Introducción a la Ciencia de Datos **Antonio Barragán**
Tarea I

Hazel Sánchez
Omar García Ramos

Ejercicio 1

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

Demostración:

Dada la matriz $X_{n\times p}$ de rango completo, tenemos que:

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

Por lo tanto es idempotente.
Por otro lado tenemos que:

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

Los elementos de la diagonal $h_i = H_{ii}$, denominados *leverages*, son la influencia de la i^{th} observación, las cuales están entre $[0,1]$ pues dado que es idempotente, $H^2 = H$, así:

$$h_{ii} = H^2_{ii} \implies h_{ii} = \sum_j h_{ij}h_{ji}$$

Pero como es simétrica:

$$h_{ii} = h^2_{ii} + \sum_{j\neq i} h^2_{ji} \implies h_{ii} \geq h^2_{ii}$$

De esta manera se tiene que podemos medir la influencia de las observaciones

Ejercicio 2

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:

Dada la matriz H de n observaciones y p parametros, tenemos que:

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

Por otro lado, dadas la propiedades cíclicas de la traza de matrices y si X es de rango p, entonces

$$tr(H) = tr(X(X^\top X)^{-1}X^\top) = tr((X^\top X)^{-1}X^\top X) = tr(\mathbf{I}_p) = p \tag{2}$$

Por lo tanto de (1) y (2), se tiene que:

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

La suma total $\sum h_{ii} = p$ indica que el modelo utiliza sus p parámetros para ajustar las n observaciones, por lo tanto valores de h_{ii} cercanos a 1 significan que casi un parámetro completo se dedica a ajustar esa observación particular y hará que la recta ajuste a este punto alejándose del resto de observaciones menos influyentes.

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.

Demostración

Dado el modelo de regresión lineal,tenemos que para p covariables y n observaciones

$$Y = X\beta + \epsilon \quad \epsilon \sim N_n(0, \sigma^2 I_n)$$

Así los residuales:

$$\mathbf{e} = (I_n - H)Y \sim N_n(0, (I_n - P)\sigma^2)$$

Ya que es una transformación lineal de Y y además:

$$\begin{aligned}
\mathbb{E}[\mathbf{e}] &= (I_n - P)\mathbf{E}[Y] = 0 \\
\mathbb{V}[e] &= (I_n - P)\mathbb{V}[Y](I_n - P)^\top = (I_n - P)\sigma^2
\end{aligned}$$

Tiene una distribución normal multivariada ya que son una transformación lineal de Y con parametros:

$$\mathbb{E}[\mathbf{e}] = (I_n - H)\mathbb{E}[Y] = (I_n - H)\mathbb{E}[Y] = 0$$

Dado que la varianza (σ^2), no es conocida usaremos el estimador de máxima verosimilitud insesgado para σ^2:

$$\begin{aligned}
\hat{\sigma^2} &= \frac{1}{n}(Y - X\hat{\beta})^\top (Y - X\hat{\beta}) \\
&= \frac{1}{n} Y^\top (I_n - H) Y \\
&= \frac{n-p}{n} S^2, \qquad S^2 \sim \chi^2_{n-p}
\end{aligned}$$

Por lo tanto dado que para $\gamma \in \mathbb{R}^p$

$$\gamma^\top \mathbf{e} \sim N(\gamma^\top 0, \gamma^\top (I_n - H)\gamma\sigma^2)$$

Tomando la base canonica, tenemos que:

$$\mathbf{e} = (e_1, .., e_n) \quad e_i \sim N(0, (1 - h_{ii})\sigma^2)$$

De esta manera definimos a:

$$\begin{aligned}
Z &= \frac{e_i}{\sqrt{1 - h_{ii}}\sigma} \sim N(0, 1) \\
S^2 &= \frac{1}{n-p}\hat{\sigma}^2 \sim \chi^2_{n-p}
\end{aligned}$$

De esta manera, dado que no son independiente Z, S^2, tenemos que el cociente entre ellas sigue aproximadamente un distribución T student

$$\begin{aligned}
T &= \frac{\frac{e_i}{\sigma\sqrt{1-h_{ii}}}}{\sqrt{\frac{(n-p)}{(n-p)\sigma^2}\hat{\sigma}^2}} \\
&= \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}} \sim t(n - p).
\end{aligned}$$

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.

