Introducción a Ciencia de Datos

Tarea 1

Jessica Rubí Lara Rosales jessica.lara@cimat.mx
Eric Ernesto Moreles Abonce eric.moreles@cimat.mx
Luis Erick Palomino Galván luis.palomino@cimat.mx

Ejercicio 1. *Redactado por Jessica Rubí Lara Rosales*

Hat Matriz y propiedades algebraicas *Demuestre que la matriz*

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

es idempotente y simétrica. Explique por qué estas propiedades son fundamentales para la interpretación de los leveranges.

Solución. Primero veamos que es idempotente

$$HH = X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T = X(X^TX)^{-1}[X^TX](X^TX)^{-1}X^T = X(X^TX)^{-1}X^T = H$$

Además, es simétrica pues

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

Estas propiedades son fundamentales para la interpretación de los leveranges ya que toda matriz simétrica e idempotente es una matriz de proyección y en particular, como H esta definida de esa manera, se tiene que H proyecta sobre el espacio $C(X)$ el espacio generado por las columnas de X el cual tiene dimensión p. Dado que los valores ajustados se definen como

$$\hat{y} = Hy = h_{ii}y_i + \sum_j h_{ij}y_j$$

Aquí se puede ver como el peso de los *leveranges* h_{ii} afecta a la proyección pues si h_{ii} fuera grande se tendría que la proyección jala a los demás puntos hacia y_i. ■

Ejercicio 2. *Redactado por Luis Erick Palomino Galván*

Lema 1

Sea cumple que $tr(AB) = tr(BA)$

(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 parámetrosz discuta su relación con el sobreajuste.

Demostración. Sea $X \in \mathbb{R}^{n \times p}$ la matriz con n observación y p parámetros, por el Lema 1, notemos que

$$\sum_{j=1}^{n} h_j j = tr(H) = tr(X(X^T X)^{-1} X^T) = tr((X^T X)^{-1} X^T X) = tr(I_p) = p.$$

Por lo que se cumple la igualdad. ■

Ejercicio 3. *Redactado por Jessica Rubí Lara Rosales*

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* $n-p-1$ *grados de libertad. Explique cómo esta propiedad justifica su uso en la detección de outliers.*

Resultado 1. Sea $Y \sim N_n(0, I_n)$ y sea A una matriz simétrica. Entonces

$$Y^T AY \sim \chi^2(r)$$

ssi A es idempotente y de rango igual a r.

Resultado 2. Si $Z \sim N(0,1)$, $U \sim \chi^2(k)$ y Z y U independientes, entonces

$$X := \frac{Z}{\sqrt{U/k}} \sim t(k).$$

t es la distribución t- student con k grados de libertad.

Solución. Por definición tenemos

$$e = Y - \hat{Y} = Y - HY = (I-H)Y = (I-H)(X\beta + \varepsilon)$$

De las hipótesis de regresión lineal, sabemos que $\varepsilon \sim N(0, \sigma^2 I_n)$ así que $e \sim N(0, \sigma^2(I-H))$ pues $I-H$ es la matriz de poryección que proyecta sobre el espacio $C(X)^\perp$ y además cumple que es idempotente y simétrica.
Así que al dividir cada residuo por su desviación estándar tenemos que

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

Como normalmente no se conoce la varianza se usa el estimador insesgado de la varianza, el cual es

$$\hat{\sigma}^2 = \frac{e^T e}{n-p}.$$

Notemos

$$\frac{ee^T}{\sigma^2} = \frac{\varepsilon^T(I-H)^T(I-H)\varepsilon}{\sigma^2} = \frac{\varepsilon^T(I-H)\varepsilon}{\sigma^2} = \left(\frac{\varepsilon}{\sigma}\right)^T (I-H)\frac{\varepsilon}{\sigma}.$$

Por el resultado 1, como $I-H$ es una matriz de proyección de dimensión $n-p$ se tiene que $\frac{e^T e}{\sigma^2} \sim \chi^2(n-p)$. Así que

