Imputacion Multiple

Felipe Molina - Subdepartamento de investigación estadística

2023-05-09

Índice

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 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 (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:

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

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

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

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:

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

La primera componente de (2) 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

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

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

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

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.

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

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

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)):

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

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.

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

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

La ecuacion anterior puede ser reescrita como

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

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.