Demostración:

Bajo la suposición **MCAR**, sean:

1. $Y = (Y_{obs}, Y_{mis})$ los datos (faltantes y observados)
2. R el patrón de faltantes, es decir la matriz indicadora de faltantes.
3. θ nuestro vector de parámetros del modelo de datos.
4. ψ el vector de parámetros del mecanismo de faltantes.

Por la definición formal de MCAR, la distribución del patrón de faltantes R es independiente de los datos Y (tanto observados como faltantes) y de los parámetros θ, condicional sólo en sus propios parámetros ψ. Esto se expresa como:

$$p(R \mid Y_{obs}, Y_{mis}, \theta, \psi) = p(R \mid \psi)$$

Partimos de la densidad conjunta de todos los elementos:

$$\begin{aligned} p(Y, R, \theta, \psi) &= p(R \mid Y, \theta, \psi)\, p(Y, \theta, \psi) \\ &= p(R \mid Y, \theta, \psi)\, p(Y \mid \theta, \psi)\, p(\theta, \psi) \end{aligned}$$

Nuestro objetivo es encontrar la densidad conjunta de los datos y el patrón de faltantes condicional en los parámetros, $p(Y, R \mid \theta, \psi)$. Por definición de probabilidad condicional:

$$p(Y, R \mid \theta, \psi) = \frac{p(Y, R, \theta, \psi)}{p(\theta, \psi)}$$

Sustituyendo la expresión anterior:

$$\begin{aligned} p(Y, R \mid \theta, \psi) &= \frac{p(R \mid Y, \theta, \psi)\, p(Y \mid \theta, \psi)\, p(\theta, \psi)}{p(\theta, \psi)} \\ &= p(R \mid Y, \theta, \psi)\, p(Y \mid \theta, \psi) \end{aligned}$$

Aplicando ahora la hipótesis de MCAR ($p(R \mid Y, \theta, \psi) = p(R \mid \psi)$):

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

Se asume la distribución de los datos depende sólo de θ, es decir, $p(Y \mid \theta, \psi) = p(Y \mid \theta)$, por lo tanto:

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

Por otro lado la inferencia se basa en la verosimilitud de los parámetros dados los datos observados, (Y_{obs}, R). La verosimilitud se obtiene integrando la densidad conjunta sobre los valores no observados Y_{mis}:, es decir obteniendo la densidad marginal.

$$\begin{aligned} L(\theta, \psi \mid Y_{obs}, R) &= p(Y_{obs}, R \mid \theta, \psi) \\ &= \int p(Y_{obs}, Y_{mis}, R \mid \theta, \psi) dY_{mis} \end{aligned}$$

Por la factorización anterior:

$$\begin{aligned} L(\theta, \psi \mid Y_{obs}, R)dY_{mis} &= \int p(Y_{obs}, Y_{mis} \mid \theta)p(R \mid \psi)dY_{mis} \\ &= p(R \mid \psi) \int p(Y_{obs}, Y_{mis} \mid \theta)dY_{miss} \end{aligned}$$

Notemos que la marginal sobre la densidad conjuta de los datos observados:

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

Por lo tanto, la verosimilitud conjunta final es:

$$L(\theta, \psi \mid Y_{obs}, R) = p(R \mid \psi)\, p(Y_{obs} \mid \theta)$$

Notemos que para estimar θ, se necesita encontrar:

$$\arg\max_{\theta} L(\theta, \psi | Y_{obs}, R) = \arg\max_{\theta} p(Y_{obs}|\theta)$$

Esto debido a que la densidad $p(R|\psi)$ no depende de θ, así la inferencia sobre el parámetro puede basarse únicamente en $p(Y_{obs}|\theta)$

Ejercicio 5

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

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

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

Demostración:

