Tarea 1

Ciencia de Datos

Brain de Jesús Salazar, César Ávila, Iván García

12 de septiembre de 2025

Solución del problema 1. Recordemos la definición de la matriz Hat:

$$H \coloneqq X(X^TX)^{-1}X^T.$$

Para ver que H es idempotente, notemos que

$$\begin{aligned} HH &= \left(X(X^TX)^{-1}X^T\right)\left(X(X^TX)^{-1}X^T\right) \\ &= X(X^TX)^{-1}\left((X^TX)(X^TX)^{-1}\right)X^T \\ &= X(X^TX)^{-1}X^T \\ &= H. \end{aligned}$$

Por consiguiente, $H^2 = H$, así que H es idempotente. Por otra parte, veamos que

$$\begin{aligned} H^T &= \left(X(X^TX)^{-1}X^T\right)^T \\ &= (X^T)^T\left((X^TX)^{-1}\right)^T X^T \\ &= X\left((X^TX)^T\right)^{-1}X^T \\ &= X(X^TX)^{-1}X^T \\ &= H. \end{aligned}$$

Por lo tanto, $H^T = H$, así que H es simétrica. Como también es idempotente, entonces H es una matriz de proyección. De hecho, es la proyección ortogonal sobre el espacio de columnas de X (que se asume que tiene rango de columnas completo), y se cumple que $\hat{\boldsymbol{Y}} = H\boldsymbol{Y}$, con la interpretación usual de $\hat{Y}$.

Además, el vector de residuales es $\boldsymbol{e} = (I-H)\boldsymbol{Y}$, y como $H(I-H) = 0$ por la idempotencia, entonces $H\boldsymbol{Y}$ es la proyección de Y sobre el espacio de columnas de X. Esto implica que h_{ii} mide la influencia de la observación i-ésima para el ajuste lineal, que es precisamente el *leverage*.

Solución del problema 2. Considere el modelo de regresión lineal $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ donde $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 I_n)$, y X es una matriz de dimensiones $n \times p$, que asumimos de rango de columnas completo (de modo que $X^T X$ es invertible). Ahora bien, notemos que $X^T X$ es una matriz de dimensiones $p \times p$, y X^T es de dimensiones $p \times n$. Puesto que

$$H = X(X^T X)^{-1} X^T,$$

entonces H es una matriz de $n \times n$, y por la propiedad cíclica de la traza se tiene que

$$\sum_{i=1}^{n} h_{ii} = tr(H) = tr\left(X(X^T X)^{-1} X^T\right) = tr\left(X^T X(X^T X)^{-1}\right) = tr(I_p) = p,$$

en donde I_p es la matriz identidad de dimensiones $p \times p$.

Observemos además que los valores ajustados están dados por $\hat{\boldsymbol{Y}} = HY$, así que, como $\mathrm{Var}[Y] = \sigma^2 I$,

$$\mathrm{Var}[\hat{\boldsymbol{Y}}] = \sigma^2 H H^T = \sigma^2 H^2 = \sigma^2 H,$$

en donde hemos usado el ejercicio anterior para ver que H es simétrica e idempotente. Así pues, la "varianza total" de la estimación es

$$\sum_{i=1}^{n} \mathrm{Var}[\hat{\boldsymbol{Y}}_i] = \sigma^2 tr(H) = \sigma^2 p;$$

es decir, es igual al número de parámetros multiplicado por σ^2 (la varianza común de los errores). Además, cada elemento h_{ii} de la diagonal de H mide la influencia de la observación i-ésima, así que el resultado anterior puede ser interpretado como que el número efectivo de parámetros es justamente p.

Por otra parte, mientras mayor es el número de parámetros utilizados, el modelo puede explicar a los datos de una mejor manera, disminuyendo el sesgo, pero esto incrementa la varianza total, así que puede llevar a un sobreajuste, disminuyendo la generalidad del modelo.

