Imputación Múltiple

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

Felipe Molina Jaque

Jefe Subdepartamento: José Bustos

Contents

1 Los fundamentos del enfoque de la imputación múltiple **2**

2 Implementación general de los métodos de imputación múltiple **3**
2.1 Consideraciones técnicas en la aplicación de los métodos de imputación 5
2.2 Número de imputaciones a realizar 7
2.3 Imputación basada en Regresión lineal con Data Augmentation 9
2.4 Imputación multivariada para datos continuos 10
2.5 Imputación de variables binarias 12
2.6 Imputación de variables categóricas no ordenadas 13
2.7 Imputación multivariada para datos categóricos 14
2.8 Imputación mixta de variables categóricas y continuas 14

3 Analisis de sensibilidad **14**

4 Consideraciones de imputar en encuestas complejas **14**

Referencias **15**

1 Los fundamentos del enfoque de la imputación múltiple

La imputación múltiple (MI, por sus siglas en inglés), introducida por 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 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 R. J. Little and Rubin (2002) 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, $\hat{\theta} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_j$.

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:

$$P(\theta|Y_{\text{obs}}) = \int P(\theta, Y_{\text{mis}}|Y_{\text{obs}})dY_{\text{mis}} \tag{1}$$

$$= \int P(\theta|Y_{\text{obs}}, Y_{\text{mis}})P(Y_{\text{mis}}|Y_{\text{obs}})dY_{\text{mis}} \tag{2}$$

$$= \lim_{M\to\infty} \frac{1}{M} \sum_{m=1}^{M} P(\theta|Y_{\text{obs}}, Y_{\text{mis}}^{(m)}) \tag{3}$$

$$\approx \frac{1}{M} \sum_{m=1}^{M} P(\theta|Y_{\text{obs}}, Y_{\text{mis}}^{(m)}) \tag{4}$$

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 (2), 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

$$E(\theta|Y_{\text{obs}}) = E(E[\theta|Y_{\text{obs}}, Y_{\text{mis}}]|Y_{\text{obs}}) \tag{5}$$

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

$$\bar{\theta} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m \tag{6}$$

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:

$$V(\theta|Y_{\text{obs}}) = E(V(\theta|Y_{\text{obs}}, Y_{\text{mis}})|Y_{\text{obs}}) + V(E(\theta|Y_{\text{obs}}, Y_{\text{mis}})|Y_{\text{obs}}) \tag{7}$$

La primera componente de (7) 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

$$\bar{U} = \frac{1}{M} \sum_{m=1}^{M} \bar{U}_m \tag{8}$$

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

$$B = \frac{1}{M-1} \sum_{m=1}^{M} (\hat{\theta}_m - \bar{\theta})(\hat{\theta}_m - \bar{\theta})' \tag{9}$$

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. 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.

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 (Rubin 1988; Van Buuren 2018; 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:

$$\frac{\theta - \bar{\theta}}{\sqrt{T}} \sim t_\nu \tag{10}$$

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

$$\bar{\theta} \pm t_{\nu,1-\alpha/2}\sqrt{T} \tag{11}$$

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

$$P_s = \Pr\left[F_{1,\nu} > \frac{(\theta_0 - \bar{\theta})^2}{T}\right] \tag{12}$$

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

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

$$\nu = (M-1)\left[1 + \frac{\bar{U}}{(1+\frac{1}{M})B}\right]^2 \tag{13}$$

La ecuacion anterior puede ser reescrita como

$$\nu = (M-1)\left[1 + \frac{1}{r_M}\right]^2 \tag{14}$$

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

$$\lambda_M = \frac{(1+m^{-1})B}{T} \tag{15}$$

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 (Rubin 1988):

$$\gamma_M = \frac{r_M + \frac{2}{\nu+3}}{1+r_M} \tag{16}$$

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 (14) 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 (14) 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

$$\nu_{obs} = \frac{\nu_{com}+1}{\nu_{com}+3}\nu_{com}(1-\lambda) \tag{17}$$

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

$$\nu = \frac{\nu_{old}\nu_{obs}}{\nu_{old}+\nu_{obs}} \tag{18}$$

Van Buuren (2018) señala que para la cantidad de la ecuación (18) siempre se tiene que $\nu \leq \nu_{com}$. Si $\nu_{com} = \infty$ entonces (18) se reduce a (14).

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

$$T_m = \left(1 + \frac{\gamma_0}{m}\right) T_\infty \tag{19}$$

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

$$P(\theta, Y_{\text{mis}}|Y_{\text{obs}}) = P(\theta|Y_{\text{mis}}, Y_{\text{obs}})P(Y_{\text{mis}}|Y_{\text{obs}}) = P(Y_{\text{mis}}|\theta, Y_{\text{obs}})P(\theta|Y_{\text{obs}}) \tag{20}$$

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 2021).

