Fundamentos y preparación de los datos

---

1. *Hat Matrix y propiedades algebraicas.*
Demuestra que la matriz

$$X = X(X^TX)^{-1}X$$

es idempotente y simétrica. Explica por qué estas propiedades son fundamentales para la interpretación de los *leverages*.

Recordemos que una matriz es simétrica cuando $H^T = H$. Entonces, utilizando propiedades del producto de matrices bajo la transpuesta, obtenemos

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

por lo tanto, H es simétrica. Para mostrar que H es idempotente veremos que $H^2 = H$

$$\begin{aligned} H^2 &= \left[X(X^TX)^{-1}X^T\right]\left[X(X^TX)^{-1}X^T\right] \\ &= X(X^TX)^{-1}\underbrace{(X^TX)(X^TX)^{-1}}_{I}X^T \\ &= X(X^TX)^{-1}X^T \\ &= H \end{aligned}$$

por lo tanto, H es idempotente.

Discución. Recordemos que H proyecta el vector y sobre el espacio generado por las columnas de X, $\hat{y} = Hy$. La idempotencia y simetría aseguran que H sea una proyección ortogonal, lo que sustenta la interpretación de que h_{ii} mide la "influencia propia" de la observación i. Además, hace que h_{ij} sea una medida recíproca de influencia entre observaciones i y j.

En específico, la idempotencia da la interpretación geométrica de leverage como "qué tanto participa la observación en su propia predicción" y explica sus límites $0 \leq h_{ii} \leq 1$. Mientras que la simetría asegura consistencia y reciprocidad en la influencia entre observaciones, lo que da sentido a la interpretación de h_{ii} como medida de influencia de la propia observación.

2. *Suma de leverages.*
Muestra que para un modelo lineal con n observaciones y p parámetros se cumple

$$\sum_{i=1}^{n} h_{ii} = p.$$

Interpreta este resultado en términos del "número efectivo de parámetros" y discuta su relación con el sobreajuste.

Supongamos que tenemos el siguiente modelo lineal

$$Y = X\beta + \varepsilon, \quad \text{con } \varepsilon \sim N(0, \sigma^2 I_n),$$

donde β es el vector de parámetros y X es la matriz de covariables de tamaño $n \times p$, de modo que $X^\top X$ es una matriz $p \times p$. Notemos que

$$\begin{aligned}\sum_{i=1}^{n} h_{ii} &= \operatorname{tr}(H) \\ &= \operatorname{tr}(X(X^\top X)^{-1}X^\top) \\ &= \operatorname{tr}(X^\top X(X^\top X)^{-1}) \qquad \operatorname{tr}(AB) = \operatorname{tr}(BA) \\ &= \operatorname{tr}(I_p).\end{aligned}$$

Por lo tanto,

$$\sum_{i=1}^{n} h_{ii} = p.$$

De acuerdo a este resultado tendríamos que el "número efectivo de parámetros" para realizar un ajuste al modelo lineal es p. El leverage promedio esta definido como:

$$\bar{h} = \frac{p}{n}.$$

Donde h_{ii} mide la influencia de la observación i en su valor ajustado, esto quiere decir que si p se acerca a n entonces tendríamos un leverage promedio cercano a 1, lo que daría señales a un sobreajuste del modelo.

3. *Distribución de los residuos estandarizados.*
Bajo el modelo lineal clásico con errores normales, demuestra que los residuos estandarizados

$$r_i = \frac{e_i}{\widehat{\sigma}\sqrt{1-h_{ii}}}$$

tienen, aproximadamente, una distribución t de Student con $n-p-1$ grados de libertad. Explica cómo esta propiedad justifica su uso en la detección de outliers.

Bajo el modelo lineal

$$Y = X\beta + \varepsilon, \quad \text{con } \varepsilon \sim N(0, \sigma^2 I_n).$$

De modo que los residuales del modelo ajustado se definen como

$$e = (I_n - H)Y.$$

Como $Y \sim N(X\beta, \sigma^2 I_n)$, al ser los residuales una transformación lineal de Y, se tendría que $e \sim N(0, (I_n - P))\sigma^2$. Por lo tanto, para cada i se cumple que

$$\mathbb{E}[e_i] = 0 \text{ y } \text{Var}(e_i) = (1 - h_{ii})\sigma^2.$$

Así, el residuo normalizado es tal que

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

Por otra parte, el estimador insesgado de σ^2 es

$$\hat{\sigma}^2 = \frac{e^\top e}{n-p} = \frac{1}{n-p}\sum (Y_i - \hat{Y}_i)^2.$$

Para el caso del modelo lineal se tiene que la suma de cuadrados residuales SS_{Res} sigue una distribución ji-cuadrada, esto es

$$\frac{SS_{Res}}{\sigma^2} = \frac{e^\top e}{\sigma^2} \sim \chi^2_{n-p}.$$

Por lo tanto,

$$U = \frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}.$$

Notemos que podemos escribir a r_i como:

$$r_i = \frac{z_i}{\sqrt{U/(n-p)}} = \frac{e_i/\sigma\sqrt{1-h_{ii}}}{\sqrt{\hat{\sigma}^2/\sigma^2}} = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}.$$

En consecuencia los residuos estandarizados son la razón de una variable aleatoria normal estándar (z_i) con la raíz cuadrada de una variable aleatoria ji-cuadrada dividida por sus grados de libertad. Pero como el numerador y el denominador no son independientes, la distribución no es exacta pero se aproxima a una t de studen con $n-p$ grados de libertad.