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

Notando que U no es independiente de Z, pues ambos estan usando la observación i-ésima. Definimos entonces $e_{(i)}$ como el vector de residuos sin el dato i por un argumento analogo al anterior tenemos que

$$\frac{e_{(i)}^T e_{(i)}}{\sigma^2} \sim \chi^2(n-p-1).$$

De ello que

$$U_{(i)} := \frac{(n-p-1)\hat{\sigma}^2_{(i)}}{\sigma^2} \sim \chi^2(n-p-1).$$

Por el resultado 2 obtenemos que

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

Por lo tanto, podemos concluir que r_i tiene aproximadamente una distribución t de Student con $n-p-1$ grados de libertad.

Esta propiedad nos permite contruir umbrales para poder detectar outliers. Pues, sabiendo que

$$t^*_i = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}} \sim t(n-p-1)$$

Planteamos una prueba de hipotesis

$$\begin{aligned} H_0 &: \text{ la observación } i \text{ es consistente con el modelo.} \\ H_a &: \text{ la observación no es consistente con el modelo (es un outlier)} \end{aligned}$$

Tomando α como el nivel de error tipo I. Consideramos los

$$|t^*_i| > t_{1-\alpha/2,n-p-1}$$

donde $t_{1-\alpha/2,n-p-1}$ es el punto critico de una t-student con $1-\alpha/2$ de probabilidad. Por convención se suele tomar $\alpha = 0.05$. Si n es suficientemente grande la regla de descarte es

$$\begin{aligned} |t^*_i| > 2 &\qquad \text{sospechoso} \\ |t^*_i| > 3 &\qquad \text{altamente probable que sea outlier} \end{aligned}$$

■

Ejercicio 4. *Redactado por Jessica Rubí Lara Rosales*

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

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

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

Solución. Por hipótesis del modelo MCAR se tiene

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

Usando esto por definición de probabilidad condicional se sigue

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

Así que la función de verosimilitud esta dada por

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

De ello que el estimador de máxima verosimilitud solo para θ en teoría frecuentista resulta ser

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

Así que el mecanismo de faltantes es ignorable para la inferencia sobre θ. ■

Ejercicio 5. *Redactado por Luis Erick Palomino Galván*

(Insesgadez bajo eliminación de casos MCAR) Sea $\hat{Y}_{obs}$ la media muestral basada solo en los casos observados. Demuestre que

$$\mathbb{E}\left[\hat{Y}_{obs}\right] = \mu,$$

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

Demostración. Sean $Y_1, \ldots, Y_n$ independientes e idénticamente distribuidos con media μ y varianza σ, definimos

$$R_j = \begin{cases} 1, & \text{si } Y_j \text{ está observado,} \\ 0, & \text{si } Y_j \text{ falta.} \end{cases} \qquad \text{y} \quad \hat{Y}_{obs} = \frac{\sum_{j=1}^{n} R_j Y_j}{n_{obs}},$$

donde $n_{obs} = \sum_{j=1}^{n} R_j$ es el número de observaciones observadas (supondremos $\mathbb{P}\left[n_{obs>0}\right] = 1$). Aplicando la descomposición de esperanza condicional, obtenemos

$$\mathbb{E}\left[\hat{Y}_{obs}\right] = \mathbb{E}\left[\mathbb{E}\left[\hat{Y}_{obs}|R_j\right]\right] = \mathbb{E}\left[\frac{\sum_{j=1}^{n} R_j Y_j}{n_{obs}}|R\right] = \frac{1}{n_{obs}} \sum_{j=1}^{n} R_j \mathbb{E}\left[Y_j|R_j\right].$$

Dado que estamos suponiendo MCAR, entonces R_j es independiente de Y_j para cada j, por lo que $\mathbb{E}\left[Y_j|R_j\right] = \mathbb{E}\left[Y_j\right] = \mu$.

$$\frac{1}{n_{obs}} \sum_{j=1}^{n} R_j \mathbb{E}\left[Y_j|\mathbb{1}_{Y_j}\right] = \frac{1}{n_{obs}} \sum_{j=1}^{n} R_j \mu = \mu \frac{n_{obs}}{n_{obs}} = \mu.$$

En consecuencia $\mathbb{E}\left[\hat{Y}_{obs}\right] = \mu$. ■

Ejercicio 6. *Redactado por Jessica Rubí Lara Rosales*

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?

