Introducción a Ciencia de Datos
Profesor: Dr. Marco Antonio Aquino López.
Alumnos: Rodrigo, Alejandra, Luz Maria

Tarea 1: Teórica
Semestre 3

CIMAT

Ejercicio 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.

Solución:

Observe que:

$$\begin{aligned} H^\top &= (X(X^\top X)^{-1}X^\top)^\top, \\ &= ((X^\top X)^{-1}X^\top)^\top X^\top, \\ &= x^{\top\top}((X^\top X)^{-1})^\top X^\top. \end{aligned}$$

Como $X^\top X$ es una matriz simétrica DP, se tiene que $(X^\top X)^{-1}$ también es simétrica, así:

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

Por lo que, la Hat Matriz es simétrica.

Ahora, se probará que es idempotente:

$$\begin{aligned} H^2 &= HH, \\ &= (X(X^\top X)^{-1}X^\top)(X(X^\top X)^{-1}X^\top), \\ &= X(X^\top X)^{-1}\left(X^\top X\right)(X^\top X)^{-1}X^\top, \\ &= XI(X^\top X)^{-1}X^\top, \text{ donde } I \text{ es la matriz identidad}, \\ &= X(X^\top X)^{-1}X^\top, \\ &= H. \end{aligned}$$

Por lo tanto, la Hat Matriz es idempotente.

En regresión, la matriz H también es conocida como "matriz de proyección" (Ver Definición 2.2) porque proyecta el vector de observaciones y en el espacio de columnas de X, obteniendo los valores ajustados:

$$\hat{y} = Hy$$

Los elementos de la diagonal de H, denotados h_{ii}, se llaman *leverages* (influencias o apalancamientos). Esotos miden cuánto influye la observación i-ésima en su propio valor ajustado.

Ahora, las propiedades de simetría e idempotencia se relacionan de la siguiente manera con los *leverages*:

- La idempotencia asegura que los valores ajustados $\hat{y}$ son invariantes a proyecciones adicionales: $H\hat{y} = H(Hy) = H^2y = Hy = \hat{y}$.
- Esto también implica que los residuos $e = (I - H)y$ son ortogonales a los valores ajustados: $\hat{y}^\top e = 0$.

- La simetría de H garantiza que los leverages h_{ii} son medidas de distancia al cuadrado en el espacio de columnas de X. Es decir, h_{ii} mide la distancia de la observación i-ésima al centro de los datos en el espacio de X.

Más información aquí, C6.

Ejercicio 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.

Solución:
Tenemos que como el modelo de regresión lineal es de n observaciones y p parámetros (incluyendo la intersección), la matriz de diseño $\mathbf{X}$ es de tamaño $n \times p$ y la matríz sombrero se define como:

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

Los elementos h_{ii} son los valores de la diagonal de $\mathbf{H}$, entonces,

$$\sum_{i=1}^{n} h_{ii} = \text{tr}(\mathbf{H}) = \text{tr}\left(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right).$$

Recordemos que una propiedad de la traza es que $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$ para matrices $\mathbf{A}$ y $\mathbf{B}$ compatibles, por lo que usando dicha propiedad (con $\mathbf{A} = \mathbf{X}$ y $\mathbf{B} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$) se tiene que:

$$\text{tr}\left(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right) = \text{tr}\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\right).$$

Pero $\mathbf{X}^T\mathbf{X}$ es una matriz $p \times p$, y $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$, la matriz identidad de tamaño $p \times p$. Por lo tanto,

$$\text{tr}\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\right) = \text{tr}(\mathbf{I}_p) = p.$$

Así,

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

Interpretación:
h_{ii} mide la influencia que tiene la i-ésima observación en la predicción, un valor alto indica que tiene un peso muy grande en la determinación de los parámetros del modelo. La suma $\sum_{i=1}^{n} h_{ii} = p$ muestra que el "número efectivo de parámetros" en el modelo es exactamente p (que determian la complejidad del modelo), por lo que si p es muy grande (es decir, si el modelo tiene muchos parámetros), entonces el modelo tiene una mayor capacidad para ajustarse a los datos, pero los valores h_{ii} tienden a ser más altos en promedio lo que implica que el modelo es muy sensible a cada observación. Esto puede llevar a sobreajuste.

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

Solución:

Consideremos el modelo de regresión lineal clásico $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$, con n observaciones, p parámetros a estimar y con vector de errores $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$.
Recordemos que el estimador de mínimos cuadrados de $\boldsymbol{\beta}$ está dado por $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, de donde tenemos que

$$\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}, \quad \text{con } \mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T.$$

Sea $\mathbf{e}$ el vector de residuales del modelo. Sabemos que $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$. Podemos ver entonces que