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

Para garantizar que los términos del numerador y del denominador sean independientes, Belsley, Kuh y Welsch (1980) propusieron estandarizar cada elemento del vector de residuales dividiéndolo por su desviación estándar,

pero asegurándose de que ésta sea calculada de forma independiente. Específicamente, para cada observación i, la desviación estándar utilizada se estima omitiendo justamente dicha observación, es decir

$$r_i^* = \frac{e_i}{s_{(i)}\sqrt{1 - v_{ii}}}$$

con

$$s_{(i)} = \sqrt{\frac{\sum_{j \neq i}(e_j - \bar{e})^2}{n - p - 1}}$$

Entonces, cada residual "Studentizado" r_i^* se distribuye con una variable aleatoria T de Student con $n - p - 1$ grados de libertad. Intuitivamente, el 1 restado se debe a la eliminación del i-esimo residual.

Como hemos visto, si el modelo está bien planteado, los residuos estandarizados r_i deben de comportarse aproximadamente como una muestra aleatoria de una distribución T de Student. Por lo tanto, una opción para detectar outliers, es calcular la probabilidad de que ocurra un valor igual o mayor a cada r_i: si éste es un valor muy pequeño, indicaría que es un posible outlier. Por lo regular se definen umbrales de acuerdo al cuantil del 0.975, así, si $|r_i|$ rebasa este umbral tendríamos evidencia de que las i-ésima observación es un outlier.

4. *Factorización bajo MCAR.*
Partiendeo de la definición de MCAR, demuestra formalmente que

$$\mathbb{P}[Y, R|\theta, \psi] = \mathbb{P}[Y|\theta]\mathbb{P}[R|\psi].$$

Concluya por qué en este caso el mecanismo de faltantes es ignorable para la inferencia sobre θ.

Sean $Y = (Y_{obs}, Y_{mis})$ el vector de observaciones, dividido en dos tipos de dato: observados y faltantes, y R la matriz indicadora de datos faltantes con θ parámetros del modelo de datos y ψ del mecanismo de faltantes.

Bajo el mecanismo MCAR tenemos que la probabilidad de que falte un dato no depende de los valores de Y, esto es

$$p(R|Y_{obs}, Y_{mis}, \theta, \psi) = p(R|\psi). \tag{1}$$

Ahora, por definición tenemos que

$$\begin{aligned} p(Y, R \mid \theta, \psi) &= \frac{p(Y, R, \theta, \psi)}{p(\theta, \psi)} \\ &= \frac{p(R \mid, Y, \theta, \psi)p(Y, \theta, \psi)}{p(\theta, \psi)} \\ &= p(R \mid, Y, \theta, \psi)p(Y \mid \theta, \psi). \end{aligned}$$

Suponiendo que los parámetros del modelo y los del mecanismo de faltantes son distintos (separabilidad), tendríamos que R no depende de θ y tomando en cuenta 1 tendríamos que

$$p(R \mid, Y, \theta, \psi) = p(R \mid \psi)$$

Del mismo modo, los datos Y dependen de los parámetros del modelo θ, pero no de los parámetros del mecanismo de faltantes, por lo que

$$p(Y \mid \theta, \psi) = p(Y \mid \theta).$$

De lo que se sigue,

$$p(R, Y \mid \theta, \psi) = p(R \mid \psi)p(Y \mid \theta).$$

Sabemos que la verosimilitud es proporcional a la densidad conjunta, es decir

$$L(\theta, \psi; Y, R) \propto p(R, Y \mid \theta, \psi) = p(R \mid \psi)p(Y \mid \theta).$$

Para realizar inferencia sobre θ, sólo necesitaríamos la parte que depende de θ que corresponde a $p(Y \mid \theta)$ ya que la otra parte no aporta información sobre θ, (al maximizar la verosimilitud este término es constante por lo que no aporta información).

Por lo tanto, podemos hacer inferencia sobre θ como si los datos estuvieran completos ignorando el mecanismo de faltantes.

$$L(\theta; Y) \propto p(Y \mid \theta).$$

5. *Insesgadez bajo eliminación de casos en MCAR.*
Sea $\overline{Y}_{\text{obs}}$ la media muestral basada sólo en los casos observados. Demuestra que

$$\mathbb{E}[\overline{Y}_{\text{obs}}] = \mu$$

bajo MCAR. Discute por qué, a pesar de ser insesgado, este estimador pierde eficiencia.

Sea $\bar{Y}_{obs}$ la media muestral basada solamente en los n_{obs} casos observados. Definamos a R_i como la variable aleatoria que indica si el dato Y_i fue observado o faltante, es decir

$$\begin{cases} R_i = 1 & \text{si } Y_i \text{ fue observado.} \\ R_i = 0 & \text{si } Y_i \text{ no fue observado.} \end{cases}$$

Con esto tendríamos que

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

Notemos que $R_i \sim \text{Bern}(P)$, (P es la probabilidad de que un dato sea observado) entonces

$$n_{obs} = \sum_{i=1}^{n} R_i \sim \text{Bin}(n, p).$$

Entonces, por la ley de la esperanza total tenemos que

$$\begin{aligned} \mathbb{E}[\bar{Y}_{\text{obs}}] &= \sum_{m=0}^{n} \mathbb{E}[\bar{Y}_{\text{obs}} | n_{obs} = m] \mathbb{P}(n_{obs} = m) \\ &= \sum_{m=0}^{n} \mathbb{E}\left[\frac{\sum_{i=1}^{n} Y_i R_i}{n_{obs}} | n_{obs} = m \right] \mathbb{P}(n_{obs} = m) \\ &= \sum_{m=0}^{n} \frac{1}{m} \mathbb{E}\left[\sum_{i=1}^{n} Y_i R_i | n_{obs} = m \right] \mathbb{P}(n_{obs} = m). \end{aligned}$$