Sea $Y_1, \dots, Y_n \sim F_\theta$ una muestra i.i.d. con esperanza $\mathbb{E}[Y_i] = \mu$ y varianza $\mathrm{Var}(Y_i) = \sigma^2$. y R_i la variable indicadora de si Y_i fue observado . De esta manera:

$$\overline{Y}_{\text{obs}} = \frac{1}{N} \sum_{i=1}^{n} R_i Y_i, \quad N = \sum_{i=1}^{n} R_i.$$

De esta manera, dada la muestra:

$$\mathbb{E}\big[\overline{Y}_{\text{obs}} \mid R\big] = \frac{1}{k} \sum_{i=1}^{n} R_i \, \mathbb{E}[Y_i \mid R].$$

Bajo **MCAR**, se cumple que $Y_i \perp R_i$, de modo que:

$$\mathbb{E}[Y_i \mid R] = \mathbb{E}[Y_i] = \mu$$

Así,

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

Por la propiedad torre, concluimos que:

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

Por lo tanto, $\overline{Y}_{\text{obs}}$ es un estimador **insesgado** de μ.
Por otro lado la varianza de $\overline{Y}_{\text{obs}}$ condicionada a $N = k$ es

$$\mathrm{Var}(\overline{Y}_{\text{obs}} \mid N = k) = \frac{\sigma^2}{k}.$$

Mientrás que la varianza del estimador completo (sin datos faltantes) es:

$$\mathrm{Var}(\overline{Y}) = \frac{\sigma^2}{n}.$$

Además, por la ley de la varianza total:

$$\mathrm{Var}(\overline{Y}_{\text{obs}}) = \mathbb{E}[\mathrm{Var}(\overline{Y}_{\text{obs}} \mid \mathbf{R})] + \mathrm{Var}(\mathbb{E}[\overline{Y}_{\text{obs}} \mid \mathbf{R}]) = \mathbb{E}\left[\frac{\sigma^2}{N}\right] + \mathrm{Var}(\mu) = \sigma^2 \mathbb{E}\left[\frac{1}{N}\right].$$

Por la desigualdad de Jensen dado que $\frac{1}{N}$ es convenxa:

$$\mathbb{E}\left[\frac{1}{N}\right] \geq \frac{1}{\mathbb{E}[N]} = \frac{1}{np},$$

donde $p = \mathbb{P}(R_i = 1) \leq 1$, entonces

$$\mathbb{E}[1/N] \geq 1/(np) \geq 1/n$$

Por tanto:

$$\mathrm{Var}(\overline{Y}_{\text{obs}}) \geq \frac{\sigma^2}{n} = \mathrm{Var}(\overline{Y}).$$

La igualdad se da solo si no hay datos faltantes , así la varianza de $\overline{Y}_{\text{obs}}$ es mayor que la del estimador completo, por lo que pierde eficiencia.

Ejercicio 6

Factorización bajo MAR. A partir de la definición de MAR, muestre que
$L(\theta; Y_{obs}, R) \propto p(Y_{obs}|\theta)$.
¿Qué suposición adicional en el prior es necesaria en el enfoque bayesiano para concluir ignorabilidad?

Demostración

Bajo la suposición **MAR** el mecanismo de faltantes satisface

$$p(R \mid Y, \theta, \psi) = p(R \mid Y_{\text{obs}}, \psi),$$

es decir, la probabilidad del patrón R puede depender de los datos observados Y_{obs} pero *no* de los valores faltantes Y_{mis} ni de θ.
De esta manera tenemos que:

$$\begin{aligned} p(Y, R|\theta, \psi) &= \frac{p(R|Y, \theta, \psi)}{p(\theta, \psi)} p(Y, \theta, \psi) \\ &= p(R|Y_{obs}, \psi)p(Y|\theta, \psi) \\ &= p(R|Y_{obs}, \psi)p(Y|\theta) \end{aligned}$$

Para la verosimilitud de los datos observados y el patrón R, debemos de marginalizar:

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

Para estimación por máxima verosimilitud de θ, tenemos que:

$$\arg\max_{\theta} L(\theta, \psi) = \arg\max_{\theta} p(Y_{\text{obs}} \mid \theta)$$

Por esto, concluimos que:

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

Notemos que en un enfoque bayesiano:

$$\begin{aligned} p(\theta, \psi|Y, R) &= \frac{p(R, Y|\theta, \psi)p(\theta, \psi)}{p(Y, R)} \\ &\propto p(R, Y|\theta, \psi)p(\theta, \psi) \end{aligned}$$

Por la factorización de la densidad:

$$p(\theta, \psi|Y, R) \propto p(R|Y_{obs}, \psi)p(Y_{obs}|\theta)p(\theta, \psi)$$

Notemos que si podemos la distribución del parámetro es independiente al mecanismo de faltantes:

$$p(\theta, \psi|Y, R) \propto p(Y|\theta)p(\theta)\left(p(\psi)p(R|Y_{obs}, \psi)\right)$$

De esta manera, la densidad a *posteriori* puede factorizarse como el producto de 2 términos, el primero que depende de θ, mientras que el segundo solamente depende de ψ, así:

$$\begin{aligned} p(\theta|Y, R) &= \int p(\theta, \psi|Y, R) d\psi \\ &\propto \int p(Y|\theta)p(\theta)\left(p(\psi)p(R|Y_{obs}, \psi)\right) d\psi \\ &\propto p(Y|\theta)p(\theta) \end{aligned}$$

Por lo que bajo esta suposición tenemos que podemos ignorar a ψ para la inferencia sobre θ

Ejercicio 7

Distancia de Cook como medida global de influencia. Partiendo de la definición
$D_i = \frac{\sum_{j=1}^{n}(y_j - \hat{y}_{j(i)})^2}{p\hat{\sigma}^2}$
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

Tenemos que la distancia de Cook [a]: se construye a partir de remover la i^{th} observación y comparar los coeficientes estimados de la muestra completa contra la muestra sin la i^{th} observación, así sean:

1. $\hat{\beta}$ los coeficientes del modelo $Y = X\beta + \epsilon$
2. V_i, el vector V si su i^{th} entrada, es decir si V es un vector de tamaño n, V_i es de tamaño $n-1$
3. Dada una matriz $X_{n\times m}$, tenemos que X_1 denota a la matríz sin la primera fila, es decir una matriz $X_{(n-1)\times m}$

Por lo tanto, sea $Y = \begin{pmatrix} Y_1 \\ Y_{(1)} \end{pmatrix}$ la partición del véctor, así:

$$Y = X\beta + \epsilon = \begin{pmatrix} x_1^\top \\ X_1 \end{pmatrix} \beta + \epsilon$$

Si quitamos la primera observación, tenemos:

$$Y_{(1)} = X_1\beta + \epsilon_{(1)} \implies \hat{\beta_{(1)}} = \left(X_1^\top X_1\right)^{-1} X_1^\top Y_{(1)}$$

Dado el primer renglón de la matriz x_1, notamos lo siguiente:

$$X = \begin{pmatrix} x_1^\top \\ X_1 \end{pmatrix} \implies X^\top X = (x_1, X_1^\top) \begin{pmatrix} x_1^\top \\ X_1 \end{pmatrix} = x_1 x_1^\top + X_1^\top X_1$$

Sustituyendo la expresión para $X_1^\top X_1$ y usando la **Proposición B.6 (Fórmula de Woodbury)**[b]: Tenemos que:

$$(X_1^\top X_1)^{-1} = (X^\top X - x_1 x_1^\top)^{-1} = (X^\top X)^{-1} + (X^\top X)^{-1} x_1 (1 - x_1^\top (X^\top X)^{-1} x_1)^{-1} x_1^\top (X^\top X)^{-1}$$

Por otro lado:

$$X^\top Y = \begin{pmatrix} x_1 & X_1^\top \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_{(1)} \end{pmatrix} \implies X_1^\top Y_{(1)} = X^\top Y - x_1 Y_1$$