Solución. Por hipótesis del modelo MAR se tiene

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

Usando esto y que θ es el parámetro del modelo propuesto para Y, por definición de probabilidad condicional se tiene que

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

Así que la función de verosimilitud esta dada por

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

De ello que el estimador de máxima verosimilitud solo para θ en teoría frecuentista es

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

La suposición adicional que se necesita para el enfoque bayesiano es que $\pi(\theta, \psi) = \pi(\theta)\pi(\psi)$, es decir, las distribuciones a priori de los parámetros θ, ψ sean independientes. De esta manera al calcular

$$\mathbb{P}\left[\theta, \psi|Y_{obs}\right] \propto L(\theta, \psi; Y_{obs}, R)\pi(\theta, \psi) \propto \mathbb{P}\left[R|Y_{obs}, \psi\right]\mathbb{P}\left[Y_{obs}|\theta\right]\pi(\theta)\pi(\psi)$$

Marginalizando se tiene

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

■

Ejercicio 7. *Redactado por Eric Ernesto Moreles Abonce*

Distancia de Cook como medida global de influencia *Partiendo de la definicion*

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

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

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

Solución. Veamos primero que:

$$\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2 = \left(\hat{\beta}_{(i)} - \hat{\beta}\right)' X'X \left(\hat{\beta}_{(i)} - \hat{\beta}\right)$$

Desarrollando el lado derecho:

$$\left(\hat{\beta}_{(i)} - \hat{\beta}\right)' X'X \left(\hat{\beta}_{(i)} - \hat{\beta}\right) = \left(X\hat{\beta}_{(i)} - X\hat{\beta}\right)' \left(X\hat{\beta}_{(i)} - X\hat{\beta}\right)' = \left(\hat{Y}_{(i)} - \hat{Y}\right)' I \left(\hat{Y}_{(i)} - \hat{Y}\right)$$

Donde lo ultimo es la forma cuadrática de $\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2$, entonces:

$$D_i = \frac{\left(\hat{\beta}_{(i)} - \hat{\beta}\right)' X'X \left(\hat{\beta}_{(i)} - \hat{\beta}\right)}{p\hat{\sigma}^2}$$

Ahora notamos que:

$$X'X = (X'_{(i)}X_{(i)}) + x_i x'_i$$

Por la formula de Sherman-Morrison:

$$(X'_{(i)}X_{(i)})^{-1} = \left((X'X) - x_i x'_i\right)^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1}x_i x'_i (X'X)^{-1}}{1 - x'_i(X'X)^{-1}x_i}$$

Donde $h_{ii} = x'_i(X'X)^{-1}x_i$. Multiplicando ambos lados por x_i por la derecha:

$$
\begin{aligned}
(X'_{(i)}X_{(i)})^{-1}x_i &= (X'X)^{-1}x_i + \frac{(X'X)^{-1}x_i x_i'(X'X)^{-1}x_i}{1 - x_i'(X'X)^{-1}x_i} \\
&= (X'X)^{-1}x_i \left(1 + \frac{x_i'(X'X)^{-1}x_i}{1 - x_i'(X'X)^{-1}x_i}\right) \\
&= (X'X)^{-1}x_i \left(\frac{1 - x_i'(X'X)^{-1}x_i + x_i'(X'X)^{-1}x_i}{1 - x_i'(X'X)^{-1}x_i}\right) \\
&= (X'X)^{-1}x_i \left(\frac{1}{1 - h_{ii}}\right) \\
&= \frac{(X'X)^{-1}x_i}{1 - h_{ii}}
\end{aligned}
$$

