Regresión múltiple y otras técnicas multivariadas

Tarea 06

Rivera Torres Francisco de Jesús
Rodríguez Maya Jorge Daniel
Samayoa Donado Víctor Augusto
Trujillo Barrios Georgina

Marzo 27, 2019

Ejercicio 1

Considerar las siguientes matrices

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \text{y} \qquad \boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

Calcular lo siguiente:

Inciso 2.a)

$$\boldsymbol{X}^T\boldsymbol{X}$$

Inciso 2.b)

$$\boldsymbol{X}^T\boldsymbol{y}$$

Inciso 2.c)

$$\left|\boldsymbol{X}^T\boldsymbol{X}\right|$$

Inciso 2.d)

$$\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1},$$ ¿qué se debe cumplir para que tal inversa exista?

Inciso 2.e)

$$\left(\boldsymbol{X}^T\boldsymbol{X}\right)\boldsymbol{X}^T\boldsymbol{y}$$

Ejercicio 2

Considerar la matriz sombrero $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$ del análisis del modelo RLM. Mostrar que las siguientes matrices son simétricas e idempotentes.

Recordemos que una matriz A es simétrica si $A^T = A$. Y una matriz B es idempotente si $BB = B$.

Recordemos que $\boldsymbol{I}_n$ es la matriz identidad de dimensión $n \times n$ y $\boldsymbol{J}_n$ es la matriz de puros "unos" de dimensión $n \times n$.

Inciso 2.a)

$\boldsymbol{I}_n - \boldsymbol{H}$

Primero demostramos que es simétrica.

Demostración.

$$\begin{aligned}
(I_n - H)^T &= \left[I_n - X\left(X^TX\right)^{-1}X^T\right]^T = (I_n)^T - \left[X\left(X^TX\right)^{-1}X^T\right]^T \\
&= I_n - \left(X^T\right)^T\left[\left(X^TX\right)^{-1}\right]^T X^T = I_n - X\left[\left(X^TX\right)^T\right]^{-1}X^T \\
&= I_n - X\left(X^T\left(X^T\right)^T\right)^{-1}X^T = I_n - X\left(X^TX\right)^{-1}X^T \\
&= I_n - H
\end{aligned}$$

por lo tanto $\boldsymbol{I}_n - \boldsymbol{H}$ es simétrica. □

Ahora se procede a demostrar que es idempotente

Demostración.

$$\begin{aligned}
(I_n - H)(I_n - H) &= I_nI_n - I_nH - HI_n + HH \\
&= I_n - 2H + HH \\
&= I_n - 2X\left(X^TX\right)^{-1}X^T + \left[X\left(X^TX\right)^{-1}X^T\right]\left[X\left(X^TX\right)^{-1}X^T\right] \\
&= I_n - 2X\left(X^TX\right)^{-1}X^T + X\left(X^TX\right)^{-1}\left(X^TX\right)\left(X^TX\right)^{-1}X^T \\
&= I_n - 2X\left(X^TX\right)^{-1}X^T + X\left(X^TX\right)^{-1}X^T \\
&= I_n - X\left(X^TX\right)^{-1}X^T \\
&= I_n - H
\end{aligned}$$

por lo tanto $\boldsymbol{I}_n - \boldsymbol{H}$ es idempotente. □

Inciso 2.b)

$\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{J}_n$

Primero demostramos que es simétrica.

Demostración.

$$\left(I_n - \frac{1}{n}J_n\right)^T = I_n^T - \left(\frac{1}{n}J_n\right)^T = I_n - \frac{1}{n}J_n$$

ya que $J_n = (1_{ik})$, donde $1_{ik} = 1$, para cualesquiera $i, k = 1, \ldots, n$. Y $(J_n)^T = (1_{ik})^T = (1_{ki})$, pero $1_{ik} = 1_{ki} = 1$. Por lo tanto, $\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{J}_n$ es simétrica. □

Ahora se procede a demostrar que es idempotente

Demostración.

