Cómputo Estadístico

Diego Godinez Bravo

Centro de Investigación en Matemáticas

Maestría en Cómputo Estadístico

Noviembre 2024

Evaluación de Pruebas de Hipótesis para Datos Faltantes Completamente al Azar (MCAR)

Muchos análisis estadísticos de datos con valores faltantes suponen que los datos se pierden de manera completamente aleatoria (*Missing Completely at Random*, MCAR); es decir, que la falta de datos no depende de ninguna de las variables en el conjunto sujeto a análisis.

Little (1988) propone un solo estadístico de prueba para evaluar el supuesto de datos faltantes de manera completamente aleatoria (MCAR). Dado que las pruebas formales definidas con anterioridad consideraban la evaluación de $(p-1)$ pruebas t, donde, para cada variable con datos faltantes, la muestra se dividía en dos casos, comparando los valores de la media para las variables restantes a través de pruebas t para dos muestras, resultando en el número de pruebas mencionado con anterioridad.

Sea y una matriz de n observaciones en p variables $(n \times p)$, y r una matriz indicadora de datos faltantes $(n \times p)$, donde $r_{ij} = 1$ si r_{ij} es dato faltante y 0 en caso contrario. El modelo para el conjunto de datos y está definido por una distribución $f(y|\theta)$, donde el parámetro θ se desconoce. Mientras que el mecanismo de datos faltantes dado y se define por la distribución $f(r|y, \psi)$, donde el parámetro ψ se desconoce.

Consideremos $y = (y_{obs}, y_{mis})$, donde y_{obs} representa los valores observados de y, y y_{mis} representa los valores faltantes. Se definen los datos faltantes como MCAR si $f(r|y_{obs}, y_{mis}, \psi) = f(r|\psi)$ para todo y_{obs} y y_{mis}. En otras palabras, los datos faltantes no dependen de los datos observados ni de los datos faltantes de y. Esta definición fue propuesta por Rubin (1976), quien además propuso una condición débil sobre el mecanismo de datos faltantes llamada datos faltantes perdidos de manera aleatoria (*Missing at Random*, MAR). Donde a diferencia del caso MCAR, los valores faltantes no dependen de y_{mis}, pero podrían depender de y_{obs}.

Evaluación del Supuesto MCAR para Datos Multivariados

La importancia de evaluar el supuesto de MCAR en el contexto del manejo de datos faltantes radica en 3 puntos esenciales.

- Métodos como el análisis restrictivo a casos completos o la eliminación por pares se basan en la suposición de MCAR para generar estimados insesgados. Si los datos no se pierden completamente al azar, estos métodos pueden introducir sesgos, conduciendo a conclusiones incorrectas.
- La estimación por Máxima Verosimilitud bajo valores faltantes no requiere estrictamente de la suposición MCAR. Sin embargo, es sensible a la especificación incorrecta del modelo cuando los datos se desvían del caso MCAR. Un ejemplo claro es el caso normal multivariado, el cual puede generar estimadores consistentes bajo la suposición MCAR, pero si los datos no siguen un patrón MCAR, pequeños errores en la especificación del modelo pueden dar lugar a estimaciones con un gran sesgo.
- La precisión de los errores estándar derivados de la matriz de información esperada depende de que los datos sean MCAR. Si no se cumple el supuesto MCAR, los errores estándar pueden no

reflejar con precisión la incertidumbre de las estimaciones de los parámetros, lo cual conduce a inferencias inválidas.

Estadístico de Prueba con Σ *Desconocida.* Considerando el caso donde se conoce Σ. Sea μ^* el estimador de máxima verosimilitud de μ, suponiendo que los datos faltantes son MCAR, Little (1988) propone el siguiente estadístico:

$$d_0^2 = \sum_j = 1^J m_j(\bar{y}_{obs} - \mu^*_{obs,j}) \sum_{obs,j}^{-1} (\bar{y}_{obs,j} - \mu^*_{obs,j})^T \tag{0.1}$$

Suponiendo que y_i sigue una distribución normal multivariada con media μ y matriz de covarianzas Σ. Si los datos siguen un patrón MCAR, entonces:

$$(y_{obs,j}|r_i)\ N(\mu_{obs,j}, \Sigma_{obs,j}) \quad i \in S_j, 1 \leq j \leq J \tag{0.2}$$

Si los datos no son MCAR, entonces las medias de las variables observadas pueden variar, por lo que:

$$(y_{obs,j}|r_i)\ N(v_{obs,j}, \Sigma_{obs,j}) \quad i \in S_j, 1 \leq j \leq J \tag{0.3}$$

donde $v_{obs,j}, j = 1, ..., J$ corresponde a los vectores $(1 \times p_j)$ de los parámetros de la media que son distintos para cada patrón j. Donde cada patrón es una configuración única de valores faltantes entre las distintas variables.

d_0^2 es el estadístico de razón de verosimilitud para evaluar el modelo **(0.2)** frente al modelo **(0.3)**. Bajo el modelo **(0.2)**, el estadístico d_0^2 sigue una distribución χ^2 con distribución $f = \sum p_j - p$ df.

Otra consideración importante es que, si los datos siguen un patrón MCAR y y_i tiene cualquier distribución con media μ y matriz de covarianzas Σ, d_0^2 es asintóticamente χ^2 con f df. Es decir, para muestras grandes, los resultados de los métodos estadísticos se vuelven menos sensibles a desviaciones de la suposición de normalidad (robustez del análisis a violaciones de normalidad).

Estadístico de Prueba con Σ *Conocida.* En el caso de no conocer Σ, Little (1988) propone el reemplazo de μ^* y Σ en **(0.1)** con $\hat{\mu}$ y $\bar{\Sigma}$ del algoritmo normal multivariado de máxima verosimilitud, dando lugar al estadístico.

$$d^2 = \sum_{j=1}^{J} m_j(\bar{y}_{obs,j} - \hat{\mu}_{obs,j})\bar{\Sigma}^{-1}_{obs,j}(\bar{y}_{obs,j} - \hat{\mu}_{obs,j})^T \tag{0.4}$$

Suponiendo que los datos observados contienen suficiente información para que cada par de variables en el conjunto de datos tenga suficientes observaciones superpuestas de manera que los valores de la media, varianza y covarianza sean posibles de estimar. Si los datos siguen un patrón MCAR y la distribución de y_i tiene momentos cuartos finitos, $\bar{\Sigma}$ es un estimador consistente de Σ. Bajo estas condiciones, d^2 seguirá aproximadamente una distribución χ^2 con f grados de libertad cuando el tamaño de muestra sea grande.

Consideraciones Finales

La prueba de Little (1988) se basa en comparar las medias y covarianzas de la muestra en presencia de diferentes patrones de datos faltantes. Si los datos son verdaderamente MCAR, las medias y covarianzas de $y_{obs,i}$ a lo largo de diferentes patrones de valores faltantes no deberían diferir de manera significativa.

En conclusión, bajo el supuesto datos faltantes de manera completamente aleatoria (MCAR), la distribución de los valores observados $y_{obs,j}$ depende únicamente de las medias y covarianzas de los datos observados, no de los datos faltantes. Lo cual implica que **los datos observados pueden ser tratados como una muestra representativa de los datos completos.**

Referencias

Little, R. J. A. (1988). *A Test of Missing Completely at Random for Multivariate Data with Missing Values*. Journal of the American Statistical Association, 83(404), 1198–1202. https://doi.org/10.1080/01621459.1988.10478722