Ahora, para $\hat{\beta}_{(i)} - \hat{\beta}$:

$$
\begin{aligned}
\hat{\beta}_{(i)} - \hat{\beta} &= \left(X'_{(i)}X_{(i)}\right)^{-1} X'_{(i)}Y_{(i)} - \hat{\beta} \\
&= \left(X'_{(i)}X_{(i)}\right)^{-1} X'Y - (X'_{(i)}X_{(i)})^{-1}x_i y_i - \hat{\beta} \\
&= \frac{(X'X)^{-1}x_i x_i'\hat{\beta}}{1 - h_{ii}} - (X'_{(i)}X_{(i)})^{(-1)}x_i y_i \\
&= \frac{(X'X)^{-1}x_i}{1 - h_{ii}}(x_i'\hat{\beta} - y_i)
\end{aligned}
$$

Sustituyendo esto en nuestra formula de D_i:

$$
\begin{aligned}
D_i &= \frac{\left(\hat{\beta}_{(i)} - \hat{\beta}\right)' X'X \left(\hat{\beta}_{(i)} - \hat{\beta}\right)}{p\hat{\sigma}^2} \\
&= \frac{\left(\frac{(X'X)^{-1}x_i}{1-h_{ii}}(x_i'\hat{\beta} - y_i)\right)' X'X \left(\frac{(X'X)^{-1}x_i}{1-h_{ii}}(x_i'\hat{\beta} - y_i)\right)}{p\hat{\sigma}^2} \\
&= \frac{1}{(1 - h_{ii})^2} \frac{(x_i'\hat{\beta} - y_i)' x_i'((X'X)^{-1})' X'X (X'X)^{-1}x_i (x_i'\hat{\beta} - y_i)}{p\hat{\sigma}^2} \\
&= \frac{1}{(1 - h_{ii})^2} \frac{(x_i'\hat{\beta} - y_i)' x_i'(X'X)^{-1}x_i (x_i\hat{\beta} y_i)}{p\hat{\sigma}^2} \\
&= \frac{1}{(1 - h_{ii})^2} \frac{e_i^2 h_{ii}}{p\hat{\sigma}^2} \\
&= \left(\frac{e_i}{\hat{\sigma}\sqrt{(1 - h_{ii})}}\right)^2 \frac{h_{ii}}{(1 - h_{ii})p} \\
&= \frac{r_i^2}{p} \frac{h_{ii}}{1 - h_{ii}}
\end{aligned}
$$

Y esta ultima igualdad es lo que queríamos. Los residuales estandarizados miden que tan bien cabe un punto con el modelo, los leverages miden que tan inusuales son los valores predichos para la observación i, y esto a su vez afecta al denominador $1 - h_{ii}$ pues si tenemos un leverage muy alto, se ve reflejado mucho en la formula. Entonces la reformulación nos dice que tanto cambiaría borrar las observaciones i. ■

Ejercicio 8. *Redactado por Luis Erick Palomino Galván*

(Invarianza afín en Min-Max) Sea $x_1, \ldots, x_n$ un conjunto de datos y defina la transformación

$$x_j^* = \frac{x_j - \text{mín}\{x\}}{\text{máx}\{x\} - \text{mín}\{x\}}.$$

Pruebe que si $y_j = ax_j + b$ con $0 < a$, entonces $y_j^* = x_j^*$.

