Tarea 1
Fundamentos y Preparación de los Datos
Parte Teórica

Rodolfo Ramírez
Guillermo Aguilar
Jesús Alonso

1. **Hat Matrix y propiedades algebraicas.**
Demuestre que la matriz
$$H = X(X^\top X)^{-1}X^\top$$
es idempotente y simétrica. Explique por qué estas propiedades son fundamentales para la interpretación de los *leverages*.

Proof. Nótese que la transpuesta de la matriz H es H
$$\begin{aligned} H^T &= (X(X^TX)^{-1}X^T)^T \\ &= (X^T)^T((X^TX)^{-1})^T(X)^T \\ &= X((X^TX)^T)^{-1}X^T \\ &= X(X^TX)^{-1}X^T \\ &= H \end{aligned}$$
por lo que H es una matriz simétrica. También es idempotente dado que
$$\begin{aligned} HH &= X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T \\ &= X(X^TX)^{-1}X^T, \quad \text{pues } (X^TX)^{-1}X^TX = I \\ &= H \end{aligned}$$
entonces $HH = H$ lo que significa que H es idempotente. Dado que sabemos que $HH = H$ (H es idempotente) entonces para toda $i \in \{1, \ldots, n\}$ se tiene que
$$h_{ii} = h_{ii}^2 + \sum_{j \neq i} h_{ij}h_{ji}$$
y dado que H es simétrica, $h_{ij} = h_{ji}$ por lo que
$$h_{ii} = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2$$
así, $h_{ii} \geq h_{ii}^2 \geq 0$ pues $\sum_{j \neq i} h_{ij}^2 \geq 0$ lo que implica que $0 \leq h_{ii} \leq 1$.

$\square$

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

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

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

Proof. Por definición, $\sum_{i=1}^{n} h_{ii} = tr(H)$, mas aun tenemos

$$\sum_{i=1}^{n} h_{ii} = tr(X(X^{\top}X)^{-1}X^{\top}) = tr((X^{\top}X)^{-1}X^{\top}X) = tr(I).$$

Donde la segunda igualdad se sustenta por propiedades de la traza. Dado que X es una matriz de $n \times p$, H es una matriz de $p \times p$, por lo que $\sum_{i=1}^{n} h_{ii} = p$. Dado que la suma siempre es constante, valores grandes de $h_{i,i}$ ocasionan la existencia de valores pequeños de $h_{i,i}$ para otras observaciones. Esto provoca que si hay una series de $h_{i,i}$'s cercanos a 1, estos obliguen a la regresión a pasar cerca de ellos. □

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

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

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

Proof. Nótese que, multiplicando por $1 = \sigma^2/\sigma^2$ podemos reescribir a r_i como:

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

dónde σ^2 es la varianza teórica de los errores. Además, sabemos que

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

y

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

así, si el numerador y el denominador fueran independientes, r_i tendría una distribución t de Student con n-p grados de libertad (t_{n-p}). Sin embargo no es el caso pues $\hat{\sigma}^2$ depende de e_i. Es así que r_i tiene una distribución casi t_{n-p}. Dado que sabemos la distribución de r_i entonces al calcular el valor podemos decir si es un valor típico dentro de la distribución teórica que tiene o estamos obteniendo valores raros dada su distribución. □

4. **Factorización bajo MCAR.**
Partiendo de la definición de MCAR, pruebe formalmente que

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

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

Proof. Sea $\mathbf{Y}$ un vector aleatorio con distribución que depende del vector de parámetros θ entonces, por el teorema de Bayes,

$$\begin{aligned} f[\mathbf{Y}, \mathbf{R}|\theta, \psi] &= \frac{f[\mathbf{Y}, \mathbf{R}, \theta, \psi]}{f[\theta, \psi]} \\ &= \frac{f[\mathbf{Y}, \mathbf{R}, \theta, \psi]}{f[\theta, \psi]} \frac{f[\theta, \psi, \mathbf{Y}]}{f[\theta, \psi, \mathbf{Y}]} \\ &= f[\mathbf{R}|\mathbf{Y}, \theta, \psi] f[\mathbf{Y}|\theta, \psi] \end{aligned}$$