$$\begin{aligned}\left(I_n - \frac{1}{n}J_n\right)\left(I_n - \frac{1}{n}J_n\right) &= I_nI_nI_n - \frac{1}{n}I_nJ_n - \frac{1}{n}J_nI_n + \frac{1}{n^2}J_nJ_n \\ &= I_n - \frac{2}{n}J_n + \frac{1}{n^2}J_nJ_n = I_n - \frac{2}{n}J_n + \frac{1}{n}J_n \\ &= I_n - \frac{1}{n}J_n\end{aligned}$$

ya que $J_nJ_n = (\sum_{l=1}^n 1_{il}1_{lk})_{ik} = (n_{ik})$. Por lo tanto, $\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{J}_n$ es idempotente. □

Inciso 2.c)

$\boldsymbol{H} - \frac{1}{n}\boldsymbol{J}_n$

Primero demostramos que es simétrica.

Demostración.

$$\begin{aligned}\left(H - \frac{1}{n}J_n\right)^T &= \left[\left(I_n - \frac{1}{n}J_n\right) - (I_n - H)\right]^T = \left(I_n - \frac{1}{n}J_n\right)^T - (I_n - H)^T \\ &= \left(I_n - \frac{1}{n}J_n\right)^T - (I_n - H)\,, \text{ ya que ambas son simétricas por incisos 2.a y 2.b} \\ &= H - \frac{1}{n}J_n\end{aligned}$$

por lo tanto $\boldsymbol{H} - \frac{1}{n}\boldsymbol{J}_n$ es simétrica. □

Ahora se procede a demostrar que es idempotente

Demostración.

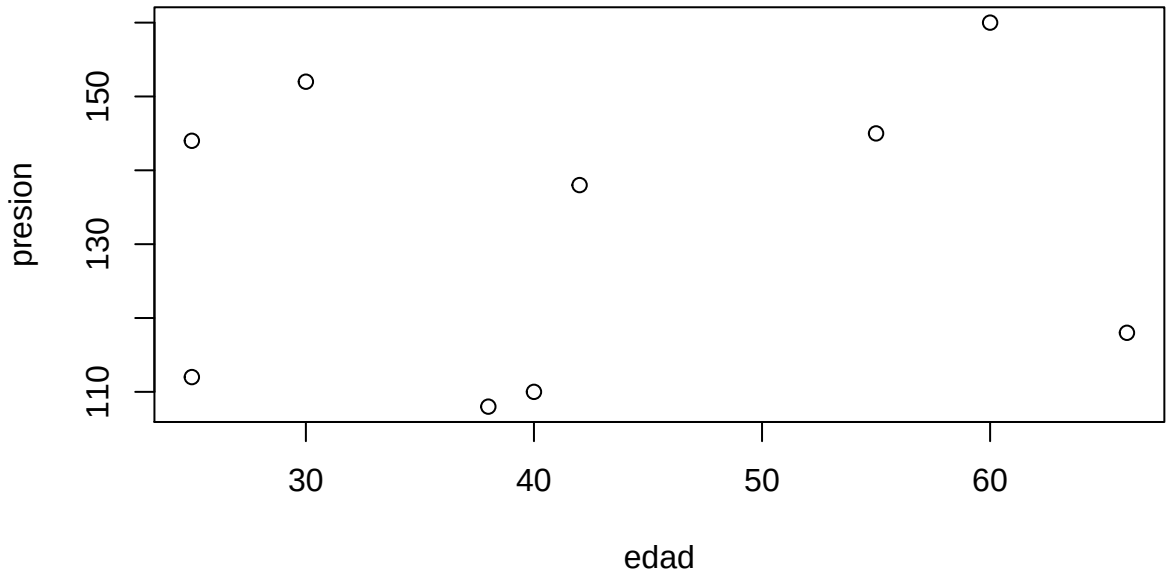
$$\begin{aligned}\left(H - \frac{1}{n}J_n\right)\left(H - \frac{1}{n}J_n\right) &= HH - \frac{1}{n}HJ_n - \frac{1}{n}J_nH + \frac{1}{n^2}J_nJ_n \\ &= H - \frac{1}{n}HJ_n - \frac{1}{n}J_nH + \frac{1}{n}J_n, \text{ ya que } H \text{ y } \frac{1}{n}J_n \text{ son idempotentes}\end{aligned}$$