Demostración. Notemoss que

$$\begin{aligned} y_j^* &= \frac{y_j - \text{mín}\{y\}}{\text{máx}\{y\} - \text{mín}\{y\}} \\ &= \frac{(ax_j + b) - (a\,\text{mín}\{x\} + b)}{(a\,\text{máx}\{x\} + b) - (a\,\text{mín}\{x\} + b)} \\ &= \frac{a(x_j - \text{mín}\{x\})}{a(\text{máx}\{x\} - \text{mín}\{x\})} \\ &= \frac{x_j - \text{mín}\{x\}}{\text{máx}\{x\} - \text{mín}\{x\}} \\ &= x_j^*. \end{aligned}$$

Por lo que en efecto $y_j^* = x_j^*$. ■

Ejercicio 9. *Redactado por Jessica Rubí Lara Rosales*

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

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

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

1. Encuentre la distribución de Y y su función de densidad.

   Solución. Tomando $y \geq \log(x_m) > 0$ se tiene

   $$\mathbb{P}[Y \leq y] = \mathbb{P}[\log(X) \leq y] = \mathbb{P}[X \leq e^y].$$

   Por el teorema de cambio de variable, la densidad de Y es

   $$f_Y(y) = f_X(e^y)e^y = \frac{\alpha x_m^\alpha}{e^{y(\alpha+1)}}e^y = \alpha x_m^\alpha e^{-y\alpha} = \alpha e^{-y\alpha+\alpha\log x_m} = \alpha e^{-\alpha(y-\log x_m)}.$$

   Así que la distribución de Y resulta ser una distribución exponencial de parámetro α trucada a la izquierda en $\log(x_m)$. ■

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

   Solución. Sabemos que la cola de la distribución Pareto es pesada pues decae a una tasa polinómica. Mientras que la cola de la distribución exponencial es ligera pues decae a una tasa exponencial. Así que el comportamiento cambia drásticamente. ■

3. Explique por que la transformación logarítmica 'acorta' colas largas y produce distribuciones más cercanas a la simetría.

   Solución. Esto se debe a que la función $\log$ reduce mucho la escala para valores muy grandes, De esta manera los valores extremos de X se mapean a valores más cercanos al centro de Y. Además, se obtiene una distribución más simétrica debido a que la compresión no se hace lineal, comprime más los valores grandes que los pequeños. ■

Ejercicio 10. *Redactado por Eric Ernesto Moreles Abonce*

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

a) Calcule la media $\overline{x}$ *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?

Solución. a) La media es:

$$\begin{aligned}\overline{x} &= \frac{1+2+3+4+M}{5} \\ &= \frac{10}{5} + \frac{M}{5} \\ &= 2 + \frac{M}{5}\end{aligned}$$

La varianza:

$$\begin{aligned}s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2 \\ &= \frac{1}{5}(1+4+9+16+M^2) - \left(2+\frac{M}{5}\right)^2 \\ &= \frac{30+M^2}{5} - \left(2+\frac{M}{5}\right)^2 \\ &= \frac{30+M^2}{5} - \frac{(10+M)^2}{5^2} \\ &= \frac{1}{5}\left(30 + M^2 - \frac{100 + -20M + M^2}{5}\right) \\ &= \frac{1}{5}\left(\frac{150 + 5M^2 - 100 + 20M - M^2}{5}\right) \\ &= \frac{1}{5}\left(\frac{4M^2 + 20M + 50}{5}\right) \\ &= \frac{4M^2 + 20M + 50}{25} \\ &= \frac{4M^2}{25} + \frac{4}{5}M + 2 \\ &= \frac{4}{5}\left(\frac{M^2}{5} + M\right) + 2\end{aligned}$$

Y con lo anterior, tenemos la media y la desviación estándar:

$$\begin{aligned}\overline{x} &= 2 + \frac{M}{5} \\ s &= \sqrt{\frac{4}{5}\left(\frac{M^2}{5} + M\right) + 2}\end{aligned}$$

b) La mediana es $m = 3$, porque no importa que tan grande sea M, el valor intermedio siempre sera 3. Para el rango intercuartílico RIQ, necesitamos la mediana de la mitad "inferior" de los datos y la de la mitad "superior":

$$Q_1 = \frac{1+2}{2}$$
$$Q_3 = \frac{4+M}{2}$$

Y entonces:

$$RIQ = Q_3 - Q_1 = \frac{4+M}{2} - \frac{1+2}{2} = \frac{M+1}{2}$$

c) Tenemos así las medidas:

$$\overline{x} = 2 + \frac{M}{5}$$
$$s = \sqrt{\frac{4}{5}\left(\frac{M^2}{5} + M\right) + 2}$$
$$m = 3$$
$$RIQ = \frac{M+1}{2}$$

Todas menos la mediana dependen de M, y cuando M aumenta, también lo hacen estas. Por lo tanto, la única medida que permanece estable es la mediana.