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{H}\mathbf{Y} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y} = (\mathbf{I}_n - \mathbf{H})\mathbf{X}\boldsymbol{\beta} - (\mathbf{I}_n - \mathbf{H})\boldsymbol{\varepsilon},$$

pero

$$(\mathbf{I}_n - \mathbf{H})\mathbf{X} = \mathbf{X} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X} = \mathbf{X} - \mathbf{X} = \mathbf{0},$$

por lo que

$$\mathbf{e} = (\mathbf{I}_n - \mathbf{H})\boldsymbol{\varepsilon}.$$

Por consiguiente,

$$\mathbb{E}(\mathbf{e}) = \mathbb{E}[(\mathbf{I}_n - \mathbf{H})\boldsymbol{\varepsilon}] = (\mathbf{I}_n - \mathbf{H})\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0},$$

y

$$Var(\mathbf{e}) = Var[(\mathbf{I}_n - \mathbf{H})\boldsymbol{\varepsilon}] = (\mathbf{I}_n - \mathbf{H})Var(\boldsymbol{\varepsilon})(\mathbf{I}_n - \mathbf{H})^T = \sigma^2(\mathbf{I}_n - \mathbf{H})(\mathbf{I}_n - \mathbf{H})^T = \sigma^2(\mathbf{I}_n - \mathbf{H}),$$

donde la última igualdad viene dada de que $\mathbf{I}_n - \mathbf{H}$ es simétrica e idempotente.
Así, vemos que, si e_i denota la i-ésima entrada del vector $\mathbf{e}$

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

con $1 - h_{ii}$ la $i-$ésima entrada diagonal de $\mathbf{I}_n - \mathbf{H}$.

Luego, de la multiplicación matricial, sabemos que $e_i = \sum_{j=1}^{n}(\mathbf{I}_n - \mathbf{H})_{ij}\varepsilon_j$, y además los ε_j's son v.a.i.i.d. Normales, por lo que e_i es una combinación lineal de v.a. Normales independientes y por ende

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

Ahora bien, consideremos el estimador insesgado de σ^2 el cual denotamos por $\hat{\sigma}^2$ y sabemos que está dado por $\hat{\sigma}^2 = \frac{\mathbf{e}^T\mathbf{e}}{n-p}$. Probaremos a continuación que $\frac{\mathbf{e}^T\mathbf{e}}{\sigma^2}$ sigue una distribución ji-cuadrada con $n-p$ grados de libertad.
Veamos que

$$\frac{\mathbf{e}^T\mathbf{e}}{\sigma^2} = \frac{[(\mathbf{I}_n - \mathbf{H})\mathbf{Y}]^T(\mathbf{I}_n - \mathbf{H})\mathbf{Y}}{\sigma^2} = \frac{\mathbf{Y}^T(\mathbf{I}_n - \mathbf{H})\mathbf{Y}}{\sigma^2} = \frac{\mathbf{Y}^T}{\sigma}(\mathbf{I}_n - \mathbf{H})\frac{\mathbf{Y}}{\sigma}.$$

Como sabemos, $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, por lo que $\frac{\mathbf{Y}}{\sigma} \sim N_n(\frac{\mathbf{X}\boldsymbol{\beta}}{\sigma}, \mathbf{I}_n)$. Además, dado que $\mathbf{I}_n - \mathbf{H}$ es idempotente, entonces su traza es igual a su rango, de donde

$$Rango(\mathbf{I}_n - \mathbf{H}) = tr(\mathbf{I}_n - \mathbf{H}) = tr(\mathbf{I}_n) - tr(\mathbf{H}) = n - \sum_{i=1}^{n} h_{ii} = n - p.$$

Así, por un resultado de formas cuádraticas tenemos que $\frac{\mathbf{Y}^T}{\sigma}(\mathbf{I}_n - \mathbf{H})\frac{\mathbf{Y}}{\sigma}$ sigue una distribución ji-cuadrada no centrada con $n-p$ grados de libertad y parámetro de no centralidad $\frac{1}{2}\frac{(\mathbf{X}\boldsymbol{\beta})^T}{\sigma}(\mathbf{I}_n - \mathbf{H})\frac{\mathbf{X}\boldsymbol{\beta}}{\sigma}$, sin embargo,

$$\frac{(\mathbf{X}\boldsymbol{\beta})^T}{\sigma}(\mathbf{I}_n - \mathbf{H})\frac{\mathbf{X}\boldsymbol{\beta}}{\sigma} = \frac{\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}}{\sigma^2} = \frac{\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}}{\sigma^2} = 0.$$

Por lo tanto, $\frac{\mathbf{e}^T\mathbf{e}}{\sigma^2} \sim \chi^2_{n-p}$.

Conociendo ya que $e_i \sim N(0, \sigma^2(1-h_{ii}))$ y $\frac{\mathbf{e}^T\mathbf{e}}{\sigma^2} \sim \chi^2_{n-p}$, veamos que

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

pero notemos que $\frac{e_i}{\sqrt{\sigma^2(1-h_{ii})}}$ es una v.a. Normal estándar y $\sqrt{\frac{\mathbf{e}^T\mathbf{e}}{\sigma^2(n-p)}}$ es la raíz cuadrada de una v.a. ji-cuadrada dividida entre sus grados de libertad, es decir que el cociente de estas tendría una distribución t de student con $n-p$ grados de libertad SI las v.a. e_i y $\mathbf{e}^T\mathbf{e}$ fuesen independientes, sin embargo, dado que la suma de cuadrados residuales sí incluye al i-ésimo residual, el numerador y denominador del cociente mencionado no son independientes. Por lo tanto los residuos estandarizados se comportan de manera similar a una variable aleatoria t de Student pero la distribución no es exacta.