Solución del problema 3. Considere el modelo de regresión lineal clásico, $\boldsymbol{Y} = X\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ donde $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 I_n)$ y a la matriz de proyección ortogonal,

$$H = X(X^T X)^{-1} X^T.$$

Ahora bien, el vector de residuos se puede expresar como

$$\boldsymbol{e} = (I_n - H)\boldsymbol{Y},$$

y como $\boldsymbol{Y}$ sigue una distribución normal multivariada con media $X\boldsymbol{\beta}$ y varianza $\sigma^2 I_n$, se tiene que,

$$\boldsymbol{e} \sim N_n\left(\mathbf{0}, \sigma^2 \left(I_n - H\right)\right).$$

De lo anterior se sigue que, para cada $i \in \{1, \ldots, n\}$,

$$e_i \sim N(0, \sigma^2(1 - h_{ii})),$$

y normalizando,

$$\frac{e_i}{\sigma\sqrt{1 - h_{ii}}} \sim N(0, 1).$$

Por otro lado, observemos que si se considera al estimador insesgado de la varianza,

$$\hat{\sigma}^2 = \frac{\mathbf{e}^\top \mathbf{e}}{n - p} = \frac{\sigma^2 \mathbf{Y}^\top (I_n - H)\mathbf{Y}}{\sigma^2(n - p)},$$

como $(I_n - H)$ es una matriz idempotente de rango $n - p$ (la idempotencia se sigue del ejercicio 1, y en una matriz idempotente el rango es igual a la traza), se tiene que

$$\frac{(n - p)}{\sigma^2}\hat{\sigma}^2 = \frac{\mathbf{Y}^\top (I - H)\mathbf{Y}}{\sigma^2} \sim \chi^2(n - p).$$

Por último, se sabe que la razón entre una variable aleatoria normal estándar sobre la raíz cuadrada de una variable aleatoria independiente con distribución ji cuadrada dividida entre sus r grados, se distribuye como una variable aleatoria t de Student con r grados de libertad. Ya que en este caso los residuos normalizados (con una distribución normal estándar) no son independientes de la ji cuadrada se tiene que la siguiente variable aleatoria cumple que aproximadamente,

$$\frac{\frac{e_i}{\sigma\sqrt{1-h_{ii}}}}{\sqrt{\frac{(n-p)}{(n-p)\sigma^2}\hat{\sigma}^2}} \sim t(n - p).$$

Simplificando la expresión de la izquierda se tiene que los residuos estandarizados tienen una distribución aproximada de:

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}} \sim t(n - p).$$

Para arreglar el problema de la independencia y obtener la distribución que pide el problema, se puede estimar a σ^2 sin usar el i-ésimo dato. Es decir, definamos

$$\hat{\sigma}_i^2 = \frac{\mathbf{e_i}^\top \mathbf{e_i}}{n-p-1} = \frac{\sigma^2 \mathbf{Y}_i^\top (I_{n-1} - H_i)\mathbf{Y_i}}{\sigma^2(n-p-1)},$$

donde el vector $\mathbf{Y}_i$ se obtiene eliminando la i-ésima entrada de Y y H_i se construye eliminando la i-ésima entrada de X. En este caso la matriz $(I_{n-1} - H_i)$ es idempotente y de rango $n-p-1$, por lo que

$$\frac{(n-1-p)}{\sigma^2}\hat{\sigma_i}^2 = \frac{\mathbf{Y_i}^\top (I_{n-1} - H_i)\mathbf{Y}_i}{\sigma^2} \sim \chi^2(n-1-p).$$

Como $\hat{\sigma}_i$ no depende de la i-ésima observación, y los errores son independientes, entonces también es independiente de e_i. De aquí se tiene la siguiente distribución exacta:

$$\frac{\frac{e_i}{\sigma\sqrt{1-h_{ii}}}}{\sqrt{\frac{(n-1-p)}{(n-1-p)\sigma^2}\hat{\sigma}_i^2}} \sim t(n-1-p).$$

Simplificando el lado derecho se concluye que, de manera exacta,

$$\frac{e_i}{\hat{\sigma}_i\sqrt{1-h_{ii}}} \sim t(n-p-1).$$

Gracias a lo demostrado anteriormente, para cada observación i se puede calcular su residuo estandarizado (o studentizado), y bajo la hipótesis nula de que la observación i-ésima es bien consistente con el modelo, dicho residuo debería pertenecer a una distribución t de Student con los grados de libertad anteriormente mencionados, así que al calcular su p-valor se puede establecer un criterio que permita la detección de outliers.