Notemos que condicionando a $n_{obs} = m$, se tiene que R_i es 1 para exactamente m índices, además como Y_i es independiente de R_i tendríamos que

$$\mathbb{E}\left[\sum_{i=1}^{n} Y_i R_i | n_{obs} = m \right] = \mathbb{E}\left[\sum_{i \in I} Y_i \right],$$

donde I es un subconjunto aleatorio de tamaño m de $\{1, 2, \ldots, n\}$, tomando que $\mathbb{E}[Y_i] = \mu$

$$\mathbb{E}\left[\sum_{i \in I} Y_i \right] = m\mu.$$

Por lo tanto,

$$\begin{aligned} \mathbb{E}[\bar{Y}_{\text{obs}}] &= \sum_{m=0}^{n} \frac{1}{m} m\mu \mathbb{P}(n_{obs} = m) \\ &= \mu \sum_{m=0}^{n} \mathbb{P}(n_{obs} = m) \\ &= \mu. \end{aligned}$$

Esto último demuestra que $\bar{Y}_{obs}$ es un estimador insesgado. Ahora analicemos la varianza de este estimador, para ello notemos que dado $n_{obs} = m$ entonces $\bar{Y}_{obs}$ es la media de m observaciones así,

$$\text{Var}(\bar{Y}_{obs}|n_{obs} = m) = \text{Var}\left(\frac{\sum_{i\in I} Y_i}{m}\right) = \frac{1}{m}\text{var}(Y_i) = \frac{\sigma^2}{m}.$$

Aplicando la ley de varianza total se sigue que

$$\begin{aligned}\text{Var}(\bar{Y}_{obs}) &= \mathbb{E}[\text{Var}(\bar{Y}_{obs}|n_{obs})] + \text{Var}(\mathbb{E}[\bar{Y}_{\text{obs}}|n_{obs}]) \\ &= \mathbb{E}\left[\frac{\sigma^2}{n_{obs}}\right] + \text{Var}[\mu] \\ &= \sigma^2\mathbb{E}\left[\frac{1}{n_{obs}}\right].\end{aligned}$$

Vamos aproximar esta esperanza, sea $f(n_{obs}) = 1/n_{obs}$ y $a = \mathbb{E}[n_{obs}] = np$ utilizando la aproximación de Taylor hasta primer orden en torno a np tenemos que

$$\frac{1}{n_{obs}} \approx \frac{1}{np} - \frac{n_{obs} - np}{(np)^2}.$$

Por lo tanto,

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n_{obs}}\right] &\approx \mathbb{E}\left[\frac{1}{np} - \frac{n_{obs} - np}{(np)^2}\right] \\ &= \mathbb{E}\left[\frac{1}{np}\right] - \mathbb{E}\left[\frac{n_{obs} - np}{(np)^2}\right] \\ &= \frac{1}{np} \approx \frac{1}{n_{obs}}.\end{aligned}$$

Entonces, la varianza del estimador insesgado $\bar{Y}_{obs}$ es aproximadamente

$$\text{Var}(\bar{Y}_{obs}) \approx \frac{\sigma^2}{n_{obs}}.$$

Además, si todos los datos fueran observados tendríamos que

$$\text{Var}(\bar{Y}_{obs}) = \text{Var}(\bar{Y}) = \frac{\sigma^2}{n}.$$

Dado que $n_{obs} \leq n$, se sigue que

$$\text{Var}(\bar{Y}_{obs}) \geq \text{Var}(\bar{Y}).$$

Entonces, aunque ambos estimadores son insesgados, el estimador $\bar{Y}_{obs}$ tiene una varianza mayor que el estimador $\bar{Y}$, por lo que es menos eficiente.

6. *Factorización bajo MAR.*
A partir de la definición de MAR, demuestra que

$$L(\theta; Y_{\mathbf{obs}}, R) \propto \mathbb{P}[Y_{\mathbf{obs}}|\theta].$$

¿Qué hipótesis adicional en la distribución *a priori* es necesaria en el enfoque Bayesiano para concluir ignorabilidad.

Sean Y el vector de datos $Y = (Y_{\text{obs}}, Y_{\text{mis}})$, R la matriz indicadora de datos faltantes, θ los parámetros del modelo de datos y ψ el mecanismo de datos faltantes. Recordemos que los datos faltantes Y_{mis} son *Missing at random* (MAR) si la probabilidad de ausencia depende de los valores observados pero no de los faltantes, es decir, si

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

Notemos que la función de densidad del modelo completo se puede reescribir como

$$\mathbb{P}[Y_{\text{obs}}, Y_{\text{mis}}, R|\theta, \psi] = \mathbb{P}[Y_{\text{obs}}, Y_{\text{mis}}|\theta]\mathbb{P}[R|Y_{\text{obs}}, Y_{\text{mis}}, \theta, \psi],$$

por la definición de probabilidad condicional $\mathbb{P}[A, B] = \mathbb{P}[B|A]\mathbb{P}[A]$. Ahora, recordemos que la función de verosimilitud de un vector de parámetros ϕ dado un conjunto de datos D es

$$L(\phi; D) = \mathbb{P}[D|\phi],$$

que en este caso corresponde a

$$L(\theta, \psi; Y_{\text{obs}}, R) = \mathbb{P}[Y_{\text{obs}}, R|\theta, \psi].$$