Recordemos que un leverage alto puede indicar que el correspondiente punto es influyente, pero al tener un valor de leverage alto, el denominador del residuo estandarizado asociado será cercano a 0 y por consiguiente tendrá un valor grande. Es decir que valores grandes de los residuales estandarizados indican puntos influyentes. Así, el uso de los residuos estandarizados en la detección de outliers está justificado por el hecho de que estos se comportan casi como una t de Student, pues, como sabemos, esta es una distribución con colas más pesadas que la normal, entonces la mayoría de los valores deberían caer en el rango esperado, por lo que hay probabilidad muy baja de que un valor de residual estandarizado sea muy alto, entonces si llegase a pasar podemos inferir que dicho dato tiene un comportamiento diferente a los demás y por ende puede ser un outlier.

Ejercicio 4
Factorización bajo MCAR.

Partiendo de la definición de MCAR, pruebe formalmente que

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

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

Solución:
Sean:

- $Y = (Y_{\text{obs}}, Y_{\text{mis}})$ la matriz de datos (observados y faltantes)
- R la matriz indicadora de faltantes ($R_{ij} = 1$ si observado, 0 si faltante)
- θ parámetros del modelo de datos
- ψ parámetros del mecanismo de faltantes

El mecanismo **MCAR** se define como:

$$\mathbb{P}(R \mid Y, \theta, \psi) = \mathbb{P}(R \mid \psi)$$

Es decir, la probabilidad de que un dato falte es independiente de Y.

Para probar la proposición, se parte de la probabilidad condicional $\mathbb{P}(Y, R \mid \theta, \psi)$:

$$\begin{aligned}\mathbb{P}(Y, R \mid \theta, \psi) &= \frac{\mathbb{P}(R, Y, \theta, \psi)}{\mathbb{P}(\theta, \psi)}, \\ &= \frac{\mathbb{P}(R \mid Y, \theta, \psi)\mathbb{P}(Y, \theta, \psi)}{\mathbb{P}(\theta, \psi)}, \\ &= \mathbb{P}(R \mid Y, \theta, \psi)\frac{\mathbb{P}(Y, \theta, \psi)}{\mathbb{P}(\theta, \psi)}, \\ &= \mathbb{P}(R \mid Y, \theta, \psi)\,\mathbb{P}(Y \mid \theta, \psi), \text{ todo esto por probabilidad condicional.}\end{aligned}$$

Por la definición de MCAR:

$$\mathbb{P}(R \mid Y, \theta, \psi) = \mathbb{P}(R \mid Y, \psi) = \mathbb{P}(R \mid \psi)$$

Además, el modelo de datos no depende de ψ:

$$\mathbb{P}(Y \mid \theta, \psi) = \mathbb{P}(Y \mid \theta)$$

Sustituyendo:

$$\mathbb{P}(Y, R \mid \theta, \psi) = \mathbb{P}(R \mid \psi)\,\mathbb{P}(Y \mid \theta)$$

¿Por qué el mecanismo de faltantes es ignorable para la inferencia sobre θ?

Un mecanismo de faltantes es **ignorable,** si los parámetros θ y ψ son distintos (espacios de parámetros separados)

Sobre la inferencia, se debe mostrar el corolario que viene en las notas:

Corolario. La *verosimilitud de datos observados* para θ es

$$L_{\text{obs}}(\theta) \propto \int p(Y_{\text{obs}}, Y_{\text{mis}} \mid \theta) dY_{\text{mis}} = p(Y_{\text{obs}} \mid \theta).$$

Así, la inferencia sobre θ puede basarse en $p(Y_{\text{obs}} \mid \theta)$, ignorando $p(R \mid \psi)$.

Mini prueba:

La verosimilitud basada en los datos observados (Y_{obs}, R) es:

$$L(\theta, \psi \mid Y_{\text{obs}}, R) \propto \mathbb{P}(Y_{\text{obs}}, R \mid \theta, \psi)$$

Integrando sobre los valores faltantes:

$$\mathbb{P}(Y_{\text{obs}}, R \mid \theta, \psi) = \int \mathbb{P}(Y_{\text{obs}}, Y_{\text{mis}}, R \mid \theta, \psi)\, dY_{\text{mis}}$$

Usando la factorización demostrada:

$$\begin{aligned}\mathbb{P}(Y_{\text{obs}}, R \mid \theta, \psi) &= \int \mathbb{P}(Y_{\text{obs}}, Y_{\text{mis}} \mid \theta)\,\mathbb{P}(R \mid \psi)\, dY_{\text{mis}}, \\ &= \mathbb{P}(R \mid \psi) \int \mathbb{P}(Y_{\text{obs}}, Y_{\text{mis}} \mid \theta)\, dY_{\text{mis}}, \\ &= \mathbb{P}(R \mid \psi)\,\mathbb{P}(Y_{\text{obs}} \mid \theta).\end{aligned}$$

Por lo tanto:

$$L(\theta, \psi \mid Y_{\text{obs}}, R) \propto \mathbb{P}(Y_{\text{obs}} \mid \theta)\,\mathbb{P}(R \mid \psi)$$

Para la inferencia sobre θ, la verosimilitud marginal es:

$$L(\theta \mid Y_{\text{obs}}, R) \propto \mathbb{P}(Y_{\text{obs}} \mid \theta)$$

El término $\mathbb{P}(R \mid \psi)$ no depende de θ y puede tratarse como constante. Si además θ y ψ son independientes *a priori*, la *posteriori* de θ es:

$$\mathbb{P}(\theta \mid Y_{\text{obs}}, R) \propto \mathbb{P}(Y_{\text{obs}} \mid \theta)\, \mathbb{P}(\theta)$$

