Tarea I

Introducción a la Ciencia de Datos

Alfredo Bistrain, Debany Hernandez, Oswaldo Bueno

13 de septiembre de 2025

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*.

Demostración. Sea $H = X(X^\top X)^{-1}X^\top$, decimos que la matriz H es idempotente si se cumple que $H^2 = HH = H$. Usando la definición de H

$$\begin{aligned} H^2 &= X(X^\top X)^{-1}X^\top X(X^\top X)^{-1}X^\top \\ &= X(X^\top X)^{-1}X^\top \\ &= H \end{aligned}$$

Por lo que H es idempotente. Ahora para ver que H es simétrica, supongamos que X es de dimensión $n \times m$. Esto implica que $X^\top X$ es de dimensión $m \times m$ y a su vez $(X^\top X)^{-1}$ tendrá esta dimensión; $X(X^\top X)^{-1}$ es de dimensión $n \times m$ por lo que $X(X^\top X)^{-1}X^\top$ será de dimensión $n \times n$. Es de aquí que podemos decir que H es simétrica $n \times n$ donde n es el número de filas de X. □

Interpretación

- Al ser simétrica e idempotente, H es la proyección ortogonal sobre el subespacio columna $\mathcal{C}(X)$. En regresión lineal, $\hat{y} = Hy$ es la proyección ortogonal de y sobre $\mathcal{C}(X)$; los residuales $e = (I - H)y$ son la proyección sobre $\mathcal{C}(X)^{\perp}$ y satisfacen $X^{\top}e = 0$.
- La idempotencia $H^2 = H$ indica que proyectar dos veces no cambia nada. La simetría garantiza que la proyección es ortogonal.

Los leverages son las diagonales de H: $h_{ii} = e_i^{\top} H e_i$.

1. Para toda proyección ortogonal P, $0 \leq e_i^{\top} P e_i \leq 1$. Por ello $0 \leq h_{ii} \leq 1$. Geométricamente, h_{ii} mide cuánto contribuye la observación i a su propio valor ajustado $\hat{y}_i$ vía la proyección.

2. Dado que
$$\operatorname{tr}(H) = \operatorname{tr}((X^{\top}X)^{-1}X^{\top}X) = \operatorname{tr}(I_p) = p.$$
Así, $\sum_i h_{ii} = p$ y el promedio de leverage es p/n. Esto da una regla práctica: valores muy por encima de p/n son altos leverages.

3. La relación
$$e_{(i),i} = \frac{e_i}{1 - h_{ii}}$$
depende de que H sea proyección ortogonal. Un h_{ii} grande amplifica el residual al dejar fuera la observación, evidenciando su potencial de influencia.

4. En notación de filas $x_i^{\top}$ de X,
$$h_{ii} = x_i^{\top}(X^{\top}X)^{-1}x_i,$$
que es (hasta factores) una distancia de Mahalanobis al centro del diseño. La simetría e idempotencia que hacen de H una proyección ortogonal legitiman esta lectura geométrica.

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.

Demostración. Sea X una matriz de tamaño $n \times p$, y sea H la matriz hat definida como,

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

Por definición, sabemos que la suma de los elementos de la diagonal de H es llamada la traza,

$$\sum_{i=1}^{n} h_{ii} = \operatorname{tr}(H).$$

Además, utilizando la propiedad cíclica de la traza, $\operatorname{tr}(ABC) = \operatorname{tr}(BCA) = \operatorname{tr}(CAB)$, esto siempre que los productos estén bien definidos, tenemos que,

$$\begin{aligned}\operatorname{tr}(H) &= \operatorname{tr}\big(X(X^\top X)^{-1}X^\top\big) \\ &= \operatorname{tr}\big((X^\top X)^{-1}X^\top X\big)\end{aligned}$$

Pero sabemos que $\big((X^\top X)^{-1}X^\top X\big) = I_p$ la cual es la matriz identidad de tamaño $p \times p$, de lo que se sigue,

$$\operatorname{tr}(H) = \operatorname{tr}(I_p) = p$$

Por lo tanto,

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

□

Interpretación.

El resultado anterior nos indica que la suma total de los leverages refleja el número de parámetros estimados en el modelo. Como lo mencionamos en el ejercicio anterior, cada observación tiene, en promedio, un leverage de,

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

Esto implica que, al incrementar p, el leverage promedio aumenta, mientras que al incrementar n, disminuye.

En términos prácticos, esta propiedad se interpreta como el *número efectivo de parámetros*, es decir, la cantidad de grados de libertad que el modelo consume para ajustarse a los datos.

Respecto al sobreajuste, a medida que el número de parámetros p crece en relación con el número de observaciones n, los leverages tienden a ser mayores, otorgando a algunas observaciones un peso excesivo en la estimación. Esto incrementa la varianza del modelo y su sensibilidad a valores atípicos.