dado que asumimos que los errores son **MCAR** entonces

$$f[\mathbf{R}|\mathbf{Y}, \theta, \psi] = f[\mathbf{R}|\psi]$$

pues se asume que los datos faltantes no dependen de los valores de $\mathbf{Y}$ y por lo tanto tampoco del vector de parámetros θ. Es prudente pensar que $\sigma(\mathbf{Y}, \theta) \perp \sigma(\psi)$ pues asumimos que la distirbución de los datos sólo depende de θ. Así,

$$f[\mathbf{Y}|\theta, \psi] = f[\mathbf{Y}|\theta].$$

así, juntando los dos resultados anteriores,

$$f[\mathbf{Y}, \mathbf{R}|\theta, \psi] = f[\mathbf{Y}|\theta] f[\mathbf{R}|\psi] \tag{1}$$

Ahora, nótese que podemos deducir la distribución de los valores missings de la siguiente manera:

$$\begin{aligned} f[\mathbf{Y}_{obs}, \mathbf{R}|\theta, \psi] &= \int_{\mathbb{R}^d} f[\mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \mathbf{R}|\theta, \psi] d\mathbf{Y}_{mis} \\ &= \int_{\mathbb{R}^d} f[\mathbf{Y}, \mathbf{R}|\theta, \psi] d\mathbf{Y}_{mis} \\ &= \int_{\mathbb{R}^d} f[\mathbf{Y}|\theta] f[\mathbf{R}|\psi] d\mathbf{Y}_{mis} \quad \text{usando (4)} \\ &= f[\mathbf{R}|\psi] \int_{\mathbb{R}^d} f[\mathbf{Y}|\theta] d\mathbf{Y}_{mis} \\ &= f[\mathbf{R}|\psi] f[\mathbf{Y}_{obs}|\theta] \end{aligned}$$

así, marginalizando $\mathbf{Y}_{obs}$,

$$f[\mathbf{Y}_{obs}|\theta, \psi] = f[\mathbf{Y}_{obs}|\theta]$$

por lo que para haver inferencia sobre θ podemos ignorar el mecanismo de faltantes ψ. □

5. **Insesgadez bajo eliminación de casos (MCAR).**
Sea Y_{obs} la media muestral basada solo en los casos observados. Demuestre que

$$\mathbb{E}[Y_{\text{obs}}] = \mu$$

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

Proof. Notemos que $\overline{Y}_{obs} = \frac{\sum_{i=1}^{n} Y_i R_i}{n_{obs}}$ y $n_{obs} = \sum_{i=1}^{n} R_i$. Donde R_i es la variable indicadora donde $R_i = 1$ si Y_i es observado y $R_i = 0$ en el caso contrario. Utilizando esperanza condicional tenemos que

$$\begin{aligned}
\mathbb{E}[\overline{Y}_{obs}] &= \mathbb{E}[\mathbb{E}[\overline{Y}_{obs}|R_1, ..., R_n]] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{\sum_{i=1}^{n} Y_i R_i}{n_{obs}}\middle| R_1, ..., R_n\right]\right] \\
&= \mathbb{E}\left[\frac{\sum_{i=1}^{n} R_i}{n_{obs}}\mathbb{E}\left[Y_i|R_1, ..., R_n\right]\right] \\
&= \mathbb{E}\left[\frac{\sum_{i=1}^{n} R_i}{n_{obs}}\mathbb{E}\left[Y_i|R_i\right]\right]
\end{aligned}$$

Dado que estamos trabajando sobre MCAR se cumple que $R_i \perp\!\!\!\perp Y_i$ entonces

$$\mathbb{E}[Y_i|R_i] = \mathbb{E}[Y_i] = \mu.$$

Luego,

$$\mathbb{E}[\overline{Y}_{obs}] = \mathbb{E}\left[\mu\frac{\sum_{i=1}^{n} R_i}{n_{obs}}\right] = \mathbb{E}[\mu] = \mu$$

Calculemos la varianza de $\overline{Y}_{\text{obs}}$. Usamos la descomposición de varianza total:

$$\text{Var}(\overline{Y}_{\text{obs}}) = \mathbb{E}\big[\text{Var}(\overline{Y}_{\text{obs}} \mid R)\big] + \text{Var}\big(\mathbb{E}[\overline{Y}_{\text{obs}} \mid R]\big)\,.$$

