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 $Y_1, \dots, Y_n$ 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 $i \in \{1, \dots, n\}$. 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 $\boldsymbol{R} = (R_1, \dots, R_n)$, 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,

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

Por otra parte,

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

por lo que

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

De lo anterior se sigue que

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

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

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

De lo anterior podemos notar que $\overline{Y}_{obs}$ es insesgado, pero que $\operatorname{Var}\left[\overline{Y}_{obs}\right] \geq \operatorname{Var}\left[\ \overline{Y}\ \right]$, 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. Dado que $a > 0$, se tiene que

$$\operatorname{mín}(y) \coloneqq \operatorname*{mín}_{1\leq i\leq n} y_i = \operatorname*{mín}_{1\leq i\leq n}(ax_i + b) = b + \operatorname*{mín}_{1\leq i\leq n}(ax_i) = b + a \operatorname*{mín}_{1\leq i\leq n} x_i = a \operatorname{mín}(x) + b.$$

De manera análoga,

$$\operatorname{máx}(y) \coloneqq \operatorname*{máx}_{1\leq i\leq n} y_i = \operatorname*{máx}_{1\leq i\leq n}(ax_i + b) = b + \operatorname*{máx}_{1\leq i\leq n}(ax_i) = b + a \operatorname*{máx}_{1\leq i\leq n} x_i = a \operatorname{máx}(x) + b.$$

Por consiguiente, para todo $i \in \{1, \ldots, n\}$, se tiene que

$$y_i^* = \frac{y_i - \operatorname{mín}(y)}{\operatorname{máx}(y) - \operatorname{mín}(y)} = \frac{(ax_i + b) - (a \operatorname{mín}(x) + b)}{(a \operatorname{máx}(x) + b) - (a \operatorname{mín}(x) + b)} = \frac{a\,(x_i - \operatorname{mín}(x))}{a\,(\operatorname{máx}(x) - \operatorname{mín}(x))} = x_i^*,$$

que es lo deseado.

Solución del problema 9. (a) Como el soporte de X es $[x_m, \infty)$, con $x_m > 0$, la transformación $Y = \log(X)$ está bien definida, y Y tiene soporte en $[\log(x_m), \infty)$. Además, la función $g(x) = \log(x)$ definida en $\mathbb{R}^+$ es uno a uno y tiene inversa $g^{-1}(y) = e^y$, la cual es una función derivable, con $\frac{d}{dy}g^{-1}(y) = e^y$. Por lo tanto, como se cumple la relación $Y = \log(X)$ y por consiguiente $X = e^Y$, por el Teorema de Cambio de Variables Y tiene densidad dada por

$$f_Y(y) = \left|\frac{dx}{dy}\right| f_X(e^y)\mathbb{1}_{[\log(x_m),\infty)}(y) = e^y \frac{\alpha x_m^\alpha}{e^{y(\alpha+1)}}\mathbb{1}_{[\log(x_m),\infty)}(y) = \alpha\left(\frac{x_m}{e^y}\right)^\alpha \mathbb{1}_{[\log(x_m),\infty)}(y).$$

Notemos que esta última expresión puede ser escrita como

$$f_Y(y) = \alpha e^{-\alpha(y-\log(x_m))}\mathbb{1}_{[0,\infty)}(y - \log(x_m)),$$

de donde se puede observar que $Y \overset{d}{=} \log(x_m) + Exp(\alpha)$, en donde $Exp(\alpha)$ es una variable aleatoria con distribución exponencial de media $\frac{1}{\alpha}$. En particular, de aquí se sigue que la función de distribución acumulada de Y es

$$F_Y(y) = \begin{cases} 0, & \text{si } y < \log(x_m), \\ 1 - e^{-\alpha(y-\log(x_m))}, & \text{si } y \geq \log(x_m). \end{cases}$$