La probabilidad de la derecha se puede obtener del modelo original al marginalizar sobre los datos faltantes, es decir,

$$\mathbb{P}[Y_{\text{obs}}, R|\theta, \psi] = \int_{\mathbb{R}^n} \mathbb{P}[Y_{\text{obs}}, Y_{\text{mis}}, R|\theta, \psi] dY_{\text{mis}}.$$

Considerando la condición de MAR en la factorización del modelo completo, se tiene que ésta toma la forma

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

Entonces, sustituyendo esta última expresión en la integral, se obtiene que

$$\mathbb{P}[Y_{\text{obs}}, R|\theta, \psi] = \mathbb{P}[R|Y_{\text{obs}}, \psi] \int_{\mathbb{R}^n} \mathbb{P}[Y_{\text{obs}}, Y_{\text{mis}}|\theta] dY_{\text{mis}}.$$

Pero notemos que esta última integral corresponde precisamente a la distribución marginal de Y_{obs}, así

$$\mathbb{P}[Y_{\text{obs}}, R|\theta, \psi] = \mathbb{P}[R|Y_{\text{obs}}, \psi]\mathbb{P}[Y_{\text{obs}}|\theta].$$

Finalmente, notemos que la dependencia de la función de verosimilitud en θ es únicamente a través del factor $\mathbb{P}[Y_{\text{obs}}|\theta]$, ya que el otro factor no depende de θ. Por lo tanto,

$$L(\theta, \psi; Y_{\text{obs},R}) = \mathbb{P}[R|Y_{\text{obs}}, \psi]\mathbb{P}[Y_{\text{obs}}|\theta],$$

y, en consecuencia,

$$L(\theta; Y_{\text{obs}}, R) \propto \mathbb{P}[Y_{\text{obs}}|\theta],$$

teniendo así el primer resultado.

Desde el enfoque Bayesiano, la distribución *posterior* de los parámetros es proporcional a la verosimilitud por la distribución *a priori*, es decir,

$$\pi(\theta, \psi | Y_{\text{obs}}, R) \propto \mathbb{P}[R | Y_{\text{obs}}, \psi] \mathbb{P}[Y_{\text{obs}} | \theta] \pi(\theta, \psi).$$

La posterior de θ se obtiene al marginalizar respecto al mecanismo de datos faltantes ψ,

$$\pi(\theta | Y_{\text{obs}}, R) = \int_{\mathbb{R}^n} \pi(\theta, \psi | Y_{\text{obs}}, R) d\psi \propto \mathbb{P}[Y_{\text{obs}} | \theta] \int_{\mathbb{R}^n} \mathbb{P}[R | Y_{\text{obs}}, \psi] \pi(\theta, \psi) d\psi.$$

De aquí, una posible hipótesis adicional para concluir ignorabilidad es que los parámetros θ y ψ sean *a priori* independientes, es decir, $\pi(\theta, \psi) = \pi(\theta)\pi(\psi)$. En este caso, la integral implica que

$$\pi(\theta | Y_{\text{obs}}, R) \propto \pi(\theta) \mathbb{P}[Y_{\text{obs}} | \theta].$$

7. *Distancia de Cook como medida global de influencia.*
Partiendo de la definición

$$D_i = \frac{1}{p\widehat{\sigma}^2}\sum_{j=1}^{n}(\widehat{y}_j - \widehat{y}_{j(i)})^2,$$

demuestra que se puede reescribir en función de los residuos estandarizados y el leverage como

$$D_i = \frac{r_i^2}{p}\frac{h_{ii}}{1-h_{ii}}.$$

Discuta la interpretación de esta forma alternativa.

Partimos de la definición de la distancia de Cook:

$$D_i = \frac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2}{p\hat{\sigma}^2}$$

donde:

- $\hat{y}_j$ es el valor ajustado para la observación j usando todo el conjunto de datos.
- $\hat{y}_{j(i)}$ es el valor ajustado para la observación j cuando se excluye la observación i.
- p es el número de parámetros en el modelo.
- $\hat{\sigma}^2$ es el estimador de la varianza del error.

Recordemos que el residuo estandarizado se define como $r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$, donde $e_i = y_i - \hat{y}_i$ es el residuo ordinario. Por otro lado, el leverage h_{ii} es el elemento i-ésimo de la diagonal de la matriz de proyección (hat matrix) $H = X(X^TX)^{-1}X^T$.
Tomando la fórmula de actualización para los coeficientes estimados cuando se elimina la observación i, obtenemos

$$\hat{\beta}_{(i)} = \hat{\beta} - \frac{(X^TX)^{-1}x_i^Te_i}{1-h_{ii}}$$

donde x_i es la fila i-ésima de la matriz X. Ahora, el valor ajustado para la observación j usando todo el conjunto de datos es:

$$\hat{y}_j = x_j\hat{\beta}$$

Del mismo modo, el valor ajustado para la observación j cuando se elimina la observación i es:

$$\hat{y}_{j(i)} = x_j\hat{\beta}_{(i)} = x_j\left[\hat{\beta} - \frac{(X^TX)^{-1}x_i^Te_i}{1-h_{ii}}\right]$$

Entonces, la diferencia entre ambos valores ajustados es

$$\hat{y}_j - \hat{y}_{j(i)} = x_j\hat{\beta} - x_j\left[\hat{\beta} - \frac{(X^TX)^{-1}x_i^Te_i}{1-h_{ii}}\right] = x_j\left[\frac{(X^TX)^{-1}x_i^Te_i}{1-h_{ii}}\right]$$