Ya vimos que $\mathbb{E}[\overline{Y}_{\text{obs}} \mid R] = \mu$, así la segunda parte es cero. Para la primera parte, dado que la muestra es independiente,

$$\text{Var}(\overline{Y}_{\text{obs}} \mid R) = \text{Var}\left(\frac{1}{n_{\text{obs}}}\sum_{i=1}^{n} Y_i \,\middle|\, R\right) = \frac{1}{n_{\text{obs}}^2}\sum_{i=1}^{n}\text{Var}(Y_i) = \frac{\sigma^2}{n_{\text{obs}}}.$$

Por tanto,

$$\text{Var}(\overline{Y}_{\text{obs}}) = \mathbb{E}\left[\frac{\sigma^2}{n_{\text{obs}}}\right] = \sigma^2\,\mathbb{E}\left[\frac{1}{n_{\text{obs}}}\right].$$

Observa que n_{obs} es aleatorio y satisface

$$\mathbb{E}[n_{\text{obs}}] = \sum_{i=1}^{n}\mathbb{E}[R_i].$$

Usando la desigualdad de Jense, dado que la función $x \mapsto 1/x$ es convexa en los reales positivos, se tiene que

$$\mathbb{E}\left[\frac{1}{n_{\text{obs}}}\right] \;\geq\; \frac{1}{\mathbb{E}[n_{\text{obs}}]} \;=\; \frac{1}{np}.$$

con $p := \mathbb{P}[R_i = 1]$ y dado que $0 < p \leq 1$ entonces $np \leq n$. Así,

$$\operatorname{Var}(\overline{Y}_{\text{obs}}) = \sigma^2\,\mathbb{E}\left[\frac{1}{n_{\text{obs}}}\right] \;\geq\; \frac{\sigma^2}{np} \geq \frac{\sigma^2}{n}.$$

Si no hubiera valores faltantes entonces la varianza de la media muestral, dada por,

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

estaría dada por

$$\operatorname{Var}(\overline{Y}) = \frac{\sigma^2}{n},$$

De esta manera,

$$\operatorname{Var}(\overline{Y}_{\text{obs}}) \geq \operatorname{Var}(\overline{Y})$$

por lo que la media muestral de los datos observados pierde eficiencia al tener una varianza más grande.

□

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

$$L(\theta; Y_{obs}, R) \propto p(Y_{obs}|\theta).$$

¿Qué suposición adicional en el prior es necesaria en el enfoque bayesiano para concluir ignorabilidad?

Proof. Missing at Random (MAR) significa que, para todo r y y,

$$p(r \mid y, \psi) = p(r \mid y_{\text{obs}}, \psi).$$

Es decir, condicional en los datos observados, el mecanismo no depende de los valores faltantes y_{fal}. El verosímil basado en (Y_{obs}, R) para (θ, ψ) se obtiene marginalizando los datos faltantes:

$$\begin{aligned} L(\theta, \psi; y_{\text{obs}}, r) &\propto p(y_{\text{obs}}, r \mid \theta, \psi) \\ &= \int p(y_{\text{obs}}, y_{\text{fal}}, r \mid \theta, \psi)\, \mathrm{d}y_{\text{fal}} \\ &= \int \underbrace{p(y_{\text{obs}}, y_{\text{fal}} \mid \theta)}_{p(y|\theta)}\ \underbrace{p(r \mid y_{\text{obs}}, y_{\text{fal}}, \psi)}_{p(r|y,\psi)}\, \mathrm{d}y_{\text{fal}}. \qquad (2) \end{aligned}$$

Bajo MAR, $p(r \mid y_{\mathrm{obs}}, y_{\mathrm{fal}}, \psi) = p(r \mid y_{\mathrm{obs}}, \psi)$ no depende de y_{fal}, por lo que puede sacarse de la integral:

$$\begin{aligned}
L(\theta, \psi; y_{\mathrm{obs}}, r) &= p(r \mid y_{\mathrm{obs}}, \psi) \int p(y_{\mathrm{obs}}, y_{\mathrm{fal}} \mid \theta)\, \mathrm{d}y_{\mathrm{fal}} \\
&= p(r \mid y_{\mathrm{obs}}, \psi)\, p(y_{\mathrm{obs}} \mid \theta).
\end{aligned}$$