Solución del problema 4. Sean $\boldsymbol{Y} = (Y_{obs}, Y_{mis})$, $\boldsymbol{R}$ el patrón de datos faltantes, θ los parámetros del modelo y ψ los parámetros del mecanismo de faltantes. Por la definición de MCAR,

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

Luego,

$$\begin{aligned}
\mathbb{P}\left[\boldsymbol{Y}, \boldsymbol{R} \,|\, \theta, \psi\right] &= \mathbb{P}\left[\boldsymbol{R} \,|\, \boldsymbol{Y}, \theta, \psi\right] \mathbb{P}\left[\boldsymbol{Y} \,|\, \theta, \psi\right] \\
&= \mathbb{P}\left[\boldsymbol{R} \,|\, \psi\right] \mathbb{P}\left[\boldsymbol{Y} \,|\, \theta, \psi\right] \\
&= \mathbb{P}\left[\boldsymbol{R} \,|\, \psi\right] \mathbb{P}\left[\boldsymbol{Y} \,|\, \theta\right],
\end{aligned}$$

pues el mecanismo de faltantes es independiente de los datos, que es lo que queríamos probar. Por lo tanto, la verosimilitud de datos observados para θ es

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

Por lo tanto, la inferencia sobre θ puede basarse únicamente en $\mathbb{P}\left[Y_{obs} \,|\, \theta\right]$, ignorando $\mathbb{P}\left[\boldsymbol{R} \,|\, \psi\right]$.

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. Sean $\hat{\boldsymbol{\beta}}$ los coeficientes estimados de la regresión completa, y $\hat{\boldsymbol{\beta}}_{(i)}$ los coeficientes estimados sin la observación i-ésima. Es un hecho conocido que para la regresión lineal se cumple que

$$\hat{\boldsymbol{\beta}}_{(i)} = \hat{\boldsymbol{\beta}} - (X^T X)^{-1}\boldsymbol{x}_i \frac{e_i}{1-h_{ii}}. \tag{1}$$

Para ver lo anterior, supongamos sin pérdida de generalidad que $i = 1$ (el resto de los argumentos son análogos). Podemos particionar la información a eliminar de la siguiente manera:

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \boldsymbol{Y}_{(1)} \end{bmatrix} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} = \begin{bmatrix} \boldsymbol{x}_1^T \\ X_1 \end{bmatrix} \boldsymbol{\beta} + \boldsymbol{\epsilon},$$

en donde (1) significa el vector sin la observación 1, y $\boldsymbol{x}_1^T$ es la fila 1 de X (la correspondiente a la observación 1). De este modo, $\hat{\boldsymbol{\beta}}_{(1)}$ es el ajuste del modelo

$$\boldsymbol{Y}_{(1)} = X_1\beta + \boldsymbol{\epsilon}_{(1)},$$

por lo que $\hat{\boldsymbol{\beta}}_{(1)} = (X_1^T X_1)^{-1} X_1^T \boldsymbol{Y}_{(1)}$. Por otro lado, veamos que

$$X^T X = [\boldsymbol{x}_1, X_1^T] \begin{bmatrix} \boldsymbol{x}_1^T \\ X_1 \end{bmatrix} = \boldsymbol{x}_1\boldsymbol{x}_1^T + X_1^T X_1, \tag{2}$$

de donde se sigue que $X_1^T X_1 = X^T X - \boldsymbol{x}_1\boldsymbol{x}_1^T$. Luego, por la identidad de Woodbury para matrices se tiene que

$$(X_1^T X_1)^{-1} = (X^T X)^{-1} + (X^T X)^{-1}\boldsymbol{x}_1(1 - \boldsymbol{x}_1^T(X^T X)^{-1}\boldsymbol{x}_1)^{-1}\boldsymbol{x}_1^T(X^T X)^{-1}.$$

De manera análoga a (2), tenemos que $X_1^T \boldsymbol{Y}_{(1)} = X^T\boldsymbol{Y} - \boldsymbol{x}_1 Y_1$, de donde se sigue que

