Prueba de Little para Identificación de MCAR (Missing Completely at Random)

Definición

La prueba de Little (Little's MCAR Test) es una prueba estadística utilizada para evaluar si los datos faltantes en un conjunto de datos son completamente al azar (MCAR).

Hipótesis

- **Hipótesis nula (H0)**: Los datos están faltando completamente al azar (MCAR).
- **Hipótesis alternativa (H1)**: Los datos no están faltando completamente al azar (no MCAR).

Cálculo del Estadístico

La prueba de Little utiliza un estadístico de prueba basado en una combinación de chi-cuadrado que compara las medias y las varianzas entre diferentes patrones de datos faltantes.

El estadístico de prueba de Little se calcula utilizando la siguiente fórmula:

$$\chi^2_{calc} = \sum_{g=1}^{G} N_g \left[\mathrm{tr}\left(\mathbf{S}_g \mathbf{S}_p^{-1} \right) + (\bar{\mathbf{x}}_g - \bar{\mathbf{x}}_p)^T \mathbf{S}_p^{-1} (\bar{\mathbf{x}}_g - \bar{\mathbf{x}}_p) - (n_g - 1) \right]$$

Donde:

- G es el número de grupos definidos por patrones de valores faltantes.
- N_g es el número de observaciones en el grupo g.
- $\mathbf{S}_g$ es la matriz de covarianza dentro del grupo g.
- $\mathbf{S}_p$ es la matriz de covarianza combinada de los grupos.
- $\bar{\mathbf{x}}_g$ es el vector de medias del grupo g.
- $\bar{\mathbf{x}}_p$ es el vector de medias combinado de los grupos.
- n_g es el número de observaciones completas en el grupo g.
- tr denota la traza de una matriz (la suma de sus elementos diagonales).

Identificación de Grupos

1. **Matriz de Datos**: Supongamos que tenemos un conjunto de datos con n observaciones y p variables, representado por una matriz de datos $\mathbf{X}$.

2. **Matriz de Indicadores de Datos Faltantes**: Construimos una matriz de indicadores de datos faltantes $\mathbf{R}$ de tamaño $n \times p$,

$$\mathbf{R} = \begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1p} \\ R_{21} & R_{22} & \cdots & R_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ R_{n1} & R_{n2} & \cdots & R_{np} \end{bmatrix}$$

   Donde el elemento R_{ij} es:

$$R_{ij} = \begin{cases} 1 & \text{si } X_{ij} \text{ está presente} \\ 0 & \text{si } X_{ij} \text{ está ausente} \end{cases}$$

3. **Identificación de Patrones Únicos**: Cada fila de la matriz $\mathbf{R}$ representa un patrón de datos faltantes para una observación. Identificamos los patrones únicos de datos faltantes agrupando las filas de $\mathbf{R}$ que son idénticas.

4. **Formación de Grupos**: Agrupamos las observaciones que comparten el mismo patrón de datos faltantes. Cada grupo corresponde a un patrón único de la matriz $\mathbf{R}$.

Supuestos

Para que la prueba de Little sea válida, se deben cumplir los siguientes supuestos:

- **Independencia de Observaciones**: Las observaciones deben ser independientes entre sí.
- **Tamaño de Muestra Adecuado**: Un tamaño de muestra suficiente es necesario para obtener resultados precisos y confiables.
- **Homogeneidad de Varianzas**: Las varianzas dentro de los grupos de patrones de datos faltantes deben ser homogéneas.
- **Distribución Multinormal (opcional)**: Aunque no es un requisito estricto, la precisión de la prueba mejora si las variables siguen una distribución aproximadamente normal dentro de cada patrón de valores faltantes.

Interpretación de los Resultados

Sabiendo que χ^2_{calc} sigue una distribución chi cuadrada el valor crítico se obtiene a traves de la tabla de la distribución, dado un nivel de significancia α, usualmente, 0.05 y unos grados de libertad que, para esta prueba, son $\mathrm{df} = \sum_{g=1}^{G}(N_g - 1)$, donde N_g es el número de observaciones en el grupo g el valor crítico $\chi^2_{\alpha,df}$ y podemos compararlo con el estadístico calculado:

- Si $X^2_{calc} > \chi^2_{\alpha,df}$ se rechaza la hipótesis nula (los datos no son MCAR).
- Si $X^2_{calc} \leq \chi^2_{\alpha,df}$ no se rechaza la hipótesis nula (los datos pueden ser MCAR).