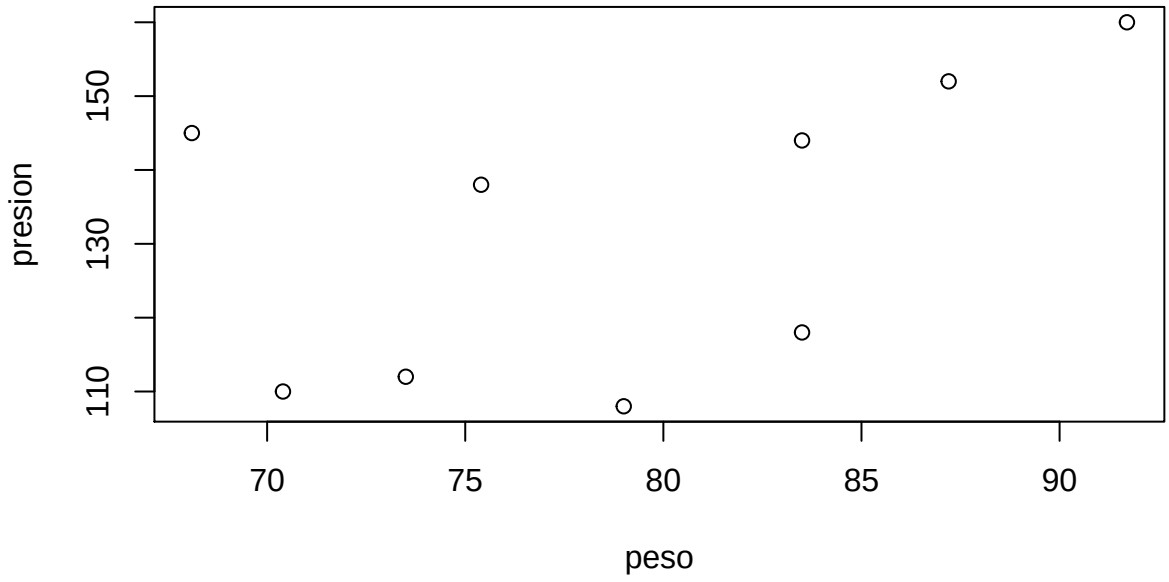
Por lo tanto $\boldsymbol{H} - \frac{1}{n}\boldsymbol{J}_n$ es idempotente. □

Ejercicio 3

En un estudio clínico se registró la edad (años), el peso (kg) y la presión sistólica (mmHg) de nueve voluntarios con estilos de vida similares. Los datos se muestran en la siguiente tabla.

a) Graficar la presión sistólica contra la edad y contra el peso. ¿Se aprecia una relación lineal entre estos pares de variables?





No parece haber una relación lineal entre presión y edad, y presión y peso.

b) Ajustar un modelo RLM para explicar la distribución de la presión sistólica como función de la edad y el peso. Interpretar las estimaciones en el contexto de los datos.

```
## (Intercept)        edad        peso
##  27.9633285   0.1162305   1.2509423
```

Como vemos el valor del intercepto es 27.9633285, el cual no puede interpretarse debido a que carece de sentido pensar en personas con peso cero. No obstante que tiene sentido hablar de edad cero. La estimación para la variable independiente *edad* es 0.1162305, la cual indica que dado un incremento de un año de edad (y manteniedo el peso constante), se espera que en promedio la presión aumente en 0.1162305 $mmHg$. Análogamente para la estimación *peso*, dado un incremento de $1kg$ en el peso de una persona (pensando su edad como constante), esperaríamos que en promedio la presión sistólica aumente en 1.2509423.

c) Estimar el valor esperado de la presión de un individuo de 35 años y 80 kg.

Esto implica que la presión sistólica esperada para un individuo de 35 años y 80 kg de peso es de 132.1067824 $mmHg$.

