo en un conjunto de datos imputados, se ignora el hecho de que la imputación estima los valores faltantes con incertidumbre. Los errores estándar de dicho modelo tienden a ser demasiado pequeños. La solución a esto es la imputación múltiple y una forma de implementarla es mediante bootstrap.

Trabajaremos con los datos `biopics`. El objetivo es utilizar la imputación múltiple mediante bootstrap y la regresión lineal para ver si, en función de los datos disponibles, las películas

biográficas con mujeres ganan menos que las de hombres.

Comencemos escribiendo una función que construya una muestra de bootstrap, la impute y ajuste un modelo de regresión lineal.

```
calc_gender_coef <- function(data, indices) {
  # Obtener una muestra bootstrap
  data_boot <- data[indices, ]
  # Imputa con imputacion kNN
  data_imp <- kNN(data_boot, k = 5)
  # Ajusta una regresion lineal
  linear_model <- lm(earnings ~ sub_sex + sub_type + year, data = data_imp)
  # Extrae y calcula coeficiente para gender
  gender_coefficient <- coef(linear_model)[2]
  return(gender_coefficient)
}
```

La función `calc_gender_coef()` toma los datos y los índices de bootstrap como entradas, y produce nuestra estadística de interés: el impacto del género en las ganancias de la regresión lineal. Ahora podemos usar esta función en el algoritmo de bootstrapping.

10.3.24 Ejecutando bootstrap

Esta función crea una muestra de bootstrap, la imputa y produce el coeficiente de regresión lineal que describe el impacto de que el tema de la película sea femenino en las ganancias de la película.

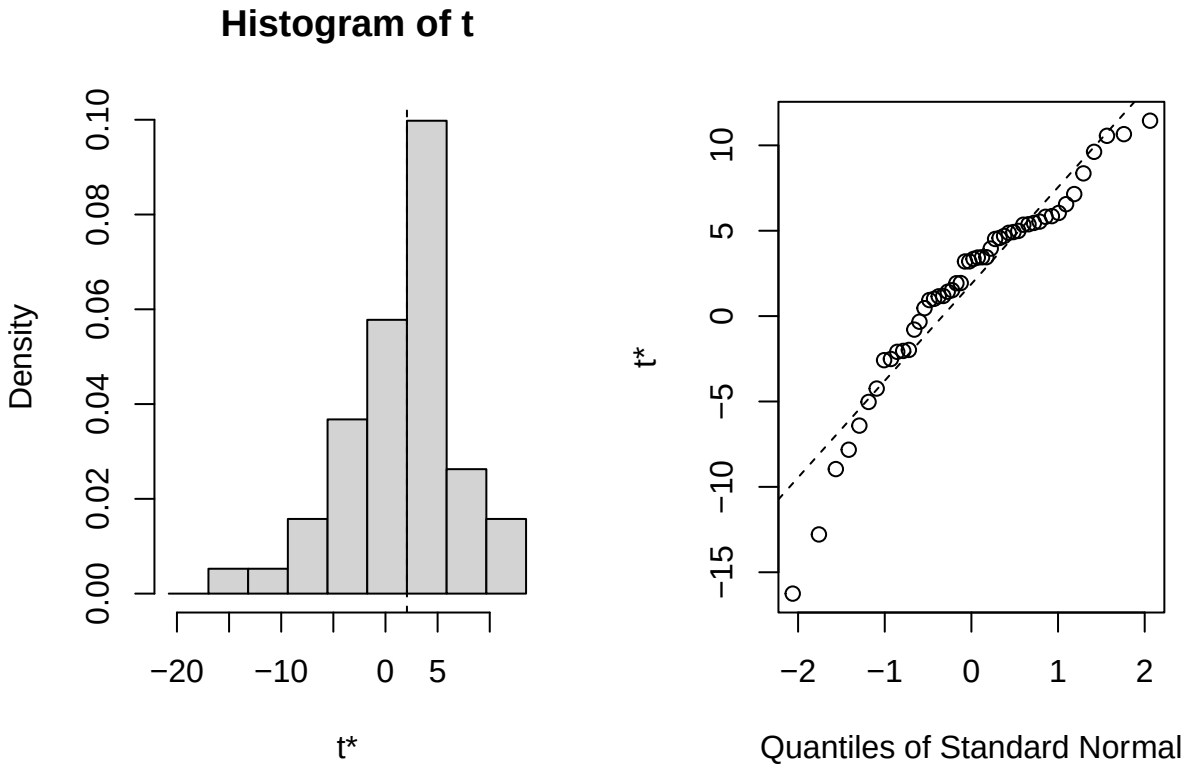
En este ejercicio, usarás el paquete `boot` para obtener una distribución de bootstrap de estos coeficientes. La propagación de esta distribución capturará la incertidumbre de la imputación. También verás cómo la distribución de bootstrap difiere de una imputación y regresión.

```
# Carga la libreria boot
library(boot)

# Ejecuta bootstrap sobre los datos biopics
boot_results <- boot(biopics, statistic = calc_gender_coef, R = 50)

# Imprime y grafica los resultados del bootstrapping
print(boot_results)
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = biopics, statistic = calc_gender_coef, R = 50)
##
##
## Bootstrap Statistics :
##     original     bias    std. error
```

```
## t1* 2.060346 -0.1781512    5.666494
```

```
plot(boot_results)
```



Si hubieramos ejecutado la imputación `kNN` y el análisis de regresión en los datos de `biopics` solo una vez, habríamos obtenido un coeficiente de `-1.45` para las películas sobre mujeres (llamado "original" en la salida de la consola), lo que sugiere que las películas sobre mujeres ganan menos. Sin embargo, al corregir la incertidumbre de la imputación, hemos obtenido una distribución que cubre tanto valores negativos como positivos.

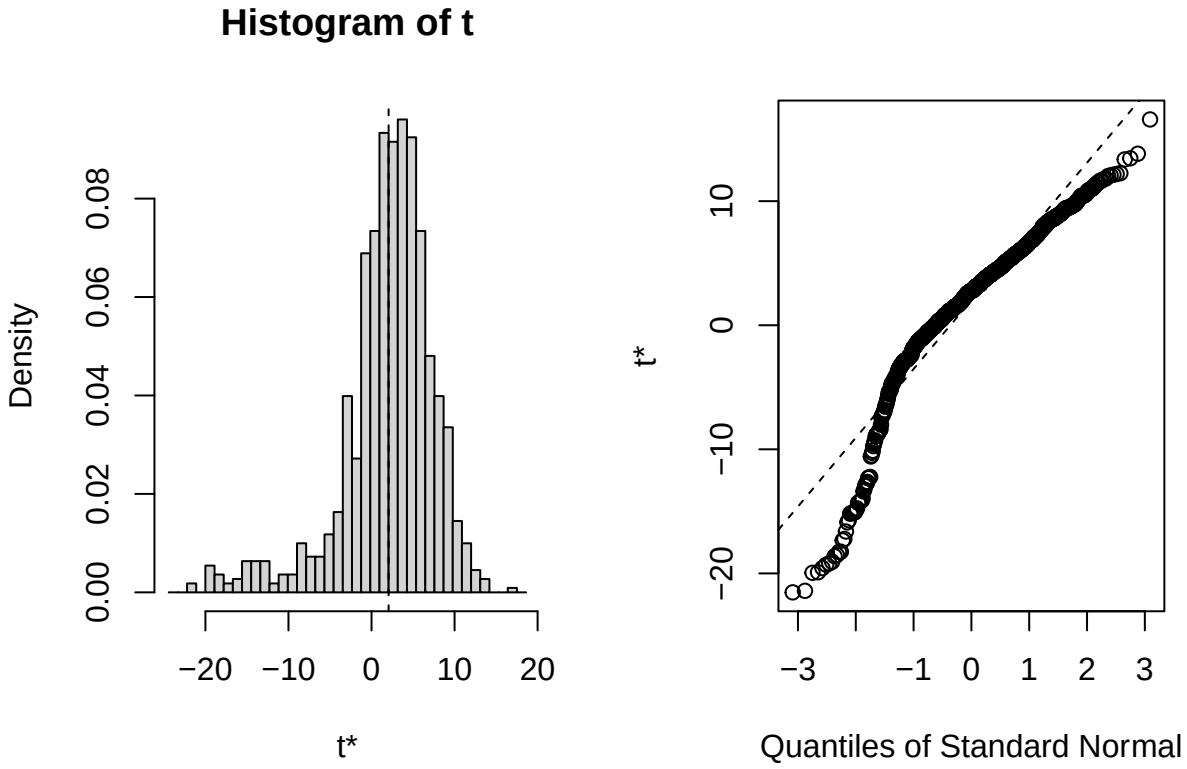
10.3.25 Bootstrapping para intervalos de confianza

