Tarea 1

Ciencia de Datos

Brain de Jesús Salazar, César Ávila, Iván García

12 de septiembre de 2025

Solución del problema 1.

Solución del problema 2. Considere el modelo de regresión lineal $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ donde $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 I)$. Del ejercicio anterior se sabe que la matriz

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

es una matriz de proyección ortogonal. Usando la descomposición espectral, se puede descomponer a H de la forma

$$H = RR^T,$$

basta con reescribir a la matriz diagonal como el producto de dos matrices donde cada matriz tiene en su diagonal a la raíz cuadrada de cada elemento. Por otro lado, observe que ya que H es idempotente se tiene que

$$H^2 = (RR^T)(RR^T) = H = RR^T,$$

por lo que

$$R^T R = I_p.$$

Por otro lado, note que por la propiedad cíclica de la traza

$$tr(H) = tr(RR^T) = tr(R^T R) = tr(I_p) = p.$$

Por lo tanto,

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

Solución del problema 3. Considere el modelo de regresión lineal clásico, $\boldsymbol{Y} = X\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ donde $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 I)$ y a la matriz de proyección ortogonal,

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

Se sabe que el vector de residuos se puede expresa como

$$\boldsymbol{e} = (I - H)\boldsymbol{Y},$$

luego ya que $\boldsymbol{Y}$ sigue una distribución normal multivariada con media $X\boldsymbol{\beta}$ y varianza $\sigma^2 I_n$, se tiene que,

$$\boldsymbol{e} \sim N_n\left(\mathbf{0}, \sigma^2\left(I_n - H\right)\right).$$

De lo anterior, se tiene que

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

y normalizando,

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

Por otro lado, observe que si se considera a el estimador insesgado de la varianza,

$$\hat{\sigma}^2 = \frac{\mathbf{e}^\top \mathbf{e}}{n - p} = \frac{\sigma^2 \mathbf{Y}^\top (I - H)\mathbf{Y}}{\sigma^2 (n - p)},$$

y se observa que $(I - H)$ es una matriz idempotente de rango $n - p$ se tiene que

$$\frac{(n - p)}{\sigma^2}\hat{\sigma}^2 = \frac{\mathbf{Y}^\top (I - H)\mathbf{Y}}{\sigma^2} \sim \chi^2(n - p).$$

Por ultimo, se sabe que una variable aleatoria t de Student con r grados de libertad se define como la razón entre una variable aleatoria normal estándar sobre la raiz cuadrada de una variable chi cuadrada dividida entre sus r grados, con las variables aleatorias independientes. Ya que en este caso la normal estandar no es independiente de la chi cuadra se tiene que la siguiente variable aleatoria cumple que aproximadamente,

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

Simplificando la expresión de la izquierda se tiene,

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

Para arreglar el problema de la independencia se propone estimar a σ^2 sin usar el i-esimo dato, es decir

$$\hat{\sigma}_i^2 = \frac{\mathbf{e_i}^\top \mathbf{e_i}}{n-1-p} = \frac{\sigma^2 \mathbf{Y}_i^\top (I - H_i)\mathbf{Y_i}}{\sigma^2(n-1-p)},$$

donde el vector $\mathbf{Y}_i$ se obtiene eliminando al i-esima entrada y H_i se construye eliminando la i-esíma entrada de X. En este caso la matriz $(I - H_i)$ es idempotente de rango n-1-p, por lo que,

$$\frac{(n-1-p)}{\sigma^2}\hat{\sigma_i}^2 = \frac{\mathbf{Y_i}^\top (I - H_i)\mathbf{Y}_i}{\sigma^2} \sim \chi^2(n-1-p).$$

Como en este caso $\hat{\sigma}_i$ no depende de e_i se tiene que es independiente de la normal estandar definida previamente, de aquí se cumple que,

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

Simplificando el lado derecho se tiene que,

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

Solución del problema 4.

Solución del problema 5. Consideremos una muestra $Y_1, \ldots, Y_n$ de variables aleatorias independientes e idénticamente distribuidas con media μ y varianza σ^2, con indicadores $R_i \in \{0, 1\}$, de tal manera que $R_i \perp Y_i$ para cada $i \in \{1, \ldots, n\}$. Dichas R_i existen pues estamos bajo el modelo MCAR, y se interpretan como $R_i = 1$ si y solo si el dato Y_i fue observado.

