Tarea 1
Ciencia de Datos

Brain de Jesús Salazar, César Ávila, Iván García

12 de septiembre de 2025

Solución del problema 1.

Solución del problema 2.

Solución del problema 3.

Solución del problema 4.

Solución del problema 5. Consideremos una muestra $Y_1, \ldots, Y_n$ de variables aleatorias independientes e idénticamente distribuidas con media μ y varianza σ^2, con indicadores $R_i \in \{0, 1\}$, de tal manera que $R_i \perp Y_i$ para cada $i \in \{1, \ldots, n\}$. Dichas R_i existen pues estamos bajo el modelo MCAR, y se interpretan como $R_i = 1$ si y solo si el dato Y_i fue observado.

Notemos que si n_{obs} representa el número de datos observados, entonces $n_{obs} = \sum_{i=1}^n R_i$. Además, por la definición de los R_i,

$$\overline{Y}_{obs} = \frac{1}{n_{obs}} \sum_{i=1}^n R_i Y_i = \frac{1}{n_{obs}} \sum_{i:R_i=1} Y_i.$$

Así pues, si $\boldsymbol{R} = (R_1, \ldots, R_n)$, entonces

$$\begin{aligned}
\mathbb{E}\left[\overline{Y}_{obs} \,\middle|\, \boldsymbol{R}\right] = \mathbb{E}\left[\frac{1}{\sum_{i=1}^n R_i} \sum_{i=1}^n R_i Y_i \,\middle|\, \boldsymbol{R}\right] &= \frac{1}{\sum_{i=1}^n R_i} \sum_{i=1}^n \mathbb{E}\left[R_i Y_i \mid \boldsymbol{R}\right] \\
&= \frac{1}{\sum_{i=1}^n R_i} \sum_{i=1}^n R_i \mathbb{E}\left[Y_i \mid \boldsymbol{R}\right] \\
&= \frac{1}{\sum_{i=1}^n R_i} \sum_{i=1}^n R_i \mathbb{E}\left[Y_i\right] \\
&= \mu,
\end{aligned}$$

en donde hemos usado que las Y_i son iid con media μ y son independientes de $\boldsymbol{R}$ (por las hipótesis del modelo MCAR). Por consiguiente,

$$\mathbb{E}\left[\overline{Y}_{obs}\right] = \mathbb{E}\left[\mathbb{E}\left[\overline{Y}_{obs} \,\middle|\, \boldsymbol{R}\right]\right] = \mu.$$

Por otra parte,

$$\overline{Y}^2_{obs} = \frac{1}{\left(\sum_{i=1}^n R_i\right)^2} \left[\sum_{i=1}^n R_i^2 Y_i^2 + \sum_{i \neq j} R_i R_j Y_i Y_j\right],$$

por lo que

$$\begin{aligned}
\mathbb{E}\left[\overline{Y}^2_{obs} \,\middle|\, \boldsymbol{R}\right] &= \frac{1}{\left(\sum_{i=1}^n R_i\right)^2} \left[\sum_{i=1}^n R_i^2 \mathbb{E}\left[Y_i^2 \,\middle|\, \boldsymbol{R}\right] + \sum_{i \neq j} R_i R_j \mathbb{E}\left[Y_i Y_j \mid \boldsymbol{R}\right]\right] \\
&= \frac{1}{\left(\sum_{i=1}^n R_i\right)^2} \left[\sum_{i=1}^n R_i^2 \mathbb{E}\left[Y_i^2\right] + \sum_{i \neq j} R_i R_j \mathbb{E}\left[Y_i Y_j\right]\right] \\
&= \frac{1}{\left(\sum_{i=1}^n R_i\right)^2} \left[(\sigma^2 + \mu^2) \sum_{i=1}^n R_i^2 + \mu^2 \sum_{i \neq j} R_i R_j\right] \\
&= \mu^2 + \sigma^2 \frac{\sum_{i=1}^n R_i^2}{\left(\sum_{i=1}^n R_i\right)^2}.
\end{aligned}$$

De lo anterior se sigue que

$$\begin{aligned}\text{Var}\left[\overline{Y}_{obs}\right] = \mathbb{E}\left[\overline{Y}_{obs}^2\right] - \left(\mathbb{E}\left[\overline{Y}_{obs}\right]\right)^2 = \mathbb{E}\left[\mathbb{E}\left[\overline{Y}_{obs}^2 \,\middle|\, \boldsymbol{R}\right]\right] - \mu^2 &= \sigma^2\mathbb{E}\left[\frac{\sum_{i=1}^n R_i^2}{\left(\sum_{i=1}^n R_i\right)^2}\right] \\ &= \sigma^2\mathbb{E}\left[\frac{\sum_{i=1}^n R_i}{\left(\sum_{i=1}^n R_i\right)^2}\right] \\ &= \sigma^2\mathbb{E}\left[\frac{1}{\sum_{i=1}^n R_i}\right] \\ &= \sigma^2\mathbb{E}\left[\frac{1}{n_{obs}}\right] \\ &\geq \frac{\sigma^2}{n} = \text{Var}\left[\ \overline{Y}\ \right],\end{aligned}$$

en donde hemos usado que $R_i^2 = R_i$, pues $R_i \in \{0, 1\}$, para todo $i \in \{1, \ldots, n\}$, y que $n_{obs} \leq n$. De hecho, siguiendo un procedimiento completamente análogo, pero ahora con varianzas condicionales, se sigue que

$$\text{Var}\left[\overline{Y}_{obs} \,\middle|\, \boldsymbol{R}\right] = \frac{\sigma^2}{n_{obs}}.$$