Por lo tanto, bajo MCAR, la verosimilitud de θ depende sólo de $\mathbb{P}(Y_{\text{obs}} \mid \theta)$, por lo tanto no es necesario modelar el mecanismo de faltantes. □

Ejercicio 5
Insesgadez bajo eliminación de casos (MCAR). Sea $\bar{Y}_{\text{obs}}$ la media muestral basada solo en los casos observados. Demuestre que

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

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

Solución:
Sea $Y_1, \ldots, Y_n$ una muestra i.i.d. con $\mathbb{E}(Y_i) = \mu$ y $\text{Var}(Y_i) = \sigma^2$. Definimos indicadores $R_i \in \{0, 1\}$ que señalan si Y_i está observado ($R_i = 1$) o faltante ($R_i = 0$). Sea $r_{\text{obs}} = \sum_{i=1}^{n} R_i$ el número de observaciones disponibles. La media de datos completos

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

Entonces

$$\mathbb{E}(\bar{Y}_{\text{obs}} \mid R) = \frac{1}{r_{\text{obs}}} \sum_{i=1}^{n} R_i\, \mathbb{E}(Y_i \mid R).$$

Bajo el supuesto de *MCAR*, los Y_i son independientes de R, de modo que $\mathbb{E}(Y_i \mid R) = \mathbb{E}(Y_i) = \mu$. Por tanto,

$$\mathbb{E}(\bar{Y}_{\text{obs}} \mid R) = \frac{1}{r_{\text{obs}}} \sum_{i=1}^{n} R_i \mu = \mu \frac{r_{\text{obs}}}{r_{\text{obs}}} = \mu.$$

Luego

$$\mathbb{E}(\bar{Y}_{\text{obs}}) = \mathbb{E}\big[\mathbb{E}(\bar{Y}_{\text{obs}} \mid R)\big] = \mathbb{E}[\mu] = \mu.$$

Así, $\bar{Y}_{\text{obs}}$ es un estimador insesgado de μ.

La pérdida de la eficiencia se puede ver con la varianza de de $\bar{Y}_{obs}$ pues

$$\text{Var}(\bar{Y}_{\text{obs}} \mid R) = \frac{1}{r_{\text{obs}}^2} \sum_{i=1}^{n} R_i\, \text{Var}(Y_i \mid R) = \frac{\sigma^2}{r_{\text{obs}}}.$$

Por la ley de la varianza total

$$\text{Var}(\bar{Y}_{\text{obs}}) = E[Var(\bar{Y}_{\text{obs}}|R)] + Var[E(\bar{Y}_{\text{obs}}|R)]. = \mathbb{E}\left[\frac{\sigma^2}{r_{\text{obs}}}\right].$$

Comparando con la varianza de la media sin faltantes,

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

se observa que, dado que $r_{\text{obs}} \leq n$, entonces $1/r_{\text{obs}} \geq 1/n$ por lo que

$$\text{Var}(\bar{Y}_{\text{obs}}) = \mathbb{E}\left[\frac{\sigma^2}{r_{\text{obs}}}\right] \geq \frac{\sigma^2}{n} = \text{Var}(\bar{Y})$$

Por lo que bajo MCAR, la eliminación de casos completos produce un estimador insesgado, pero menos eficiente debido a la pérdida de tamaño muestral efectivo y, en consecuencia, mayor varianza.

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

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

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

Solución:

Sean $\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$ el vector de valores observados y valores faltantes, respectivamente, $\mathbf{R}$ el patrón de datos faltantes, y θ y y ψ los vectores de parámetros del modelo de datos y del mecanismo de faltantes, respectivamente.

Por definición del mecanismo MAR, tenemos que

$$P(\mathbf{R}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \theta, \psi) = P(\mathbf{R}|\mathbf{Y}_{obs}, \psi),$$

es decir que la probabilidad de ausencia solo depende de los valores observados y los parámetros del mecanismo de faltantes, no de los valores faltantes ni parámetros del modelo de datos.

Luego, en el modelo de selección podemos ver que

$$p(\mathbf{Y}, \mathbf{R}|\theta, \psi) = \frac{p(\mathbf{Y}, \mathbf{R}, \theta, \psi)}{p(\theta, \psi)} = p(\mathbf{R}|\mathbf{Y}, \theta, \psi)\frac{p(\mathbf{Y}, \theta, \psi)}{p(\theta, \psi)} = P(\mathbf{R}|\mathbf{Y}_{obs}, \psi)p(\mathbf{Y}|\theta).$$

Ahora bien, si L es la función de verosimilitud de θ, tenemos que

$$L(\theta|\mathbf{Y}, \mathbf{R}) = p(\mathbf{Y}, \mathbf{R}|\theta, \psi) = P(\mathbf{R}|\mathbf{Y}_{obs}, \psi)p(\mathbf{Y}|\theta),$$

entonces, si queremos conocer la verosimilitud de θ sin considerar los datos faltantes, solo necesitamos marginalizar la función L dada arriba, de manera que

$$L(\theta|\mathbf{Y}_{obs}, \mathbf{R}) = \int p(\mathbf{R}|\mathbf{Y}_{obs}, \psi)p(\mathbf{Y}, \theta)d\mathbf{Y}_{mis} = p(\mathbf{R}|\mathbf{Y}_{obs}, \psi)p(\mathbf{Y}_{obs}|\theta).$$