Notemos que si n_{obs} representa el número de datos observados, entonces $n_{obs} = \sum_{i=1}^{n} R_i$. Además, por la definición de los R_i,

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

Así pues, si $\boldsymbol{R} = (R_1, \ldots, R_n)$, entonces

$$\begin{aligned}
\mathbb{E}\left[\overline{Y}_{obs} \,\middle|\, \boldsymbol{R}\right] = \mathbb{E}\left[\frac{1}{\sum_{i=1}^{n} R_i} \sum_{i=1}^{n} R_i Y_i \,\middle|\, \boldsymbol{R}\right] &= \frac{1}{\sum_{i=1}^{n} R_i} \sum_{i=1}^{n} \mathbb{E}\left[R_i Y_i \mid \boldsymbol{R}\right] \\
&= \frac{1}{\sum_{i=1}^{n} R_i} \sum_{i=1}^{n} R_i \mathbb{E}\left[Y_i \mid \boldsymbol{R}\right] \\
&= \frac{1}{\sum_{i=1}^{n} R_i} \sum_{i=1}^{n} R_i \mathbb{E}\left[Y_i\right] \\
&= \mu,
\end{aligned}$$

en donde hemos usado que las Y_i son iid con media μ y son independientes de $\boldsymbol{R}$ (por las hipótesis del modelo MCAR). Por consiguiente,

$$\mathbb{E}\left[\overline{Y}_{obs}\right] = \mathbb{E}\left[\mathbb{E}\left[\overline{Y}_{obs} \mid \boldsymbol{R}\right]\right] = \mu.$$

Por otra parte,

$$\overline{Y}_{obs}^2 = \frac{1}{\left(\sum_{i=1}^{n} R_i\right)^2} \left[\sum_{i=1}^{n} R_i^2 Y_i^2 + \sum_{i \neq j} R_i R_j Y_i Y_j\right],$$

por lo que

$$\begin{aligned}
\mathbb{E}\left[\overline{Y}_{obs}^2 \,\middle|\, \boldsymbol{R}\right] &= \frac{1}{\left(\sum_{i=1}^{n} R_i\right)^2} \left[\sum_{i=1}^{n} R_i^2 \mathbb{E}\left[Y_i^2 \mid \boldsymbol{R}\right] + \sum_{i \neq j} R_i R_j \mathbb{E}\left[Y_i Y_j \mid \boldsymbol{R}\right]\right] \\
&= \frac{1}{\left(\sum_{i=1}^{n} R_i\right)^2} \left[\sum_{i=1}^{n} R_i^2 \mathbb{E}\left[Y_i^2\right] + \sum_{i \neq j} R_i R_j \mathbb{E}\left[Y_i Y_j\right]\right] \\
&= \frac{1}{\left(\sum_{i=1}^{n} R_i\right)^2} \left[(\sigma^2 + \mu^2) \sum_{i=1}^{n} R_i^2 + \mu^2 \sum_{i \neq j} R_i R_j\right] \\
&= \mu^2 + \sigma^2 \frac{\sum_{i=1}^{n} R_i^2}{\left(\sum_{i=1}^{n} R_i\right)^2}.
\end{aligned}$$

De lo anterior se sigue que

$$\begin{aligned}
\operatorname{Var}\left[\overline{Y}_{obs}\right] = \mathbb{E}\left[\overline{Y}_{obs}^2\right] - \left(\mathbb{E}\left[\overline{Y}_{obs}\right]\right)^2 = \mathbb{E}\left[\mathbb{E}\left[\overline{Y}_{obs}^2 \,\middle|\, \boldsymbol{R}\right]\right] - \mu^2 &= \sigma^2 \mathbb{E}\left[\frac{\sum_{i=1}^n R_i^2}{\left(\sum_{i=1}^n R_i\right)^2}\right] \\
&= \sigma^2 \mathbb{E}\left[\frac{\sum_{i=1}^n R_i}{\left(\sum_{i=1}^n R_i\right)^2}\right] \\
&= \sigma^2 \mathbb{E}\left[\frac{1}{\sum_{i=1}^n R_i}\right] \\
&= \sigma^2 \mathbb{E}\left[\frac{1}{n_{obs}}\right] \\
&\geq \frac{\sigma^2}{n} = \operatorname{Var}\left[\ \overline{Y}\ \right],
\end{aligned}$$

en donde hemos usado que $R_i^2 = R_i$, pues $R_i \in \{0,1\}$, para todo $i \in \{1, \dots, n\}$, y que $n_{obs} \leq n$. De hecho, siguiendo un procedimiento completamente análogo, pero ahora con varianzas condicionales, se sigue que

$$\operatorname{Var}\left[\overline{Y}_{obs} \,\middle|\, \boldsymbol{R}\right] = \frac{\sigma^2}{n_{obs}}.$$