En particular, la *verosímilitud para* θ (tratando ψ como parametro auxiliar) verifica

$$L(\theta; y_{\mathrm{obs}}, r) \;\propto\; p(y_{\mathrm{obs}} \mid \theta)$$

pues el factor $p(r \mid y_{\mathrm{obs}}, \psi)$ *no* depende de θ.
La *ignorabilidad* para la inferencia sobre θ exige que el posterior $p(\theta \mid y_{\mathrm{obs}}, r)$ coincida con el que se obtendría ignorando el mecanismo (i.e., usando sólo $p(y_{\mathrm{obs}} \mid \theta)$). Para ello, además de MAR, se requiere una suposición sobre el prior.

El prior se factoriza como

$$\pi(\theta, \psi) = \pi(\theta)\, \pi(\psi),$$

y los espacios paramétricos son *distintos* (la factorización no impone restricciones cruzadas sobre (θ, ψ)). El posterior conjunto es

$$\begin{aligned}
p(\theta, \psi \mid y_{\mathrm{obs}}, r) &\propto \pi(\theta, \psi)\, L(\theta, \psi; y_{\mathrm{obs}}, r) \\
&\propto \pi(\theta)\pi(\psi)\; \underbrace{p(r \mid y_{\mathrm{obs}}, \psi)}_{\text{indep. de } \theta}\; \underbrace{p(y_{\mathrm{obs}} \mid \theta)}_{\text{indep. de } \psi}.
\end{aligned}$$

Al marginalizar ψ,

$$\begin{aligned}
p(\theta \mid y_{\mathrm{obs}}, r) &= \int p(\theta, \psi \mid y_{\mathrm{obs}}, r)\, \mathrm{d}\psi \\
&\propto \pi(\theta)\, p(y_{\mathrm{obs}} \mid \theta)\; \underbrace{\int \pi(\psi)\, p(r \mid y_{\mathrm{obs}}, \psi)\, \mathrm{d}\psi}_{\text{constante en } \theta}.
\end{aligned}$$

El término entre llaves no depende de θ, por lo que

$$p(\theta \mid y_{\mathrm{obs}}, r) \;\propto\; \pi(\theta)\, p(y_{\mathrm{obs}} \mid \theta)$$

que es exactamente el posterior que se obtendría *ignorando* el mecanismo de faltantes. Por consiguiente, bajo **MAR** y **prior factorizado** (distinción de parámetros), el mecanismo es **ignorable** para la inferencia bayesiana en θ.

□

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

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

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

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

Discuta la interpretación de esta forma alternativa.

Proof. Sea $\hat{y}_j$ la predicción de y_j obtenida mediante el modelo completo y $\hat{y}_{j(i)}$ la predicción de y obtenida bajo el modelo dónde se ha omitido la observación i-ésima, entonces. Podemos reescribir la expresión dada en términos matriciales, más aún, en términos de las betas de los dos modelos:

$$\begin{aligned} D_i :&= \frac{\sum_{i=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2}{ps^2} \\ &= \frac{(\widehat{\mathbf{Y}}_{(i)} - \widehat{\mathbf{Y}})^T(\widehat{\mathbf{Y}}_{(i)} - \widehat{\mathbf{Y}})}{ps^2} \\ &= \frac{\left(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}}\right)^T (X^TX) \left(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}}\right)}{ps^2} \end{aligned}$$

dónde $\widehat{\boldsymbol{\beta}}_{(i)}$ y $\widehat{\boldsymbol{\beta}}$ son las betas estimadas bajo el modelo sin la observación i-ésima y bajo el modelo completo respectivamente. Se tiene además que la relación entre $\widehat{\boldsymbol{\beta}}_{(i)}$ y $\widehat{\boldsymbol{\beta}}$ está dada por

$$\widehat{\boldsymbol{\beta}}_{(i)} = \widehat{\boldsymbol{\beta}} - (X^TX)^{-1}\mathbf{x}_i \frac{e_i}{1-h_{ii}}$$

dónde $\mathbf{x}_i$ es el vector de variables dependientes de la i-ésima observación. Así,