De esta forma tenemos:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{(1)} &= (X_1^\top X_1)^{-1} X_1^\top \boldsymbol{Y}_{(1)} \\
&= (X^\top X)^{-1} X^\top \boldsymbol{Y} + (X^\top X)^{-1} \boldsymbol{x}_1 \left[1 - \boldsymbol{x}_1^\top (X^\top X)^{-1} \boldsymbol{x}_1\right]^{-1} \boldsymbol{x}_1^\top (X^\top X)^{-1} X^\top \boldsymbol{Y} \\
&\quad - (X^\top X)^{-1} \boldsymbol{x}_1 \boldsymbol{Y}_1 - (X^\top X)^{-1} \boldsymbol{x}_1 \left[1 - \boldsymbol{x}_1^\top (X^\top X)^{-1} \boldsymbol{x}_1\right]^{-1} \boldsymbol{x}_1^\top (X^\top X)^{-1} \boldsymbol{x}_1 \boldsymbol{Y}_1
\end{aligned}$$

Notemos que:

$$H = X^\top (X^\top X)^{-1} X \implies x_1^\top (X^\top X)^{-1} x_1 = h_{11}$$

Por lo tanto:

$$\begin{aligned}
\hat{\beta_{(1)}} = &\ \widehat{\boldsymbol{\beta}} + (X^\top X)^{-1}\boldsymbol{x}_1(1-h_{11})^{-1}\boldsymbol{x}_1^\top\widehat{\boldsymbol{\beta}} - (X^\top X)^{-1}\boldsymbol{x}_1\boldsymbol{Y}_1 - (X^\top X)^{-1}\boldsymbol{x}_1(1-h_{11})^{-1}h_{11}\boldsymbol{Y}_1 \\
= &\ \widehat{\boldsymbol{\beta}} + \frac{(X^\top X)^{-1}}{1-h_{11}}\left[\boldsymbol{x}_1\widehat{\boldsymbol{Y}}_1 - (1-h_{11})\boldsymbol{x}_1 Y_1 - \boldsymbol{x}_1 h_{11} Y_1\right] \\
= &\ \hat{\beta} + \frac{(X^\top X)^{-1}}{1-h_{11}}\left[x_1(\hat{Y_1} - Y_1)\right]
\end{aligned}$$

Recordemos que:

$$\mathbf{e} = Y - \hat{Y} \quad \Longrightarrow \mathbf{e_i} = (Y_i - \hat{Y_i})$$

Así concluimos que:

$$\hat{\beta_{(1)}} = \widehat{\boldsymbol{\beta}} - (X^\top X)^{-1}\boldsymbol{x}_1\frac{e_1}{1-h_{11}}.$$

Notemos lo siguiente:

$$\begin{aligned}
D_i = \frac{\sum_{j=1}^n(\hat{y}_j - \hat{y}_{j(i)})}{p\hat{\sigma}^2} &= \frac{1}{p\hat{\sigma}^2}(\hat{Y} - \hat{Y}_{(i)})^\top(\hat{Y} - \hat{Y}_{(i)}) \\
&= \frac{1}{p\hat{\sigma}^2}(X(\hat{\beta} - \hat{\beta}_i))^\top(X(\hat{\beta} - \hat{\beta}_i)) \\
&= \frac{1}{p\hat{\sigma}^2}(\hat{\beta} - \hat{\beta}_i)^\top X^\top X(\hat{\beta} - \hat{\beta}_i))
\end{aligned}$$

Mientras que de la expresión obtenida tenemos que:

$$\hat{\beta_{(i)}} - \hat{\beta} = \frac{(X^\top X)^{-1}}{1-h_{ii}}x_i(\hat{Y_i} - Y_i)$$

Lo cual nos lleva a lo siguiente:

$$\begin{aligned}
(\hat{\beta} - \hat{\beta}_i)^\top X^\top X(\hat{\beta} - \hat{\beta}_i)) &= \left(\frac{\hat{Y_i - Y_i}}{1-h_{ii}}\right)x_i^\top(X^\top X)^{-1}(X^\top X)\left(\frac{\hat{Y_i - Y_i}}{1-h_{ii}}\right)(X^\top X)^{-1}x_i \\
&= \left(\frac{\hat{Y_i} - Y_i}{1-h_{ii}}\right)^2 x_i^\top(X^\top X)^{-1}x_i \\
&= \left(\frac{\hat{Y_i} - Y_i}{1-h_{ii}}\right)^2 h_{ii}
\end{aligned}$$

