Regresión Lineal Múltiple

Luca Mazzarello, Ignacio Pardo

April 21, 2023

0.1 Primera parte

El objetivo de esta sección es deducir una fórmula para la solución óptima β^* siguiendo los pasos a continuación:

a) Mostrar que el espacio columna de la matriz X es un subespacio vectorial de $\mathbb{R}^n$:

$$Col(X) = \{b \text{ en } \mathbb{R}^n \text{ tales que b = X con variando en } \mathbb{R}^p\}$$

Para mostrar que $Col(X)$ es un subespacio vectorial de $\mathbb{R}^n$, debemos demostrar que $Col(X)$ cumple con las siguientes propiedades:

1. El vector cero pertenece al espacio $Col(X)$.

$$Col(X) = Gen\{x_1, x_2, ..., x_p\}$$

El vector 0 está en $Col(X)$ porque $0 * x_1 + 0 * x_2$ es combinación lineal de los vectores de $Col(X)$.

2. La suma de dos vectores pertenecientes a $Col(X)$ también pertenece a $Col(X)$.

Sean u, v vectores en $Col(X)$, s_1, s_2 y t_1, t_2 escalares tales que:

$$u = s_1x_1 + s_2x_2 \iff u \in Col(X)$$

Por ser combinación lineal de los vectores de $Col(X)$.

$$v = t_1x_1 + t_2x_2 \iff v \in Col(X)$$

Por ser combinación lineal de los vectores de $Col(X)$.

Entonces:

$$u + v = s_1x_1 + s_2x_2 + t_1x_1 + t_2x_2$$
$$u + v = (s_1 + t_1)x_1 + (s_2 + t_2)x_2$$

Llamamos $c_1 = s_1 + t_1$ y $c_2 = s_2 + t_2$.

Entonces:

$$u + v = c_1x_1 + c_2x_2$$

$$c_1x_1 \in Col(X) \land c_2x_2 \in Col(X) \implies c_1x_1 + c_2x_2 \in Col(X) \implies u + v \in Col(X)$$

3. El producto de un escalar por un vector perteneciente a $Col(X)$ también pertenece a $Col(X)$.

Sea u un vector en $Col(X)$, s_1, s_2 escalares tales que:

$$u = s_1x_1 + s_2x_2 \iff u \in Col(X)$$

Llamemos c al escalar.

Entonces:

$$cu = c * (s_1x_1 + s_2x_2)$$
$$cu = (c * s_1)x_1 + (c * s_2)x_2$$

Llamemos $c_1 = c * s_1$ y $c_2 = c * s_2$

Entonces:

$$cu = c_1x_1 + c_2x_2$$

$$c_1x_1 \in Col(X) \land c_2x_2 \in Col(X) \implies c_1x_1 + c_2x_2 \in Col(X) \implies cu \in Col(X)$$

Por lo tanto $Col(X)$ es un subespacio vectorial de $\mathbb{R}^n$.

b) Supongamos que cuando hablamos de vectores en $\mathbb{R}^n$ nos referimos a vectores columna de $\mathbb{R}^{n\times 1}$. Mostrar en ese caso que el producto escalar entre dos vectores u, v en $\mathbb{R}^n$ puede calcularse como:

$$u \cdot v = v^\top u$$

donde operación en el lado derecho de la igualdad es el producto de matrices usual.

Sean u, v vectores columna en $\mathbb{R}^n$.

$$u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}, v = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

$$v^\top = \begin{bmatrix} v_1 & v_2 & ... & v_n \end{bmatrix}$$

Entonces:

$$u \cdot v = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} * \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = \sum_{i=1}^{n} u_i v_i$$

$$v^\top u = \begin{bmatrix} v_1 & v_2 & ... & v_n \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} = \sum_{i=1}^{n} v_i u_i$$

Ademas, como el producto vectorial es conmutativo:

$$\begin{aligned} u \cdot v = v \cdot u &\implies \sum_{i=1}^{n} u_i v_i = \sum_{i=1}^{n} v_i u_i \\ &\implies u \cdot v = v^\top u \end{aligned}$$

c) Aplicando el teorema tomando como subespacio S el subespacio del ítem (a), el punto y de $\mathbb{R}^n$ como el vector de la variable dependiente, y el vector b como $b = X\beta^*$, convertir esta ecuación de optimalidad

$$||y - X\beta^*|| = \min_{\beta \text{ en } \mathbb{R}^p} ||y - X\beta||$$

en la condición de ortogonalidad que corresponde a la equivalencia 2 del teorema.

$$\forall \beta \in \mathbb{R}^p, (\mathrm{y} - \mathrm{X}^*) \cdot \mathrm{X} = 0$$

Llamemos $t = X\beta^*$ y $s = X\beta$ donde s son todos los valores en el subespacio $S = X\beta$ con β variando en $\mathbb{R}^p$, y t es el vector $X\beta^*$ que es el vector que minimiza la distancia entre y y $X\beta$. t es la *proyección ortogonal* de y sobre el subespacio S.