Ejemplo

Paso 1: Definir los Datos

Observación	X1	X2
1	10	NA
2	3	6
3	NA	2
4	7	NA
5	2	4

Paso 2: Definir los Grupos de Datos Faltantes

1. **Grupo 1**: Observaciones con datos completos en X1 y X2 (Observaciones 2 y 5).
2. **Grupo 2**: Observaciones con X2 faltante (Observaciones 1 y 4).
3. **Grupo 3**: Observaciones con X1 faltante (Observación 3).

Paso 3: Calcular las Medias y Matrices de Covarianza para Cada Grupo

Grupo 1 (Datos completos en X1 y X2)

- **Observaciones**: 2 y 5
- **Media**:

$$\bar{\mathbf{x}}_1 = \left[\frac{3+2}{2}, \frac{6+4}{2}\right] = [2.5, 5.0]$$

- **Matriz de Covarianza**:

$$\mathbf{S}_1 = \begin{bmatrix} \frac{(3-2.5)^2+(2-2.5)^2}{1} & \frac{(3-2.5)(6-5)+(2-2.5)(4-5)}{1} \\ \frac{(3-2.5)(6-5)+(2-2.5)(4-5)}{1} & \frac{(6-5)^2+(4-5)^2}{1} \end{bmatrix} = \begin{bmatrix} 0.5 & 1.0 \\ 1.0 & 2.0 \end{bmatrix}$$

Grupo 2 (X2 faltante)

- **Observaciones**: 1 y 4
- **Media**:

$$\bar{\mathbf{x}}_2 = \left[\frac{10+7}{2}, \mathrm{NA}\right] = [8.5, \mathrm{NA}]$$

- **Matriz de Covarianza** (Solo se puede calcular para X1):

$$\mathbf{S}_2 = \begin{bmatrix} \frac{(10-8.5)^2+(7-8.5)^2}{1} & \mathrm{NA} \\ \mathrm{NA} & \mathrm{NA} \end{bmatrix} = \begin{bmatrix} 4.5 & \mathrm{NA} \\ \mathrm{NA} & \mathrm{NA} \end{bmatrix}$$

Grupo 3 (X1 faltante)

- **Observación**: 3
- **Media**:

$$\bar{\mathbf{x}}_3 = [\mathrm{NA}, 2]$$

- **Matriz de Covarianza**: No se puede calcular con una sola observación.

Paso 4: Calcular la Media y Matriz de Covarianza Combinada

La media combinada $\bar{\mathbf{x}}_p$ y la matriz de covarianza combinada $\mathbf{S}_p$ se calculan combinando todas las observaciones completas:

- **Media combinada**:

$$\bar{\mathbf{x}}_p = \left[\frac{3+2}{2}, \frac{6+4}{2}\right] = [2.5, 5.0]$$

- **Matriz de Covarianza Combinada**:

$$\mathbf{S}_p = \begin{bmatrix} 0.5 & 1.0 \\ 1.0 & 2.0 \end{bmatrix}$$

Paso 5: Calcular el Estadístico de Prueba

Usamos la fórmula del estadístico de prueba de Little:

$$\chi^2_{\text{calc}} = \sum_{g=1}^{G} N_g \left[\mathrm{tr}\left(\mathbf{S}_g \mathbf{S}_p^{-1}\right) + (\bar{\mathbf{x}}_g - \bar{\mathbf{x}}_p)^T \mathbf{S}_p^{-1} (\bar{\mathbf{x}}_g - \bar{\mathbf{x}}_p) - (n_g - 1)\right]$$

Grupo 1 (Datos completos en X1 y X2)