También, recordemos que el elemento (j,i) de la matriz hat H es $h_{ji} = x_j(X^TX)^{-1}x_i^T$, por lo tanto:

$$\hat{y}_j - \hat{y}_{j(i)} = \frac{h_{ji}e_i}{1-h_{ii}}$$

así,

$$\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2 = \sum_{j=1}^{n}\left(\frac{h_{ji}e_i}{1-h_{ii}}\right)^2 = \frac{e_i^2}{(1-h_{ii})^2}\sum_{j=1}^{n}h_{ji}^2.$$

Notemos que como H es idempotente y simétrica, ésta cumple que

$$\sum_{j=1}^{n}h_{ji}^2 = h_{ii}$$

pues la diagonal de H^2 es igual a la diagonal de H. Así,

$$\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2 = \frac{e_i^2}{(1-h_{ii})^2}h_{ii}.$$

Por lo que, al sustuir en D_i obtenemos

$$D_i = \frac{1}{p\hat{\sigma}^2}\cdot\frac{e_i^2 h_{ii}}{(1-h_{ii})^2}.$$

Por otro lado, recordemos que $r_i = \dfrac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$, entonces $e_i^2 = r_i^2\hat{\sigma}^2(1-h_{ii})$. Sustituyendo,

$$D_i = \frac{1}{p\hat{\sigma}^2}\cdot\frac{[r_i^2\hat{\sigma}^2(1-h_{ii})]h_{ii}}{(1-h_{ii})^2} = \frac{r_i^2}{p}\cdot\frac{h_{ii}}{1-h_{ii}}$$

Interpretación de la forma alternativa: La distancia de Cook D_i ahora se expresa como el producto de dos partes importantes:

1. $\dfrac{r_i^2}{p}$: depende del residuo estandarizado al cuadrado y se escala por el número de parámetros p.
2. $\dfrac{h_{ii}}{1-h_{ii}}$: depende del leverage h_{ii} (que mide la influencia de la observación i en el espacio de las variables explicativas).

D_i combina la información sobre cuán atípica es la observación en términos de su residuo (r_i^2) y cuán influyente es en el ajuste (h_{ii}). Si una observación tiene un residuo grande y alto leverage (alto h_{ii}), entonces D_i será grande, indicando que la observación tiene alta influencia en el modelo. Si el residuo es pequeño, incluso con alto leverage, D_i será pequeña. Similarmente, si el leverage es bajo, D_i puede no ser muy alta.

Esta forma muestra que la distancia de Cook es una medida del efecto de ser outlier en y y en x. Es útil para detectar observaciones que afectan significativamente las estimaciones de los coeficientes y las predicciones.

8. *Invarianza afín en Min-Max.*
Sea $x = \{x_1, \ldots, x_n\}$ un conjunto de datos y definimos la transformación

$$x_i^* = \frac{x_i - \min(x)}{\max(x) - \min(x)}.$$

Demuestra que si $y_i = ax_i + b$ con $a > 0$, entonces $y_i^* = x_i^*$.

Sea $y_i = ax_i + b$ con $a > 0$. Notemos que, $\min(x) = x_k \leq x_i$, para todo i. Así,

$$\min(ax) = a\min(x) = ax_k \leq ax_i \quad \forall\, i$$

además;

$$\min(y) = \min(ax + b) = a\min(x) + b = ax_k + b \leq ax_i + b \quad \forall\, i$$

Análogamente se obtiene que $\max(y) = a\max(x) + b$ esto muestra que multiplicar por $a > 0$ y sumar b conserva el orden.

Por otro lado, la transformación normalizada de y_i es

$$y_i^* = \frac{y_i - \min(y)}{\max(y) - \min(y)} = \frac{ax_i + b - \big(a\min(x) + b\big)}{a\max(x) + b - \big(a\min(x) + b\big)}.$$

Al simplificar, se concluye que:

$$y_i^* = \frac{a\big(x_i - \min(x)\big)}{a\big(\max(x) - \min(x)\big)} = \frac{x_i - \min(x)}{\max(x) - \min(x)} = x_i^*.$$

9. *Transformación logarítmica y reducción de colas.*
Considera una variable aleatoria X tal que $X \sim$ Pareto(α, x_m), es decir, X tiene función de densidad

$$f_X(x) = \frac{\alpha x_m^\alpha}{x^{\alpha+1}}, \quad x \geq x_m > 0, \quad \alpha > 0.$$

Definimos variable aleatoria Y por la transformación $Y = \ln(X)$.

a) Encuentra la función distribución de Y y su función de densidad.

Primero notemos que la función de distribución de X está dada por

$$F_Y(y) = \mathbb{P}[Y \leq y] = \mathbb{P}[\ln(X) \leq y] = \mathbb{P}[X \leq e^y] = F_X(e^y),$$

que al evaluar directamente en la distribución de X, significa que

$$F_Y(y) = F_X(e^y) = 1 - x_m^\alpha e^{-\alpha y} = 1 - e^{-\alpha(y - \ln(x_m))}, \quad y \geq \ln x_m.$$

Esta distribución corresponde a la de una variable aleatoria exponencial con tiempo de vida garantizado $\ln(x_m)$. Derivando, obtenemos la función de densidad de Y

$$f_Y(y) = \frac{dF_Y}{dy}(y) = \alpha e^{-\alpha(y - \ln(x_m))}, \quad y \geq \ln(x_m).$$

b) Discuta cómo cambia el comportamiento de la cola al pasar de X a Y.

