



Actividad 5.1 Matriz HAT

Integrantes

A01068244 - Jared Andrés Silva Villa

A00227869 - Paola Félix Torres

Fecha: 15 de Agosto del 2024

Índice

1. Identificar el porcentaje de datos faltantes.	3
2. Identificar el mecanismo que ocasiona datos faltantes (MCAR, MAR, NMAR)}	3
3. Obtener estadísticas descriptivas de los datos (histograma, media, desviación estándar, mediana, moda, etc).	4
4. Utilizar el método de imputación adecuado para cada una de las variables con datos faltantes.	5
5. Realizar un boxplot e interpretarlo.	7

1. Para el siguiente conjunto de datos, obtener la Matriz HAT, H , los valores predichos, \hat{y} , y los coeficientes de la ecuación de regresión β .

X_1	y
2	5
3	8
5	7
7	10
9	12

X	y
2	5
3	8
5	7
7	10
9	12

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

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 3 & 5 & 7 & 9 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 3 & 5 & 7 & 9 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 5 \\ 1 & 7 \\ 1 & 9 \end{bmatrix} = \begin{bmatrix} 5 & 26 \\ 26 & 168 \end{bmatrix}$$

$\begin{matrix} a & c \\ b & d \end{matrix}$

$$(X^T X)^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \frac{1}{169} \begin{bmatrix} 168 & -26 \\ -26 & 5 \end{bmatrix} = \begin{bmatrix} 1.02 & -0.158 \\ -0.158 & 0.03 \end{bmatrix}$$

$$X (X^T X)^{-1} = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 5 \\ 1 & 7 \\ 1 & 9 \end{bmatrix} \begin{bmatrix} 1.02 & -0.158 \\ -0.158 & 0.03 \end{bmatrix} = \begin{bmatrix} 0.7 & -0.09 \\ 0.5 & -0.06 \\ 0.2 & -0.008 \\ -0.08 & 0.05 \\ -0.4 & 0.11 \end{bmatrix}$$

$$X (X^T X)^{-1} X^T = \begin{bmatrix} 0.7 & -0.09 \\ 0.5 & -0.06 \\ 0.2 & -0.008 \\ -0.08 & 0.05 \\ -0.4 & 0.11 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 3 & 5 & 7 & 9 \end{bmatrix} = \begin{bmatrix} 0.5 & 0.4 & 0.2 & 0.07 & -0.1 \\ 0.38 & 0.32 & 0.2 & 0.08 & -0.04 \\ 0.18 & 0.17 & 0.16 & 0.14 & 0.12 \\ 0.02 & 0.07 & 0.17 & 0.27 & 0.37 \\ -0.18 & -0.07 & 0.15 & 0.37 & 0.59 \end{bmatrix} \Rightarrow H$$

$$\hat{y} = Hy$$

$$Hy = \begin{bmatrix} 0.5 & 0.4 & 0.2 & 0.07 & -0.1 \\ 0.38 & 0.32 & 0.2 & 0.08 & -0.04 \\ 0.18 & 0.17 & 0.16 & 0.14 & 0.12 \\ 0.02 & 0.07 & 0.17 & 0.27 & 0.37 \\ -0.18 & -0.07 & 0.15 & 0.37 & 0.59 \end{bmatrix} \begin{bmatrix} 5 \\ 8 \\ 7 \\ 10 \\ 12 \end{bmatrix} = \begin{bmatrix} 6.6 \\ 6.18 \\ 6.22 \\ 8.99 \\ 10.37 \end{bmatrix} \Rightarrow \hat{y}$$

$$\beta = (X^T X)^{-1} X^T y$$

$$X^T y = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 3 & 5 & 7 & 9 \end{bmatrix} \begin{bmatrix} 5 \\ 8 \\ 7 \\ 10 \\ 12 \end{bmatrix} = \begin{bmatrix} 42 \\ 247 \end{bmatrix}$$

$$\beta = \begin{bmatrix} 1.02 & -0.158 \\ -0.158 & 0.03 \end{bmatrix} \begin{bmatrix} 42 \\ 247 \end{bmatrix} = \begin{bmatrix} 3.81 \\ 0.77 \end{bmatrix} \Rightarrow \beta$$