Así, como $p(\mathbf{R}|\mathbf{Y}_{obs}, \psi)$ es constante con respecto a θ, tenemos que

$$L(\theta|\mathbf{Y}_{obs}, \mathbf{R}) \propto p(\mathbf{Y}_{obs}|\theta).$$

En el enfoque Bayesiano es necesaria la supocisión de que la función a priori, $\pi(\theta, \psi)$, se factorice como $\pi(\theta, \psi) = \pi(\theta)\pi(\psi)$ para que el mecanismo sea ignorable en la inferencia de θ. Veamos a continuación por qué:

Teniendo la función de verosimilitud $L(\theta|\mathbf{Y}_{obs}, \mathbf{R}) = p(\mathbf{R}|\mathbf{Y}_{obs}, \psi)p(\mathbf{Y}_{obs}|\theta)$ y la distribución a priori $\pi(\theta, \psi) = \pi(\theta)\pi(\psi)$, por Teorema de Bayes, la distribución posterior es tal que

$$\pi(\theta, \psi|\mathbf{Y}_{obs}, \mathbf{R}) \propto p(\mathbf{R}|\mathbf{Y}_{obs}, \psi)p(\mathbf{Y}_{obs}|\theta)\pi(\theta)\pi(\psi).$$

Luego, para obtener la porterior de θ integramos sobre ψ, de modo que

$$\pi(\theta|\mathbf{Y}_{obs}, \mathbf{R}) \propto p(\mathbf{Y}_{obs}, \theta)\pi(\theta) \int p(\mathbf{R}|\mathbf{Y}_{obs}, \psi)\pi(\psi)d\psi,$$

pero notemos que la integral obtenida es constante con respecto a θ, por lo que podemos escribir

$$\pi(\theta|\mathbf{Y}_{obs}, \mathbf{R}) \propto p(\mathbf{Y}_{obs}, \theta)\pi(\theta),$$

donde ya no encontramos que la distribución porterior dependa del mecanismo de faltantes.
Por lo tanto, si la distribución a priori es tal que $\pi(\theta, \psi) = \pi(\theta)\pi(\psi)$, podemos concluir ignorabilidad del mecanismo para la inferencia de θ.

Ejercicio 7
Distancia de Cook como medida global de influencia.

Partiendo de la definición

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

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

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

Discuta la interpretación de esta forma alternativa.

Antes, de poder dar la solución, daré el contexto completo de la distancia de Cook, debido a que en la clase y en las notas solo se mencionó de manera rápida.

Extraído de la página de notas de Modelos Estadísticos, Capítulo 3.2.1:

La distancia de Cook o D de Cook es una estimación comúnmente utilizada de la influencia de un punto de datos al realizar un análisis de regresión por mínimos cuadrados. En un análisis práctico de mínimos cuadrados ordinarios, la distancia de Cook puede usarse de varias maneras: para indicar puntos de datos influyentes que vale la pena verificar por su validez, o para indicar regiones del espacio de diseño donde sería bueno obtener más puntos de datos.

La distancia de Cook se calcula removiendo el $i-$ésimo dato del modelo y recalculando la regresión. Resume que tanto los valores del modelos de regresión cambian cuando la $i-$ésima observación no se considera.

Consideremos ahora que removemos la primera observación. Queremos comparar $\widehat{\boldsymbol{\beta}}_{(1)}$ (los coeficientes estimados de regresión cuando no consideramos la observación 1) contra $\widehat{\boldsymbol{\beta}}$. Particionamos la información a eliminar:

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

Se quiere $\widehat{\boldsymbol{\beta}}_{(1)}$, obtenido del ajuste de

$$\boldsymbol{Y}_{(1)} = \boldsymbol{X}_1\boldsymbol{\beta} + \boldsymbol{\epsilon}_{(1)}$$

entonces $\widehat{\boldsymbol{\beta}}_{(1)} = (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top \boldsymbol{Y}_{(1)}$.

La ecuación anterior implica que si queremos calcular el impacto de cada dato, se requerirían ajustar $\boldsymbol{n}+1$ regresiones. En realidad es suficiente con hacer una sola.

Note que:

$$\boldsymbol{X}^\top \boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{X}_1^\top] \begin{bmatrix} \boldsymbol{x}_1^\top \\ \boldsymbol{X}_1 \end{bmatrix} = \boldsymbol{x}_1\boldsymbol{x}_1^\top + \boldsymbol{X}_1^\top \boldsymbol{X}_1,$$

entonces $\boldsymbol{X}_1^\top \boldsymbol{X}_1 = \boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{x}_1\boldsymbol{x}_1^\top$, y por la Identidad de Woodbury, se tiene:

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

Similarmente,

$$\boldsymbol{X}_1^\top \boldsymbol{Y}_{(1)} = \boldsymbol{X}^\top \boldsymbol{Y} - \boldsymbol{x}_1\boldsymbol{Y}_1,$$

entonces:

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

En general, si se elimina la observación i, se obtiene:

$$\hat{\beta}_{(i)} = \hat{\beta} - (X^T X)^{-1}x_i\frac{e_i}{1-h_{ii}},$$

esto es, los impactos individuales de cada observación pueden obtenerse del ajuste de regresión original.