Entonces por el Teroema. Sea y un vector cualquiera de $\mathbb{R}^n$ y S un subespacio de $\mathbb{R}^n$. El vector de S que minimiza la distancia del subespacio S al vector y es aquel b de S tal que $y - b$ es ortogonal a todo vector s de S. Es decir, las siguientes dos condiciones son equivalentes:

$$\begin{aligned} ||y - t|| &= \min_{s \in S} ||y - s|| \\ \implies (y - t) \cdot s &= 0, \forall s \in S \\ \implies (y - X\beta^*) \cdot X\beta &= 0, \forall \beta \in \mathbb{R}^p \end{aligned}$$

d) A la ecuación obtenida en el ítem (c), aplicarle la identidad del producto escalar vista en el item (b), para llegar a la ecuación:

$$X^\top (y - X\beta^*) \cdot \beta = 0$$

$$(y - X\beta^*) \cdot X\beta = 0$$

Llamemos $u = y - X\beta^*$ y $v = X\beta$ con $u \in \mathbb{R}^n$ y $v \in \mathbb{R}^n$.

Entonces por la propiedad del producto escalar del item (b):

$$\begin{aligned} u \cdot v = v^\top u &\iff (X\beta)^\top (y - X\beta^*) = 0 \\ &\iff \beta^\top X^\top (y - X\beta^*) = 0 \end{aligned}$$

Ahora llamamos $w^\top = \beta^\top$ y $z = X^\top(y - X\beta^*)$, y por la misma propiedad del item (b):

$$\begin{aligned} w^\top z = z \cdot w &\iff 0 = \beta^\top (X^\top (y - X\beta^*)) \\ &\qquad = (X^\top (y - X\beta^*)) \cdot \beta \end{aligned}$$

$$\iff X^\top (y - X\beta^*) \cdot \beta = 0$$

e) Se sabe que el único vector que es ortogonal a todo vector v de $\mathbb{R}^n$ es el vector nulo. Es decir, si u es un vector fijo tal que $u \cdot v = 0$ para todo v en $\mathbb{R}^n$, entonces $u = 0$. Usando esto y la ecuación obtenida en el ítem (d), llegar a la fórmula:

$$X^\top X\beta^* = X^\top y$$

Partimos de la igualdad obtenida en el ítem (d):

$$\begin{aligned} 0 &= X^\top (y - X\beta^*) \cdot \beta \\ &= (X^\top y - X^\top X\beta^*) \cdot \beta \end{aligned}$$

Fijamos $(X^\top y - X^\top X\beta^*) = 0$ ya que es el único vector que es ortogonal a todo vector β de $\mathbb{R}^p$.

$$\begin{aligned} &(X^\top y - X^\top X\beta^*) = 0 \\ &\iff X^\top X\beta^* = X^\top y \end{aligned}$$

f) Finalmente, suponiendo que las columnas de X son linealmente independientes, se tiene que la matriz $X^\top X$ es invertible. Despejar β^* de la ecuación del ítem (e) para llegar a la fórmula de la solución óptima al problema de regresión.

$$\begin{aligned} X^\top X\beta^* &= X^\top \\ (X^\top X)^{-1} X^\top X\beta^* &= (X^\top X)^{-1} X^\top \\ \beta^* &= (X^\top X)^{-1} X^\top y \end{aligned}$$

0.2 Segunda parte.

En esta sección la idea es realizar regresión lineal en $\mathbb{R}^2$ y analizar como se comportan las soluciones obtenidas.

0.2.1 1. Usando los datos del archivo ejercicio_1.csv:

a) Graficar todos los puntos en el plano xy. Nota: La primer columna del archivo marca el valor de x y la segunda el valor de y de cada punto. Recomendamos usar la biblioteca pandas para leer los archivos con la función read_csv.