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, $Y^{(m)}_{\text{mis}}$.
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 2021).

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.

$$Y_{\text{obs}}|X;\boldsymbol{\beta} \sim \text{N}\left(X\boldsymbol{\beta}, \sigma^2\right) \tag{21}$$

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

$$\sigma^* = \hat{\sigma}\sqrt{(n_{\text{obs}} - k)/\text{g}} \tag{22}$$

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

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

$$\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}} + \frac{\sigma^*}{\hat{\sigma}}\mathbf{u_1}V^{\frac{1}{2}} \tag{23}$$

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

$$y_i^* = \boldsymbol{\beta}\mathbf{x}_i + u_{2i}\sigma^* \tag{24}$$

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

2.4 Imputación multivariada para datos continuos

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:

$$f(Y_1|Y_2) = \mathcal{N}(\beta_{01} + \beta_{11}Y_2, \tau_1^2), \tag{25}$$

$$f(Y_2|Y_1) = \mathcal{N}(\beta_{02} + \beta_{12}Y_1, \tau_2^2). \tag{26}$$

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 (25) y (26), 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.

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

$$Y_i|Z_i, \mu, \Sigma \sim \mathcal{N}(\mu_{Z_i}, \Sigma_{Z_i}), \tag{27}$$

$$Z_i|\pi \sim \text{Multinomial}(\pi_1, \ldots, \pi_K). \tag{28}$$

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

$$f(Y_i|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \Sigma_k). \tag{29}$$

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) $(k = 1, \dots K)$ 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 (Gelman et al. 2013). Consideremos además, $\pi(\pi_1, \dots, \pi_K) \sim$ Dirichlet$(1, \dots, 1)$, 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 (27) y (28), 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 2020, 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 2021).

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 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 2020, 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.

2.5 Imputación de variables binarias

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

$$\text{logit}P(Y_{\text{obs}} = 1|\mathbf{x}\beta) = \boldsymbol{\beta x} \tag{30}$$

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

$$y_i^* = \begin{cases} 1 & \text{si } \ u_i < p_i^* \\ \\ 0 & \text{En otro caso.} \end{cases} \tag{31}$$

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

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 $l = 1$)

$$P(y_{\text{obs}} = l|\mathbf{x}, \boldsymbol{\beta}) = \left[\sum_{l'=1}^{L} \exp(\beta_{l'}\mathbf{x})\right]^{-1} \exp(\boldsymbol{\beta}_l\mathbf{x}) \tag{32}$$

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

$$y_i^* = 1 + \sum_{l'=1}^{L} I(u_i > c_{il}) \tag{33}$$

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.

2.7 Imputación multivariada para datos categóricos

2.8 Imputación mixta de variables categóricas y continuas

3 Analisis de sensibilidad

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.

Referencias

Barnard, John, and Donald B Rubin. 1999. "Miscellanea. Small-Sample Degrees of Freedom with Multiple Imputation." *Biometrika* 86 (4): 948–55.

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.

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.

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.

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. 2021. *Multiple Imputation of Missing Data in Practice: Basic Theory and Analysis Strategies*. CRC Press.

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.

Kish, Leslie. 1965. *Survey Sampling*. John Wiley & Sons.

Little, Roderick J. A., and Donald B. Rubin. 2020. *Statistical Analysis with Missing Data*. 3rd ed. New York: Wiley.

Little, Roderick JA, and Donald B Rubin. 2002. "Bayes and Multiple Imputation." *Statistical Analysis with Missing Data*, 200–220.

Liu, Chuanhai. 1995. "Missing Data Imputation Using the Multivariate t Distribution." *Journal of Multivariate Analysis* 53: 139–58.

Marron, JS, and MP Wand. 1992. "Exact Mean Integrated Squared Error." *The Annals of Statistics* 20: 712–36.

Medina, Fernando, and Marco Galván. 2007. *Imputación de Datos: Teoría y Práctica*. Cepal.

Molenberghs, Geert, Garrett Fitzmaurice, Michael G Kenward, Anastasios Tsiatis, and Geert Verbeke. 2014. *Handbook of Missing Data Methodology*. CRC Press.

Priebe, Carey E. 1994. "Adaptive Mixtures." *Journal of the American Statistical Association* 89: 796–806.

Royston, Patrick. 2004. "Multiple Imputation of Missing Values." *The Stata Journal* 4 (3): 227–41.

Rubin, Donald B. 1988. "An Overview of Multiple Imputation." In *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 79:84. Citeseer.

———. 2004. *Multiple Imputation for Nonresponse in Surveys*. Vol. 81. John Wiley & Sons.

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.

Tukey, John W. 1977. "Modern Techniques in Data Analysis." *NSF-Sponsored Regional Research Conference at Southeastern Massachusetts University, North Dartmouth, MA*.

Van Buuren, Stef. 2018. *Flexible Imputation of Missing Data*. CRC press.