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{(1)} =& \ (X_1^T X_1)^{-1} X_1^T \boldsymbol{Y}_{(1)} \\
=& \ (X^T X)^{-1} X^T \boldsymbol{Y} + (X^T X)^{-1}\boldsymbol{x}_1 \left[1 - \boldsymbol{x}_1^T (X^T X)^{-1}\boldsymbol{x}_1\right]^{-1} \boldsymbol{x}_1^T (X^T X)^{-1} X^T \boldsymbol{Y} \\
& - (X^T X)^{-1}\boldsymbol{x}_1 Y_1 - (X^T X)^{-1}\boldsymbol{x}_1 \left[1 - \boldsymbol{x}_1^T (X^T X)^{-1}\boldsymbol{x}_1\right]^{-1} \boldsymbol{x}_1^T (X^T X)^{-1}\boldsymbol{x}_1 Y_1 \\
=& \ \hat{\boldsymbol{\beta}} + (X^T X)^{-1}\boldsymbol{x}_1(1-h_{11})^{-1}\boldsymbol{x}_1^T\hat{\boldsymbol{\beta}} - (X^T X)^{-1}\boldsymbol{x}_1 Y_1 - (X^T X)^{-1}\boldsymbol{x}_1(1-h_{11})^{-1}h_{11}Y_1 \\
=& \ \hat{\boldsymbol{\beta}} + \frac{(X^T X)^{-1}}{1-h_{11}}\left[\boldsymbol{x}_1\hat{Y}_1 - (1-h_{11})\boldsymbol{x}_1 Y_1 - \boldsymbol{x}_1 h_{11} Y_1\right] \\
=& \ \hat{\boldsymbol{\beta}} - \frac{(X^T X)^{-1}}{1-h_{11}}\boldsymbol{x}_1 e_1 \\
=& \ \hat{\boldsymbol{\beta}} - (X^T X)^{-1}\boldsymbol{x}_1 \frac{e_1}{1-h_{11}}.
\end{aligned}$$

Por consiguiente, como $i = 1$ fue tomado sin pérdida de generalidad, se concluye que (1) es válida.

Por otra parte,

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{p\hat{\sigma}^2} = \frac{\left(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}}\right)^T \left(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}}\right)}{p\hat{\sigma}^2} = \frac{(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})^T (X^T X)\, (\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})}{p\hat{\sigma}^2}.$$

Por consiguiente,

$$\begin{aligned}
D_i &= \frac{1}{p\hat{\sigma}^2}(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})^T (X^T X)(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}}) \\
&= \frac{1}{p\hat{\sigma}^2}\left(\frac{e_i}{1-h_{ii}}\right)^2 \boldsymbol{x}_i^T (X^T X)^{-1}(X^T X)(X^T X)^{-1}\boldsymbol{x}_i \\
&= \frac{1}{p\hat{\sigma}^2}\left(\frac{e_i}{1-h_{ii}}\right)^2 \boldsymbol{x}_i^T (X^T X)^{-1}\boldsymbol{x}_i \\
&= \frac{1}{p\hat{\sigma}^2}\left(\frac{e_i}{1-h_{ii}}\right)^2 h_{ii} \\
&= \left(\frac{e_i}{\hat{\sigma}\sqrt{(1-h_{ii})}}\right)^2 \left(\frac{h_{ii}}{p\,(1-h_{ii})}\right) \\
&= \frac{r_i^2}{p} \cdot \frac{h_{ii}}{1-h_{ii}}.
\end{aligned}$$

La distancia de Cook con la expresión anterior tiene la siguiente interpretación: el término r_i^2 mide qué tan alejada se encuentra la observación i-ésima del resto de los datos, y esta se multiplica por el factor $\frac{h_{ii}}{1-h_{ii}}$, que mientras más grande la influencia de la observación i-ésima (h_{ii}), es mayor. Por consiguiente, esta distancia combina la influencia de la observación con su discrepancia con el resto de los datos, dándole mayor peso a observaciones con influencias mayores. Por último, el factor $\frac{1}{p}$ indica que, mientras más complejo es el modelo, las distancias de Cook serán menores, porque el espacio parametral es más amplio.

Solución del problema 8. Dado que $a > 0$, se tiene que

$$\text{mín}(y) \coloneqq \min_{1\leq i\leq n} y_i = \min_{1\leq i\leq n} (ax_i + b) = b + \min_{1\leq i\leq n} (ax_i) = b + a \min_{1\leq i\leq n} x_i = a\,\text{mín}(x) + b.$$

De manera análoga,

$$\text{máx}(y) \coloneqq \max_{1\leq i\leq n} y_i = \max_{1\leq i\leq n} (ax_i + b) = b + \max_{1\leq i\leq n} (ax_i) = b + a \max_{1\leq i\leq n} x_i = a\,\text{máx}(x) + b.$$

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

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

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 $x > x_m$,

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

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

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

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

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

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. (a) La media $\overline{x}$ está dada por

$$\overline{x} = \frac{1+2+3+4+M}{5} = 2 + \frac{M}{5},$$

mientras que la desviación estándar de la muestra es

$$\begin{aligned} s = \sqrt{\sum_{i=1}^{5} (x_i - \overline{x})^2} &= \sqrt{\left(1 + \frac{M}{5}\right)^2 + \left(\frac{M}{5}\right)^2 + \left(1 - \frac{M}{5}\right)^2 + \left(2 - \frac{M}{5}\right)^2 + \left(2 - \frac{4M}{5}\right)^2} \\ &= \sqrt{\frac{4}{5}M^2 - 4M + 10}. \end{aligned}$$

(b) Dado que $M \to \infty$, los datos en x están ordenados, y por consiguiente su mediana es 3 (el de en medio).