[ ]:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from typing import Callable, List, Tuple
from IPython.display import display, HTML, Latex
```

[ ]:
```
data_ej1: pd.DataFrame = pd.read_csv('data/ejercicio_1.csv', sep=',')
x_key: pd.Index = data_ej1.keys()[0]
y_key: pd.Index = data_ej1.keys()[1]

x_values: pd.DataFrame = data_ej1[x_key]
y_values: pd.DataFrame = data_ej1[y_key]
```

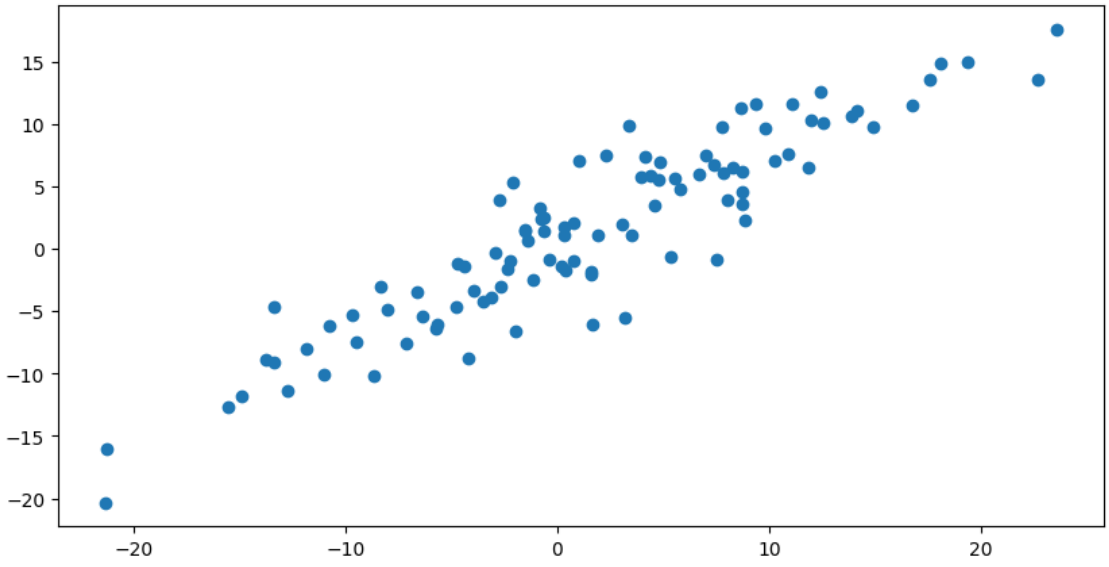
[ ]:
```
display(data_ej1)
```

	X	Y
0	-4.406280	-1.383344
1	-2.722675	3.918974
2	4.610505	3.513816
3	3.510524	1.139324
4	8.767744	4.575692
..	…	…
95	-3.938697	-3.316558
96	18.097452	14.914027
97	7.836851	6.051094
98	-7.145026	-7.592345
99	-9.524871	-7.480263

[100 rows x 2 columns]

[ ]:
```
plt.figure(figsize=(10, 5))
plt.plot(x_values, y_values, 'o')
```

[ ]:
```
[<matplotlib.lines.Line2D at 0x7fbb4267dfc0>]
```



b) Utilizando los conceptos teóricos desarrollados en la primera parte, hallar la recta que mejor aproxima a los datos.

$$||y - X\beta^*|| = \min_{\beta \text{ en } \mathbb{R}^p} ||y - X\beta|| \beta^* = (X^\top X)^{-1} X^\top y$$

[ ]:
```
X: np.ndarray = np.array(x_values).reshape(-1, 1)
y: np.ndarray = np.array(y_values).reshape(-1, 1)
X.shape, y.shape
```

[ ]:
```
((100, 1), (100, 1))
```

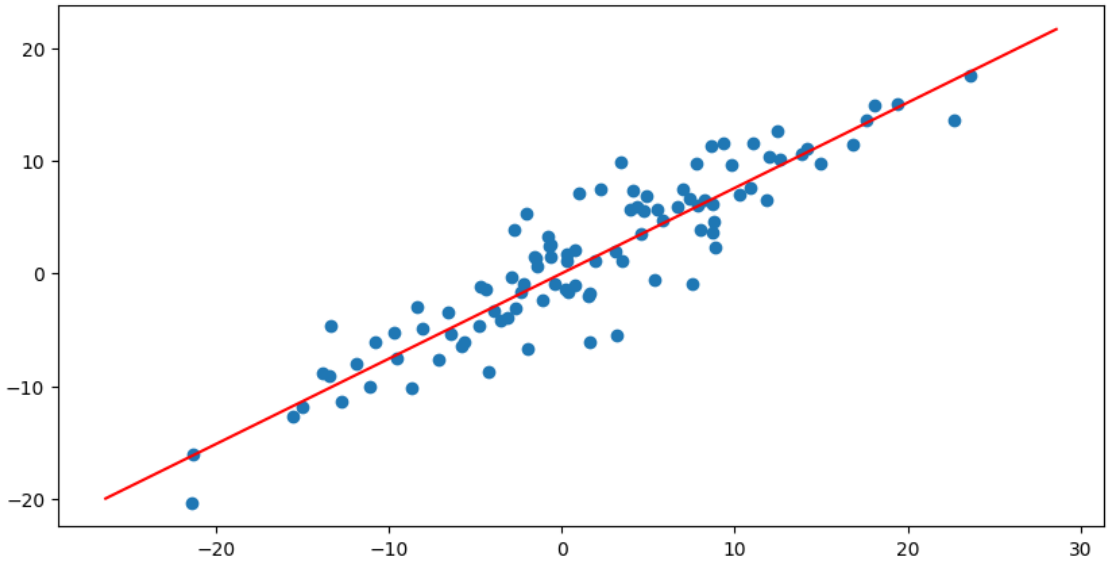
[ ]:
```
def linear_regression(X: np.ndarray, y: np.ndarray) -> np.ndarray:
    """
    Args:
        X (np.ndarray): X values
        y (np.ndarray): y values