Como lo mencionamos en clase, una regla práctica común es considerar posibles observaciones influyentes aquellas cuyo leverage supere $2p/n$ o incluso $3p/n$. Además, esta propiedad explica por qué la estimación de la varianza residual utiliza $n - p$ grados de libertad.

Ejercicio 3.

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

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

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*.

Demostración. Bajo el modelo lineal clásico con errores normales tenemos que $\hat{Y} = HY$, donde $H = X(X^T X)^{-1} X^T$ es la hat-matrix. Además, sabemos que los residuales están dados por $e = Y - \hat{Y}$, de manera que

$$e = Y - HY = (I - H)Y.$$

Dado que H es la matriz de proyección sobre el espacio columna de X, se tiene que $I - H$ es la matriz de proyección sobre el espacio ortogonal al espacio columna de X. Haciendo uso

de que $Y = X\beta + \varepsilon$, tenemos

$$e = (I - H)(X\beta + \varepsilon) = (I - H)X\beta + (I - H)\epsilon = (I - H)\varepsilon, \tag{1}$$

donde hemos usado que $(I - H)X\beta = 0$, pues $X\beta$ es un elemento del espacio columna de X.

De (1), y recordando que $\varepsilon \sim N(0, \sigma^2 I)$, se sigue

$$e \sim N(0, (I - H)\sigma^2 I(I - H)^T) = N(0, \sigma^2(I - H)),$$

puesto que $I - H$ es una matriz simétrica e idempotente al ser una matriz de proyección. Dicho esto resulta que $e_i \sim N(0, \sigma^2(1 - h_{ii}))$, donde h_{ii} es el i-ésimo elemento diagonal de H.

Denotaremos $\hat{\sigma}^2 = \frac{e^T e}{n-p}$, donde p es el número de parámetros en el modelo. Se cumple que $\hat{\sigma}^2$ es un estimador insesgado de σ^2, y además $\frac{e^T e}{\sigma}$ sigue una distribución χ^2_{n-p}, lo cual puede verse fácilmente pues

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

Ahora, dado que $Y \sim N(X\beta, \sigma^2 I)$, entonces $Y\sigma \sim N(X\beta/\sigma, I)$. Dado que $I - H$ es una matriz idempotente, se tiene que su rango es igual a su traza, de manera que

$$Rango(I - H) = tr(I) - tr(H) = n - rango(H) = n - p,$$

donde hemos usado que H es de rango p, al ser X de rango p. Por resultado de formas cuadráticas visto en Métodos Estádisticos, se tiene que la forma cuadrática anterior sigue una distribución $\chi^2_{n-p}(\lambda)$, donde λ es el parámetro de no centralidad, el cual está dado mediante

$$\lambda = \frac{(X\beta/\sigma)^T(I - H)(X\beta/\sigma)}{2} = 0,$$

pues $(I - H)X\beta = 0$, como habíamos mencionado anteriormente. Se sigue que $\frac{e^T e}{\sigma^2} \sim \chi^2_{n-p}$.

Consideremos ahora el i-ésimo residuo estandarizado,

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

Sustituyendo $\hat{\sigma}^2 = \frac{e^T e}{n-p}$, notemos que podemos reescribir lo anterior como

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

Como $e_i \sim N(0, \sigma^2(1-h_{ii}))$, entonces $e_i/\sqrt{\sigma^2(1-h_{ii})} \sim N(0,1)$. Además, como $\frac{e^Te}{\sigma^2} \sim \chi^2_{n-p}$, tenemos que lo anterior es el cociente de una normal estándar dividida por la raíz cuadrada de una ji-cuadrada divida por sus grados de libertad. Si tanto el numerador como el denominador fueran independientes, el cociente seguiría exactamente una distribución t de student de $n-p$ grados de libertad. Sin embargo, e_i y e^Te no son independientes. Debido a esta dependencia podemos únicamente decir que el residuo estandarizado se comporta de manera similar a una distribución t de student con $n-p$ grados de libertad, pero no exactamente (para ello se definen los residuos studentizados, en los cuales se elimina el factor e_i^2 de e^Te).

Los residuos estandarizados son útiles para detectar outliers, ya que un outlier tiene un leverage alto (es decir, un valor grande de h_{ii} y cercano a 1), lo cual tiene como consecuencia que $\sqrt{1-h_{ii}}$ se aproxime a cero. De esta manera, un leverage alto implica que el cociente $1/\sqrt{1-h_{ii}}$ se vuelve grande, y por lo tanto el residuo estandarizado en sí se vuelve grande. □

Ejercicio 4.

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

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

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

Demostración. Sea $Y = (Y_{\text{obs}}, Y_{\text{mis}})$ el vector (o matriz) de datos completos, R el indicador de faltantes con la misma dimensión que Y (toma valor 1 si el dato se observa y 0 si falta), θ el parámetro del modelo de datos y ψ el parámetro del mecanismo de faltantes. Supondremos *distinción de parámetros*, esto es, el modelo de datos no depende de ψ: $p(Y \mid \theta, \psi) = p(Y \mid \theta)$.