Por un lado, la cola de la variable aleatoria X es

$$\overline{F}_X(x) = \mathbb{P}[X > x] = \int_x^\infty f_X(u)du = \left(\frac{x_m}{x}\right)^\alpha.$$

Por el otro lado, como Y es una variable aleatoria exponencial de parámetro $\alpha > 0$ con tiempo de vida garantizado $\ln(x_m)$, la variable aleatoria Z definida como $Z = Y - \ln(x_m)$ es una variable aleatoria exponencial de parámetro $\alpha > 0$. Con esta información, podemos dar la cola de Y utilizando la de Z como

$$\overline{F}_Y(y) = \mathbb{P}[Y > y] = \mathbb{P}[Z > y - \ln(x_m)] = e^{-\alpha(y - \ln(x_m))} = ke^{-\alpha y},$$

donde $k = x_m^\alpha$. Sólo para compararlas funcionalmente (porque en sentido estricto habría que considerar la transformación al sustituir en el límite, además de que representan exactamente el mismo problema pero reescalado), notemos que

$$\lim_{z\to\infty} \frac{\overline{F}_Y(z)}{\overline{F}_X(z)} = \lim_{z\to\infty} \frac{e^{-\alpha z}}{z^{-\alpha}} = \lim_{z\to\infty} \frac{z^\alpha}{e^{\alpha z}} = \left(\lim_{z\to\infty} \frac{z}{e^z}\right)^\alpha = 0,$$

resultado clásico sobre que el crecimiento exponencial supera al polinomial. Intuitivamente, esto quiere decir que la cola de Y es mucho más ligera que la de X.

c) Explica por qué la transformación logarítmica "acorta" colas largas y produce distribuciones más cercanas a la simetría.

Dividimos la respuesta en las dos afirmaciones, aunque de fondo la razón es la misma.

- Sobre el "peso" de las colas antes y después de la transformación. La justificación es la presentada en el inciso anterior. Recordemos que una distribución es de colas pesadas si su función generadora de momentos no existe, y el que ésta no exista sugiere que a partir de cierto k, los momentos de orden superior a k no existen. El que no existan momentos de orden superior a algún k, significa que hay cierta dominación polinomial en la función de densidad. Al aplicar la transformación logaritmo, el comportamiento polinomial es dominado por el exponencial, lo que hace que las colas sean más ligeras.

- Sobre la simetría antes y después de la transformación. Continuando la discusión del punto anterior y tomando como ejemplo este inciso, podemos destacar que existen más momentos para la variable aleatoria transformada, en el caso del ejemplo Y sí tiene todos sus momentos. La escala logarítmica también hace que puntos distantes se acerquen, por lo que tiende a equilibrar distancias.

10. *Robustez de la mediana vs la media.*
Considera $x = \{1, 2, 3, 4, M\}$, con $M \to \infty$. a) Calcula la media $\overline{x}$ y la desviación estándar s_x como función de M.

Recordemos que para un conjunto de datos $y = \{y_1, y_2, \ldots, y_m\}$, la media y la desviación estándar del conjunto son

$$\overline{y} = \frac{1}{m}\sum_{i=1}^{m} y_i \qquad \text{y} \qquad s_y = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \overline{y})^2}.$$

En este caso, se tiene que el tamaño de muestra es $n = 5$, entonces, sustituyendo directamente en las definiciones, se tiene que la media es

$$\overline{x}(M) = \frac{1}{5}(1 + 2 + 3 + 4 + M) = \frac{1}{5}(10 + M) = 2 + \frac{1}{5}M,$$

y la desviación estándar es

$$\begin{aligned} s_x^2(M) &= \frac{1}{5}\left[\left(1 - 2 - \frac{1}{5}M\right)^2 + \left(2 - 2 - \frac{1}{5}M\right)^2 + \left(3 - 2 - \frac{1}{5}M\right)^2 \right. \\ &\qquad \left. + \left(4 - 2 - \frac{1}{5}M\right)^2 + \left(M - 2 - \frac{1}{5}M\right)^2\right] \\ &= \frac{1}{5}\left[\frac{4}{5}M^2 - 4M + 10\right] = \frac{4}{25}M^2 - \frac{4}{5}M + 2, \\ s_x(M) &= \sqrt{\frac{4}{25}M^2 - \frac{4}{5}M + 2}. \end{aligned}$$

Podríamos utilizar la versión insesgada de la varianza muestral dividiendo entre $5 - 1$ en lugar de entre 5, pero para los fines de este problema ese detalle no influye.

b) Calcula la mediana m y el rango intercuartílico RIQ.

Recordemos que para un conjunto de datos **ordenado** $y = \{y_1, y_2, \ldots, y_m\}$, con m impar (como es el caso de este ejercicio), la mediana de y es el dato que se encuentra justo a la mitad, es decir

$$m_y = y_{\frac{m+1}{2}}.$$

En este caso, como la muestra ya está ordenada, la mediana es

$$m_x = 3.$$

El rango intercuartil depende de cómo se calcule:

- Con una metodología que incluye a la mediana, se toma la mediana del conjunto $y_1 = \{y_1, \ldots, m\}$ como el primer cuartil y se toma la mediana del conjunto $y_2 = \{m, \ldots, y_m\}$ como el tercer cuartil. Con este criterio, se tiene que

$$Q_1 = 2 \quad \text{y} \quad Q_3 = 4, \quad \text{de modo que} \quad RIQ = Q_3 - Q_1 = 2.$$

- Con una metodología que excluye a la mediana, se toma la mediana del conjunto $y_1' = \{y_1, \ldots, y_{k-1}\}$ como el primer cuartil y se toma la mediana del conjunto $y_2' = \{y_{k+1}, \ldots, y_m\}$ como el tercer cuartil, donde $k = \dfrac{m+1}{2}$ es la posición de la mediana en la lista de datos. Con este criterio, se tiene que

$$Q_1' = \frac{3}{2} \quad \text{y} \quad Q_3' = 2 + \frac{M}{2}, \quad \text{de modo que} \quad RIQ' = Q_3' - Q_1' = \frac{M+1}{2}.$$

c) ¿Qué medidas permanecen estables y cuáles se distorsionan al crecer M?