Después de haber generado la distribución del coeficiente del efecto femenino en el último ejercicio, ahora podemos usarla para estimar un intervalo de confianza. Esto permitirá hacer la siguiente evaluación sobre los datos: "Dada la incertidumbre de la imputación, estamos 95% seguros de que el efecto femenino en las ganancias se encuentra entre `a` y `b`", donde `a` y `b` son los límites inferior y superior del intervalo.

En el último ejercicio, ejecutamos la técnica de bootstrapping con `R = 50` réplicas. Sin embargo, en la mayoría de las aplicaciones esto no es suficiente. En este ejercicio, puedes utilizar los `boot_results` que se prepararon utilizando 1000 réplicas. Primero, verás si la distribución de bootstrapping parece normal. Si es así, entonces podrás confiar en la distribución normal para calcular el intervalo de confianza.

```r
# Ejecuta bootstrap sobre los datos biopics y mide el tiempo de ejecucion
boot_results <- boot(biopics, statistic = calc_gender_coef, R = 1000)
```

```r
# Plot and print boot_results
plot(boot_results)
```




```r
print(boot_results)
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = biopics, statistic = calc_gender_coef, R = 1000)
##
##
## Bootstrap Statistics :
##     original      bias    std. error
## t1* 2.060346 -0.04651124    5.542859
```

```r
# Calculate and print confidence interval
boot_ci <- boot.ci(boot_results, conf = 0.95, type = "norm")
print(boot_ci)
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_results, conf = 0.95, type = "norm")
##
## Intervals :
## Level      Normal
## 95%   (-8.757, 12.971 )
## Calculations and Intervals on Original Scale
```

A pesar de que la tendencia general parece ser una relación negativa, las réplicas de bootstrap muestran que algunas películas con protagonistas femeninas en realidad ganan más. Al tener en cuenta la incertidumbre de la imputación, no se puede estar al 100% seguro acerca de la dirección de esta relación, aunque un análisis único sugiera lo contrario.

10.3.26 El flujo de MICE: mice - with - pool

El flujo de `MICE` (imputación múltiple por ecuaciones encadenadas) nos permite estimar la incertidumbre de la imputación mediante la imputación de un conjunto de datos varias veces mediante la imputación basada en modelos, mientras se extrae de las distribuciones condicionales. De esta manera, cada conjunto de datos imputados es ligeramente diferente. Luego, se realiza un análisis en cada uno de ellos y se combinan los resultados, obteniendo las cantidades de interés junto con sus intervalos de confianza que reflejan la incertidumbre de la imputación.

En este ejercicio, practicaremos el flujo típico de la imputación con `MICE`: `mice()` - `with()` - `pool()`. Realizaremos un análisis de regresión en los datos de `biopics` para ver qué tipo de ocupación de sujeto, `sub_type`, está asociada con mayores ingresos en las películas.

```r
# Carga el paquete mice
library(mice)

# Imputa biopics con mice usando 5 imputaciones
biopics_multiimp <- mice(biopics, m = 5, seed = 3108)
```

```
##
##  iter imp variable
##   1   1  earnings  sub_race
##   1   2  earnings  sub_race
##   1   3  earnings  sub_race
##   1   4  earnings  sub_race
##   1   5  earnings  sub_race
##   2   1  earnings  sub_race
##   2   2  earnings  sub_race
##   2   3  earnings  sub_race
##   2   4  earnings  sub_race
##   2   5  earnings  sub_race
##   3   1  earnings  sub_race
```

```
##   3   2  earnings  sub_race
##   3   3  earnings  sub_race
##   3   4  earnings  sub_race
##   3   5  earnings  sub_race
##   4   1  earnings  sub_race
##   4   2  earnings  sub_race
##   4   3  earnings  sub_race
##   4   4  earnings  sub_race
##   4   5  earnings  sub_race
##   5   1  earnings  sub_race
##   5   2  earnings  sub_race
##   5   3  earnings  sub_race
##   5   4  earnings  sub_race
##   5   5  earnings  sub_race
```

```r
# Ajusta una regresion lineal para cada set de datos imputados
lm_multiimp <- with(biopics_multiimp, lm(earnings ~ year + sub_type))

# Combina las estimaciones por las reglas de Rubin (pool)
lm_pooled <- pool(lm_multiimp)
summary(lm_pooled, conf.int = TRUE, conf.level = 0.95)
```

```
##                              term     estimate  std.error   statistic
## 1                     (Intercept) -287.8866750 490.062824 -0.58744851
## 2                            year    0.1612176   0.239277  0.67376974
## 3  sub_typeAcademic (Philosopher)  -32.8423364  39.593926 -0.82947915
## 4                 sub_typeActivist  -17.4281787  16.052579 -1.08569339
## 5                    sub_typeActor  -30.7740287  19.378762 -1.58802862
## 6                  sub_typeActress  -27.3033456  21.249448 -1.28489670
## 7       sub_typeActress / activist   16.0962907  39.192849  0.41069458
## 8                  sub_typeArtist  -27.4779956  18.148136 -1.51409465
## 9                 sub_typeAthlete   -4.2336944  12.503822 -0.33859202
## 10     sub_typeAthlete / military   79.1943734  38.055447  2.08102595
## 11                 sub_typeAuthor  -23.6540827  18.471621 -1.28056347
## 12          sub_typeAuthor (poet)  -23.4506193  19.985477 -1.17338304
## 13               sub_typeComedian  -22.9330598  21.576033 -1.06289508
## 14               sub_typeCriminal   -2.6478096  16.754147 -0.15803906
## 15             sub_typeGovernment   -5.0221576  21.879188 -0.22954040
## 16             sub_typeHistorical   -9.3734433  19.183957 -0.48860845
## 17             sub_typeJournalist  -26.5071032  29.005345 -0.91386960
## 18                  sub_typeMedia  -11.5302020  26.461566 -0.43573393
## 19               sub_typeMedicine    4.7788761  15.006237  0.31845932
## 20               sub_typeMilitary   10.9417685  19.325322  0.56618816
## 21    sub_typeMilitary / activist   37.2248141  39.752579  0.93641257
## 22               sub_typeMusician  -19.7931797  17.320867 -1.14273611
## 23                  sub_typeOther  -15.5895605  16.044509 -0.97164460
## 24             sub_typePolitician  -13.6751859  39.752579 -0.34400752
```

```
## 25                 sub_typeSinger    0.6677979  17.844592  0.03742298
## 26                sub_typeTeacher   51.1575084  39.268925  1.30274786
## 27           sub_typeWorld leader    5.9976436  16.882980  0.35524793
##             df    p.value        2.5 %       97.5 %
## 1    3.983993 0.58858265 -1650.677928 1074.9045785
## 2    4.030421 0.53712441    -0.501149    0.8235843
## 3  139.785256 0.40824758  -111.122704   45.4380311
## 4   10.576004 0.30174143   -52.933253   18.0768961
## 5   10.847500 0.14097812   -73.499669   11.9516115
## 6    7.876739 0.23532207   -76.438456   21.8317642
## 7  169.869958 0.68181408   -61.271473   93.4640548
## 8    8.280382 0.16719720   -69.082290   14.1262992
## 9   15.246283 0.73953513   -30.847513   22.3801244
## 10 336.810947 0.03818668     4.338081  154.0506662
## 11   7.047583 0.24087818   -67.272814   19.9646489
## 12  10.612273 0.26630514   -67.635233   20.7339945
## 13  16.686775 0.30297041   -68.519689   22.6535693
## 14   7.275709 0.87872330   -41.962758   36.6671385
## 15  81.573759 0.81902349   -48.550237   38.5059216
## 16   6.172715 0.64199292   -55.998343   37.2514567
## 17 113.284343 0.36272632   -83.970362   30.9561559
## 18   6.128774 0.67795894   -75.950965   52.8905609
## 19 247.938063 0.75040464   -24.777080   34.3348320
## 20   6.645835 0.58986247   -35.253693   57.1372305
## 21 130.043524 0.35079655   -41.420661  115.8702894
## 22   6.996299 0.29074124   -60.754913   21.1685537
## 23   7.168854 0.36286255   -53.348394   22.1692725
## 24 130.043524 0.73139625   -92.320661   64.9702894
## 25  10.377953 0.97085786   -38.897026   40.2326216
## 26 163.415623 0.19449369   -26.382404  128.6974204
## 27  14.000056 0.72769899   -30.212733   42.2080205
```