(b) Primero veamos que, dado $x > x_m$,

$$\mathbb{P}\left[X \geq x\right] = \int_x^\infty \frac{\alpha x_m^\alpha}{t^{\alpha+1}}dt = x_m^\alpha\left[-t^{-\alpha}\right]_{t=x}^\infty = \left(\frac{x_m}{x}\right)^\alpha.$$

De este modo, la cola de X decae de forma polinomial, del orden $x^{-\alpha}$. Por otro lado, si $y > \log(x_m)$,

$$\mathbb{P}\left[Y \geq y\right] = e^{\alpha\log(x_m)}e^{-\alpha y} = x_m^\alpha e^{-\alpha y},$$

de donde podemos ver que la cola de Y decae de forma polinomial, del orden $e^{-\alpha y}$ (más rápidamente que el decaimiento polinomial). Es decir, X tiene colas más pesadas, y al transformarse a Y, cambia a colas más ligeras.

(c) Notemos que, como $Y = \log(X)$, para todo $y \in \mathbb{R}$ se cumple que

$$\mathbb{P}\left[Y > y\right] = \mathbb{P}\left[X > e^y\right],$$

de modo que, como e^y crece más rápido que y, las colas de Y decaen más rápidamente de las de X, como lo visto con la distribución Pareto, en donde un decaimiento polinomial se convierte en uno exponencial. Además, como la función logaritmo es creciente y $\log(x) \leq \log(x+1) \leq x$ para todo $x > 0$, por lo general Y tiene un soporte más grande que X.

Más aún, por las propiedades de la función logarítmica, los cambios grandes en X se reflejan en cambios más chicos de Y. Por ejemplo, si un valor de X se duplica, en la transformación logarítmica el valor de Y solo incrementa en $\log 2$ (cambios multiplicativos se transforman en cambios aditivos). Por consiguiente, si X tiene colas muy pesadas, Y tiende a distribuir el peso a lo largo de los reales y no tan concentrado en las colas; es decir, se "acortan" las colas largas. Además esto produce, por lo general, distribuciones más cercanas a la simetría, en especial cuando hay errores multiplicativos, que se convierten en errores aditivos al aplicar logaritmo, y el Teorema del Límite Central explica dicha simetría.

Solución del problema 10.

Solución del problema 11. (a) Sea $x > 0$, y veamos que

$$\lim_{\lambda \to 0}(x^{\lambda} - 1) = 1 - 1 = 0,$$

mientras que $\lim_{\lambda \to 0} \lambda = 0$. Además, la función $\lambda \mapsto \lambda$ es derivable, con derivada igual a $1 (\neq 0)$. Por lo tanto, ya que el siguiente límite existe, por la Regla de l'Hôpital se tiene que

$$\log(x) = \lim_{\lambda \to 0} \frac{\log(x) x^{\lambda}}{1} = \lim_{\lambda \to 0} \frac{x^{\lambda} - 1}{\lambda} = \lim_{\lambda \to 0} y(\lambda).$$

(b) Consideremos a la sucesión $(x_n)_{n \in \mathbb{N}}$, en donde $x_n = 2^n$ para todo $n \in \mathbb{N}$. Dicha sucesión toma valores muy dispersos cuando n es muy grande, pues sus primeros valores son

$$2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192, 16384, 32768, 65536, 131072, 262144, \ldots.$$

La sucesión correpondiente a la transformación de Box y Cox con $\lambda = 1$ es la misma pero recorrida en 1, así que sigue siendo igual de dispersa. Sin embargo, con la transformación logarítmica ($\lambda = 0$), se convierte en $(y_n)_{n \in \mathbb{N}}$, en donde $y_n = n \log(2)$ para todo $n \in \mathbb{N}$, que es mucho menos dispersa. A manera de ilustración, sus primeros valores son aproximadamente iguales a:

$$0.6931, 1.3863, 2.0794, 2.7726, 3.4657, 4.1589, 4.852, 5.5452, 6.2383, 6.9315, 7.6246, 8.3178, 9.0109, \ldots$$

Solución del problema 12.

Solución del problema 13.