Centro de Investigación en Matemáticas (CIMAT).
Unidad Monterrey

Computo Estadístico
Tarea 5

Marcelo Alberto Sanchez Zaragoza

16 de noviembre de 2021

Como punto de partida, es útil distinguir entre patrones de datos faltantes y mecanismos de datos faltantes. Un patrón de datos faltantes se refiere a la configuración de los datos observados y faltantes en un conjunto de datos. Los mecanismos de datos faltantes, describen posibles relaciones entre variables medidas y la probabilidad de datos faltantes. Aunque los mecanismos de datos faltantes no ofrecen una explicación de la causa de datos faltantes, si representan relaciones matemáticas genéricas entre los datos observados y los ausentes.

1. MAR

Los datos se consideran faltantes en forma aleatoria(MAR), cuando la probabilidad de datos faltantes en la variable Y está relacionada o otra variables(o variables) en el módelo de análisis, pero no a los valores de Y en si misma.
Dicho de otro modo, no existe relación entre la propensión a tener datos faltantes en Y y en los valores de Y, luego de haber dejado fuera otras variables. El término 'datos faltantes en forma aleatoria' es un poco confuso, ya que implica que los datos están faltando de una manera azarosa. MAR significa en realidad que existe una relación sistemática entre una o más variables y la probabilidad de tener datos faltantes.

Dado que en MAR se supone que la distribución de los datos faltante depende sólo de

la observación de datos entonces:

$$Pr\left(R|Y_m, Y_o, X\right) = Pr\left(R|Y_o, X\right) \tag{1.1}$$

2. MCAR

El mecanismo de datos faltantes en forma completamente aleatorio(MCAR) es lo que los investigadores toman como datos ausentes en forma puramente azarosa. La definición formal de MCAR requiere que la probabilidad de los datos faltantes en la variable Y no esté relacionada con otras variables y que esté relacionada con los valores de Y en si misma. Dicho de otro modo, los puntos de datos observados son una muestra simple aleatoria de los scores que serían analizados en el casos que los datos estuvieran completos. Nótese que el mecanismo MCAR determina una condición más restrictiva que el mecanismo MAR porque asume que la ausencia de datos no está relacionada en absoluto con los datos.
Para MCAR partimos de que los datos faltantes son completamente independientes tanto de los errores como de los datos observados.

$$Pr\left(R|Y_m, Y_o, X\right) = Pr(R) \tag{2.1}$$

En la mayor parte de los casos los faltantes no cumplen MCAR. Por ejemplo, la tasa de no respuesta suele diferir entre blancos y negros, esto es un indicador de que la pregunta ingresos no cumple los supuestos de MCAR.
El supuesto de MAR es más general, por ejemplo si observamos genero, raza, educación y edad para todos los encuestados, enotnces ingresos es MAR si la porbabilidad de no-respuesta para esta pregunta depende únicamente de estas variables completamente observada.

3. MCAR

La prueba de MCAR se modela basandose de una distribución normal miltivariada p-dimensional con vector de media μ y matriz de covarianza σ. En algunas ocasiones existen consideraciones especiales que no satisfacen la normalidad pero la prueba de Little todavía funciona para los vectores cuantitativos y_i.
El estadístico de Little's X tiene la siguiente forma:

$$d_0^2 = \sum_{j=1}^{J} n_j \left(\bar{y}_{oj} - \mu_{oj}\right)^T \Sigma_{oj}^{-1} \left(\bar{y}_{oj} - \mu_{oj}\right) \tag{3.1}$$

Partiendo de que los datos son MCAR, se cumple la siguiente hipótesis nula:

$$H_0 : \ y_{o,i}|r_i \sim N\left(\mu_o, \Sigma_{oj}\right) \ \ i \in I_j, 1 \leq j \leq J \tag{3.2}$$

donde μ_{0j} es un subvector del vector medio μ.
Cabe descatar que en la practica como μ y σ son usualmente desconocidos Little propusó reemplazar los parámetros desconocidos con los estimadores insesgados $\hat{\mu}$ y $\hat{\sigma}$. De este modo Σ_{oj} es reemplazado por la submatrix $\tilde{\Sigma_{oj}}$ de $\tilde{\Sigma}$, que sustituido es:

$$d^2 = \sum_{j=1}^{J} n_j \left(\bar{y_{oj}} - \hat{\mu}_{oj}\right)^T \tilde{\Sigma_{oj}}^{-1} \left(\bar{y_{oj}} - \hat{\mu}_{oj}\right) \tag{3.3}$$

Cabe aclarar que los estimadores $\hat{\mu}$ y $\hat{\Sigma}$ puden obtenerse con el algoritmo EM visto en clase sobre los datos observado Y_o.