—

Bajo este contexto, se puede definir **la distancia de Cook**:

El indicador de influencia D de Cook cuantifica el grado de disparidad entre $\hat{\beta}$, estimando una distancia estadística entre ellos:

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T(X^T X)(\hat{\beta}_{(i)} - \hat{\beta})}{pMS_{Res}} = \frac{(\hat{Y}_{(i)} - \hat{Y})^T(\hat{Y}_{(i)} - \hat{Y})}{ps^2}$$

o bien, como se definió en el enunciado:

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

Solución:

Ya se puede demostrar lo requerido. Defina:

$$G := \hat{\beta}_{(i)} - \hat{\beta} = \hat{\beta} - (X^T X)^{-1}x_i\frac{e_i}{1-h_{ii}} - \hat{\beta} = -(X^T X)^{-1}x_i\frac{e_i}{1-h_{ii}}.$$

Sustituyendo,

$$\begin{aligned}
D_i &= \frac{1}{ps^2}G^T(X^T X)G, \\
&= \frac{1}{ps^2}\left(-(X^T X)^{-1}x_i\frac{e_i}{1-h_{ii}}\right)^T(X^T X)\left(-(X^T X)^{-1}x_i\frac{e_i}{1-h_{ii}}\right), \\
&= \frac{1}{ps^2}\left(\frac{e_i}{1-h_{ii}}\right)^2 x_i^T(X^T X)^{-1}(X^T X)(X^T X)^{-1}x_i, \\
&= \frac{1}{ps^2}\left(\frac{e_i}{1-h_{ii}}\right)^2 x_i^T(X^T X)^{-1}x_i, \\
&= \frac{1}{ps^2}\left(\frac{e_i}{1-h_{ii}}\right)^2 h_{ii}, \\
&= \left(\frac{e_i}{s\sqrt{(1-h_{ii})}}\right)^2\left(\frac{h_{ii}}{p(1-h_{ii})}\right).
\end{aligned}$$

Esta forma alternativa, sirve para determinar si una observación tendrá una D de Cook grande si su residual es grande (esto es, si $\hat{Y}_i$ esta lejos de $\hat{Y}$) y el correspondiente h_{ii} es cercano a 1 (esto es si la x_i esta alejada del centroide de las x's).

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

$$x_i^* = \frac{x_i - \text{mín}(x)}{\text{máx}(x) - \text{mín}(x)}.$$

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

Solución:
Sea $y_i = ax_i + b$ con $a > 0$ y definamos

$$y_i^* = \frac{y_i - \text{mín}(y)}{\text{máx}(y) - \text{mín}(y)}.$$

Notemos que

$$\text{mín}(y) = a\,\text{mín}(x) + b, \qquad \text{máx}(y) = a\,\text{máx}(x) + b,$$

Entonces

$$\begin{aligned} y_i^* &= \frac{ax_i + b - (a\,\text{mín}(x) + b)}{a\,\text{máx}(x) + b - (a\,\text{mín}(x) + b)} \\ &= \frac{a\big(x_i - \text{mín}(x)\big)}{a\big(\text{máx}(x) - \text{mín}(x)\big)} \\ &= \frac{x_i - \text{mín}(x)}{\text{máx}(x) - \text{mín}(x)} \\ &= x_i^*. \end{aligned}$$

Por lo tanto $y_i^* = x_i^*$.

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

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

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.

Solución:

Sea X una v.a. con distribución Pareto de parámetros α y x_m, con función de densidad

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

Definamos a la variable Y como $Y = \log(X)$. Tenemos entonces que $X = e^Y$.

a) Sea g la función de densidad de la v.a. Y. Entonces g toma valores en el conjunto $[\log(x_m), \infty)$. Ahora bien, por el Teorema de Cambio de Variable, la función g está dada por

$$g(y) = f(e^y)\left|\frac{dX}{dY}\right| = \frac{\alpha x_m^\alpha}{e^{y(\alpha+1)}}e^y = \frac{\alpha x_m^\alpha}{e^{y\alpha}}, \quad y \geq \log(x_m).$$

Por tanto, la función de densidad de Y es $g(y) = \alpha x_m^\alpha e^{-\alpha y} 1_{\{y \geq \log(x_m)\}}$.

Luego, G la función de distribución de Y y sea $y \geq \log(x_m)$. Entonces tenemos que,

$$G(y) = \int_{\log(x_m)}^{y} \alpha x_m^\alpha e^{-\alpha s} ds = x_m^\alpha \int_{\log(x_m)}^{y} \alpha e^{-\alpha s} ds,$$

pero notemos que nuestro integrando corresponde precisamente a la función de densidad de una v.a. exponencial de parámetro α, por lo que

$$\int_{\log(x_m)}^{y} \alpha e^{-\alpha s} ds = 1 - e^{-\alpha y} - [1 - e^{-\alpha \log(x_m)}] = e^{\log(x_m^{-\alpha})} - e^{\alpha y} = x_m^{-\alpha} - e^{-\alpha y}.$$

Por lo tanto, la función de distribución de Y es $G(y) = (1 - x_m^\alpha e^{-\alpha y}) 1_{\{y \geq \log(x_m)\}}$.