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}} &= \widehat{\boldsymbol{\beta}} - (X^TX)^{-1}\mathbf{x}_i \frac{e_i}{1-h_{ii}} - \widehat{\boldsymbol{\beta}} \\ &= (X^TX)^{-1}\mathbf{x}_i \frac{e_i}{1-h_{ii}} \end{aligned}$$

así,

$$\begin{aligned} D_i &= \frac{1}{ps^2}\left(\frac{e_i}{1-h_{ii}}\right)^2 \mathbf{x}_i^\top (X^\top X)^{-1}(X^\top X)(X^\top X)^{-1}\mathbf{x}_i \\ &= \frac{1}{ps^2}\left(\frac{e_i}{1-h_{ii}}\right)^2 \mathbf{x}_i^\top (X^\top X)^{-1}\mathbf{x}_i \\ &= \frac{1}{ps^2}\left(\frac{e_i}{1-h_{ii}}\right)^2 h_{ii} \\ &= \left(\frac{e_i}{s\sqrt{1-h_{ii}}}\right)^2 \left(\frac{h_{ii}}{p(1-h_{ii})}\right) \\ &= \frac{r_i^2}{p}\left(\frac{h_{ii}}{p(1-h_{ii})}\right) \end{aligned}$$

dónde

$$r_i := \frac{e_i}{s\sqrt{1-h_{ii}}}$$

son los residuos estandarizados. De las úlitmas dos expresiones podemos ver que entre más grande el leverage i.e. h_{ii} está más cerca de 1, la distancia de cook de la i-ésima observación es más grande. □

8. **Invarianza afín en Min–Max** Sea $x_1, \ldots, x_n$ un conjunto de datos y defina la transformación

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

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

Proof. Ya que por hipótesis $y_i = ax_i + b$, se cumple $\min(y) = a\min(x) + b$ y $\max(y) = a\max(x) + b$ esto pues $a > 0$ i.e. $\min(ax) = a\min(x)$ y $\max(ax) = a\max(x)$. Así que por definición,

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

Lo que termina con el ejercicio. □

9. **Transformación logarítmica y reducción de colas** Considere $X \sim \text{Pareto}(\alpha, x_m)$ con densidad

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

Defina la transformación $Y = \log(X)$.

(a) Encuentre la distribución de Y y su función de densidad.

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

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

Proof. (a) Sea $X \sim \text{Pareto}(\alpha, x_m)$, con función de densidad

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

Definimos la transformación

$$Y = \log(X).$$

La densidad de Y se obtiene aplicando la fórmula de cambio de variable:

$$\begin{aligned} f_Y(y) &= f_X(y) \left| \frac{dx}{dy} \right|_{x=e^y} \\ &= \frac{\alpha x_m^\alpha}{(e^y)^{\alpha+1}} \cdot e^y \\ &= \alpha x_m^\alpha e^{-\alpha y}, \qquad y \geq \log(x_m). \end{aligned}$$

Es decir, Y tiene densidad exponencial desplazada:

$$f_Y(y) = \alpha \exp\left(-\alpha(y - \log(x_m))\right), \qquad y \geq \log(x_m).$$

(b) Para $X \sim \text{Pareto}(\alpha, x_m)$, la cola se caracteriza por

$$\Pr(X > t) = \left(\frac{x_m}{t}\right)^{\alpha}, \qquad t \geq x_m,$$

que decae como una ley de potencia (*cola pesada*).
En cambio, para Y,

$$\Pr(Y > y) = \Pr(X > e^y) = \left(\frac{x_m}{e^y}\right)^{\alpha} = \exp\big(-\alpha(y - \log(x_m))\big),$$

lo cual corresponde a un decaimiento *exponencial*, es decir, una cola mucho más ligera.

(c) La transformación logarítmica acorta colas largas porque comprime los valores grandes de la variable más que los pequeños. Esto reduce la influencia de los extremos y suele acercar la distribución a la simetría.

□

10. **Robustez de la mediana vs. la media** Considere $x = \{1, 2, 3, 4, M\}$ con $M \to \infty$.

(a) Calcule la media $\bar{x}$ y la desviación estándar s como función de M.

(b) Calcule la mediana m y el rango intercuartílico RIQ.

(c) Analice: ¿qué medidas permanecen estables y cuáles se distorsionan al crecer M?

Proof. Asumiremos que la v.a. que toma valores en $1, 2, 3, 4, M$ asigna probabilidades a estos elementos de manera uniforme.