    Returns:
        np.ndarray: B values for the linear regression, B = (X^T X)^-1 X^T y,␣
 ↪without 1s in the first column
    """
    X_t = X.T
    X_t_X = np.dot(X_t, X)
    X_t_X_inv = np.linalg.inv(X_t_X)
    X_t_X_inv_X_t = np.dot(X_t_X_inv, X_t)
    B = np.dot(X_t_X_inv_X_t, y)
    return B
```

[ ]:
```
b = linear_regression(X, y)
b
```

[ ]:
```
array([[0.75785414]])
```

$$y = X\beta^*$$
$$y = \hat{\beta}_0 \times x$$
$$y = 0.75785414 \times x$$

[ ]:
```
def plot_regression_line(X: np.ndarray, y: np.ndarray, reg_func: Callable[[np.
↪ndarray, np.ndarray], np.ndarray] = linear_regression):
    """
    Args:
        X (_type_): _description_
        y (_type_): _description_
        reg_func (_type_, optional): _description_. Defaults to␣
↪linear_regression.
    """

    b = reg_func(X, y).flatten()

    plt.figure(figsize=(10, 5))

    plt.plot(X, y, 'o')

    min_x = np.min(X) - 5
    max_x = np.max(X) + 5

    line_x = np.linspace(min_x, max_x, 1000)

    if len(b) == 1:
        line_y = b[0] * line_x
    else:
        line_y = b[0] + b[1] * line_x

    plt.plot(line_x, line_y, 'r')

    plt.show()