b) Escribamos la funciones f y g como

$$f(x) = \alpha x_m^\alpha x^{-(\alpha+1)} \quad \text{y} \quad g(y) = \alpha x_m^\alpha e^{-\alpha y}, \quad \text{para} \quad x \geq x_m,\ y \geq \log(x_m).$$

Es fácil notar ahora que la función de Y decae mucho más rápido que la de X, pues lo hace a una escala exponencial mientras la segunda lo hace a una escala polinómica. Así, al pasar de X a Y dejamos las colas pesadas de la Pareto y conseguimos colas más ligeras.

c) Al aplicar la transformación logarítimica a una variable tomamos los valores (positivos) de esta y, si son pequeños, la transformación no los cambia en una medida relevante, pero, si son valores grandes, el logaritmo los reduce considerablemente, de manera que la cola derecha de la distibución que estemos transformando parece ser acortada, pues decrece mucho más rápido que la original, consiguiendo una cola derecha mucho menos pesada y haciendo ver la distribución más centrada al comprimir los valores grandes, produciendo así una distribución un poco más simétrica.

Ejercicio 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?

Solución:

a)

La media muestral es:

$$\bar{x} = \frac{1}{5}\sum_{i=1}^{5} x_i = \frac{1+2+3+4+M}{5} = \frac{10+M}{5} = 2 + \frac{M}{5}.$$

La varianza muestral es:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{4}\sum_{i=1}^{5}(x_i - \bar{x})^2.$$

Calculamos las desviaciones al cuadrado:

$$
\begin{aligned}
(1-\bar{x})^2 &= \left(1-2-\frac{M}{5}\right)^2 = \left(-1-\frac{M}{5}\right)^2 = \left(1+\frac{M}{5}\right)^2,\\
(2-\bar{x})^2 &= \left(2-2-\frac{M}{5}\right)^2 = \left(-\frac{M}{5}\right)^2 = \frac{M^2}{25},\\
(3-\bar{x})^2 &= \left(3-2-\frac{M}{5}\right)^2 = \left(1-\frac{M}{5}\right)^2,\\
(4-\bar{x})^2 &= \left(4-2-\frac{M}{5}\right)^2 = \left(2-\frac{M}{5}\right)^2,\\
(M-\bar{x})^2 &= \left(M-2-\frac{M}{5}\right)^2 = \left(\frac{4M}{5}-2\right)^2.
\end{aligned}
$$

Sumando:

$$
\begin{aligned}
s^2 &= \sum_{i=1}^{5}(x_i-\bar{x})^2,\\
&= \left(1+\frac{M}{5}\right)^2 + \frac{M^2}{25} + \left(1-\frac{M}{5}\right)^2 + \left(2-\frac{M}{5}\right)^2 + \left(\frac{4M}{5}-2\right)^2,\\
&= 2+\frac{2M^2}{25}+\frac{M^2}{25}+4-\frac{4M}{5}+\frac{M^2}{25}+\frac{16M^2}{25}+4-\frac{16M}{5},\\
&= 10+\frac{20M^2}{25}-\frac{20M}{5}.
\end{aligned}
$$

Por lo tanto:

$$
\begin{aligned}
s^2 &= 10+\frac{4M^2}{5}-4M.\\
s &= \sqrt{10+\frac{4M^2}{5}-4M}.
\end{aligned}
$$

b)

Los datos ordenados son: $\{1,2,3,4,M\}$.

La mediana es el valor central, $Q_{.50} = m = 3$.

Para el rango intercuartílico:

- Primer cuartil, es el cuantil $Q_{0.25}$, con esto el primer cuartil es 2.
- Tercer cuartil, es el cuantil $Q_{0.75}$, con esto el tercer cuartil es 4.

$$RIQ = 4-2 = 2.$$

c)

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

Observe que:

- Media $\bar{x}$: Se distorsiona, por el factor $\frac{M}{5}$, cuando $M \to \infty$.
- Desviación estándar s: Se distorsiona, ya que $\frac{4M^2}{\sqrt{5}}$ es superior a $4M$ cuando $M \to \infty$.
- Mediana m: Permanece estable en $m = 3$, independiente de M.
- Rango intercuartílico RIQ: Permanece estable en $RIQ = 2$, independiente de M.

La mediana y el rango intercuartílico son medidas robustas que no son afectadas por outliers.

Ejercicio 11
Propiedades de la transformación Box–Cox Sea $y^{(\lambda)}$ la transformación de Box–Cox definida como:

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

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).

Solución:

a) Para $\lambda \neq 0$, la transformación de Box-Cox es:

$$y^{(\lambda)} = \frac{x^\lambda - 1}{\lambda}.$$

Notamos que tanto numerador como denominador tienden a 0 cuando $\lambda \to 0$, por lo que aplicamos la regla de L'Hôpital:

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

Por lo tanto, $\lim_{\lambda \to 0} y^{(\lambda)} = \log(x)$.

b) Sea el conjunto de datos $x = \{1,\ 5,\ 500,\ 2000\}$

Caso 1: $\lambda = 1$

$$y^{(1)} = \frac{x^1 - 1}{1} = x - 1,$$

entonces

$$y^{(1)} = \{0,\ 4,\ 499,\ 1999\}$$

Caso 2: $\lambda = 0$

$$y^{(0)} = \log(x),$$

entonces

$$y^{(0)} = \{\log(1),\ \log(5),\ \log(500),\ \log(2000)\} \approx \{0,\ 1.609,\ 6.215,\ 7.601\}$$

Vemos que con $\lambda = 1$ los datos siguen muy dispersos (de 0 hasta 1999), mientras que con $\lambda = 0$ la dispersión se reduce considerablemente (de 0 hasta 7.6).