En este caso, hemos seguido el flujo "mice-with-pool" para imputar, modelar y agrupar los resultados. Ahora, echemos un vistazo a la salida en la consola: algunos `sub_types` tienen un impacto positivo en las ganancias. Sin embargo, al tener en cuenta la incertidumbre de la imputación con una confianza del 95%, nunca estamos seguros de estos efectos, ya que los límites inferiores son negativos. Con una excepción: para `sub_typeAthlete / military`, tanto los límites inferiores como los superiores son positivos. Lo que podemos decir con seguridad es que las películas sobre atletas militares son populares.

10.3.26.1 Selección de modelos por defecto MICE crea un modelo de imputación separado para cada variable en los datos. El tipo de modelo depende del tipo de variable en cuestión. Una forma popular de especificar los tipos de modelos que queremos usar es establecer un modelo predeterminado para cada uno de los cuatro tipos de variables.

Podemos hacer esto usando el argumento defaultMethod en la función `mice()`, que debe ser un vector de longitud 4 que contenga los métodos de imputación predeterminados para:

Variables continuas, Variables binarias, Variables categóricas (factores no ordenados), Variables factoriales (factores ordenados).

En este caso, aprovecharemos la documentación de `mice` para ver la lista de métodos disponibles y seleccionar los deseados para que el algoritmo los use.

```
# Imputa biopics usando los metodos especificados en la instruccion
biopics_multiimp <- mice(biopics, m = 20,
                         defaultMethod = c("cart", "lda", "pmm", "polr"))
```

```
##
##  iter imp variable
##   1   1  earnings  sub_race
##   1   2  earnings  sub_race
##   1   3  earnings  sub_race
##   1   4  earnings  sub_race
##   1   5  earnings  sub_race
##   1   6  earnings  sub_race
##   1   7  earnings  sub_race
##   1   8  earnings  sub_race
##   1   9  earnings  sub_race
##   1   10  earnings  sub_race
##   1   11  earnings  sub_race
##   1   12  earnings  sub_race
##   1   13  earnings  sub_race
##   1   14  earnings  sub_race
##   1   15  earnings  sub_race
##   1   16  earnings  sub_race
##   1   17  earnings  sub_race
##   1   18  earnings  sub_race
##   1   19  earnings  sub_race
##   1   20  earnings  sub_race
##   2   1  earnings  sub_race
##   2   2  earnings  sub_race
##   2   3  earnings  sub_race
##   2   4  earnings  sub_race
##   2   5  earnings  sub_race
##   2   6  earnings  sub_race
##   2   7  earnings  sub_race
##   2   8  earnings  sub_race
##   2   9  earnings  sub_race
##   2   10  earnings  sub_race
##   2   11  earnings  sub_race
##   2   12  earnings  sub_race
##   2   13  earnings  sub_race
##   2   14  earnings  sub_race
##   2   15  earnings  sub_race
##   2   16  earnings  sub_race
```

```
##   2   17  earnings  sub_race
##   2   18  earnings  sub_race
##   2   19  earnings  sub_race
##   2   20  earnings  sub_race
##   3   1  earnings  sub_race
##   3   2  earnings  sub_race
##   3   3  earnings  sub_race
##   3   4  earnings  sub_race
##   3   5  earnings  sub_race
##   3   6  earnings  sub_race
##   3   7  earnings  sub_race
##   3   8  earnings  sub_race
##   3   9  earnings  sub_race
##   3   10  earnings  sub_race
##   3   11  earnings  sub_race
##   3   12  earnings  sub_race
##   3   13  earnings  sub_race
##   3   14  earnings  sub_race
##   3   15  earnings  sub_race
##   3   16  earnings  sub_race
##   3   17  earnings  sub_race
##   3   18  earnings  sub_race
##   3   19  earnings  sub_race
##   3   20  earnings  sub_race
##   4   1  earnings  sub_race
##   4   2  earnings  sub_race
##   4   3  earnings  sub_race
##   4   4  earnings  sub_race
##   4   5  earnings  sub_race
##   4   6  earnings  sub_race
##   4   7  earnings  sub_race
##   4   8  earnings  sub_race
##   4   9  earnings  sub_race
##   4   10  earnings  sub_race
##   4   11  earnings  sub_race
##   4   12  earnings  sub_race
##   4   13  earnings  sub_race
##   4   14  earnings  sub_race
##   4   15  earnings  sub_race
##   4   16  earnings  sub_race
##   4   17  earnings  sub_race
##   4   18  earnings  sub_race
##   4   19  earnings  sub_race
##   4   20  earnings  sub_race
##   5   1  earnings  sub_race
##   5   2  earnings  sub_race
```

```
##   5   3  earnings  sub_race
##   5   4  earnings  sub_race
##   5   5  earnings  sub_race
##   5   6  earnings  sub_race
##   5   7  earnings  sub_race
##   5   8  earnings  sub_race
##   5   9  earnings  sub_race
##   5   10  earnings  sub_race
##   5   11  earnings  sub_race
##   5   12  earnings  sub_race
##   5   13  earnings  sub_race
##   5   14  earnings  sub_race
##   5   15  earnings  sub_race
##   5   16  earnings  sub_race
##   5   17  earnings  sub_race
##   5   18  earnings  sub_race
##   5   19  earnings  sub_race
##   5   20  earnings  sub_race
```

```
# Imprime biopics_multiimp
print(biopics_multiimp)
```

```
## Class: mids
## Number of multiple imputations:  20
## Imputation methods:
##   country      year  earnings   sub_num  sub_type  sub_race non_white   sub_sex
##        ""        ""    "cart"        ""        ""     "pmm"        ""        ""
## PredictorMatrix:
##          country year earnings sub_num sub_type sub_race non_white sub_sex
## country        0    1        1       1        1        1         1       1
## year           1    0        1       1        1        1         1       1
## earnings       1    1        0       1        1        1         1       1
## sub_num        1    1        1       0        1        1         1       1
## sub_type       1    1        1       1        0        1         1       1
## sub_race       1    1        1       1        1        0         1       1
## Number of logged events:  200
##   it im      dep meth                              out
## 1  1  1 earnings cart sub_typeAcademic (Philosopher)
## 2  1  1 sub_race  pmm    sub_typeMilitary / activist
## 3  1  2 earnings cart sub_typeAcademic (Philosopher)
## 4  1  2 sub_race  pmm    sub_typeMilitary / activist
## 5  1  3 earnings cart sub_typeAcademic (Philosopher)
## 6  1  3 sub_race  pmm    sub_typeMilitary / activist
```

La capacidad de especificar modelos de imputación puede resultar útil cuando se observa que algunos métodos específicos no funcionan bien. Otro factor que influye en cómo funcionan los métodos de imputación es el conjunto de predictores que utilizan. En el siguiente ejercicio, veremos cómo establecer estos predictores.