Definición (MCAR). Decimos que el mecanismo es *Missing Completely At Random* (MCAR) si

$$p(R = r \mid Y = y, \theta, \psi) = p(R = r \mid \psi) \quad \text{para todo } (y, r),$$

es decir, el patrón de faltantes es independiente de los valores (observados o no) de Y.

Factorización. Aplicando la regla de la cadena y usando distinción de parámetros,

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

Bajo MCAR, $p(R \mid Y, \theta, \psi) = p(R \mid \psi)$. Por lo tanto,

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

que es la factorización dada. La verosimilitud con datos observados (y_{obs}, r) es, integrando/-sumando sobre los faltantes y_{mis},

$$L(\theta, \psi \mid y_{\text{obs}}, r) \;\propto\; \int p(y_{\text{obs}}, y_{\text{mis}}, r \mid \theta, \psi)\, dy_{\text{mis}} \;=\; \int p(y_{\text{obs}}, y_{\text{mis}} \mid \theta)\, p(r \mid \psi)\, dy_{\text{mis}}.$$

El factor $p(r \mid \psi)$ no depende de θ ni de y_{mis}, así que

$$L(\theta, \psi \mid y_{\text{obs}}, r) \;\propto\; p(r \mid \psi) \underbrace{\int p(y_{\text{obs}}, y_{\text{mis}} \mid \theta)\, dy_{\text{mis}}}_{= p(y_{\text{obs}} \mid \theta)}.$$

En consecuencia, para inferir θ (por máxima verosimilitud), basta maximizar $p(y_{\text{obs}} \mid \theta)$; el término $p(r \mid \psi)$ es constante en θ y puede ignorarse. En el enfoque Bayesiano, si además la previa factoriza como $\pi(\theta, \psi) = \pi(\theta)\pi(\psi)$, entonces

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

independiente de $p(r \mid \psi)$. Por lo tanto, bajo MCAR (y distinción de parámetros), el mecanismo de faltantes es *ignorable* para la inferencia sobre θ. □

Ejercicio 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.

Demostración. Consideremos una muestra aleatoria $Y_1, \dots Y_n$ i.i.d con $\mathbb{E}[Y_i] = \mu$ y $\mathbb{V}[Y_i] = \sigma^2 < \infty$. Para cada unidad i, definimos el indicador de observación,

$$R_i = \begin{cases} 1, & \text{si } Y_i \text{ está observado,} \\ 0, & \text{si } Y_i \text{ está faltante.} \end{cases}$$

Sea $n_{\text{obs}} = \sum_{i=1}^{n} R_i$ el número de observaciones efectivas (suponemos $n_{\text{obs}} > 0$). La media por casos completos se define como,

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

Bajo MCAR, el mecanismo de valores faltante satisface que la distribución de $R = (R_1, \ldots, R_n)$ es independiente de los valores de Y, esto es,

$$\mathbb{P}[R \mid Y] = \mathbb{P}[R] \quad \text{o equivalentemente} \quad Y \perp R.$$

Ahora, para demostrar que es insesgado, utilizamos la ley de la esperanza condicionada,

$$\mathbb{E}[\bar{Y}_{\text{obs}}] = \mathbb{E}\big[\, \mathbb{E}[\bar{Y}_{\text{obs}} \mid R] \,\big],$$

entonces,

$$\mathbb{E}[\bar{Y}_{\text{obs}} \mid R] = \mathbb{E}\left[\frac{1}{n_{\text{obs}}} \sum_{i=1}^{n} R_i Y_i \;\middle|\; R \right] = \frac{1}{n_{\text{obs}}} \sum_{i=1}^{n} R_i \, \mathbb{E}[Y_i \mid R].$$

Bajo MCAR, Y_i es independiente de R, por lo que $\mathbb{E}[Y_i \mid R] = \mathbb{E}[Y_i] = \mu$. Entonces,

$$\mathbb{E}[\bar{Y}_{\text{obs}} \mid R] = \frac{1}{n_{\text{obs}}} \sum_{i=1}^{n} R_i \, \mu = \frac{n_{\text{obs}} \, \mu}{n_{\text{obs}}} = \mu.$$

Finalmente, de lo anterior, tenemos que,

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

Por lo tanto, Y_{obs} es un estimador insesgado de μ bajo MCAR. □

¿Por qué el estimador pierde eficiencia?