■

Ejercicio 11. *Redactado por Luis Erick Palomino Galván*

(Propiedades de la transformación Box - Cox) Sea $y(\lambda)$ la transformación de Box - Cox definida para $0 < x$ como

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

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

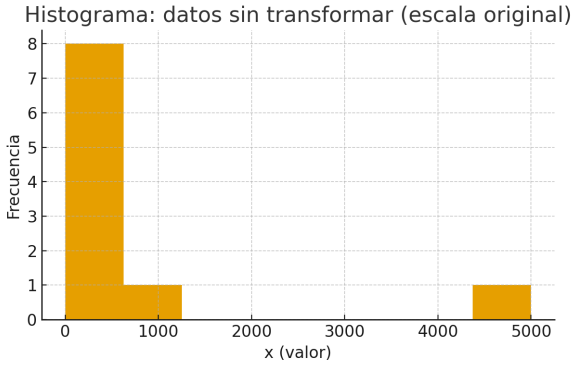
   Demostración. Notemos que

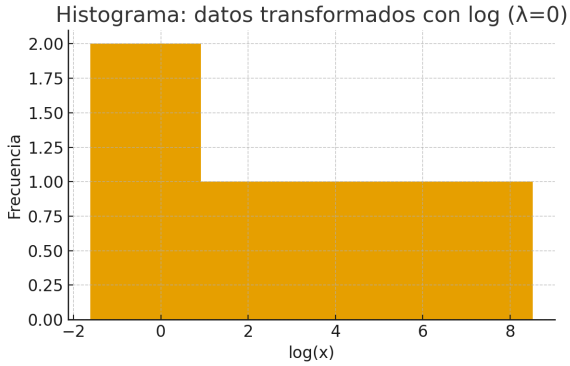
   $$\begin{aligned} \lim_{\lambda \to 0} \frac{x^\lambda - 1}{\lambda} &= \lim_{\lambda \to 0} \frac{e^{\ln(x)\lambda} - 1}{\lambda} \\ &= \lim_{\lambda \to 0} \frac{x^\lambda \ln(x)}{1} \\ &= x^0 \ln(x) = \ln(x). \end{aligned}$$

   Por lo que se cumple $\lim_{\lambda \to 0} y(\lambda) = \ln(x)$. ∎

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

   Solución. Veamos los siguientes histogramas:





   Donde los datos están muy sesgados a la derecha, la mayoría son de valores pequeños y unos pocos enormes que dominan la escala. Notemos que, en la escala logarítmica la distribución se comprime y se aproxima a una forma menos sesgada, donde la diferencia entre la mediana y la media disminuye. Vemos que el logaritmo reduce la influencia de valores extremos. ∎

Ejercicio 12. *Redactado por Jessica Rubí Lara Rosales*

Propiedades del histograma *Sea $x_1, \ldots, x_n$ una muestra iid de una variable aleatoria continua con densidad $f(x)$. Considere el histograma con k intervalos de ancho h y estimador:*

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

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

   Solución. Sea $x \in Dom(X)$ donde X es una distribución con densidad f. Como $n, h > 0$ y la función $\mathbb{1}$ regresa valores cero o uno, se cumple que $\hat{f}_h(x) \geq 0$. ■

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

   Solución. Primero notemos que

   $$\int_{-\infty}^{\infty} \hat{f}_h(x)dx = \sum_{j=1}^{k} \int_{I_j} \frac{1}{nh} \sum_{i=1}^{n} \mathbb{1}\{x_i \in I_j\}dx$$