10.3.26.2 Usando una matriz predictora Se trata de tomar decisiones importantes cuando se utiliza la imputación basada en modelos, como por ejemplo, qué variables deben incluirse como predictores y en qué modelos. En `mice()`, esto se rige por la matriz de predictores y, por defecto, todas las variables se utilizan para imputar todas las demás.

En caso de tener muchas variables en los datos o poco tiempo para realizar una selección adecuada del modelo, puede utilizar la funcionalidad de `mice` para crear una matriz de predictores basada en las correlaciones entre las variables. Esta matriz se puede incorporar a `mice()`. En este ejercicio, practicaremos exactamente esto: primero construiremos una matriz de predictores de modo que cada variable se imputa utilizando las variables más correlacionadas con ella; luego, usará una matriz de predictores con la función de imputación.

```
# Crea una matriz predictora con correlacion minima de 0.1
pred_mat <- quickpred(biopics, mincor = 0.1)

# Imputa biopics con mice
biopics_multiimp <- mice(biopics,
                         m = 10,
                         predictorMatrix = pred_mat,
                         seed = 3108)
```

```
##
##  iter imp variable
##   1   1  earnings  sub_race
##   1   2  earnings  sub_race
##   1   3  earnings  sub_race
##   1   4  earnings  sub_race
##   1   5  earnings  sub_race
##   1   6  earnings  sub_race
##   1   7  earnings  sub_race
##   1   8  earnings  sub_race
##   1   9  earnings  sub_race
##   1   10  earnings  sub_race
##   2   1  earnings  sub_race
##   2   2  earnings  sub_race
##   2   3  earnings  sub_race
##   2   4  earnings  sub_race
##   2   5  earnings  sub_race
##   2   6  earnings  sub_race
##   2   7  earnings  sub_race
##   2   8  earnings  sub_race
##   2   9  earnings  sub_race
##   2   10  earnings  sub_race
##   3   1  earnings  sub_race
##   3   2  earnings  sub_race
##   3   3  earnings  sub_race
##   3   4  earnings  sub_race
##   3   5  earnings  sub_race
```

```
##   3   6  earnings  sub_race
##   3   7  earnings  sub_race
##   3   8  earnings  sub_race
##   3   9  earnings  sub_race
##   3   10  earnings  sub_race
##   4   1  earnings  sub_race
##   4   2  earnings  sub_race
##   4   3  earnings  sub_race
##   4   4  earnings  sub_race
##   4   5  earnings  sub_race
##   4   6  earnings  sub_race
##   4   7  earnings  sub_race
##   4   8  earnings  sub_race
##   4   9  earnings  sub_race
##   4   10  earnings  sub_race
##   5   1  earnings  sub_race
##   5   2  earnings  sub_race
##   5   3  earnings  sub_race
##   5   4  earnings  sub_race
##   5   5  earnings  sub_race
##   5   6  earnings  sub_race
##   5   7  earnings  sub_race
##   5   8  earnings  sub_race
##   5   9  earnings  sub_race
##   5   10  earnings  sub_race
```

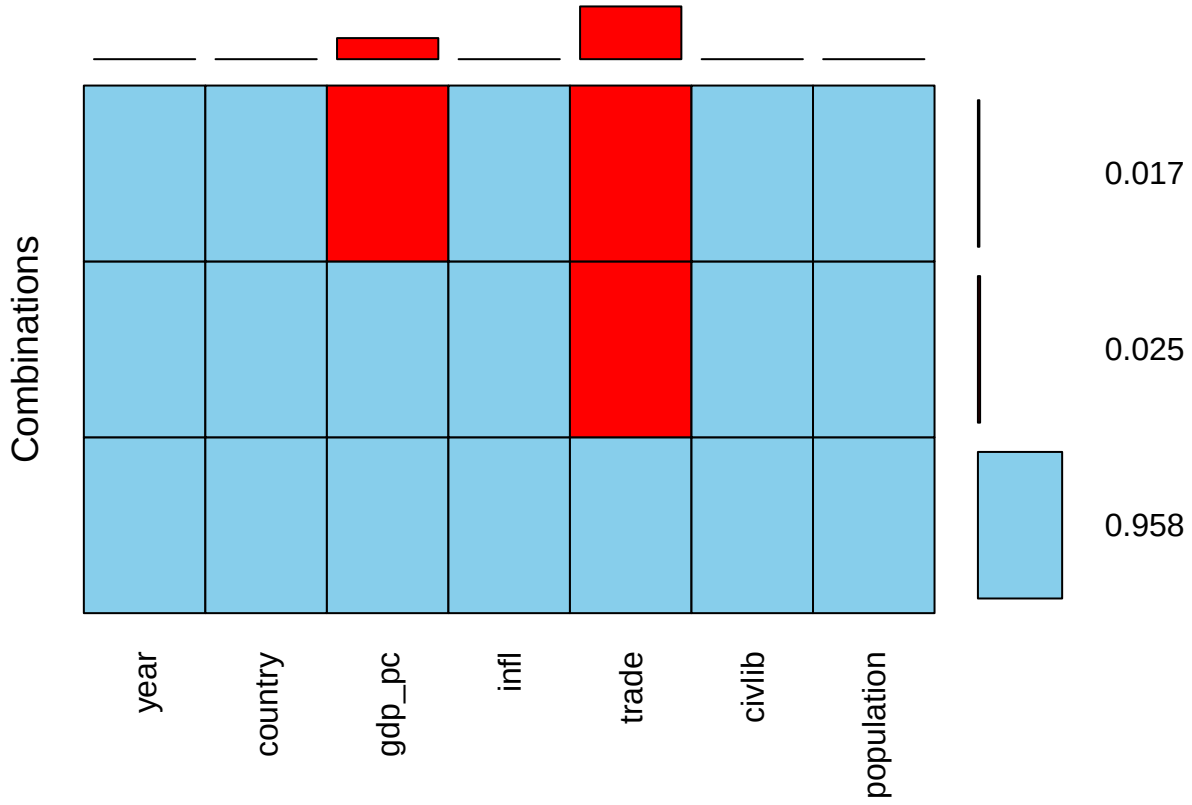
```
# Imprime biopics_multiimp
# print(biopics_multiimp)
```