De lo anterior podemos notar que $\overline{Y}_{obs}$ es insesgado, pero que $\text{Var}\left[\overline{Y}_{obs}\right] \geq \text{Var}\left[\ \overline{Y}\ \right]$, por lo que $\overline{Y}_{obs}$ tiene menor eficiencia (posee más varianza, pues la eliminación de datos hace que haya menos de ellos para poder estimar a la media de Y).

Solución del problema 6. Sean $\boldsymbol{Y} = (Y_{obs}, Y_{mis})$ y $\boldsymbol{R}$ el patrón de datos faltantes. Bajo la definición del MAR,

$$\mathbb{P}\left[\boldsymbol{R} \,|\, Y_{obs}, Y_{mis}, \theta, \psi\right] = \mathbb{P}\left[\boldsymbol{R} \,|\, Y_{obs}, \psi\right].$$

Así pues, bajo este modelo,

$$\mathbb{P}\left[\boldsymbol{Y}, \boldsymbol{R} \,|\, \theta, \psi\right] = \mathbb{P}\left[\boldsymbol{Y} \,|\, \theta\right] \mathbb{P}\left[\boldsymbol{R} \,|\, \boldsymbol{Y}, \psi\right] = \mathbb{P}\left[\boldsymbol{Y} \,|\, \theta\right] \mathbb{P}\left[\boldsymbol{R} \,|\, Y_{obs}, \psi\right].$$

Por consiguiente, la verosimilitud de θ está dada por

$$\begin{aligned} L\left(\theta; Y_{obs}, \boldsymbol{R}\right) = \int \mathbb{P}\left[\boldsymbol{Y}, \boldsymbol{R} \,|\, \theta, \psi\right] dY_{mis} &= \int \mathbb{P}\left[\boldsymbol{Y} \,|\, \theta\right] \mathbb{P}\left[\boldsymbol{R} \,|\, Y_{obs}, \psi\right] dY_{mis} \\ &= \mathbb{P}\left[\boldsymbol{R} \,|\, Y_{obs}, \psi\right] \int \mathbb{P}\left[\boldsymbol{Y} \,|\, \theta\right] dY_{mis} \\ &= \mathbb{P}\left[\boldsymbol{R} \,|\, Y_{obs}, \psi\right] \mathbb{P}\left[Y_{obs} \,|\, \theta\right]. \end{aligned}$$

Ya que el factor $\mathbb{P}\left[\boldsymbol{R} \,|\, Y_{obs}, \psi\right]$ no depende de θ, se sigue que $L\left(\theta; Y_{obs}, \boldsymbol{R}\right) \propto \mathbb{P}\left[Y_{obs} \,|\, \theta\right]$. Para ver las condiciones *a priori* que garantizan ignorabilidad bajo el enfoque bayesiano, notemos que

$$\begin{aligned} \mathbb{P}\left[\theta \,|\, Y_{obs}, \boldsymbol{R}\right] = \int \mathbb{P}\left[\theta, \psi \,|\, Y_{obs}, \boldsymbol{R}\right] d\psi &\propto \int \mathbb{P}\left[Y_{obs}, \boldsymbol{R} \,|\, \theta, \psi\right] \pi\left(\theta, \psi\right) d\psi \\ &= \int \mathbb{P}\left[\boldsymbol{R} \,|\, Y_{obs}, \psi\right] \mathbb{P}\left[Y_{obs} \,|\, \theta\right] \pi\left(\theta, \psi\right) d\psi \\ &= \mathbb{P}\left[Y_{obs} \,|\, \theta\right] \int \mathbb{P}\left[\boldsymbol{R} \,|\, Y_{obs}, \psi\right] \pi\left(\theta, \psi\right) d\psi. \end{aligned}$$

Para concluir ignorabilidad buscamos que $L\left(\theta \,|\, Y_{obs}, \boldsymbol{R}\right) \propto \pi(\theta)\mathbb{P}\left[Y_{obs} \,|\, \theta\right]$, y ya que la integral anterior depende de θ solamente a través del factor $\pi\left(\theta, \psi\right)$, dicha ignorabilidad se logra cuando $\pi(\theta, \psi) = \pi(\theta)\pi(\psi)$. Es decir, si en la *a priori* se pide indistinguibilidad de los parámetros (i.e. que θ y ψ sean independientes), entonces el mecanismo es ignorable para inferir θ.

Solución del problema 7.

Solución del problema 8.

Solución del problema 9.

Solución del problema 10.

Solución del problema 11.

Solución del problema 12.

Solución del problema 13.