De lo anterior podemos notar que $\overline{Y}_{obs}$ es insesgado, pero que $\operatorname{Var}\left[\overline{Y}_{obs}\right] \geq \operatorname{Var}\left[\ \overline{Y}\ \right]$, por lo que $\overline{Y}_{obs}$ tiene menor eficiencia (posee más varianza, pues la eliminación de datos hace que haya menos de ellos para poder estimar a la media de Y).

Solución del problema 6. Sean $\boldsymbol{Y} = (Y_{obs}, Y_{mis})$ y $\boldsymbol{R}$ el patrón de datos faltantes. Bajo la definición del MAR,

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

Así pues, bajo este modelo,

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

Por consiguiente, la verosimilitud de θ está dada por

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

Ya que el factor $\mathbb{P}\left[\boldsymbol{R} \,|\, Y_{obs}, \psi\right]$ no depende de θ, se sigue que $L\left(\theta; Y_{obs}, \boldsymbol{R}\right) \propto \mathbb{P}\left[Y_{obs} \,|\, \theta\right]$. Para ver las condiciones *a priori* que garantizan ignorabilidad bajo el enfoque bayesiano, notemos que

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

Para concluir ignorabilidad buscamos que $L\left(\theta \,|\, Y_{obs}, \boldsymbol{R}\right) \propto \pi(\theta)\mathbb{P}\left[Y_{obs} \,|\, \theta\right]$, y ya que la integral anterior depende de θ solamente a través del factor $\pi\left(\theta, \psi\right)$, dicha ignorabilidad se logra cuando $\pi(\theta, \psi) = \pi(\theta)\pi(\psi)$. Es decir, si en la *a priori* se pide indistinguibilidad de los parámetros (i.e. que θ y ψ sean independientes), entonces el mecanismo es ignorable para inferir θ.

Solución del problema 7.

Solución del problema 8. Dado que $a > 0$, se tiene que

$$\text{mín}(y) \coloneqq \min_{1\leq i\leq n} y_i = \min_{1\leq i\leq n}(ax_i + b) = b + \min_{1\leq i\leq n}(ax_i) = b + a \min_{1\leq i\leq n} x_i = a\,\text{mín}(x) + b.$$

De manera análoga,

$$\text{máx}(y) \coloneqq \max_{1\leq i\leq n} y_i = \max_{1\leq i\leq n}(ax_i + b) = b + \max_{1\leq i\leq n}(ax_i) = b + a \max_{1\leq i\leq n} x_i = a\,\text{máx}(x) + b.$$

Por consiguiente, para todo $i \in \{1, \ldots, n\}$, se tiene que

$$y_i^* = \frac{y_i - \text{mín}(y)}{\text{máx}(y) - \text{mín}(y)} = \frac{(ax_i + b) - (a\,\text{mín}(x) + b)}{(a\,\text{máx}(x) + b) - (a\,\text{mín}(x) + b)} = \frac{a\,(x_i - \text{mín}(x))}{a\,(\text{máx}(x) - \text{mín}(x))} = x_i^*,$$

que es lo deseado.

Solución del problema 9. (a) Como el soporte de X es $[x_m, \infty)$, con $x_m > 0$, la transformación $Y = \log(X)$ está bien definida, y Y tiene soporte en $[\log(x_m), \infty)$. Además, la función $g(x) = \log(x)$ definida en $\mathbb{R}^+$ es uno a uno y tiene inversa $g^{-1}(y) = e^y$, la cual es una función derivable, con $\frac{d}{dy}g^{-1}(y) = e^y$. Por lo tanto, como se cumple la relación $Y = \log(X)$ y por consiguiente $X = e^Y$, por el Teorema de Cambio de Variables Y tiene densidad dada por

$$f_Y(y) = \left|\frac{dx}{dy}\right| f_X(e^y)\mathbb{1}_{[\log(x_m),\infty)}(y) = e^y \frac{\alpha x_m^\alpha}{e^{y(\alpha+1)}}\mathbb{1}_{[\log(x_m),\infty)}(y) = \alpha\left(\frac{x_m}{e^y}\right)^\alpha \mathbb{1}_{[\log(x_m),\infty)}(y).$$