Aunque $\bar{Y}_{\text{obs}}$ es insesgado bajo MCAR, su varianza es, en general, mayor que la varianza de la media muestral completa $\bar{Y} = (1/n) \sum_{i=1}^{n} Y_i$ calculada con todos los datos observados. Estos son los argumentos que discutimos,

1. **Expresión alternativa de la varianza (descomposición condicional).** Por la fórmula de la varianza total:

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

Usando la insesgadez condicional ($\mathbb{E}[\bar{Y}_{\text{obs}} \mid R] = \mu$) la segunda componente se anula y queda

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

Para una realización de R con $n_{\text{obs}} > 0$,

$$\text{Var}[\bar{Y}_{\text{obs}} \mid R] = \text{Var}\left[\frac{1}{n_{\text{obs}}}\sum_{i=1}^{n} R_i Y_i \;\middle|\; R\right] = \frac{\sigma^2}{n_{\text{obs}}},$$

si asumimos independencia entre observaciones y homocedasticidad. Por lo tanto

$$\text{Var}[\bar{Y}_{\text{obs}}] = \mathbb{E}\left[\frac{\sigma^2}{n_{\text{obs}}}\right].$$

2. **Comparación con la media completa.** Si todos los datos estuvieran observados, la varianza de la media muestral sería $\text{Var}[\bar{Y}] = \sigma^2/n$. Pero bajo faltado MCAR con probabilidad de observación $\pi = \mathbb{E}[R_i]$, el número medio de observaciones es $\mathbb{E}[n_{\text{obs}}] = \pi n$. Aproximadamente (para un n grande y si n_{obs} no varía mucho),

$$\text{Var}[\bar{Y}_{\text{obs}}] \approx \frac{\sigma^2}{\pi n},$$

que es mayor que σ^2/n siempre que $\pi < 1$. Es decir, la pérdida de observaciones incrementa la varianza del estimador en un factor aproximado $1/\pi$.

3. **Intuición:** eliminar casos equivale a reducir el tamaño muestral efectivo. Menos datos implican estimadores con mayor incertidumbre (mayor varianza). Además, n_{obs} es aleatorio, por lo que la expectativa de $1/n_{\text{obs}}$ es mayor que $1/\mathbb{E}[n_{\text{obs}}]$ en virtud de la convexidad de la función $x \mapsto 1/x$ que aumenta aún más la varianza esperada.

Ejercicio 6.

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

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

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

Bajo el mecanismo MAR se supone que la probabilidad de ausencia depende de los valores observados, pero no de los valores faltantes. Es decir, condicionando en lo que sí vemos ($\mathbf{Y}_{obs}$), el mecanismo de faltantes deja de depender de los datos que no vemos ($\mathbf{Y}_{mis}$). Formalmente, esto se expresa como

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

donde $\mathbf{R}$ es el patrón de faltantes, θ son los parámetros del modelo de los datos y ψ es el mecanismo de faltantes.

Ahora, dado que $\mathbf{Y}$ no depende de ψ, ni $\mathbf{R}$ de θ, podemos escribir

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

Desarrollando la verosimilitud de datos observados para θ, resulta

$$\begin{aligned} L(\theta; \mathbf{Y}_{obs}, \mathbf{R}) = p(\mathbf{Y}_{obs}, \mathbf{R} \mid \theta) &= \int p(\mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \mathbf{R} \mid \theta) d\mathbf{Y}_{mis} \\ &= \int p(\mathbf{R} \mid \mathbf{Y}_{obs}, \psi) p(\mathbf{Y} \mid \theta) d\mathbf{Y}_{mis} \\ &= p(\mathbf{R} \mid \mathbf{Y}_{obs}, \psi) \int p(\mathbf{Y}_{obs}, \mathbf{Y}_{mis} \mid \theta) d\mathbf{Y}_{mis} \\ &= p(\mathbf{R} \mid \mathbf{Y}_{obs}, \psi) p(\mathbf{Y}_{obs} \mid \theta). \end{aligned}$$

Observemos que $p(\mathbf{R} \mid \mathbf{Y}_{obs}, \psi)$ es un factor constante (con respecto a θ), por lo que la maximización de la verosimilitud de datos observados no depende de $p(\mathbf{R} \mid \mathbf{Y}_{obs}, \psi)$. De esta manera resulta que

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

Para concluir ignorabilidad desde el enfoque bayesiano es necesario suponer adicionalmente que la prior se factoriza como $\pi(\theta, \psi) = \pi(\theta)\pi(\psi)$, esto pues la posterior de θ, ψ es tal que

$$\begin{aligned}\pi(\theta, \psi \mid \mathbf{Y}_{obs}, \mathbf{R}) &\propto p(\mathbf{Y}_{obs}, \mathbf{R} \mid \theta, \psi)\pi(\theta, \psi) \\ &= p(\mathbf{R} \mid \mathbf{Y}_{obs}, \psi)p(\mathbf{Y}_{obs} \mid \theta)\pi(\theta)\pi(\psi).\end{aligned}$$

Luego, marginalizando sobre ψ resulta,

$$\begin{aligned}\pi(\theta \mid \mathbf{Y}_{obs}, \mathbf{R}) &\propto p(\mathbf{Y}_{obs} \mid \theta)\pi(\theta) \int p(\mathbf{R} \mid \mathbf{Y}_{obs}, \psi)\pi(\psi)d\psi \\ &\propto p(\mathbf{Y}_{obs} \mid \theta)\pi(\theta),\end{aligned}$$

donde hemos usado que la integral es constante con respecto a θ. Así, si la distribución a priori se factoriza como $\pi(\theta, \psi) = \pi(\theta)\pi(\psi)$, entonces la posterior de θ no depende del mecanismo de faltantes, y por lo tanto el mecanismo de faltantes es ignorable para la inferencia de θ.