Recordando que:

$$\mathbf{e_i} = \hat{Y_i} - Y_i$$

Se sigue de lo anterior:

$$D_i = \frac{h_{ii}}{p(1-h_{ii})}\left(\frac{\mathbf{e}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}\right)^2 = \frac{\mathbf{r_i}^2}{p}\left(\frac{h_{ii}}{1-h_{ii}}\right)$$

Dado que los *leverages* miden la influencia de una observación en el ajuste del modelo, una observación altamente influyente hará que la **Distancia de Cook** aumente, permitiéndonos así detectar posibles *outliers*. Además, si una observación está mal ajustada, es decir, si presenta un residual alto, esto también contribuirá a un mayor valor de la Distancia de Cook, lo que ayuda a identificar otro tipo de *outliers*.

[a] Ramírez, L. (s.f.). *Residuos Studentizados*. Notas del curso Modelos Estadísticos I. Centro de Investigación en Matemáticas (CIMAT). Recuperado de `http://personal.cimat.mx:8181/~leticia.ramirez/Modelos_Estadisticos_I/Cap2.html#residuos-studentizados`

[b] Ramírez, L. (s.f.). *Normal multivariada*. Notas del curso Modelos Estadísticos I. Centro de Investigación en Matemáticas (CIMAT). Recuperado de `http://personal.cimat.mx:8181/~leticia.ramirez/Modelos_Estadisticos_I/Normal_multivariada.html`

Ejercicio 8

Invarianza afín en Min–Max Sea $x_1, \ldots, x_n$ un conjunto de datos y defina la transformación $x_i^* = \frac{x_i - \min(x)}{\max(x) - \min(x)}$.
Pruebe que si $y_i = ax_i + b$ con $a > 0$, entonces $y_i^* = x_i^*$.

Demostración:

Sean $x_1, .., x_n$ un conjunto de datos, definimos el re-escalamiento:

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

Para $y_i = ax_i + b$, tenemos que:

$$y_i^* = \frac{y_i - \min(y)}{\max(y) - \min(y)}$$

Tenemos que:

$$\begin{aligned}
\max(y) &= \max\{y_i | i = 1, .., n\} \\
&= \max\{ax_i + b\} = \max\{ax_i\} + b \\
&= a \max\{x_i\} + b \quad (a > 0)
\end{aligned}$$

Así concluimos que:

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

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.

Demostración:

Sea $X \sim \text{Pareto}(\alpha, x_m)$ con densidad

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

Definimos $Y = \log X$, dado el cambio de variable $x = e^y$. El soporte de Y es

$$y = \log x \geq \log x_m.$$

La derivada es $dx/dy = e^y$. Por la fórmula de cambio de variable,

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

Sea: $y_0 = \log x_m$,

$$f_Y(y) = \alpha e^{-\alpha(y-y_0)}, \qquad y \geq y_0 = \log x_m.$$

Por tanto, es una exponencial truncada:

$$Y \sim \text{Exp}(\alpha, y_0),$$

Por otro lado, el comportamientos de las colas puede estudiarse por medio de:

$$\mathbb{P}(Y > y) = \int_y^\infty f_Y(t)\, dt = e^{-\alpha(y - \log x_m)}, \qquad y \geq \log x_m.$$

Mientras que para la variable X

$$\mathbb{P}(X > x) = \left(\frac{x_m}{x} \right)^\alpha$$

Notemos lo siguiente:

$$\mathbb{P}(Y > y) = \mathbb{P}(X > e^y) = \left(\frac{x_m}{e^y} \right)^\alpha = e^{-\alpha(y - \log x_m)}.$$

En la distribución Pareto, la cola de X decrece de manera polinómica:

$$P(X > x) = x_m^\alpha x^{-\alpha}$$

lo cual corresponde a una cola pesada. Sin embargo, al transformar $Y = \log(X)$, la función de supervivencia pasa a ser

$$P(Y > y) = P(X > e^y) = (x_m e^{-y})^\alpha, \quad y \geq \log(x_m),$$

En consecuencia, Y tiene colas más ligeras que X.
Por úíltimo dado que $\ln(x)$ transforma multiplicaciones en sumas.