plot_regression_line(X, y)
```



c) Realizar nuevamente los incisos (a) y (b) pero considerando los puntos

$$(x_i, y_i + 12) \text{ con } i = 1 \dots n$$

donde (x_i, y_i) eran los puntos originales. ¿Es buena la aproximación realizada?, ¿cuál es el problema?

[ ]:
```
data_12 = data_ej1.copy()
data_12["Y + 12"] = data_12[y_key] + 12
display(data_12)
```

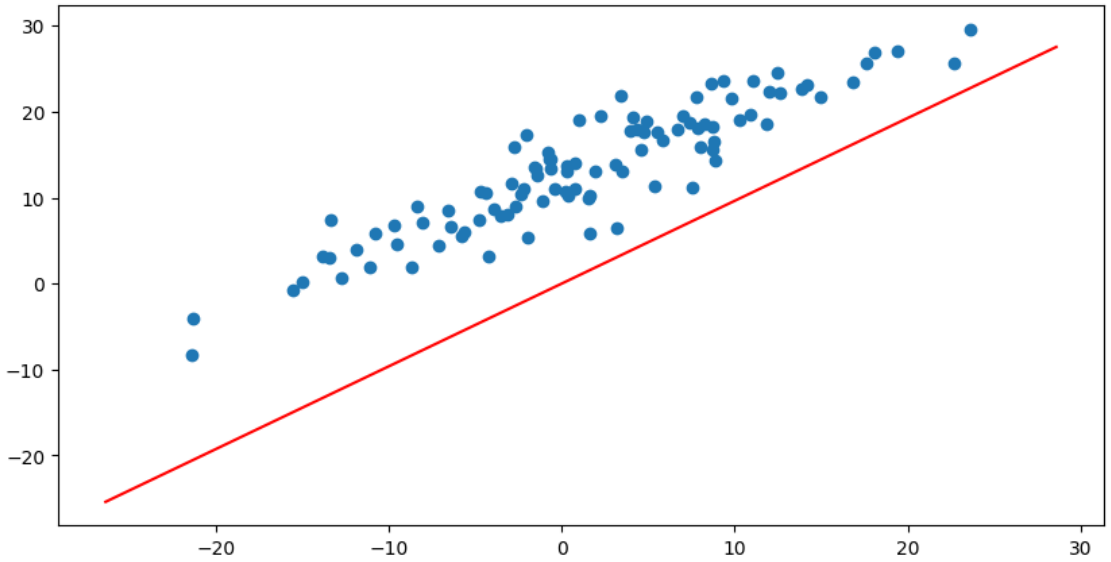
	X	Y	Y + 12
0	-4.406280	-1.383344	10.616656
1	-2.722675	3.918974	15.918974
2	4.610505	3.513816	15.513816
3	3.510524	1.139324	13.139324
4	8.767744	4.575692	16.575692
..	…	…	…
95	-3.938697	-3.316558	8.683442
96	18.097452	14.914027	26.914027
97	7.836851	6.051094	18.051094
98	-7.145026	-7.592345	4.407655
99	-9.524871	-7.480263	4.519737

[100 rows x 3 columns]

[ ]:
```
linear_regression(X, y + 12)
```

[ ]:
```
array([[0.96332903]])
```

[ ]:
```
plot_regression_line(X, y + 12)
```



La recta aproximada no es buena ya que los puntos están alejados de la recta aproximada por 12 unidades. Esto es porque la regresión lineal no tiene en cuenta la ordenada al origen, solo la pendiente.

d) ¿Cómo se podría extender el modelo para poder aproximar cualquier recta en el plano?

Para resolverlo, se puede agregar una columna de 1's a la matriz X y agregar un coeficiente β_0 a la ecuación de la recta. De esta forma, se puede aproximar cualquier recta en el plano.

[ ]:
```
def ones_column(X: np.ndarray) -> np.ndarray:
    """
    Args:
        X (np.ndarray): X values

    Returns:
        np.ndarray: X with 1s in the first column
    """
    ones = np.ones((X.shape[0], 1))
    new_X = np.concatenate((ones, X), axis=1)
    return new_X

def linear_regression_ones(X: np.ndarray, y: np.ndarray) -> np.ndarray:
    """
    Args:
        X (np.ndarray): X values
        y (np.ndarray): y values
```

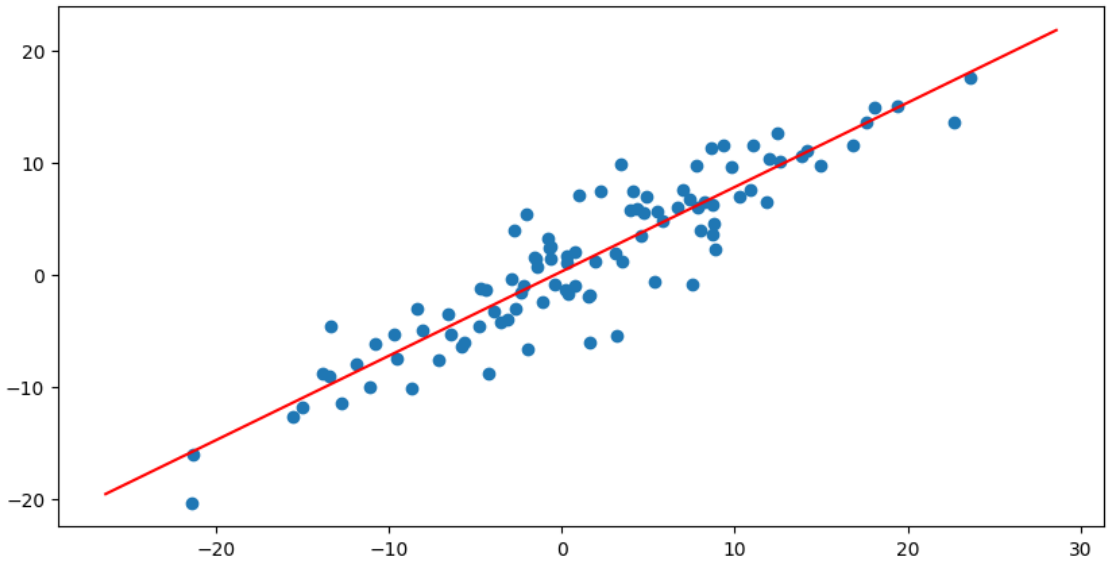
```
    Returns:
        np.ndarray: B values for the linear regression, B = (X^T X)^-1 X^T y,␣
↪with 1s in the first column
    """