Ahora bien, el cuartil 1 ($Q_{0.25}$) es la mediana de $\{1, 2\}$, que es $\frac{1+2}{2}$. Por otro lado, el tercer cuartil ($Q_{0.75}$) es la mediana de $\{4, M\}$, que es $\frac{4+M}{2}$. Luego, el rango intercuartílico es

$$RIQ = Q_{0.75} - Q_{0.25} = \frac{4+M}{2} - \frac{3}{2} = \frac{1+M}{2}.$$

(c) Cuando se hace $M \to \infty$, la media, la desviación estándar y el rango intercuartílico tienden a ∞, pero la mediana permanece constante. Como interpretación, M se convierte en un "outlier" cuando se vuelve suficientemente grande, y afecta las medidas de tendencia central y de dispersión mencionadas anteriormente, con excepción de la mediana, por lo que esta es más robusta que la media (no es tan sensible a valores extremos).

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. a) Con las hipótesis del enunciado, observemos que la función $\hat{f}_h(x)$ es una suma de funciones indicadoras, que cuenta el número de observaciones x_i que están en el mismo conjunto que x, I_j. Luego, ya que $1\{x_i \in I_j\} \geq 0$ y $nh > 0$ se tiene que $\hat{f}_h(x) \geq 0$.

b) Para dar respuesta a este inciso, notemos que la función $\hat{f}_h(x)$ se puede escribir como

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{j=1}^{k} 1\{x \in I_j\} \sum_{i=1}^{n} 1\{x_i \in I_j\},$$

en donde hemos considerado que los intervalos $I_1, \ldots, I_k$ son ajenos. Luego, la integral buscada se puede ver como

$$\begin{aligned}
\int_{-\infty}^{\infty} \hat{f}_h(x)dx &= \int_{-\infty}^{\infty} \frac{1}{nh} \sum_{j=1}^{k} 1\{x \in I_j\} \sum_{i=1}^{n} 1\{x_i \in I_j\} dx \\
&= \frac{1}{nh} \sum_{j=1}^{k} \int_{\{x \in I_j\}} \sum_{i=1}^{n} 1\{x_i \in I_j\} dx \\
&= \frac{1}{nh} \sum_{j=1}^{k} \sum_{i=1}^{n} 1\{x_i \in I_j\} \int_{\{x \in I_j\}} dx \\
&= \frac{1}{nh} \sum_{j=1}^{k} \sum_{i=1}^{n} 1\{x_i \in I_j\} h \\
&= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} 1\{x_i \in I_j\} \\
&= \frac{1}{n} \sum_{i=1}^{n} 1 \\
&= \frac{n}{n} = 1,
\end{aligned}$$

en donde la igualdad en el penúltimo renglón se tiene ya que cada x_i pertenece a algún conjunto I_j.

c) Observemos que cuando h es grande, los intervalos contendrán más datos, esto nos llevará a que no se aprecie si hay algún patrón en el comportamiento de los datos, es decir si los datos tienen preferencia por ciertos intervalos. Esto tiene la ventaja que la varianza es menor, pero el sesgo es mayor. Por otro lado, cuando h es muy pequeño, los intervalos no alcanzarán a contener muchos datos, lo cual hace que haya un sesgo menor, pero mayor varianza; es decir, un sobreajuste. Un caso extremo de ver esto es hacer a h muy cercano a cero de tal forma que cada intervalo contenga a lo más un dato, y en este caso, solo se verán barras de la misma altura alrededor de cada dato.