La media muestral de x en función de M está dada por:

$$\bar{x}(M) := \frac{1}{5}(1 + 2 + 3 + 4 + M)$$

de dónde podemos ver que si $M \to \infty$ entonces $\bar{x}(M) \to \infty$. De igual manera, la desviación estandar de x en función de M esta dada por:

$$s(M) := \sqrt{\frac{(1-\bar{x}(M))^2 + (2-\bar{x}(M))^2 + (3-\bar{x}(M))^2 + (4-\bar{x}(M))^2 + (M-\bar{x}(M))^2}{5}}$$

y dado que si $M \to \infty$ entonces $\bar{x}(M) \to \infty$ entonces $s(M) \to \infty$.

(b) Para obtener la mediana nótese que, ordenando los números en orden ascendente y considerando que $M \to \infty$, se tiene que el número que queda en medio es 3, por lo tanto, la mediana de x en función de M está dada por:

$$Q_2(M) := 3$$

El rango intercuartílico esta dado por la diferencia del primer cuartíl Q_1 y el tercer cuartíl Q_3 que están dados por los cuantíles 0.25 y 0.75 , en este caso,

$$Q_1(M) = 2 \quad \text{y} \quad Q_3(M) = 4$$

esto pues

$$F(1) = \frac{1}{5} = 0.20 < 0.25 \quad , \quad F(2) = \frac{2}{5} = 0.4 > 0.25$$

y

$$F(3) = \frac{3}{5} = 0.6 < 0.75 \quad , \quad F(4) = \frac{4}{5} = 0.8 > 0.75$$

así, el rango intercuartílico está dado por:

$$RIQ(M) = Q_3(M) - Q_1(M) = 2$$

Claramente, dado que Q_2 y RIQ no dependen de M se tiene que son robustos ante valores de M muy grandes.

(c) Cómo ya se ha mencionado, la media $\bar{x}(M)$ y la desviación estandar s se distorcionan para valores grandes de M mientras que Q_2 y RIQ permanecen estables.

□

11. **Propiedades de la transformación Box-Cox**

Sea $y(\lambda)$ la transformación de Box-Cox definida como:

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

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

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

Proof. a) Definamos a $f(\lambda) = x^\lambda$. Notemos que

$$\lim_{\lambda \to 0} y(\lambda) = \lim_{\lambda \to 0} \frac{x^{0+\lambda} - x^0}{\lambda} = f'(0).$$

Entonces usando la derivada para funciones exponenciales tenemos $f'(0) = \ln(x)x^0$, por lo tanto

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

b) Definimos la serie dada por:

$$x_n = e^n$$

para $n = 1, .., 20$ cuyos valores son:

$$\{x_n\}_{n=1}^{10} = \{2.72, 7.39, 20.09, 54.6148.41, 403.43, 1096.63, 2980.96, 8103.08, 22026.47\}$$

que claramente están bastante dispersos pues estos crecen exponencialmente sin hacerles alguna transformación ($\lambda = 1$) por otro lado, haciendo la transformación ($\lambda = 0$) i.e. logarítmica,

$$\{y_n\}_{n=1}^{10} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$$

serie que crece de forma lineal.

□

12. **Propiedades del histograma**

Sea $x_1, \dots, x_n$ una muestra i.i.d. de una variable aleatoria continua con densidad $f(x)$. Considere el histograma con k intervalos de ancho h y estimador:

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

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

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

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

Proof. a) Nótese que las indicadoras de la función $\hat{f}_h(x)$ son no negativasy dado que $nh > 0$ entonces $\hat{f}_h(x) \geq 0$.

b) Supongamos que los intervalos son del mismo ancho h y llamemos n_j al número de x_i's que caen en el intervalo I_j, dónde $\{I_j\}_{j=1}^{k}$ es una partición finita luego

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

c) Para un valor fijo de particiones, si h es demasiado grande entonces estamos suavizando mucho la estimación de la densidad pues habrá muchos x_i's que caigan en pocas particiones, mientras que pocas x_i's caerán en muchas particiones, eso no nos permitirá detectar agrupaciones de las observaciones y entonces si volviéramos a observar otra muestra de datos no cambiaría mucho la estimación i.e. la varianza sería pequeña pero el sesgo grande. El otro extremo ocurre si h es muy pequeña, entonces con cada muestra nueva observada, diferentes cantidades de x_i's caerán en cada partición, lo que provocará que la estimación cambie mucho con cada observación (la estimación tiene una alta varianza), pero el sesgo es pequeño. □