Ejercicio 7.

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

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

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.

Demostración. Nótese que D_i puede ser matricialmente escrito como

$$D_i = \frac{\left(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}}\right)^\top \left(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}}\right)}{p\,\hat{\sigma}^2},$$

donde $\widehat{\boldsymbol{Y}}_{(i)} = X\widehat{B}_{(i)}$ y $\widehat{\boldsymbol{Y}} = X\widehat{B}$. De esto último se sigue la expresión

$$\begin{aligned} D_i &= \frac{\left(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}}\right)^\top \left(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}}\right)}{p\,\hat{\sigma}^2} \\ &= \frac{(X\widehat{\boldsymbol{\beta}}_{(i)} - X\widehat{\boldsymbol{\beta}})^\top \, (X\widehat{\boldsymbol{\beta}}_{(i)} - X\widehat{\boldsymbol{\beta}})}{p\,\hat{\sigma}^2} \\ &= \frac{(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})^\top (X^\top X) \, (\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})}{p\,\hat{\sigma}^2}. \end{aligned}$$

Un resultado importante de recordar es que

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{(i)} =& \ (X_i^\top X_i)^{-1} X_i^\top \boldsymbol{Y}_{(i)} \\
=& \ (X^\top X)^{-1} X^\top \boldsymbol{Y} + (X^\top X)^{-1} \boldsymbol{x}_i \left[1 - \boldsymbol{x}_i^\top (X^\top X)^{-1} \boldsymbol{x}_i\right]^{-1} \boldsymbol{x}_i^\top (X^\top X)^{-1} X^\top \boldsymbol{Y} \\
& - (X^\top X)^{-1} \boldsymbol{x}_i \boldsymbol{Y}_i - (X^\top X)^{-1} \boldsymbol{x}_i \left[1 - \boldsymbol{x}_i^\top (X^\top X)^{-1} \boldsymbol{x}_i\right]^{-1} \boldsymbol{x}_i^\top (X^\top X)^{-1} \boldsymbol{x}_i \boldsymbol{Y}_i \\
=& \ \widehat{\boldsymbol{\beta}} + (X^\top X)^{-1} \boldsymbol{x}_i (1 - h_{ii})^{-1} \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}} - (X^\top X)^{-1} \boldsymbol{x}_i \boldsymbol{Y}_i - (X^\top X)^{-1} \boldsymbol{x}_i (1 - h_{ii})^{-1} h_{ii} \boldsymbol{Y}_i \\
=& \ \widehat{\boldsymbol{\beta}} + \frac{(X^\top X)^{-1}}{1 - h_{ii}} \left[\boldsymbol{x}_i \widehat{\boldsymbol{Y}}_i - (1 - h_{ii}) \boldsymbol{x}_i Y_i - \boldsymbol{x}_i h_{ii} Y_i\right] \\
=& \ \widehat{\boldsymbol{\beta}} - \frac{(X^\top X)^{-1}}{1 - h_{ii}} \boldsymbol{x}_i e_i \\
=& \ \widehat{\boldsymbol{\beta}} - (X^\top X)^{-1} \boldsymbol{x}_i \frac{e_i}{1 - h_{ii}}.
\end{aligned}$$

Por lo que definimos

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

Así

$$
\begin{aligned}
D_i &= \frac{1}{p\,\hat{\sigma}^2} G^\top (X^\top X) G \\
&= \frac{1}{p\,\hat{\sigma}^2} \left(\frac{e_i}{1-h_{ii}} \right)^2 \boldsymbol{x}_i^\top (X^\top X)^{-1} (X^\top X) (X^\top X)^{-1} \boldsymbol{x}_i \\
&= \frac{1}{p\,\hat{\sigma}^2} \left(\frac{e_i}{1-h_{ii}} \right)^2 \boldsymbol{x}_i^\top (X^\top X)^{-1} \boldsymbol{x}_i \\
&= \frac{1}{p\,\hat{\sigma}^2} \left(\frac{e_i}{1-h_{ii}} \right)^2 h_{ii} \\
&= \left(\frac{e_i}{\hat{\sigma}\sqrt{(1-h_{ii})}} \right)^2 \left(\frac{h_{ii}}{p\,(1-h_{ii})} \right) \\
&= \frac{r_i^2}{p} \cdot \frac{h_{ii}}{1-h_{ii}} \quad \text{con } r_i = \frac{e_i}{\hat{\sigma}\sqrt{(1-h_{ii})}}.
\end{aligned}
$$

□

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 $y_i^* = x_i^*$.