Notemos que esta última expresión puede ser escrita como

$$f_Y(y) = \alpha e^{-\alpha(y-\log(x_m))}\mathbb{1}_{[0,\infty)}(y - \log(x_m)),$$

de donde se puede observar que $Y \stackrel{d}{=} \log(x_m) + Exp(\alpha)$, en donde $Exp(\alpha)$ es una variable aleatoria con distribución exponencial de media $\frac{1}{\alpha}$. En particular, de aquí se sigue que la función de distribución acumulada de Y es

$$F_Y(y) = \begin{cases} 0, & \text{si } y < \log(x_m), \\ 1 - e^{-\alpha(y-\log(x_m))}, & \text{si } y \geq \log(x_m). \end{cases}$$

(b) Primero veamos que, dado $x > x_m$,

$$\mathbb{P}[X \geq x] = \int_x^\infty \frac{\alpha x_m^\alpha}{t^{\alpha+1}}dt = x_m^\alpha\left[-t^{-\alpha}\right]_{t=x}^\infty = \left(\frac{x_m}{x}\right)^\alpha.$$

De este modo, la cola de X decae de forma polinomial, del orden $x^{-\alpha}$. Por otro lado, si $y > \log(x_m)$,

$$\mathbb{P}[Y \geq y] = e^{\alpha\log(x_m)}e^{-\alpha y} = x_m^\alpha e^{-\alpha y},$$

de donde podemos ver que la cola de Y decae de forma polinomial, del orden $e^{-\alpha y}$ (más rápidamente que el decaimiento polinomial). Es decir, X tiene colas más pesadas, y al transformarse a Y, cambia a colas más ligeras.

(c) Notemos que, como $Y = \log(X)$, para todo $y \in \mathbb{R}$ se cumple que

$$\mathbb{P}[Y > y] = \mathbb{P}[X > e^y],$$

de modo que, como e^y crece más rápido que y, las colas de Y decaen más rápidamente de las de X, como lo visto con la distribución Pareto, en donde un decaimiento polinomial se convierte en uno exponencial. Además, como la función logaritmo es creciente y $\log(x) \leq \log(x+1) \leq x$ para todo $x > 0$, por lo general Y tiene un soporte más grande que X.

Más aún, por las propiedades de la función logarítmica, los cambios grandes en X se reflejan en cambios más chicos de Y. Por ejemplo, si un valor de X se duplica, en la transformación logarítmica el valor de Y solo incrementa en $\log 2$ (cambios multiplicativos se transforman en cambios aditivos). Por consiguiente, si X tiene colas muy pesadas, Y tiende a distribuir el peso a lo largo de los reales y no tan concentrado en las colas; es decir, se "acortan" las colas largas. Además esto produce, por lo general, distribuciones más cercanas a la simetría, en especial cuando hay errores multiplicativos, que se convierten en errores aditivos al aplicar logaritmo, y el Teorema del Límite Central explica dicha simetría.

Solución del problema 10.

Solución del problema 11. (a) Sea $x > 0$, y veamos que

$$\lim_{\lambda \to 0}(x^\lambda - 1) = 1 - 1 = 0,$$

mientras que $\lim_{\lambda \to 0} \lambda = 0$. Además, la función $\lambda \mapsto \lambda$ es derivable, con derivada igual a $1(\neq 0)$. Por lo tanto, ya que el siguiente límite existe, por la Regla de l'Hôpital se tiene que

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

(b) Consideremos a la sucesión $(x_n)_{n \in \mathbb{N}}$, en donde $x_n = 2^n$ para todo $n \in \mathbb{N}$. Dicha sucesión toma valores muy dispersos cuando n es muy grande, pues sus primeros valores son

$$2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192, 16384, 32768, 65536, 131072, 262144, \dots.$$

La sucesión correpondiente a la transformación de Box y Cox con $\lambda = 1$ es la misma pero recorrida en 1, así que sigue siendo igual de dispersa. Sin embargo, con la transformación logarítmica ($\lambda = 0$), se convierte en $(y_n)_{n \in \mathbb{N}}$, en donde $y_n = n\log(2)$ para todo $n \in \mathbb{N}$, que es mucho menos dispersa. A manera de ilustración, sus primeros valores son aproximadamente iguales a:

$$0.6931, 1.3863, 2.0794, 2.7726, 3.4657, 4.1589, 4.852, 5.5452, 6.2383, 6.9315, 7.6246, 8.3178, 9.0109, \dots$$