Ejercicio 12
Propiedades del histograma

Sea $x_1, \ldots, x_n$ 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:

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

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.

Solución:

Sea $x_1, x_2, ..., x_n$ una muestra de v.a.i.i.d. continuas con función de densidad $f(x)$.
Sea $h > 0$ y $\hat{f}_h(x)$ el estimador del histograma con k intervalos de ancho h, dado por

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

a) Notemos que como $n \in \mathbb{N}$, $h > 0$ y la función indicadora es no negativa, entonces $\hat{f}_h(x)$ es una suma de términos no negativos.

Por lo tanto $\hat{f}_h(x) \geq 0$ para todo x.

b) Sea $x \in I_j$. Como el histograma está conformado por k intervalos,

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

Luego, como cada intervalo es de longitud h, tenemos que

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

Ahora bien, notemos que, para cada j, la función $\sum_{i=1}^{n} 1_{\{x_i \in I_j\}}$ está contando cúantas observaciones hay en cada intervalo I_j, por lo que al contar en los k intervalos vamos a tener el total de nuestras observaciones, que son n. Así

$$\sum_{j=1}^{k} \sum_{i=1}^{n} 1_{\{x_i \in I_j\}} = n.$$

Por lo tanto,

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

c) La elección del ancho de banda del histograma, o bien, el valor de h, afecta significativamente al estimador de histograma en términos del sesgo y la varianza, pues, si escogemos un valor de h muy chico tendríamos más ruido y problemas con la varianza, o sea, una varianza grande. Mientras que escoger un valor de h muy grande nos lleva a problemas de sesgo. Por lo cual necesitamos una consistencia en el valor de h, es decir, escoger el óptimo que equilibre el sesgo y la varianza del estimador. Para ello debemos buscar el valor tal que $nh \to \infty$ cuando $n \to \infty$ pero al mismo tiempo $h \to 0$, de esta manera podemos equilibrar los valores de la varianza y el sesgo, respectivamente.

Ejercicio 13
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
$$E\{\hat{f}_h(x)\} - f(x) = \frac{h^2}{2}\mu_2(K)f''(x) + o(h^2).$$

Solución:

Defina el estimador Kernel,
$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$
con las condiciones ya mencionadas. Hay que demostrar que cumple:

- **Normalización**:
$$\begin{aligned}
\int_{-\infty}^{\infty} \hat{f}_h(x)dx &= \int_{-\infty}^{\infty} \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) dx, \\
&= \frac{1}{nh}\sum_{i=1}^{n} \int_{-\infty}^{\infty} K\left(\frac{x - x_i}{h}\right) dx, \text{ pues la integral puede entrar en una suma finita.} \\
&= \frac{1}{nh}\sum_{i=1}^{n} \int_{-\infty}^{\infty} K(u)\cdot h du \quad \text{(haciendo el cambio de variable} u = \frac{x - x_i}{h} \to h\cdot du = dx). \\
&= \frac{1}{nh}\sum_{i=1}^{n} h\int_{-\infty}^{\infty} K(u)du, \\
&= \frac{1}{n}\sum_{i=1}^{n} 1 = 1.
\end{aligned}$$

- **No negatividad**:

  Suponga que $K(u) \geq 0$ para todo u, entonces cada término en la suma es no negativo:
$$K\left(\frac{x - x_i}{h}\right) \geq 0 \quad \text{para todo } i = 1, \ldots, n.$$
  Además, $n > 0$ y $h > 0$, por lo que:
$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \geq 0.$$

- **Sesgo puntual**:

  El *hint* es usar la expansión de Taylor de f alrededor de x, donde f es la densidad de x_i, primero considere la esperanza del estimador KDE:
$$\mathbb{E}\{\hat{f}_h(x)\} = \frac{1}{h}\mathbb{E}\left[K\left(\frac{x - X}{h}\right)\right] = \frac{1}{h}\int_{-\infty}^{\infty} K\left(\frac{x - y}{h}\right) f(y)dy.$$

Considere el cambio de variable $u = \frac{x-y}{h}$, entonces $y = x - hu$, luego $dy = -h \cdot du$, sustituyendo:

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

Hay que utilizar el *hint,* desarrollando la serie de Taylor de $f(x-hu)$ alrededor de x,:

$$\begin{aligned} f(x-hu) &= \sum_{n=0}^{\infty} \frac{f^{(n)}(x)}{n!}(x-hu-x)^n, \\ &= \sum_{n=0}^{\infty} \frac{f^{(n)}(x)}{n!}(-hu)^n, \\ &= f(x) - huf'(x) + \frac{h^2u^2}{2}f''(x) + o(h^2). \end{aligned}$$

Sustituyendo en la esperanza del estimador:

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

Por hipótesis se tiene que:

$$\int_{-\infty}^{\infty} K(u)du = 1, \quad \int_{-\infty}^{\infty} uK(u)du = 0, \quad \int_{-\infty}^{\infty} u^2K(u)du = \mu_2(K),$$

Se obtiene:

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

Por lo tanto, el sesgo es:

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