Ejercicio 4

El siguiente conjunto de datos sobre trasplantes de corazón relaciona el tiempo de supervivencia (en días) de pacientes que recibieron un trasplante con su edad (en años) al momento del trasplante y un llamado puntaje de incompatibilidad o discrepancia que se usa como indicador de qué tan bien recibido será el corazón trasplantado por el receptor.

Tabla 1: Datos ejercicio 4

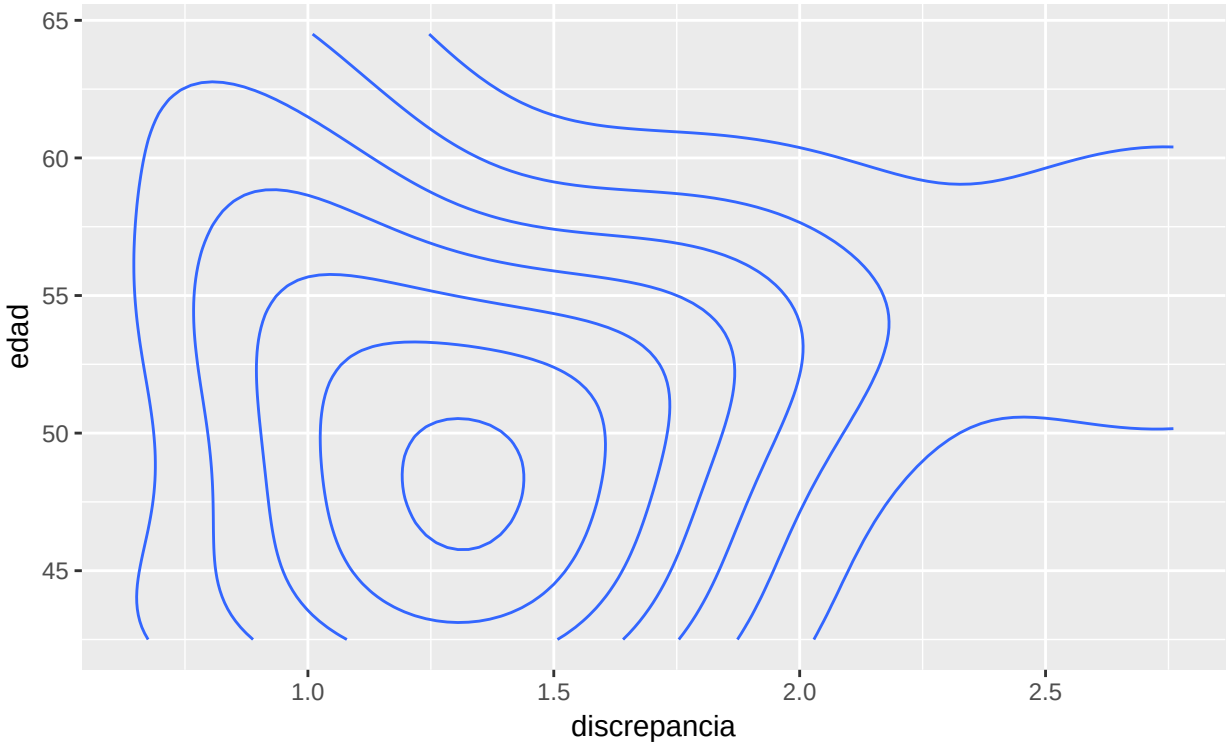
Tiempo de supervivencia	Puntaje de discrepancia	Edad en años
624	1.32	51.0
46	0.61	42.5
64	1.89	54.6
1350	0.87	54.1
280	1.12	49.5
10	2.76	55.3
1024	1.13	43.4
39	1.38	42.8
730	0.96	58.4
136	1.62	52.0
836	1.58	45.0
60	0.69	64.5

Ajustar un modelo RLM para explicar la distribución del logaritmo del tiempo de supervivencia como función de la edad y el puntaje de discrepancia.

Inciso 4.a)

Reportar las estimaciones de los coeficientes del modelo e interpretarlas en el contexto de los datos.