1. **Número de Observaciones en el Grupo 1**:

$$N_1 = 2$$

2. **Matriz inversa de covarianza combinada $\mathbf{S}_p^{-1}$**:

$$\mathbf{S}_p^{-1} = \begin{bmatrix} 4 & -2 \\ -2 & 1 \end{bmatrix}$$

3. **Media del Grupo 1 ($\bar{\mathbf{x}}_1$)**:

$$\bar{\mathbf{x}}_1 = [2.5, 5.0]$$

4. **Media combinada ($\bar{\mathbf{x}}_p$)**:

$$\bar{\mathbf{x}}_p = [2.5, 5.0]$$

5. $\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_p$:

$$\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_p = [2.5 - 2.5, 5.0 - 5.0] = [0, 0]$$

6. $(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_p)^T \mathbf{S}_p^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_p)$:

$$[0, 0] \begin{bmatrix} 4 & -2 \\ -2 & 1 \end{bmatrix} [0, 0]^T = 0$$

7. **Trazar** $\mathrm{tr}(\mathbf{S}_1 \mathbf{S}_p^{-1})$:

$$\mathrm{tr}\left(\mathbf{S}_1 \mathbf{S}_p^{-1}\right) = 0.5 \cdot 4 + 1.0 \cdot (-2) + 1.0 \cdot (-2) + 2.0 \cdot 1 = 2$$

8. **Estadístico para el Grupo 1**:

$$N_1 \left[\mathrm{tr}(\mathbf{S}_1 \mathbf{S}_p^{-1}) + 0 - (2 - 1) \right] = 2\,[2 + 0 - 1] = 2$$

Grupo 2 (X2 faltante)

1. **Número de Observaciones en el Grupo 2**:

$$N_2 = 2$$

2. **Media del Grupo 2 ($\bar{\mathbf{x}}_2$)**:

$$\bar{\mathbf{x}}_2 = [8.5, \mathrm{NA}]$$

3. $\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_p$:

$$\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_p = [8.5 - 2.5, \mathrm{NA}] = [6.0, \mathrm{NA}]$$

4. **Trazar** $\mathrm{tr}(\mathbf{S}_2 \mathbf{S}_p^{-1})$ (Solo X1):

$$\mathrm{tr}\left(\mathbf{S}_2 \mathbf{S}_p^{-1}\right) = 4.5 \cdot 4 = 18$$

5. **Estadístico para el Grupo 2**:

$$N_2\,[18 + 0 - (2 - 1)] = 2\,[18 - 1] = 34$$

Grupo 3 (X1 faltante)

1. **Número de Observaciones en el Grupo 3**:

$$N_3 = 1$$

2. **Media del Grupo 3 ($\bar{\mathbf{x}}_3$)**:

$$\bar{\mathbf{x}}_3 = [\mathrm{NA}, 2]$$

3. $\bar{\mathbf{x}}_3 - \bar{\mathbf{x}}_p$:

$$\bar{\mathbf{x}}_3 - \bar{\mathbf{x}}_p = [\mathrm{NA}, 2 - 5] = [\mathrm{NA}, -3]$$

4. **Trazar** $\mathrm{tr}(\mathbf{S}_3 \mathbf{S}_p^{-1})$ (No hay datos suficientes para calcular $\mathbf{S}_3$).

5. **Estadístico para el Grupo 3**:

   - No se puede calcular con solo una observación.

Paso 6: Sumar los Estadísticos de Todos los Grupos

$$\chi^2_{\text{calc}} = \chi^2_{\text{Grupo 1}} + \chi^2_{\text{Grupo 2}} = 2 + 34 = 36$$

Paso 7: Comparar con el Valor Crítico

Consultar una tabla de distribución chi-cuadrado con el nivel de significancia deseado (por ejemplo, 0.05) y los grados de libertad calculados. Para este ejemplo simplificado, supongamos que el valor crítico es 5.99 para 2 grados de libertad.

Paso 8: Interpretación

- Si $\chi^2_{\text{calc}} < \chi^2_{\text{crit}}$: No se rechaza la hipótesis nula, los datos podrían ser MCAR.
- Si $\chi^2_{\text{calc}} \geq \chi^2_{\text{crit}}$: Se rechaza la hipótesis nula, los datos no son MCAR.

Conclusión

En este ejemplo simplificado, $\chi^2_{\text{calc}} = 36$ es mayor que el valor crítico de 5.99, por lo que se rechaza la hipótesis nula. Esto sugiere que los datos faltantes no están completamente al azar (no MCAR).