$$\log(1000 x_m) = \log x_m + \log 1000,$$

Así comprime las colas y por ende reduce la influencia de valores extremos,además, al comprimir la escala para valores grandes de X, la distribución transformada Y tiende a mostrar un comportamiento más cercano a la simetría..
$\square$

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

Sea la muestra
$$x = \{1, 2, 3, 4, M\}, \qquad M \to \infty, \quad n = 5.$$
La media es
$$\overline{x} = \frac{1+2+3+4+M}{5} = \frac{10+M}{5}.$$
Para la varianza muestral:
$$\sum_{i=1}^{n} x_i^2 = 1^2 + 2^2 + 3^2 + 4^2 + M^2 = 30 + M^2,$$
y
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\Big(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\Big).$$
Por lo tanto:
$$\begin{aligned} s^2 &= \frac{1}{4}\Big(30 + M^2 - 5\Big(\frac{10+M}{5}\Big)^2\Big) \\ &= \frac{1}{5}M^2 - M + \frac{5}{2}. \end{aligned}$$
Así la desviación estándar muestral es
$$s = \sqrt{\frac{1}{5}M^2 - M + \frac{5}{2}}.$$
Para la mediana, ordenando la muestra: $1, 2, 3, 4, M$, dado que n es impar tenemos que
$$m = \text{median}(x) = 3,$$
independiente de M.
Para el rango intercuartílico, recordemos que el cuantil muestral de orden p se define como el mínimo valor x en la muestra ordenada que satisface $F(x) \geq p$, donde F es la función de distribución empírica. Es decir:
$$Q_p = \min\{x \in \text{muestra ordenada} : F(x) \geq p\}$$
Para Q_1 (cuantil 0.25):
$$F(1) = \frac{1}{5} = 0.20 < 0.25, \quad F(2) = \frac{2}{5} = 0.40 \geq 0.25$$
por lo tanto $Q_1 = 2$.
Para Q_3 (cuantil 0.75):
$$F(3) = \frac{3}{5} = 0.60 < 0.75, \quad F(4) = \frac{4}{5} = 0.80 \geq 0.75$$
por lo tanto $Q_3 = 4$.
El rango intercuartílico está dado por:
$$RIQ = Q_3 - Q_1 = 4 - 2 = 2.$$
Así la mediana $m = 3$ y el rango intercuartílico $RIQ = 2$ no dependen de M, por lo que son robustos ante valores grandes de M.
Mientras que la media $\bar{x}(M) = \frac{10+M}{5}$ y la desviación estándar $s = \sqrt{\frac{1}{5}M^2 - M + \frac{5}{2}}$ pierden su interpretabilidad ante valores grandes de M, ya que ambas divergen a infinito cuando $M \to \infty$.

Ejercicio 11

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

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

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

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

Demostración

a) Sea $y > 0$, entonces tenemos que $y^\lambda = e^{\lambda \ln y}$, por lo cual podemos notar que

$$\begin{aligned} y^{(\lambda)} &= \frac{y^\lambda - 1}{\lambda} = \frac{e^{\lambda \ln y} - 1}{\lambda} = \frac{1}{\lambda}\left(\sum_{k=0}^{\infty} \frac{(\lambda \ln y)^k}{k!} - 1\right) = \frac{1}{\lambda}\left(1 + \lambda \ln y + \sum_{k=2}^{\infty} \frac{(\lambda \ln y)^k}{k!} - 1\right) \\ &= \ln y + \frac{1}{\lambda}\left(\sum_{k=2}^{\infty} \frac{(\lambda \ln y)^k}{k!}\right) = \ln y + \left(\sum_{k=2}^{\infty} \frac{\lambda^{k-1}(\ln y)^k}{k!}\right), \end{aligned}$$

luego, por convergencia dominada tenemos que

$$\begin{aligned} \lim_{\lambda\to 0} y^{(\lambda)} &= \lim_{\lambda\to 0}\left(\ln y + \left(\sum_{k=2}^{\infty} \frac{\lambda^{k-1}(\ln y)^k}{k!}\right)\right) = \ln y + \lim_{\lambda\to 0}\left(\sum_{k=2}^{\infty} \frac{\lambda^{k-1}(\ln y)^k}{k!}\right) \\ &= \ln y + \left(\sum_{k=2}^{\infty} \lim_{\lambda\to 0} \frac{\lambda^{k-1}(\ln y)^k}{k!}\right) \\ &= \ln y, \end{aligned}$$

como queremos.