Demostración. Definimos

$$x_{\text{mín}} := \min_{1 \le i \le n} x_i, \qquad x_{\text{máx}} := \max_{1 \le i \le n} x_i,$$

y para la muestra transformada $y_i = ax_i + b$,

$$y_{\text{mín}} := \min_{1 \le i \le n} y_i, \qquad y_{\text{máx}} := \max_{1 \le i \le n} y_i.$$

Como $a > 0$, entonces la transformación $x \mapsto ax + b$ es estrictamente creciente, por lo que las posiciones de mínimo y máximo se conservan. En particular,

$$y_{\text{mín}} = a\, x_{\text{mín}} + b, \qquad y_{\text{máx}} = a\, x_{\text{máx}} + b.$$

Ahora, calculando la transformación min-max de y_i,

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

Por lo tanto, para $a > 0$, la transformación min-max es invariante frente a transformaciones de escala positiva y traslación. □

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.

a) Consideremos la transformación $Y = \log(X)$, o equivalentemente $X = e^Y$. Como X es una variable aleatoria continua (Pareto), podemos obtener la función de densidad de Y a través del teorema de cambio de variable. Así,

$$f_Y(y) = f_X(e^y)\left|\frac{d}{dy}e^y\right| = \frac{\alpha x_m^\alpha}{(e^y)^{\alpha+1}}e^y = \frac{\alpha x_m^\alpha}{e^{\alpha y}}, \quad y \geq \log(x_m).$$

La función de distribución de Y la podemos calcular como

$$\begin{aligned} F_Y(y) &= \int_{\log x_m}^{y} f_Y(s)ds = \int_{\log x_m}^{y} \frac{\alpha x_m^\alpha}{e^{\alpha s}}ds \\ &= \alpha x_m^\alpha \int_{\log x_m}^{y} e^{-\alpha s}ds = \alpha x_m^\alpha \left(-\frac{1}{\alpha}e^{-\alpha s}\right)\Big|_{\log x_m}^{y} \\ &= \alpha x_m^\alpha \left(\frac{1}{\alpha}x_m^{-\alpha} - \frac{1}{\alpha}e^{-\alpha y}\right) \\ &= 1 - \left(\frac{x_m}{e^y}\right)^\alpha, \quad y \geq \log(x_m). \end{aligned}$$

b) Observemos que

$$f_X(x) = \alpha x_m^\alpha \left(\frac{1}{x}\right)^{\alpha+1}, \quad x \geq x_m$$

$$f_Y(y) = \alpha x_m^\alpha \left(\frac{1}{e^y}\right)^{\alpha+1}, \quad y \geq \log(x_m),$$

y notemos que la función de densidad de Y decrece más rápido que la de X, puesto que la función exponencial crece más rápido que la función identidad. De esta manera, la transformación logarítmica acorta la cola de la distribución Pareto.

c) La función logarítmica acorta las colas de una distribución, ya que la función logaritmo comprime los valores grandes de la variable original. En el caso particular en el que nos encontramos, la distribución Pareto tiene una cola pesada, de manera que los valores grandes de X son considerablemente probables. Al aplicar la transformación logarítmica, acortamos estos valores grandes, haciendo que la distribución transformada tenga una cola más ligera.

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 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?

a) Dado $x = \{1, 2, 3, 4, M\}$ el número de datos $n = 5$. La media estará dada por

$$\bar{x}(M) = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{10 + M}{5} = 2 + \frac{M}{5}.$$

La desviación estándar la calculamos mediante

$$\sigma(M) = \sqrt{\frac{1}{n}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$
$$= \sqrt{\frac{1}{5}\left[\left(1 + \frac{M}{5}\right)^2 + \left(\frac{M}{5}\right)^2 + \left(1 - \frac{M}{5}\right)^2 + \left(2 - \frac{M}{5}\right)^2 + 4\right]}.$$

b) La mediana se calcula usando:

- Si n es impar, la mediana es el valor que está en la posición

$$\frac{n+1}{2}.$$

- Si n es par, la mediana es el promedio de los dos valores que están en las posiciones

$$\frac{n}{2} \quad \text{y} \quad \frac{n}{2}+1.$$

Al ser n=5 caemos en el primer caso y $\frac{n+1}{2} = 3$ por lo cual la mediana es $x_{(3)} = 3$ de acuerdo a los valores dados. En el caso del rango intercuantílico seguiremos los siguientes pasos:

1. Ordenar los datos de menor a mayor.
2. Encontrar Q_1 (cuantil del 25 %):
   - Es el valor que deja el 25 % de los datos por debajo.
   - Si el número de datos n es tal que $0.25(n+1)$ no es entero, se interpola entre las dos posiciones.
3. Encontrar Q_3 (cuantil del 75 %):
   - Es el valor que deja el 75 % de los datos por debajo.
   - Se calcula igual, usando la posición $0.75(n+1)$.
4. Restar:

$$\text{IQR} = Q_3 - Q_1.$$

De acuerdo a nuestros datos

$$0.25(n+1) = 0.25(6) = 1.5 \quad \text{y} \quad 0.75(n+1) = 0.75(6) = 4.5$$

por lo que Q_1 y Q_3 se encuentran entre el primer y segundo dato y entre el cuarto y quinto dato, respectivamente. Interpolando

$$Q_1 = 1 + 0.5(2-1) = 1.5 \quad y \quad Q_3 = 4 + 0.5(M-4) = 2 + \frac{M}{2}.$$

Obteniendo el valor del IQR

$$\text{IQR} = Q_3 - Q_1 = \frac{M+1}{2}.$$

c) Podemos considerar estables a las medidas que no dependan de M. Basado en las expresiones obtenidas, tenemos que sólo la mediana es estable. Por otro lado, es sabido que la mediana y desviación estándar son susceptibles a valores atípicos y es evidente bajo las expresiones obtenidas. El rango intercuantílico es considerado una medida que es resiliente a valores atipicos pero en este caso por el tamaño del conjunto de datos estuvo involucrado al definir el Q_3 y por ende aparece en el IQR.

Ejercicio 11.

Propiedades de la transformación Box–Cox.
Sea $y(\lambda)$ la transformación de Box–Cox definida como:

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

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).

a) Sea $x > 0$ fijo. Si reescribimos x^λ, de manera conveniente, en términos de la función exponencial, tenemos que,

$$x^\lambda = e^{\lambda \log(x)},$$

sustituyendo este valor en la ecuación inicial,

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

Ahora, si definimos $u = \lambda \log(x)$, cuando $\lambda \to 0$ se cumple que también $u \to 0$, y

$$\frac{e^{\lambda \log(x)} - 1}{\lambda} = \frac{e^u - 1}{u} \cdot \log(x).$$

Como $\lim\limits_{u \to 0} \dfrac{e^u - 1}{u} = 1$, entonces,

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

b) Tomemos una muestra de valores x con dispersión pronunciada (todos los valores mayores a cero):

$$x = \{0.1,\ 1,\ 10,\ 100,\ 1000\}.$$

Caso $\lambda = 1$ (sin transformación, $y(1) = x - 1$).

$$y(1) = \{\,0.1 - 1,\ 1 - 1,\ 10 - 1,\ 100 - 1,\ 1000 - 1\,\} = \{-0.9,\ 0,\ 9,\ 99,\ 999\}.$$

Para este conjunto transformado la escala y la dispersión siguen siendo enormes: la mediana es 9 y el rango es 999.9.

Caso $\lambda = 0$ (logaritmo natural, $y(0) = \log x$).

$$y(0) = \{\log(0.1),\ \log(1),\ \log(10),\ \log(100),\ \log(1000)\} \approx \{-2.3026,\ 0,\ 2.3026,\ 4.6052,\ 6.9078\}.$$

Aquí la dispersión queda drásticamente reducida: la mediana y la media coinciden aproximadamente en 2.3026 para este conjunto particular (estos x forman una progresión geométrica, por lo que sus logaritmos son una progresión aritmética). El rango es aproximadamente $6.9078 - (-2.3026) = 9.2104$, muchísimo menor que en el caso $\lambda = 1$.

Podemos concluir que la transformación Box–Cox con λ cercano a 0 "acorta" colas largas y reduce la asimetría de distribuciones sesgadas a la derecha, transformando efectos multiplicativos en efectos aditivos. El ejemplo numérico que propusimos nos ayuda a ver que la transformación logarítmica comprime órdenes de magnitud y produce valores más homogéneos y menos dominados por observaciones extremas.

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:

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

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.

a) Es claro que la función $\hat{f}_h$ es una función no negativa ya que esta es una suma de funciones indicadoras (las cuales son no negativas), multiplicadas por una constante estrictamente positiva.

b) Dado que la función $\hat{f}_h$ está definida en los intervalos disjuntos $I_1, I_2, \ldots, I_k$, podemos escribir

$$\int_{-\infty}^{\infty} \hat{f}_h(x)dx = \sum_{j=1}^{k} \int_{I_j} \hat{f}_h(x)dx. \tag{2}$$

Ahora, dado que $\hat{f}_h(x)$ es constante en cada intervalo I_j (de longitud h), e igual a

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

tenemos que (2) toma la forma

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

Observemos que esta doble suma cuenta el número de observaciones que hay en cada intervalo I_j, y suma sobre todos los intervalos. Dicho esto, la doble suma cuenta el número total de observaciones, el cual es igual a n. Por lo tanto,

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

c) Elegir un h muy grande incrementa el número de rectángulos, lo cual hace que el histograma sea muy irregular y con mucha variabilidad, es decir, que tenga mucho ruido; esto tiene como consecuencia que se incremente la varianza. Por otro lado, elegir un h muy pequeño hace que el histograma sea muy suave, evitando que se aprecien detalles importantes acerca de la distribución, y esto incrementa el sesgo de la estimación.