De los cálculos en puntos anteriores, notemos que

$$\lim_{M\to\infty} \overline{x}(M) = \lim_{M\to\infty} s_x(M) = \infty,$$

que esto justifica que la media y la desviación estándar sean muy sensibles a valores grandes.
La mediana y el rango intercuartil de la primera metodología son constantes, por lo que estos valores son estables ante valores extremos en la muestra. Del rango intercuartil se destaca la sensibilidad de una estadística ante añadir un dato más, donde el simple hecho de quitar un dato de la muestra hace que el tercer cuartil sea sensible a valores extremos.

11. *Propiedades de la transformación Box-Cox.*
Sea $y(\lambda)$ la transformación de Box-Cox definida por

$$y(\lambda) = \begin{cases} \dfrac{x^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \ln(x), & \lambda = 0, \end{cases} \qquad x > 0.$$

a) Demuestra que $\lim_{\lambda\to 0} y(\lambda) = \ln(x)$.

Sea $x > 0$, de modo que $\ln(x)$ está bien definido. Notemos que

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

Cuando hacemos el cambio de variable $u = \lambda \ln(x)$, se tiene que

$$\lim_{\lambda\to 0} \frac{e^{\lambda \ln(x)} - 1}{\lambda} = \ln(x) \lim_{u\to 0} \frac{e^u - 1}{u}.$$

Este límite ya corresponde a un límite popular, el cual se sabe que

$$\lim_{u\to 0} \frac{e^u - 1}{u} = 1.$$

La manera de demostrar esto es considerando el cambio de variable $v = e^u - 1$, el cual lleva a

$$\begin{aligned} \lim_{u\to 0} \frac{e^u - 1}{u} &= \lim_{v\to 0} \frac{v}{\ln(1+v)} = \lim_{v\to 0} \frac{1}{\ln\left[(1+v)^{\frac{1}{v}}\right]} \\ &= \frac{1}{\ln\left[\lim_{v\to 0} (1+v)^{\frac{1}{v}}\right]} = \frac{1}{\ln e} = 1. \end{aligned}$$

Por lo tanto,

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

b) Propón un ejemplo numérico donde x toma valores muy dispersos y compara el efecto de $\lambda = 1$ (sin transformación) frente a $\lambda = 0$ (logaritmo).

Sea $x = \{1, e^1, e^2, \ldots, e^n\}$. Para comparar el efecto de $\lambda = 1$ con el de $\lambda = 0$ se revisarán la media y la mediana, para lo cual consideramos n par, por comodidad. Notemos que la media y la mediana originales son

$$\overline{x} = \frac{1}{n+1} \sum_{k=0}^{n} e^k = \frac{1}{n+1} \frac{e^{n+1} - 1}{e - 1} \quad \text{y} \quad m_x = e^{\frac{n}{2}}.$$

- Los datos transformados con $\lambda = 1$ son

$$y_1 = \{0, e^1 - 1, e^2 - 1, \ldots, e^n - 1\}.$$

Por un lado, la mediana de la muestra transformada es

$$m_{y_1} = e^{\frac{n}{2}} - 1,$$

lo cual es natural ya que la transformación de Box-Cox es creciente y, por lo tanto, preserva el orden de los datos. Por otro lado, la media es

$$\overline{y}_1 = \frac{1}{n+1} \frac{e^{n+1} - 1}{e - 1} - 1.$$

- Los datos transformados con $\lambda = 0$ son

$$y_2 = \{0, 1, 2, \ldots, n\}.$$

Por un lado, la mediana de la muestra transformada es

$$m_{y_2} = \frac{n}{2},$$

lo cual es natural ya que la transformación de Box-Cox es creciente y, por lo tanto, preserva el orden de los datos. Por otro lado, la media es

$$\overline{y}_2 = \frac{1}{n+1}\sum_{k=0}^{n} k = \frac{1}{n+1}\frac{n(n+1)}{2} = \frac{n}{2}.$$

La discusión sobre este ejemplo va en el mismo sentido que la discusión del ejercicio 9. La transformación con $\lambda = 0$ hace que la media de los datos transformados crezca mucho más lento que con la transformación lineal $\lambda = 1$, otra vez teniendo un efecto lineal vs. uno exponencial. Esto ejemplifica uno de los criterios que se utilizan en la práctica cuando uno tiene datos positivamente sesgados con colas largas: elegir un valor de λ cercano a 0 para la transformación Box-Cox.

12. *Propiedades del histograma.*
Sea $x = \{x_1, \ldots, x_n\}$ una muestra de variables aleatorias independientes e idénticamente distribuidas con densidad f. Considera el histograma con k intervalos de ancho h y estimador

$$\widehat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} \mathbb{1}_{\{x_i \in I_j\}}, \quad x \in I_j.$$

a) Demuestra que $\widehat{f}_h(x) \geq 0$ para todo x.

