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 Imputación Múltiple

1.1 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 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, $\hat{\theta} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_j$.

1.2 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 $\hat{Q}$ 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:

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

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 dis-

tribución 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 (1), 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

$$E(\theta|Y_{\text{obs}}) = E(E[\theta|Y_{\text{obs}}, Y_{\text{mis}}]|Y_{\text{obs}}) \tag{2}$$

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

$$\bar{\theta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m \tag{3}$$

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:

$$V(\theta|Y_{\text{obs}}) = E(V(\theta|Y_{\text{obs}}, Y_{\text{mis}})|Y_{\text{obs}}) + V(E(\theta|Y_{\text{obs}}, Y_{\text{mis}})|Y_{\text{obs}}) \tag{4}$$

La primera componente de (4) 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

$$\bar{U} = \frac{1}{M}\sum_{m=1}^{M}\bar{U}_m \tag{5}$$

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

$$B = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\theta}_m - \bar{\theta})(\hat{\theta}_m - \bar{\theta})' \tag{6}$$

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.

1.2.1 Ignorabilidad en la imputación múltiple

En la sección anterior, se describió la imputación múltiple bajo el supuesto de "ignorabilidad". Note que la sección (INTRODUCIR SECCION EM DONDE SE TOCA ESTE TEMA) trata el tema de los datos faltantes a partir de la distribución de $P(Y_{\text{mis}}|Y_{\text{obs}}, R)$, lo que quiere decir que la distribución condicional de Y_{mis} depende de los datos observados Y_{obs} y del mecanismo de respuesta R. En otras palabras, los datos faltantes no solo dependen de los datos observados, si no que también depende de cúales valores son faltantes y cuáles fueron observados. En las siguientes secciones, se utilizarán paquetes estadísticos para ejemplificar procesos de imputación múltiple, dichos paquetes asumen la "ignorabilidad" en el sentido que omiten la matriz R en los cálculos de las distribuciones (Van Buuren (2018)). Asumir lo anterior, para la mayoría de los casos prácticos es equivalente al concepto MAR descrito en este documento.

1.2.2 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:

$$
\frac{\theta - \bar{\theta}}{\sqrt{T}} \sim t_\nu \tag{7}
$$

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

$$
\bar{\theta} \pm t_{\nu,1-\alpha/2}\sqrt{T} \tag{8}
$$

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

$$P_s = \Pr\left[F_{1,\nu} > \frac{(\theta_0 - \bar{\theta})^2}{T}\right] \tag{9}$$

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:

$$\nu = (M-1)\left[1 + \frac{\bar{U}}{(1+\frac{1}{M})B}\right]^2 \tag{10}$$

La ecuacion anterior puede ser reescrita como

$$\nu = (M-1)\left[1 + \frac{1}{r_M}\right]^2 \tag{11}$$

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

$$\lambda_M = \frac{(1+m^{-1})B}{T} \tag{12}$$

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)):

$$\gamma_M = \frac{r_M + \frac{2}{\nu+3}}{1+r_M} \tag{13}$$

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.

1.2.3 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

$$T_m = \left(1 + \frac{\gamma_0}{m}\right) T_\infty \tag{14}$$

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 (**???**), 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. (**???**) sugiere un criterio para determinar m basado en el coeficiente de confianza $t_\nu\sqrt{T}$, y propone que el coeficiente de variación de $\log\left(t_\nu\sqrt{T}\right)$ 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, (**???**) 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 seleccionar m, de tal forma que el ancho del intervalo de confianza del 95% no excediera en más del 10% su valor verdadero. Bajo este criterio, Bodner propuso recomendaciones específicas para m en relación con diferentes valores de γ_0. Para $\gamma_0 = (0.05, 0.1, 0.2, 0.3, 0.5, 0.7, 0.9)$, Bodner recomendó los siguientes valores de m: $m = (3, 6, 12, 24, 59, 114, 258)$ respectivamente.

1.3 Métodos de imputación

1.3.1 Imputación basada en Regresión lineal

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

$$Y_{\text{obs}}|X;\boldsymbol{\beta} \sim \text{N}\left(X\boldsymbol{\beta}, \sigma^2\right) \tag{15}$$

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

$$\sigma^* = \hat{\sigma}\sqrt{(n_{\text{obs}} - k)/\text{g}} \tag{16}$$

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

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

$$\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}} + \frac{\sigma^*}{\hat{\sigma}}\mathbf{u_1}V^{\frac{1}{2}} \tag{17}$$

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

$$y_i^* = \boldsymbol{\beta}\mathbf{x}_i + u_{2i}\sigma^* \tag{18}$$

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

1.3.2 Imputación de variables binarias

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

$$\text{logit}P(Y_{\text{obs}} = 1|\mathbf{x}\beta) = \boldsymbol{\beta}\boldsymbol{x} \tag{19}$$

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

$$y_i^* = \begin{cases} 1 & \text{si } \ u_i < p_i^* \\ \\ 0 & \text{En otro caso.} \end{cases} \tag{20}$$

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

1.3.3 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 multinumal, donde cada categoría tiene una regresión logistica que se compara con otra categoría determinada (digamos $l = 1$)

$$P(y_{\text{obs}} = l|\mathbf{x}, \boldsymbol{\beta}) = \left[\sum_{l'=1}^{L} \exp(\beta_{l'} \mathbf{x})\right]^{-1} \exp(\boldsymbol{\beta}_l \mathbf{x}) \tag{21}$$

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

$$y_i^* = 1 + \sum_{l'=1}^{L} I(u_i > c_{il}) \tag{22}$$

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.

Referencias

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

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

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). Elsevier: 351–65.

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