10.3.27 Analizando los patrones de datos faltantes

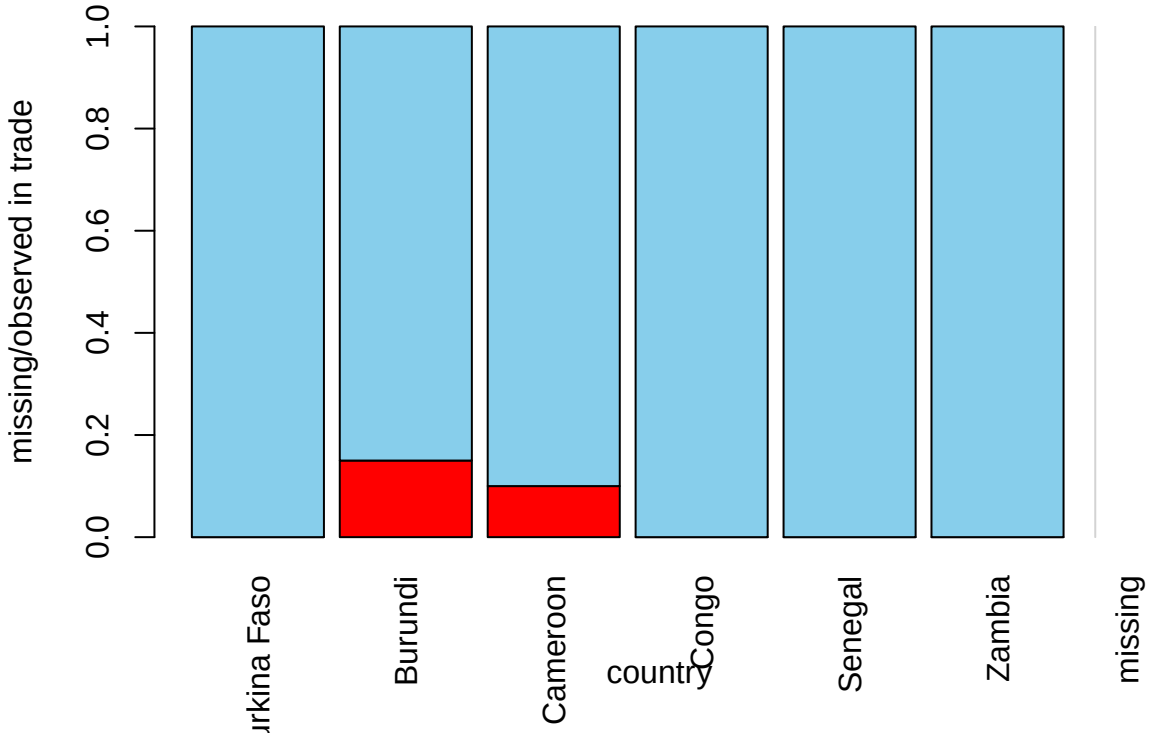
El primer paso para trabajar con datos incompletos es obtener información sobre los patrones de ausencia de datos, y una buena manera de hacerlo es mediante visualizaciones. Comenzarás tu análisis de los datos de `Áfrıca` empleando el paquete `VIM` para crear dos visualizaciones: el gráfico de agregación y el gráfico de Mosaico. Te dirán cuántos datos faltan, en qué variables y configuraciones, y si podemos decir algo sobre el mecanismo de ausencia de datos. ¡Comencemos con algunas gráficas!

```
africa <- read.csv("handing/africa.csv", sep = ";")
```

```
# carga el paquete VIM
library(VIM)



# Crea un grafico de agregación combinada del set de datos africa
africa %>%
  aggr(combined = TRUE, numbers = TRUE)
```




```
# Crea un grafico spine plot de pais vs trade
africa %>%
  select(country, trade) %>%
  spineMiss()
```



Observamos que no hay tantos valores faltantes. Además, observe en el gráfico de Mosaico para los datos de `africa` parecen ser MAR - al menos con respecto al PIB y al país, lo que significa que se pueden imputar.

10.3.28 Imputando e inspeccionando resultados

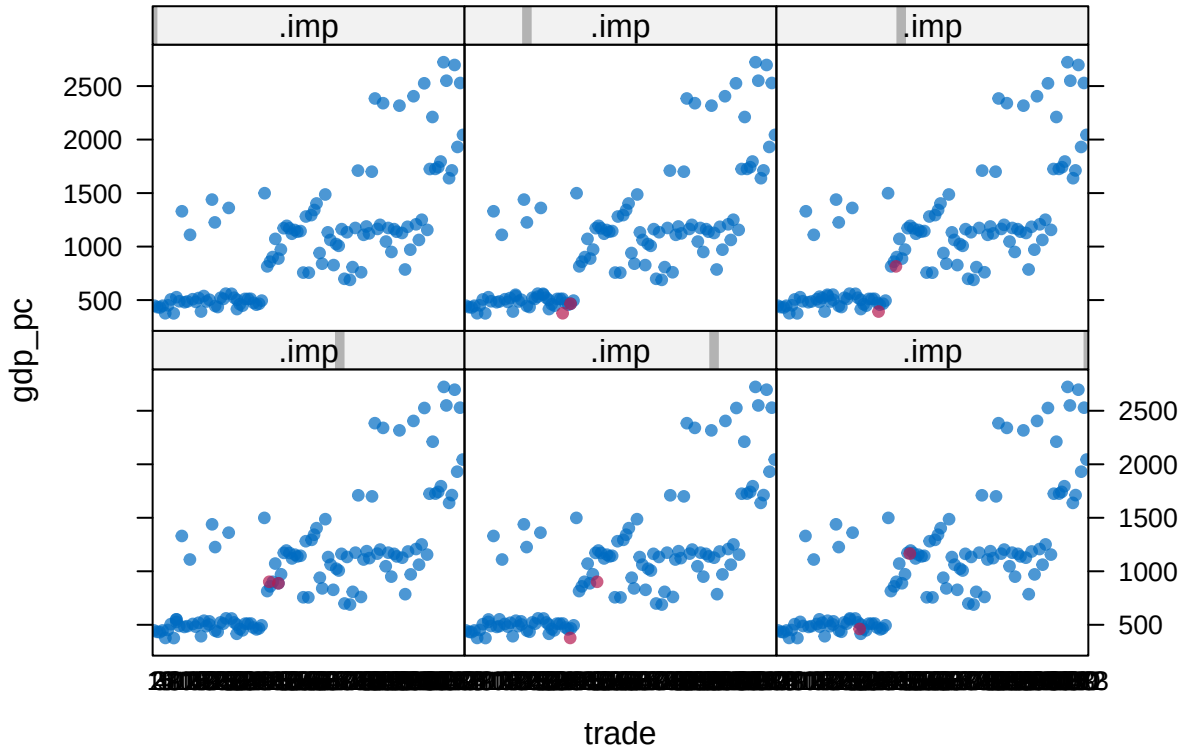
Hemos descubierto que hay algunos datos faltantes en el PIB, `gdp_pc`, y en `trade` como porcentaje del PIB. Además, se sospecha que los datos son MAR, por lo que es posible que sean imputados. En este caso, haremos uso de la imputación múltiple del paquete `mice` para imputar los datos de `africa`. Luego, crearemos un gráfico para `gdp_pc` vs `trade` para ver si los datos imputados no rompen la relación entre estas variables.

```
# Carga mice
library(mice)

# Imputa africa con mice
africa_multiimp <- mice(africa, m = 5, defaultMethod = "cart", seed = 3108)
```

```
##
##  iter imp variable
##   1   1  gdp_pc  trade
##   1   2  gdp_pc  trade
##   1   3  gdp_pc  trade
##   1   4  gdp_pc  trade
```

```
##   1   5  gdp_pc  trade
##   2   1  gdp_pc  trade
##   2   2  gdp_pc  trade
##   2   3  gdp_pc  trade
##   2   4  gdp_pc  trade
##   2   5  gdp_pc  trade
##   3   1  gdp_pc  trade
##   3   2  gdp_pc  trade
##   3   3  gdp_pc  trade
##   3   4  gdp_pc  trade
##   3   5  gdp_pc  trade
##   4   1  gdp_pc  trade
##   4   2  gdp_pc  trade
##   4   3  gdp_pc  trade
##   4   4  gdp_pc  trade
##   4   5  gdp_pc  trade
##   5   1  gdp_pc  trade
##   5   2  gdp_pc  trade
##   5   3  gdp_pc  trade
##   5   4  gdp_pc  trade
##   5   5  gdp_pc  trade
```

```r
# Crea un stripplot of gdp_pc versus trade
stripplot(africa_multiimp, gdp_pc ~ trade | .imp, pch = 20, cex = 1)
```



Se observa que la imputación funciona bien: hay pequeños grupos en los gráficos de dispersión, que probablemente corresponden a diferentes países. Cada punto de datos imputado encaja en uno de los grupos, en lugar de ser un valor atípico en algún lugar entre los grupos. Después de haber realizado la imputación, podemos proceder con el modelado.