    X = ones_column(X)
    X_t = X.T
    X_t_X = np.dot(X_t, X)
    X_t_X_inv = np.linalg.inv(X_t_X)
    X_t_X_inv_X_t = np.dot(X_t_X_inv, X_t)
    B = np.dot(X_t_X_inv_X_t, y)
    return B
```

[ ]:
```
linear_regression_ones(X, y)
```

[ ]:
```
array([[0.28565184],
       [0.75296295]])
```

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$
$$y = 0.28565184 + 0.75296295 \times x$$

[ ]:
```
plot_regression_line(X, y, linear_regression_ones)
```



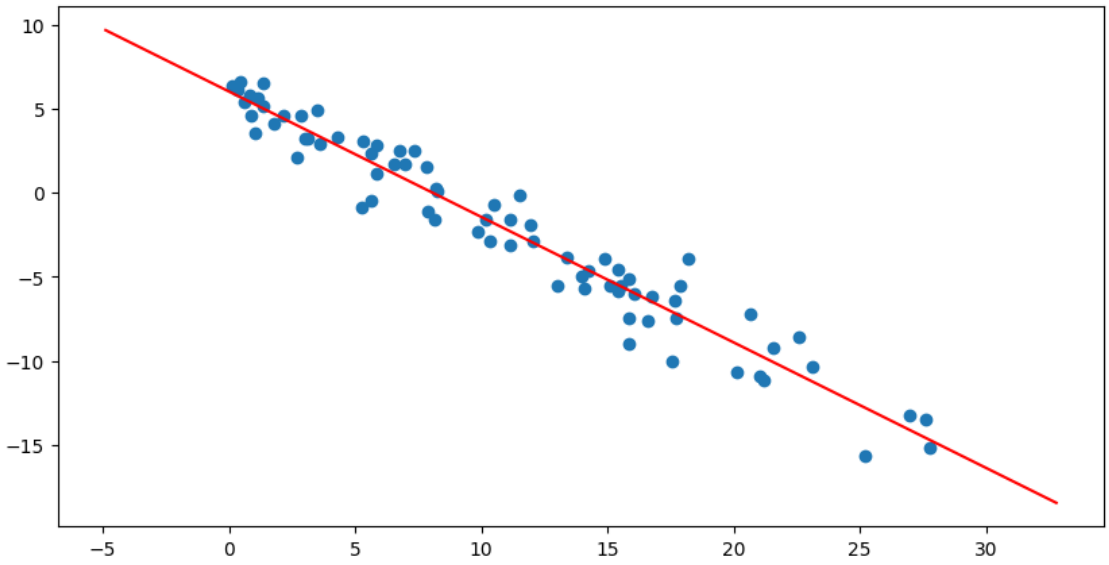
0.2.2 2. Usando los datos del archivo ejercicio_2.csv:

a) Graficar y aproximar los puntos con una recta.

[ ]:
```
data_ej2: pd.DataFrame = pd.read_csv('data/ejercicio_2.csv', sep=',')

x_2_key: pd.Index = data_ej2.keys()[0]
y_2_key: pd.Index = data_ej2.keys()[1]

x_2_values: pd.DataFrame = data_ej2[x_2_key]
y_2_values: pd.DataFrame = data_ej2[y_2_key]

X2: np.ndarray = np.array(x_2_values).reshape(-1, 1)
y2: np.ndarray = np.array(y_2_values).reshape(-1, 1)

display(data_ej2)
```

```
            X         Y
0   15.440218 -5.852276
1   14.256445 -4.605640
2   20.664278 -7.242094
3   17.864908 -5.551524
4    9.858580 -2.320108
..        …       …
70  16.034861 -6.017502
71  11.510946 -0.118044
72   7.839347  1.565136
73   6.522023  1.684089
74  15.850538 -5.161490