Solución del problema 13. Normalización) Con las hipótesis del enunciado, observemos que la integral se puede escribir como

$$\begin{aligned}\int_{-\infty}^{\infty} \hat{f}_h(x)dx &= \int_{-\infty}^{\infty} \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)dx \\ &= \frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{\infty} K\left(\frac{x-x_i}{h}\right)dx \\ &= \frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{\infty} K(u)hdu \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} K(u)du \\ &= \frac{1}{n}\sum_{i=1}^{n} 1 = 1,\end{aligned}$$

en donde hemos usado el Teorema de Fubini para el intercambio de la suma y la integral para la segunda igualdad, y la tercera igualdad se tiene haciendo el cambio de variables $u = \frac{x-x_i}{h}$. Además, se ha considerado que K integra 1 para la primera igualdad del último renglón.

No negatividad) Para este inciso basta observar por hipótesis $K(u) \geq 0$ y que $nh > 0$, y como la integral de funciones no negativas es no negativas, se concluye que $\hat{f}_h(x) \geq 0$.

Sesgo puntual) Observemos que, como la muestra $x_1, \ldots, x_n$ corresponde a variables aleatorias iid, el sesgo puntual se puede escribir como

$$\begin{aligned}\mathbb{E}\left[\hat{f}_h(x)\right] - f(x) &= \mathbb{E}\left[\sum_{i=1}^{n}\frac{1}{nh}K\left(\frac{x-x_i}{h}\right)\right] - f(x) \\ &= \frac{1}{nh}\sum_{i=1}^{n}\mathbb{E}\left[K\left(\frac{x-x_i}{h}\right)\right] - f(x) \\ &= \frac{1}{h}\mathbb{E}\left[K\left(\frac{x-x_1}{h}\right)\right] - f(x) \\ &= \frac{1}{h}\int_{-\infty}^{\infty} K\left(\frac{x-x_1}{h}\right) f(x_1)\, dx_1 - f(x).\end{aligned}$$

Haciendo el cambio de variables $u = \frac{x-x_1}{h}$ y considerando que la integral del kernel es igual a uno, se tiene que la expresión anterior es igual a

$$\begin{aligned}\mathbb{E}\left[\hat{f}_h(x)\right] - f(x) &= \frac{1}{h}\int_{-\infty}^{\infty} hK(u)f(x-uh)du - \int_{-\infty}^{\infty} K(u)f(x)du \\ &= \int_{-\infty}^{\infty} K(u)[f(x-uh) - f(x)]du.\end{aligned} \tag{3}$$

Por otro lado, asumiendo que f tiene derivadas de segundo orden, por el Teorema de Taylor con errores de Peano, se tiene la siguiente aproximación de orden dos de $f(x-uh)$ alrededor de x:

$$f(x-uh) = f(x) + f'(x)(x-uh-x) + \frac{f''(x)(x-uh-x)^2}{2!} + h_2(x-uh)(uh)^2,$$

en donde h_2 es una función tal que

$$\lim_{h\to 0} h_2(x-uh) = 0.$$

Sustituyendo lo anterior en (3) se tiene que

$$\begin{aligned}\mathbb{E}\left[\hat{f}_h(x)\right] - f(x) &= \int_{-\infty}^{\infty} K(u)\left[f'(x)(-uh) + \frac{f''(x)(uh)^2}{2!} + h_2(x-uh)(uh)^2\right]du \\ &= \frac{f''(x)}{2!}h^2\int_{-\infty}^{\infty} u^2K(u)du + h^2\int_{-\infty}^{\infty} K(u)h_2(x-uh)u^2du,\end{aligned}$$

donde esta última igualdad se tiene ya que $\int uK(u)du = 0$ y $\mu_2(K)$ es finito. Además, ya que h_2 está acotada alrededor de x (es decir, cuando h es pequeño), entonces para h suficientemente chico se tiene que

$$|h_2(x-uh)| \leq 1,$$

lo que implica que

$$\left|K(u)h_2(x-uh)u^2\right| \leq \left|u^2K(u)\right|,$$

y dicha función es integrable por hipótesis, pues $\mu_2(K)$ existe y es finito. Luego, por el Teorema de Convergencia Dominada se tiene que

$$\lim_{h\to 0}\int_{-\infty}^{\infty} K(u)h_2(x-uh)u^2du = \int_{-\infty}^{\infty}\lim_{h\to 0} K(u)h_2(x-uh)u^2du = 0.$$

Por lo tanto,

$$h^2\int_{-\infty}^{\infty} K(u)h_2(x-uh)u^2du = o(h^2).$$

En conclusión,

$$\mathbb{E}\left[\hat{f}_h(x)\right] - f(x) = \frac{h^2}{2}\mu_2(K)f''(x) + o(h^2).$$