10.3.28.1 Inferencia con datos imputados En este último caso, hemos utilizado `mice` para imputar los datos de `africa`. En este, implementaremos los otros dos pasos del flujo de "mice - with - pool" que hemos usado anteriormente. El modelo de interés es una regresión lineal que explica el PIB, `gdp_pc`, con otras variables. Nos interesa particularmente el coeficiente de libertades civiles, `civlib`. ¿Está asociado tener valores más altos en `civlib` en función a un mayor crecimiento económico una vez que incorporamos la incertidumbre de la imputación?

```
# Ajusta im regresion lineal a cada data set  imputado
lm_multiimp <- with(africa_multiimp, lm(gdp_pc ~ country + year + trade + infl + civlib)

# Combina las estimaciones por las reglas de Rubin (pool)
lm_pooled <- pool(lm_multiimp)

# Summarize pooled results
summary(lm_pooled, conf.int = TRUE, conf.level = 0.9)
```

```
##                term      estimate   std.error  statistic        df      p.value
## 1       (Intercept) -31703.576601 6031.455164 -5.2563728  92.53952 9.387968e-07
```

```
## 2   countryBurundi     63.739453   65.610134  0.9714879 103.18378 3.335772e-01
## 3  countryCameroon    622.343351   66.226798  9.3971530  56.28736 3.941785e-13
## 4     countryCongo   1303.940067  119.821885 10.8823197 101.65641 9.508463e-19
## 5   countrySenegal    516.442634   82.535746  6.2571995 106.90957 8.275032e-09
## 6    countryZambia    396.326840   87.729106  4.5176209 106.82235 1.620019e-05
## 7             year     16.156955    3.036095  5.3216230  92.06641 7.199791e-07
## 8            trade      5.468655    1.663523  3.2873939 105.04858 1.375742e-03
## 9             infl     -4.389418    1.031399 -4.2557883 107.91316 4.455825e-05
## 10          civlib    -84.852311  147.925470 -0.5736153  66.50333 5.681630e-01
##               5 %          95 %
## 1  -41724.760108 -21682.393094
## 2     -45.157324    172.636231
## 3     511.587065    733.099637
## 4    1105.037965   1502.842168
## 5     379.496718    653.388550
## 6     250.762899    541.890781
## 7      11.112260     21.201650
## 8       2.708058      8.229252
## 9      -6.100609     -2.678226
## 10   -331.605424    161.900803
```

Basándose en el resumen de los resultados de la regresión agrupada que acabamos de imprimir. Podemos decir que, dado que los límites inferior y superior tienen signos diferentes, no podemos estar seguros de la dirección del efecto.

Amemiya, T. 1985. *Advanced Econometrics*. Oxford: Basil Blackwell.

Amemiya, Takeshi. 1985. *Advanced Econometrics*. Harvard university press.

Andridge, Rebecca R, and Roderick J Little. 2009. "The Use of Sample Weights in Hot Deck Imputation." *Journal of Official Statistics* 25 (1): 21.

Barnard, John, and Donald B Rubin. 1999. "Miscellanea. Small-Sample Degrees of Freedom with Multiple Imputation." *Biometrika* 86 (4): 948–55.

Bethlehem, Jelke. 2009. *Applied Survey Methods: A Statistical Perspective*. Vol. 558. John Wiley & Sons.

Bethlehem, Jelke G, and Wouter J Keller. 1987. "Linear Weighting of Sample Survey Data." *Journal of Official Statistics* 3 (2): 141–53.

Binder, DA. 1996. "Comment to Articles by Rao, Fay and Rubin." *Journal of the American Statistical Association* 91: 510–12.

Binder, DA, and W Sun. 1996. "Frequency Valid Multiple Imputation for Surveys with Complex Designs, Bussines Survey Methods Division." *Statistics, Canada*.

Bodner, Todd E. 2008. "What Improves with Increased Missing Data Imputations?" *Structural Equation Modeling: A Multidisciplinary Journal* 15 (4): 651–75.

Böhning, D., W. Seidel, M. Alfó, B. Garel, V. Patilea, G. Walther, M. DiZio, U. Guarnera, and O. Luzi. 2007. "Imputation Through Finite Gaussian Mixture Models." *Computational Statistics and Data Analysis* 51: 5305–16.

Burnham, KP, and David R Anderson. 2002. "Model Selection and Multimodel Inference, 2nd Edn New York." *NY: Springer*.

Chambers, Ray L. 2004. *Evaluation Criteria for Statistical Editing and Imputation*. 28. Office for National Statistics.

Cotton, Cathy. 1991. "Functional Description of the Generalized Edit and Imputation System." *Statistics Canada, Business Survey Methods Division, July* 25.

Cox, David Roxbee, and David Victor Hinkley. 1979. *Theoretical Statistics*. CRC Press.

Cruz, Sonia Milena Cifuentes et al. 2013. "Compilación y Síntesis de Las Metodologías Internacionales Aplicadas a Procedimientos de Retropolación." Departamento Administrativo Nacional de Estadística-DANE.

Daniels, M. J., and J. W. Hogan. 2008. *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. Vol. 1. Chapman; Hall/CRC. https://doi.org/10.1201/9781420011180.

De la Fuente Moreno, A. 2014. "A" Mixed" Splicing Procedure for Economic Time Series." *Estadística Española* 56 (183): 107–21.

De Waal, Ton. 2000. "A Brief Overview of Imputation Methods Applied at Statistics Netherlands." *Netherlands Official Statistics* 15: 23–27.

De Waal, Ton, Jeroen Pannekoek, and Sander Scholtus. 2011. *Handbook of Statistical Data Editing and Imputation*. Vol. 563. John Wiley & Sons.

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

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

Deville, Jean-Claude, and Carl-Erik Särndal. 1994. "Variance Estimation for the Regression Imputed Horvitz-Thompson Estimator." *Journal of Official Statistics-Stockholm* 10: 381–81.

Draper, Norman R, and Harry Smith. 1998. *Applied Regression Analysis*. Vol. 326. John Wiley & Sons.

Durrant, Gabriele B et al. 2005. "Imputation Methods for Handling Item-Nonresponse in the Social Sciences: A Methodological Review." *ESRC National Centre for Research Methods and Southampton Statistical Sciences Research Institute. NCRM Methods Review Papers NCRM/002*.

Durrant, Gabriele B, and Chris Skinner. 2006. "Using Missing Data Methods to Correct for Measurement Error in a Distribution Function." *Survey Methodology* 32 (1): 25.