2. Verificar los resultados utilizando Python.

```
[86] H = X @ np.linalg.inv(X.T @ X) @ X.T
      print(H)
```

```
⇒ [[ 0.51219512  0.41463415  0.2195122  0.02439024 -0.17073171]
    [ 0.41463415  0.34756098  0.21341463  0.07926829 -0.05487805]
    [ 0.2195122  0.21341463  0.20121951  0.18902439  0.17682927]
    [ 0.02439024  0.07926829  0.18902439  0.29878049  0.40853659]
    [-0.17073171 -0.05487805  0.17682927  0.40853659  0.6402439 ]]
```

```
[87] B = np.linalg.inv(X.T @ X) @ X.T @ y
      print(B)
```

```
[ 3.86585366  0.87195122]
```

```
Y_hat = H @ y
print(Y_hat)
```

```
[ 5.6097561  6.48170732  8.22560976  9.9695122  11.71341463]
```

```
error = y - Y_hat
print(error)
```

```
[-0.6097561  1.51829268 -1.22560976  0.0304878  0.28658537]
```

3.Utilizando los Datos “Cirugía de Hígado” obtener la matriz HAT, los valores predichos, \hat{y} y los coeficientes de la ecuación de regresión β utilizando el método de matrices. (Puede realizarse con cualquier paquete).

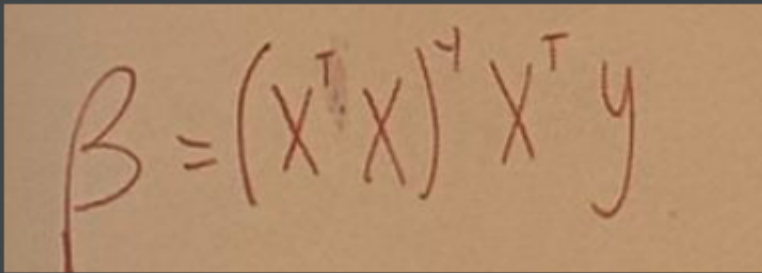
Estandarizar las variables con MinMax Scaler antes de realizar las operaciones con matrices.

Problema en colab

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

```
▶ H = X_scal @ np.linalg.inv(X_scal.T @ X_scal) @ X_scal.T
print(H)
```

```
⇒ [[ 0.03714196  0.01278513  0.03153001 ... -0.0107616  0.03002929
    0.00244981]
   [ 0.01278513  0.06106075  0.04827415 ...  0.02342793  0.0143413
    0.03831749]
   [ 0.03153001  0.04827415  0.09511866 ...  0.02121118  0.02055136
    0.03079561]
   ...
   [-0.0107616  0.02342793  0.02121118 ...  0.10349403  0.00417997
    0.03037262]
   [ 0.03002929  0.0143413  0.02055136 ...  0.00417997  0.06246334
   -0.00802438]
   [ 0.00244981  0.03831749  0.03079561 ...  0.03037262 -0.00802438
    0.05864247]]
```



A photograph of a piece of brown paper with the equation $\beta = (X^T X)^{-1} X^T y$ handwritten in brown ink.

```
[72] B = np.linalg.inv(X_scal.T @ X_scal) @ X_scal.T @ Y_scal
print(B)
```

```
⇒ [[-0.03374218]
   [ 0.08630899]
   [ 0.20125292]
   [ 0.38437348]
   [-0.06065628]
   [-0.00613695]
   [-0.04239579]
   [ 0.07147603]]
```

$$\hat{y} = Hy$$

```
Y_hat = H @ Y  
print(Y_hat)
```

```
⇒ Sobrevivencia (días)  
0      706.256237  
1      430.822921  
2      732.229832  
3      425.039576  
4      1454.587552  
..      ...  
103     566.515055  
104     568.334910  
105     584.367939  
106     385.909891  
107     459.719981  
  
[108 rows x 1 columns]
```

```
Error = Y - Y_hat  
print(Error)
```

```
⇒ Sobrevivencia (días)  
0      -11.256237  
1      -27.822921  
2      -22.229832  
3      -76.039576  
4       888.412448  
..      ...  
103      22.484945  
104      30.665090  
105      70.632061  
106      -8.909891  
107     182.280019
```

4. Concluir sobre el significado de los valores de apalancamiento y formular la ecuación de regresión.

Los valores de apalancamiento de cada observación en un modelo de regresión se calculan utilizando la matriz HAT. El valor de apalancamiento es una medida de la

influencia que tiene un punto de datos sobre el ajuste del modelo. Cuanto mayor sea el valor de apalancamiento, más influye ese punto en el resultado final.

En el problema 1, notamos que el punto h5 tiene un valor alto en comparación al promedio, lo que indica que esa observación tiene un impacto significativo en la ecuación de regresión.

En el problema 2, el análisis de los valores de apalancamiento es más complejo debido a las dimensiones más grandes de los datos.

Para el problema 1 la ecuación sería la siguiente:

$$Y = 3.86585366 + 0.87195122X_1$$

Para el problema 2 la ecuación sería la siguiente:

$$Y = -0.0337 + 0.0863X_1 + 0.2013X_2 + 0.3844X_3 - 0.0607X_4 - 0.0061X_5 - 0.0424X_6 + 0.0715X_7$$