Solución del problema 12. a)- Con las hipótesis del enunciado, observe que la función $\hat{f}_h(x)$ es una suma de funciones indicadoras, que cuenta el número de observaciones x_i que están en el mismo conjunto que x, I_j. Luego ya que $1\{x_i \in I_j\} \geq 0$ y $nh > 0$ se tiene que $\hat{f}_h(x) \geq 0$.

b)- Para dar respuesta a este inciso, observe que la función $\hat{f}_h(x)$ se puede escribir como,

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

Luego la integral buscada se puede ver como,

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

Donde la segunda igualdad se tiene de considerar la partición $I_1, ..., I_k$, la penúltima igualdad se tiene ya que cada x_i pertenece a algún conjunto I_j.

c)-Observe que cuando h es grande, los intervalos contendrán más datos, esto nos llevará a que no se aprecie si hay algún patrón en el comportamiento de los datos, es decir si los datos tienen preferencia por ciertos intervalos. Por otro lado, cuando h es muy pequeño, los intervalos no alcanzarán a contener muchos datos, un caso extremo de ver esto es hacer a h muy cercano a cero de tal forma que cada intervalo contenga a lo más un dato, en este caso, solo se verán barras de tamaño $\frac{1}{n}$ en cada dato.

Solución del problema 13. Normalización)- Con las hipótesis del enunciado, observe que la integral se puede escribir como,

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

Donde la tercer igualdad se tiene haciendo el cambio de variable $u = \frac{x-x_i}{h}$ y la penúltima igualdad se tiene gracias a que la integral del kernel K es uno.

No negatividad)- Para este inciso basta observar por hipotesis $K(u) \geq 0$ y que $nh > 0$ por lo que $\hat{f}_h(x) \geq 0$.

Sesgo puntual)- Observe que el sesgo puntual se puede escribir como,

$$\begin{aligned}
\mathbb{E}\left[\hat{f}_h(x)\right] - f(x) = \mathbb{E}\left[\sum_{i=1}^{n}\frac{1}{nh}K\left(\frac{x-x_i}{h}\right)\right] - f(x) \\
=&\sum_{i=1}^{n}\frac{1}{nh}\mathbb{E}\left[K\left(\frac{x-x_i}{h}\right)\right] - f(x) \\
=&\frac{1}{h}\mathbb{E}\left[K\left(\frac{x-x_1}{h}\right)\right] - f(x) \\
=&\frac{1}{h}\int_{-\infty}^{\infty} K\left(\frac{x-x_1}{h}\right) f\left(x_1\right) dx_1 - f(x).
\end{aligned}$$

Haciendo el cambio de variable $u = \frac{x-x_1}{h}$ y considerando que la integral del kernel es igual a uno, se tiene que la expresión anterior es igual a,

$$\begin{aligned}
\frac{1}{h}\int_{-\infty}^{\infty} hK(u)f(x-uh)du - \int_{-\infty}^{\infty} K(u)f(x)du \\
=\int_{-\infty}^{\infty} K(u)[f(x-uh) - f(x)]du.
\end{aligned}$$

Por otro lado, observe que si se desarrolla la serie de Taylor de orden dos de $f(x-uh)$ al rededor de x con error de Peano se tiene la siguiente expresión,

$$f(x-uh) = f(x)+f'(x)(x-uh-x)+\frac{f''(x)(x-vh-x)^2}{2!}+h_2(x-uh)(uh)^2, \quad \lim_{x-uh\to x} h_2(x-uh) = 0,$$

de donde se puede despejar la expresión $f(x-uh)-f(x)$ y sustituyendo en la integral se tiene,

$$\begin{aligned}&\int_{-\infty}^{\infty} K(u)\left[f'(x)(-uh)+\frac{f''(x)(uh)^2}{2!}+h_2(x-uh)(uh)^2\right]du\\ =&\frac{f''(x)}{2!}h^2\int_{-\infty}^{\infty} u^2K(u)du + h^2\int_{-\infty}^{\infty} K(u)h_2(x-uh)u^2du,\end{aligned}$$

donde esta última igualdad se tiene ya que el primer momento es cero. Además, observe que si se divide entre h^2 a la expresión de la derecha y se hace tender a h a cero se tiene,