Eggen, Theo JHM, Wim J van der Linden, and Sebie J Oosterloo. 1983. "Book Review: Discrete Statistical Models with Social Science Applications Erling b. Andersen Amsterdam: North Holland Publishing Company, 1980, 383 Pp." *Applied Psychological Measurement* 7 (1): 119–21.

Elliott, Michael R, and Nicolas Stettler. 2007. "Using a Mixture Model for Multiple Imputation in the Presence of Outliers: The 'Healthy for Life' Project." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 56: 63–78.

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

Fay, Robert E. 1996. "Alternative Paradigms for the Analysis of Imputed Survey Data." *Journal of the American Statistical Association* 91 (434): 490–98.

Fine, T. L. 1999. *Feedforward Neural Network Methodology*. Vol. 1. Springer, New York.

Fraley, Chris, and Adrian E Raftery. 2002. "Model-Based Clustering, Discriminant Analysis, and Density Estimation." *Journal of the American Statistical Association* 97: 611–31.

Gelman, Andrew E, John B Carlin, Hal S Stern, and Donald B Rubin. 2013. *Bayesian Data*

Analysis, 3rd Edition. London: Chapman; Hall.
Gelman, Andrew, John B Carlin, Hal S Stern, and Donald B Rubin. 2004. "Bayesian Data Analysis Chapman & Hall." *CRC Texts in Statistical Science* 136.
Gelman, Andrew, and Xiao-Li Meng. 2004. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. John Wiley & Sons.
Gonzalez, Jeffrey M, and John L Eltinge. 2007. "Multiple Matrix Sampling: A Review." In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 3069–75. American Statistical Association Alexandria, VA.
Graham, John W. 2012. *Missing Data: Analysis and Design*. Springer Science & Business Media.
Granquist, John G, Leopold & Kovar. 1997. "Editing of Survey Data: How Much Is Enough?" *Survey Measurement and Process Quality*, 415–35.
Granquist, L. 1984. "Data Editing and Its Impact on the Further Processing of Statistical Data." In *Workshop on Statistical Computing*. Vol. 1217.
Granquist, Leopold. 1995. "Improving the Traditional Editing Process." *Business Survey Methods*, 385–401.
———. 1997. "The New View on Editing." *International Statistical Review* 65 (3): 381–87.
Groot, Wouter de, and Ron Dekker. 2001. *The Dutch System of Official Social Surveys*. Mannheim Centre for European Social Research (MZES), Eurodata Research Archive.
Haziza, D. 2006. "The Generalized Simulation System (GENESIS)." *Proceedings of the Section on Survey Research Methods*.
He, Yulei, and Trivellore E Raghunathan. 2012. "Multiple Imputation Using Multivariate Gh Transformations." *Journal of Applied Statistics* 39: 2177–98.
He, Yulei, Guangyu Zhang, and Chiu-Hsieh Hsu. 2021a. *Multiple Imputation of Missing Data in Practice: Basic Theory and Analysis Strategies*. CRC Press.
———. 2021b. *Multiple Imputation of Missing Data in Practice: Basic Theory and Analysis Strategies*. CRC Press.
Heerschap, N., and A. Van der Graaf. 1999. "A Test of Imputation by Means of a Neural Network on Data of the Structure of Earnings Survey (in Dutch)." *Report, Statistics Netherlands*.
Johansen, Søren. 1988. "Statistical Analysis of Cointegration Vectors." *Journal of Economic Dynamics and Control* 12 (2-3): 231–54.
Kalton, G. 1983. "Compensating for Missing Survey Data. An Arbor: Institute for Social Research." *Center for Social Research of The University of Michigan*.
Kalton, Graham. 1986. "The Treatment of Missing Survey Data." *Survey Methodology* 12: 1–16.
Kalton, Graham, and Leslie Kish. 1984. "Some Efficient Random Imputation Methods." *Communications in Statistics-Theory and Methods* 13 (16): 1919–39.
Kaufman, S, and F Scheuren. 1997. "Applying Mass Imputation Using the Schools and Staffing Survey Data." In *Proceedings of the American Statistical Association*.
Kim, Hyoung-Jean, Jerome P Reiter, Qingfeng Wang, Lawrence H Cox, and Alan F Karr. 2014. "Multiple Imputation of Missing or Faulty Values Under Linear Constraints." *Journal of Business and Economic Statistics* 32: 375–86.
Kim J. K., and W. Fuller. 2008. "Parametric Fractional Imputation for Missing Data Analysis." *Proceedings of the Section on Survey Research Methods, Joint Statistical Meeting*, 158–69.
Kim, Jae Kwang, and Wayne Fuller. 2004. "Fractional Hot Deck Imputation." *Biometrika* 91

(3): 559–78.
Kirchgässner, Gebhard, Jürgen Wolters, Gebhard Kirchgässner, and Jürgen Wolters. 2007. "Vector Autoregressive Processes." *Introduction to Modern Time Series Analysis*, 125–51.
Kish, Leslie. 1965. *Survey Sampling*. John Wiley & Sons.
Knottnerus, Paul. 2003. *Sample Survey Theory: Some Pythagorean Perspectives*. Springer Science & Business Media.
Krotki, S. Black, K., and D. Creel. 2005. "Mass Imputation." In *ASA Section of Survey Research Methods*.
Lee, E. Rancourt, H., and C.-E. Särndal. 2002. *Variance Estimation from Survey Data with Imputed Values. In: Survey Non-Response, r.m. Groves, d.a. Dillman, j.l. Eltinge, and r.j.a. Little, Eds. John Wiley & Sons, Hoboken, NJ, Pp. 315–328.* Vol. 1.
Little, Roderick J. 2013. "In Praise of Simplicity Not Mathematistry! Ten Simple Powerful Ideas for the Statistical Scientist." *Journal of the American Statistical Association* 108 (502): 359–69.
Little, Roderick J. A., and Donald B. Rubin. 2020a. *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. John Wiley & Sons.
———. 2020b. *Statistical Analysis with Missing Data*. 3rd ed. New York: Wiley.
Little, Roderick JA, and Donald B Rubin. 2002a. "Bayes and Multiple Imputation." *Statistical Analysis with Missing Data*, 200–220.
———. 2002b. *Statistical Analysis with Missing Data*. Vol. 793. John Wiley & Sons.
———. 2002c. "Statistical Analysis with Missing Data." *Statistical Analysis with Missing Data*.
Liu, Chuanhai. 1995. "Missing Data Imputation Using the Multivariate t Distribution." *Journal of Multivariate Analysis* 53: 139–58.
Lucas Jr, Robert E. 1976. "Econometric Policy Evaluation: A Critique." In *Carnegie-Rochester Conference Series on Public Policy*, 1:19–46. North-Holland.
Maharaj, Elizabeth Ann, Pierpaolo D'Urso, and Jorge Caiado. 2019. *Time Series Clustering and Classification*. Chapman; Hall/CRC.
Marron, JS, and MP Wand. 1992. "Exact Mean Integrated Squared Error." *The Annals of Statistics* 20: 712–36.
McCullagh, Peter. 2019. *Generalized Linear Models*. Routledge.
McCullagh, Peter, and JA Nelder. 1989. "Generalized Linear Models Second Edition." In *Boca Raton, London, New-York, Washington, DC, Chapman & Hall/CRC*.
McLachlan, Geoffrey J, and David Peel. 2000. *Finite Mixture Models*. New York: Wiley.
Medina, Fernando, and Marco Galván. 2007. *Imputación de Datos: Teoría y Práctica*. Cepal.
Model, Business Process. n.d. "The Generic Statistical Business Process Model."
Molenberghs, Geert, Garrett Fitzmaurice, Michael G Kenward, Anastasios Tsiatis, and Geert Verbeke. 2014. *Handbook of Missing Data Methodology*. CRC Press.
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.
Monroe, Robert J, and Francis E McVay. 1980. "In Memoriam: Gertrude Mary Cox 1900–1978." *The American Statistician* 34 (1): 48–48.
Montenegro, Roberto. 2010. "Medición de La Volatilidad En Series de Tiempo Financieras. Una Evaluación a La Tasa de Cambio Representativa Del Mercado (TRM) En Colombia." *Revista Finanzas y Política Económica* 2 (1): 125–32.