Antes de ajustar el modelo, se procede a gráficar la densidad de las variables explicativas para entender los intervalos de valores para cada variable explicativa.



Se observa que la variable edad tiene valores entre los 40 y 65 años. Mientras que la variable discrepancia tiene valoes entre 0 y 3.

```
##
## Call:
## lm(formula = log(supervivencia) ~ discrepancia + edad, data = .)
##
## Coefficients:
##  (Intercept)  discrepancia          edad
##       7.9567       -1.2047       -0.0225
```

Lo anterior implica que el modelo RLM:

$$\log(\text{supervivencia}) = 7.9566593 - 1.2046532\text{discrepancia} - 0.0225039\text{edad}$$

Por el análisis anterior, los datos observados no existe ningún paciente que tenga 0 años (o una edad cercana a esto). Sin embargo, considerando el contexto del problema, es posible pensar en un bebe (0 años) que necesita un transplante de corazon. En este caso el coeficiente $\hat{\beta}_0 = 7.9566593$ se interpreta como el logaritmo del tiempo de supervivencia (en días) para un paciente de 0 años

con una discrepancia de 0. En otras palabras $\exp(\hat{\beta}_0) = 2854.5208877$ es el tiempo, promedio, de supervivencia (en días) para un paciente de 0 años con una discrepancia de 0.

EL coeficiente $\hat{\beta}_1 = -1.2046532$ se interpreta como el decremento del logaritmo del tiempo, promedio, de supervivencia por cada unidad de discrepancia que aumenta en un paciente.

EL coeficiente $\hat{\beta}_2 = -0.0225039$ se interpreta como el decremento del logaritmo del tiempo, promedio, de supervivencia por cada año (en edad) que aumenta en un paciente.

Inciso 4.b)

Si se sabe que el índice de discrepancia involucra en su cómputo, entre otros factores, a la edad del paciente, ¿tiene sentido la interpretación que acaba de dar sobre el coeficiente β_j de la edad? ¿y del coeficiente β_j del propio índice? Argumente y justifique su respuesta.

En el caso del índice de discrepancia, la interpretación anterior sobre el coeficiente $\hat{\beta}_1$ no es necesariamente válida. Esto debido a que al afirmar que "... un incremento de una unidad del índice de discrepancia ...", se asume que el valor de la variable edad no se modifica.

Sin embargo, debido a que el cálculo de la variable discrepancia involucra la edad, es posible que el incremento de una unidad solamente se de al modificar la variable edad.

Lo anterior no es válido si la relación es tal que se pueda mantener constante la variable edad y aún así lograr cambios en el índice de discrepancia. En tal caso la interpretación de $\hat{\beta}_1$ sigue siendo válida.

En el caso de de la variable edad, la interpretación anterior sobre el coeficiente $\hat{\beta}_2$ ya no es válida. Esto debido a que sin lugar a dudas una variación en la variable edad implicara una variación en el índice de discrepancia y por ende no podemos hablar de "... un incremente en la variable edad, mientras la variable discrepancia se mantiene constante".

Inciso 4.c)

Reportar la estimación de σ^2.

La estimación de la varianza está dada por $\sigma^2 = 2.4535825$.

Inciso 4.d)

Estimar la media del tiempo de supervivencia de un paciente que recibió un trasplante de corazón a los 46 años y tenía un índice de discrepancia de 1.43.

La media del logaritmo del tiempo de supervivencia está dada por

$$\log(\text{supervivencia}) = 7.9566593 - 1.2046532\text{discrepancia} + -0.0225039\text{edad}$$

por ende, la media del tiempo de supervivencia está dada por

$$\text{supervivencia} = \exp(7.9566593 - 1.2046532\text{discrepancia} - 0.0225039\text{edad})$$
$$\bar{y} = \exp(7.9566593 - 1.2046532(1.43) - 0.0225039(46))$$
$$\bar{y} = \exp(7.9566593 - 1.722654 - 1.0351816)$$
$$\bar{y} = \exp(5.1988236)$$
$$\bar{y} = 181.0591203$$