[75 rows x 2 columns]
```

[ ]:
```
plot_regression_line(X2, y2, linear_regression_ones)
```



b) Imaginemos que los datos forman parte de mediciones de algún tipo, como por ejemplo la temperatura de un procesador a lo largo del tiempo), y queremos predecir cuál va a ser la temperatura en el futuro. ¿Es buena la aproximación que realizamos?, ¿cuál fue el problema en este caso?

La aproximación realizada puede ser buena para los datos que se tienen, pero no es buena para predecir la temperatura en el futuro ya que los datos pueden no seguir una recta. Por ejemplo, la escala de temperatura no esta definida para menos de 0 grados Kelvin ($-273{,}15$ ºC), por lo que predecir una temperatura en el futuro que sea menor a 0 grados Kelvin no tiene sentido. La temperatura del procesador puede aumentar o disminuir, por lo que no es posible predecir una temperatura en el futuro que sea menor a la temperatura actual.

0.3 Tercera parte. Regresión lineal en datos reales.

En esta sección utilizaremos el conjunto de datos provisto en Machine Learning Repository. Este consiste en datos de ventas de 414 casas en Taiwan. La información provista por casa es (en orden):

i) La fecha en que se realizó la transacción. Expresada en formato

$$\text{año} + \frac{\text{numero\_mes}}{12}$$

ii) La edad de la casa en años.

iii) La distancia a la estación de tren o subte más cercana en metros.

iv) La cantidad de almacenes alcanzables a pie.

v) La latitud en grados.

vi) La longitud en grados.

vii) El precio por Ping. La cual es una unidad utilizada en Taiwan que representa 3,3 metros cuadrados.

Vamos a dividir este conjunto de datos en dos:

i) Datos de entrenamiento: usamos los datos desde la observación 1 a la 315 inclusive.

ii) Datos de test: usamos los datos desde la observación 316 a la 414 inclusive.

[ ]:
```
real_estate: pd.DataFrame = pd.read_excel('data/Real estate valuation data set.
↪xlsx')
real_estate.drop(['No'], axis=1, inplace=True)

short = real_estate.copy()
short.columns = short.columns.map(lambda x: x.split(' ')[0])
display(short)
```

	X1	X2	X3	X4	X5	X6	Y
0	2012.916667	32.0	84.87882	10	24.98298	121.54024	37.9
1	2012.916667	19.5	306.59470	9	24.98034	121.53951	42.2
2	2013.583333	13.3	561.98450	5	24.98746	121.54391	47.3

```
3    2013.500000  13.3   561.98450   5  24.98746  121.54391  54.8
4    2012.833333   5.0   390.56840   5  24.97937  121.54245  43.1
..           …    …         …  ..       …         …   …
409  2013.000000  13.7  4082.01500   0  24.94155  121.50381  15.4
410  2012.666667   5.6    90.45606   9  24.97433  121.54310  50.0
411  2013.250000  18.8   390.96960   7  24.97923  121.53986  40.6
412  2013.000000   8.1   104.81010   5  24.96674  121.54067  52.5
413  2013.500000   6.5    90.45606   9  24.97433  121.54310  63.9

[414 rows x 7 columns]
```

[ ]:
```
X_matrix: np.ndarray = np.array(real_estate.drop(['Y house price of unit␣
 ↪area'], axis=1))
y_matrix: np.ndarray = np.array(real_estate['Y house price of unit area']).
 ↪reshape(-1, 1)

X_matrix.shape, y_matrix.shape
```

[ ]:
```
((414, 6), (414, 1))
```

[ ]:
```
X_train: np.ndarray = X_matrix[0:315]
y_train: np.ndarray = y_matrix[0:315]

X_test: np.ndarray = X_matrix[315:]
y_test: np.ndarray = y_matrix[315:]

X_train.shape, y_train.shape, X_test.shape, y_test.shape
```

[ ]:
```
((315, 6), (315, 1), (99, 6), (99, 1))
```

1. Teniendo en cuenta la teoría desarrollada en la primer parte del trabajo práctico y usando los datos de entrenamiento:

a) Estimar los parámetros $\hat{\beta}$ que minimizan el error cuadrático medio para este problema.

[ ]:
```
b_top: np.ndarray = linear_regression_ones(X_train, y_train)
b_top, b_top.shape
```

[ ]:
```
(array([[-1.48772734e+04],
        [ 4.96499612e+00],
        [-2.79856106e-01],
        [-4.29355898e-03],
        [ 1.11458810e+00],
        [ 2.63777794e+02],
        [-1.36678356e+01]]),
 (7, 1))
```

b) Encontrar $\hat{y}$ la estimación de la variable de respuesta.

[ ]:
```
X_train_ones: np.ndarray = ones_column(X_train)
y_pred: np.ndarray = np.dot(X_train_ones, b_top)
y_pred.shape
```

[ ]:
```
(315, 1)
```

c) ¿Cuánto vale el error cuadrático medio?

Definimos error cuadrático medio como

$$\text{ECM} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

donde y_i son observaciones de una variable y $\hat{y}_i$ estimaciones de las mismas.

[ ]:
```
def ECM(y: np.ndarray, y_pred: np.ndarray) -> float:
    """
    Args:
        y (np.ndarray): Valores reales
        y_pred (np.ndarray): Valores predichos