   Definamos $n_j = \sum_{i=1}^{n} \mathbb{1}\{x_i \in I_j\}$ con $1 \leq j \leq k$ el numero total de datos de la muestra que caen en I_j. Notemos que $\sum_{j=1}^{k} n_j = n$ y $n_j \geq 0$. De ello

   $$\int_{-\infty}^{\infty} \hat{f}_h(x)dx = \frac{1}{nh} \sum_{j=1}^{k} \int_{I_j} n_j dx = \frac{1}{nh} \sum_{j=1}^{k} n_j |I_j| = \frac{1}{nh} nh = 1.$$

   Pues por hipótesis $|I_j| = h$ para todo j. ■

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

   Solución. Podemos notar que al hacer cada vez más grande el h se cumple que k va decreciendo y los n_j son cada vez más grandes. De ello que su varianza disminuya pues la mayoría va a estar agrupada en los mismo bloques o en bloques muy cercanos a ellos. Pero el sesgo incrementa pues va a tender a ser mas grande en un grupo que no necesariamente es tan grande.

   Por otro lado, al hacer h muy pequeño la varianza va a aumentar, pues la mayoría de los datos van a estar separados en distintas columnas a menos que tengan el mismo valor Pero el sesgo aquí no es problema porque se toma en cuenta cada dato. ■

Ejercicio 13. *Redactado por Eric Ernesto Moreles Abonce*

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 $\mu_2(K) = \int u^2 K(u)du$.

- ***Normalización****: Demuestre que $\int_{-\infty}^{\infty} \hat{f}_h(x)dx = 1$.*
- ***No negatividad****: Muestre que $\hat{f}_h(x) \geq 0$ si $K(u) \geq 0$ para todo u.*
- ***Sesgo puntual****: Usando expansión de Taylor de f alrededor de x, derive que:*

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

Solución. a) Veamos que:

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

Sustituyendo la definición en la integral:

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

Haciendo el cambio de variable $u = \frac{x-x_i}{h}$, tenemos que $du = \frac{1}{h}dx$. Procedemos:

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

Pero por hipótesis, $\int_{-\infty}^{\infty} K(u)du = 1$, por lo tanto:

$$\frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} K(u)du = \frac{1}{n}\sum_{i=1}^{n} 1 = 1$$

b) Si $K(u) \leq 0$ para todo u, se sigue que:

$$K\left(\frac{x - x_i}{h}\right) \geq 0$$

Para todo x, y en particular:

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

Pues estamos sumando cosas no negativas. Al final multiplicamos por $(nh)^{-1}$, ambos números positivos, y la desigualdad no cambia de sentido, y concluimos así que:

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

c) Por hipótesis, x_i iid, entonces:

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

Hacemos el mismo cambio de variable que en el inciso a), tal que:

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

Si expandimos $f(x - hu)$ en su serie de Taylor alrededor de x, tenemos que:

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

Sustituyendo en la integral:

$$\int_{-\infty}^{\infty} K(u) f(x - hu)du = \int_{-\infty}^{\infty} K(u)\left(f(x) - uhf'(x) + \frac{u^2h^2}{2}f''(x) + o(h^2)\right)du$$

$$= f(x)\int_{-\infty}^{\infty} K(u)du - hf'(x)\int_{-\infty}^{\infty} uK(u) + \frac{h^2 f''(x)}{2}\int_{-\infty}^{\infty} u^2K(u)du + o(h^2)$$

Pero por hipótesis, tenemos que:

$$\int_{-\infty}^{\infty} K(u)du = 1$$

$$\int_{-\infty}^{\infty} uK(u) = 0$$

$$\int_{-\infty}^{\infty} u^2K(u)du = \mu_2(K)$$

Sustituyendo en la igualdad, llegamos a que:

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

Y restando $f(x)$ de ambos lados, llegamos a lo que queremos, pues:

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

■