Ejercicio 12

Propiedades del histograma. Sea $x_1, \ldots, x_n$ una muestra i.i.d. de una variable aleatoria continua con densidad $f(x)$. Con[...]re 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.

Demostración:

Sea $x_1, \ldots, x_n$ i.i.d. con densidad f, $\{I_j\}_{j=1}^k$ una partición de ancho h
Para un x fijo, el número de observaciones en cada I_j es no negativo, por lo tanto $\hat{f}_h(x) \geq 0$

Por otro lado, tenemos que integrando $\hat{f}_h$ sobre $\mathbb{R}$

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

Dado un intervalo I_j, ya conocemos cuantas observaciones están en ese intervalo, digamos N_j, así:

$$\int_{I_j} \frac{1}{nh} \sum_{i=1}^{n} \mathbf{1}\{x_i \in I_j\}\, dx = \frac{1}{nh} \int_{I_j} N_j dx$$

Notemos que esta integral es sobre un intervalo $I_j = (a_j, b_j)$ de ancho h:

$$|I_j| = b_j - a_j = h \quad j = 1, 2, .., k$$

Entonces tenemos que:

$$\int_{I_j} \frac{1}{nh} \sum_{i=1}^{n} \mathbf{1}\{x_i \in I_j\}\, dx = \frac{1}{nh} N_j h$$

Por lo que:

$$\int_{-\infty}^{\infty} \hat{f}_h(x)\, dx = \sum_{j=1}^{k} \frac{N_j}{n} = \frac{1}{n} \sum_{j=1}^{k} N_j = \frac{n}{n} = 1,$$

Para analizar el efecto del parámetro h, observe que cuando h es grande, los intervalos I_j contienen muchas observaciones. En el caso extremo $h \to \infty$, todas las observaciones quedarían en un solo intervalo, y el histograma se aproximaría a una distribución uniforme, lo que induce conclusiones erróneas sobre la forma de la verdadera densidad.
Por otro lado, si $h \to 0$, cada intervalo contendrá a lo más una observación, produciendo un histograma extremadamente irregular y con gran variabilidad. Así, h demasiado grande incrementa el sesgo, mientras que un h demasiado pequeño incrementa la varianza. De ahí la importancia de elegir un valor de h que logre un equilibrio entre sesgo y varianza.

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

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

Demostración:

Dada la estimación de densidad:

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

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

Como se cumple que:

$$\int_{\mathbb{R}} K(u)du = 1$$

Concluimos que:

$$\int_{-\infty}^{\infty} \hat{f}_h(x)dx = \frac{1}{nh}\sum_{i=1}^{n} h = 1$$

Es decir la función $f(x)$ esta normalizada. Por otro lado tenemos que dado a que:

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

Así $\hat{f}(x)_h$ es no negativa.
Ahora notemos que, usando la expansión de Taylor de f alrededor de x

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

La esperanza del estimador:

$$\begin{aligned}
\mathbb{E}[\hat{f}_h(x)] &= \frac{1}{h}\mathbb{E}\left[K\left(\frac{x-X}{h}\right)\right] \\
&= \frac{1}{h}\int K\left(\frac{x-t}{h}\right) f(t)dt \\
&= \int K(u)f(x-hu)du
\end{aligned}$$

Sustituyendo:

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

Usando las propiedades del kernel:

$$\begin{aligned}
\int K(u)du = 1, \\
\int uK(u)du = 0, \\
\mu_2(K) = \int u^2K(u)du,
\end{aligned}$$

obtenemos:

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

Por lo tanto, el sesgo es:

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