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

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

- *Normalización.* Suponga $h > 0$. Entonces

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

donde intercambiamos suma finita e integral. Para cada i, se lleva a cabo el cambio de variable $u = (x - x_i)/h$ (de donde $dx = h\,du$). Así,

$$\int K\left(\frac{x - x_i}{h}\right) dx = \int K(u)\,h\,du = h \int K(u)\,du = h \cdot 1 = h,$$

porque $\int K(u)\,du = 1$ por hipótesis. Sustituyendo,

$$\int \hat{f}_h(x)\,dx = \frac{1}{nh}\sum_{i=1}^{n} h = \frac{1}{n}\sum_{i=1}^{n} 1 = 1.$$

Por lo tanto, $\int \hat{f}_h(x)\,dx = 1$. □

- *No negatividad.* Suponga $h > 0$ y que $K(u) \geq 0$ para todo $u \in \mathbb{R}$. Entonces, para cada $x \in \mathbb{R}$ y cada $i = 1, \ldots, n$,

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

Como $\frac{1}{nh} > 0$ y la suma tiene $n < \infty$ términos no negativos,

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

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

- *Sesgo puntual del KDE.* Suponga que $X_1, \ldots, X_n$ son i.i.d. con densidad f, que K es un kernel integrable con $\int K(u)\,du = 1$, $\int uK(u)\,du = 0$ y segundo momento finito $\mu_2(K) = \int u^2 K(u)\,du < \infty$. Sea $h > 0$, por linealidad de la esperanza,

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

Con el cambio $u = (x - t)/h$ (así $t = x - hu$, $dt = -h\,du$),

$$\mathbb{E}[\hat{f}_h(x)] = \int K(u)\, f(x - hu)\,du.$$

Para cada u fijo, por Taylor de segundo orden en x,

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

donde $r_h(u) \to 0$ cuando $h \to 0$ para cada u fijo. Sustituyendo en la expresión del valor esperado e integrando término a término,

$$\mathbb{E}[\hat{f}_h(x)] = f(x)\int K(u)\,du - hf'(x)\int uK(u)\,du + \frac{h^2}{2}f''(x)\int u^2K(u)\,du + h^2\int u^2 r_h(u)K(u)\,du.$$

Usando los momentos del kernel, $\int K = 1$, $\int uK = 0$, $\int u^2 K = \mu_2(K)$,

$$\mathbb{E}[\hat{f}_h(x)] \;=\; f(x) + \frac{h^2}{2}\,\mu_2(K)\,f''(x) \;+\; h^2 \underbrace{\int u^2 r_h(u) K(u)\,du}_{\to 0}.$$

Para justificar que $\int u^2 r_h(u) K(u)\,du \to 0$, nótese que $r_h(u) \to 0$ punto a punto por continuidad de f'' en x. Además, usando la forma de Lagrange del resto,

$$r_h(u) = \frac{f''(x - \theta h u) - f''(x)}{2} \quad \text{para algún } \theta = \theta(h,u) \in (0,1).$$

Como f'' es continua en x, dado $\varepsilon > 0$ existe $\delta > 0$ tal que $|f''(t) - f''(x)| \le \varepsilon$ si $|t - x| \le \delta$. Para h suficientemente pequeño, si $|u| \le \delta/h$ entonces $|x - \theta h u - x| \le h|u| \le \delta$ y por tanto $|r_h(u)| \le \varepsilon/2$. Separando la integral en $\{|u| \le \delta/h\}$ y su complemento,

$$\left| \int u^2 r_h(u) K(u)\,du \right| \le \frac{\varepsilon}{2} \int_{\{|u| \le \delta/h\}} u^2 |K(u)|\,du + \sup_{|t-x| \le \delta} |f''(t) - f''(x)| \int_{\{|u| > \delta/h\}} u^2 |K(u)|\,du.$$

El primer término es $\le \frac{\varepsilon}{2} \int u^2 |K(u)|\,du$. En el segundo, como $\int u^2 |K(u)|\,du < \infty$, entonces $\int_{\{|u| > \delta/h\}} u^2 |K(u)|\,du \to 0$ cuando $h \to 0$. En consecuencia, $\int u^2 r_h(u) K(u)\,du \to 0$. Esto prueba que ese término es $o(1)$ y, por tanto,

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

□