13. **Ejercicio: Estimación de densidad kernel (KDE)**

Sea

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

con kernel K integrable, $\int K(u)\,du = 1$, $\int uK(u)\,du = 0$, y segundo momento finito $\mu_2(K) = \int u^2 K(u)\,du$.

i) **Normalización**: Demuestre que $\int_{-\infty}^{\infty} \hat{f}_h(x)\,dx = 1$.

ii) **No negatividad**: Muestre que $\hat{f}_h(x) \geq 0$ si $K(u) \geq 0$ para todo u.

iii) **Sesgo puntual**: Usando expansión de Taylor de f alrededor de x, derive que

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

Proof. Notemos que $\hat{f}_h(x)$ satisface las siguientes propiedades

- Integra 1 sobre los reales,

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

  bajo el cambio de variable $u := \frac{x-x_i}{h}$, así,

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

  usando el hecho de que $\int_{-\infty}^{\infty} K(u)du = 1$.

- Si $K(u) \geq 0$ para toda $u \in \mathbb{R}$ entonces $\hat{f}_h(x) \geq 0, \forall x \in \mathbb{R}$ esto pues:

$$K\left(\frac{x-x_i}{h}\right) \geq 0, \forall x \in \mathbb{R}$$

  y entonces

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

- Sean $X_1, \ldots, X_n$ una sucesión de v.a. independientes e idénticamente distribuidas con densidad f entonces se tiene que:

$$\begin{aligned}\mathbb{E}[\hat{f}_h(x)] &= \frac{1}{nh}\sum_{i=1}^{n}\mathbb{E}\left[K\left(\frac{x-X_i}{h}\right)\right] \\ &= \frac{n}{nh}\mathbb{E}\left[K\left(\frac{x-X_1}{h}\right)\right] \quad \text{pues son ident. dist.} \\ &= \frac{1}{h}\int_{-\infty}^{\infty} f(y)K\left(\frac{x-y}{h}\right)dy \\ &= \int_{-\infty}^{\infty} f(x-hu)K(u)\,du\end{aligned}$$

Ahora, usando la expansión de taylor de $f(y)$,

$$f(y) = f(x) + \frac{f'(x)}{1}(y-x) + \frac{f''(x)}{2!}(y-x)^2 + g(y)(y-x)^2$$

dónde g es una función tal que $\lim_{y\to x} g(y) = 0$. Así, haciendo el cambio de variable $u = \frac{x-y}{h}$ entonces $y = x - hu$ y se tiene que

$$f(x-hu) = f(x) - \frac{f'(x)}{1}hu + \frac{f''(x)}{2}(hu)^2 + g(x-hu)(hu)^2$$

así,

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

pero nótese que

$$\frac{h^2u^2\int_{-\infty}^{\infty} g(x-hu)K(u)du}{h^2} = u^2\int_{-\infty}^{\infty} g(x-hu)K(u)du$$

y dado que para una h lo suficientemente cercana a 0, $x - hu \to x$ entonces

$$\lim_{h\to 0} g(x-hu) = \lim_{y\to x} g(y) = 0$$

entonces en una vecindad de x la función está acotada, digamos por M, entonces

$$|g(x-hu)K(u)| \leq MK(u)$$

y dado que $MK(u)$ es integrable, por convergencia dominada,

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

por lo anterior, la función

$$h^2u^2\int_{-\infty}^{\infty} g(x-hu)K(u)du$$

es $o(h^2)$ i.e.

$$\begin{aligned}
\mathbb{E}[\hat{f}_h(x)] - f(x) &= \frac{h^2 f''(x)}{2}\int_{-\infty}^{\infty} u^2K(u)du + o(h^2) \\
&= \frac{h^2}{2}\mu_2(K)f''(x) + o(h^2)
\end{aligned}$$

□

References

[1] L. Leticia Ramírez Ramírez (2024). *Modelos Estadísticos I. Modelos Lineales y Lineales Generalizados.*