Mueller, Dennis C et al. 1997. *Perspectives on Public Choice: A Handbook*. Cambridge University Press.

Nagelkerke, Nico JD et al. 1991. "A Note on a General Definition of the Coefficient of Determination." *Biometrika* 78 (3): 691–92.

Nicholson, Walter. 2005. *Teoría Microeconómica. Principios Básicos y Ampliaciones: Principios Básicos y Ampliaciones*. Ediciones Paraninfo, SA.

Nordbotten, Svein. 1955. "Measuring the Error of Editing the Questionnaires in a Census." *Journal of the American Statistical Association* 50 (270): 364–69.

Nordholt, Eric Schulte. 1997. "Imputation in the New Dutch Structure of Earnings Survey (SES)." *Report, Statistics Netherlands*.

———. 1998. "Imputation: Methods, Simulation Experiments and Practical Examples." *International Statistical Review* 66 (2): 157–80.

Olsson, Ola. 2013. *Essentials of Advanced Macroeconomic Theory*. Routledge.

Persson, Torsten, and Guido Tabellini. 2001. "Political Institutions and Policy Outcomes: What Are the Stylized Facts?"

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

Priebe, Carey E. 1994. "Adaptive Mixtures." *Journal of the American Statistical Association* 89: 796–806.

Qin, Yongsong, JNK Rao, and Qunshu Ren. 2008. "Confidence Intervals for Marginal Parameters Under Fractional Linear Regression Imputation for Missing Data." *Journal of Multivariate Analysis* 99 (6): 1232–59.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rao, JNK. 1996. "On Variance Estimation with Imputed Survey Data." *Journal of the American Statistical Association* 91 (434): 499–506.

Rao, Jon NK, and Jun Shao. 1992. "Jackknife Variance Estimation with Survey Data Under Hot Deck Imputation." *Biometrika* 79 (4): 811–22.

Rao, Jon NK, and RR Sitter. 1995. "Variance Estimation Under Two-Phase Sampling with Application to Imputation for Missing Data." *Biometrika* 82 (2): 453–60.

Resche-Rigon, Matthieu, and Ian R White. 2018. "Multiple Imputation by Chained Equations for Systematically and Sporadically Missing Multilevel Data." *Statistical Methods in Medical Research* 27 (6): 1634–49.

Rotemberg, Julio J. 1982. "Sticky Prices in the United States." *Journal of Political Economy* 90 (6): 1187–1211.

Royston, Patrick. 2004. "Multiple Imputation of Missing Values." *The Stata Journal* 4 (3): 227–41.

Rubin, DB. 1987. "Multiple Imputation for Non-Response in Surveys John Wiley." *New York*.

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

———. 1978. "Multiple Imputations in Sample Surveys-a Phenomenological Bayesian Approach to Nonresponse." In *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 1:20–34. American Statistical Association Alexandria, VA, USA.

———. 1987. "Multiple Imputation for Nonresponse in Surveys."

———. 1988. "An Overview of Multiple Imputation." In *Proceedings of the Survey Research*

Methods Section of the American Statistical Association, 79:84. Citeseer.

———. 1996. "Multiple Imputation After 18+ Years." *Journal of the American Statistical Association* 91 (434): 473–89.

———. 2004. *Multiple Imputation for Nonresponse in Surveys*. Vol. 81. John Wiley & Sons.

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

Rudin, Walter. 2012. *Análisis Funcional*. Reverté.

Sande, Innis G. 1982. "Imputation in Surveys: Coping with Reality." *The American Statistician* 36 (3a): 145–52.

Särndal, C. n.d. "4E.(1992). Methods for Estimating the Precision of Survey Estimates When Imputation Has Been Used." *Survey Methodology* 18: 2414252.

Särndal, Carl-Erik. 1992. "Methods for Estimating the Precision of Survey Estimates When Imputation Has Been Used." *Survey Methodology* 18 (2): 241–52.

Särndal, Carl-Erik, and Sixten Lundstrom. 2005. *Estimation in Surveys with Nonresponse*. John Wiley & Sons.

Särndal, Carl-Erik, Bengt Swensson, and Jan Wretman. 2003. *Model Assisted Survey Sampling*. Springer Science & Business Media.

Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data*. London: Chapman; Hall.

Schulte Nordholt, E. 1996. "The Used Techniques for the Imputation of Wave 1 Data of the ECHP." Research report doc. PAN 66/96, Eurostat, Luxembourg.

Schulte Nordholt, E, and J Hooft Van Huijsduijnen. 1997. "The Treatment of Item Nonresponse During the Editing of Survey Results." In *New Techniques and Technologies for Statistics II: Proceedings of the Second Bonn Seminar*, 55–62.

Shao, Jun. 2002. "Replication Methods for Variance Estimation in Complex Surveys with Imputed Data." *Survey Nonresponse*, 303–14.

Shao, Jun, and Randy R Sitter. 1996. "Bootstrap for Imputed Survey Data." *Journal of the American Statistical Association* 91 (435): 1278–88.

Shlomo, N, T De Waal, and J Pannekoek. 2009. "Mass Imputation for Building a Numerical Statistical Database." In *UNECE Statistical Data Editing Workshop, Neuchatel*.

Skinner, Chris J, David Holt, and TM Fred Smith. 1989. *Analysis of Complex Surveys*. Wiley.

Solow, Robert M. 1956. "A Contribution to the Theory of Economic Growth." *The Quarterly Journal of Economics* 70 (1): 65–94.

Sonquist, John A, Elizabeth Lauh Baker, and James N Morgan. 1971. *Searching for Structure*. Survey Research Center, Inst. for Social Research, University of Michigan.

SPSS. 1998. *AnswerTree 2.0 User's Guide*. Chicago.

Stanger, Michael et al. 2007. "Empalme Del PIB y de Los Componentes Del Gasto: Series Anuales y Trimestrales 1986-2002, Base 2003." Central Bank of Chile.

Steele, Russell J, Naisyin Wang, and Adrian E Raftery. 2010. "Inference from Multiple Imputation for Missing Data Using Mixtures of Normals." *Statistical Methodology* 7 (3): 351–65.

Stuart, A., and J. K. Ord. 1991. *Kendall's Advanced Theory of Statistics*. Fifth. Vol. 2. Oxford University Press, New York.

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.

Tukey, John W. 1977. "Modern Techniques in Data Analysis." *NSF-Sponsored Regional*

Research Conference at Southeastern Massachusetts University, North Dartmouth, MA.

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

———. 2018. *Flexible Imputation of Missing Data*. CRC press.

White, IR, G Molenberghs, G Fitzmaurice, and M Kenward. 2015. "Handbook of Missing Data Methodology." *Molenberghs, G., Fitzmaurice, G., Kenward, MG, Tsiastis, A., & Verbeke, G.(Eds.)*, 471–89.

Wilks, Samuel Stanley. 1964. "Mathematical Statistics." J. Wiley.

Willeboordse, Ad. 1998. *Handbook on the Design and Implementation of Business Surveys*. EUR-OP.

Wolter, KM. 1985. "Introduction to Variance Estimation. Springer-Verlag." *New York, Inc.*