Notemos que el número de intervalos y el ancho de cada uno no puede ser negativo, entonces, $k, h > 0$. Además, como

$$\mathbf{1}\{x_i \in I_j\} = \begin{cases} 1 & \text{si } x_i \in I_j \\ 0 & \text{o.c.} \end{cases}$$

entonces

$$\sum_{i=1}^{n} \mathbf{1}\{x_i \in I_j\} \geq 0.$$

Como $\hat{f}_h(x)$ es el producto de números no negativos tendríamos para todo x que

$$\hat{f}_h(x) \geq 0.$$

b) Demuestra que

$$\int_{\mathbb{R}} \widehat{f}_h(x) dx = 1.$$

Notemos que, por evaluación directa,

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

Por lo tanto,

$$\int_{-\infty}^{\infty} \hat{f}_h(x)\, dx = 1.$$

c) Discute cómo afecta al histograma elegir h muy grande o muy pequeño en términos de sesgo y varianza.

Algunos de los fenómenos que podrían ocurrir al elegir h son los siguientes.

- Si tomamos un valor de h muy grande, tendríamos poco intervalos. Entonces, para valores más grandes de h, habrían más datos en cada intervalo. Esto haría que el histograma se veiera mas suavizado, lo que daría una varianza baja. A su vez, el sesgo seria alto al perder información sobre la densidad.

- Si tomamos un valor de h muy pequeño, tendríamos más intervalos. Esto ocasionaría que en cada intervalo se tengan menos datos, y posiblemente muchos 0. Como en cada intervalo habría poco información, entonces la varianza seria alta. A su vez, el sesgo seria bajo.

13. ***Estimación de densidad Kernel (KDE).***
Sea

$$\widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

con kernel K integrable, tal que

$$\int_{\mathbb{R}} K(u)du = 1, \quad \int_{\mathbb{R}} uK(u)du \quad \textbf{y} \quad \mu_2(K) = \int_{\mathbb{R}} u^2 K(u)du < \infty.$$

a) ***Normalización.*** **Demuestra que**

$$\int_{\mathbb{R}} \widehat{f}_h(x)dx = 1.$$

Al la integral, directamente obtenemos que

$$\int_{-\infty}^{\infty} \hat{f}_h(x)dx = \int_{-\infty}^{\infty} \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) dx = \frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{\infty} K\left(\frac{x - x_i}{h}\right) dx.$$

Hacemos el cambio de variable $u = \frac{x-x_i}{h}$, así, $\frac{du}{dx} = \frac{1}{h}$, así

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

donde la tercera igualdad se da por la primera propiedad del Kernel.

b) ***No negatividad.*** **Demuestra que $\widehat{f}_h(x) \geq 0$ si $K(u) \geq 0$ para todo u.**
Si $K(u) \geq 0$ para todo u, entonces cada término $K(\frac{x - x_i}{h}) \geq 0$ y por lo tanto

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \geq 0.$$

c) ***Sesgo puntual.*** **Utilizando expansión de Taylor de f alrededor de x, demuestra que**

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

Al calcular directamente la esperanza, tenemos que

$$\mathbb{E}[\hat{f}_h(x)] = \mathbb{E}\left[\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)\right] = \frac{1}{nh}\sum_{i=1}^{n}\mathbb{E}\left[K\left(\frac{x - X_i}{h}\right)\right].$$

Dado que las X_i son independientes e idénticamente distribuidas,

$$\mathbb{E}[\hat{f}_h(x)] = \frac{1}{h}\mathbb{E}\left[K\left(\frac{x-X}{h}\right)\right].$$

Ahora,

$$\mathbb{E}\left[K\left(\frac{x-X}{h}\right)\right] = \int_{-\infty}^{\infty} K\left(\frac{x-t}{h}\right) f(t)\, dt.$$

Hacemos el cambio de variable:

$$u = \frac{x-t}{h} \quad \Rightarrow \quad t = x - hu \quad \Rightarrow \quad dt = -h\, du.$$

Cuando $t \to -\infty$, $u \to +\infty$; cuando $t \to +\infty$, $u \to -\infty$. Entonces,

$$\begin{aligned}\int_{-\infty}^{\infty} K\left(\frac{x-t}{h}\right) f(t)\, dt &= \int_{+\infty}^{-\infty} K(u) f(x-hu)(-h)\, du \\ &= h\int_{-\infty}^{\infty} K(u) f(x-hu)\, du.\end{aligned}$$

Por lo tanto,

$$\mathbb{E}[\hat{f}_h(x)] = \frac{1}{h}\cdot h\int_{-\infty}^{\infty} K(u) f(x-hu)\, du = \int_{-\infty}^{\infty} K(u) f(x-hu)\, du.$$

Expandimos $f(x-hu)$ alrededor de x usando Taylor:

$$f(x-hu) = f(x) - huf'(x) + \frac{(hu)^2}{2} f''(x) + o(h^2).$$

Sustituyendo esta expansión, tenemos que:

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

Usando las propiedades del kernel:

$$\begin{aligned}\int_{-\infty}^{\infty} K(u)\, du &= 1, \\ \int_{-\infty}^{\infty} uK(u)\, du &= 0, \\ \mu_2(K) = \int_{-\infty}^{\infty} u^2 K(u)\, du &< \infty,\end{aligned}$$

obtenemos:

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

Por lo tanto el sesgo es

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