    Returns:
        float: Error Cuadrático Medio (ECM) entre "y" e "y_pred"
    """
    return np.sum((y - y_pred)**2) / len(y)

ECM(y_train, y_pred)
```

[ ]:
```
83.16321956992414
```

2. Utilizando los datos de test, analizar cuál es el error cuadrático medio al utilizar los parámetros $\widehat{\beta}$ estimados en el punto anterior.

[ ]:
```
X_test_ones: np.ndarray = ones_column(X_test)
y_eval: np.ndarray = np.dot(X_test_ones, b_top)
ECM(y_test, y_eval), y_eval.shape, X_test_ones.shape
```

[ ]:
```
(58.66059708585313, (99, 1), (99, 7))
```

a) ¿Es la estimación mejor que sobre los datos originales?, ¿a qué se debe la discrepancia?

La estimación es mejor que sobre los datos originales ya que el error cuadrático medio es menor. Esto se debe a que los datos de entrenamiento son más cercanos a los datos de test que a los datos "originales".

b) ¿Qué sucede con el ECM del segundo conjunto de casas si se realiza la regresión sobre todos los datos al mismo tiempo (es decir, las 414 casas)?

[ ]:
```
b_: np.ndarray = linear_regression_ones(X_matrix, y_matrix)
b_, b_.shape
```

```
[ ]: (array([[-1.44419817e+04],
             [ 5.14901681e+00],
             [-2.69696734e-01],
             [-4.48750826e-03],
             [ 1.13332498e+00],
             [ 2.25470141e+02],
             [-1.24290622e+01]]),
      (7, 1))
```

```
[ ]: y_pred_test: np.ndarray = np.dot(X_test_ones, b_)
     ECM(y_test, y_pred_test)
```

```
[ ]: 57.38793541970973
```

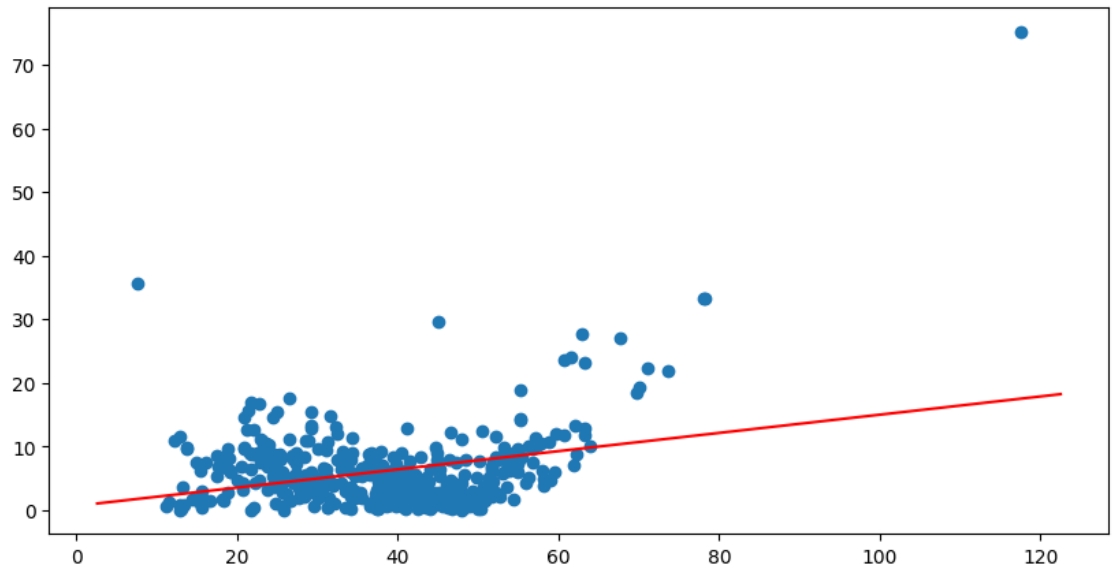
3. Graficar el error cometido por cada casa. Es decir el valor absoluto de la diferencia entre el precio por Ping real y el estimado.

```
[ ]: X_matrix_ones: np.ndarray = ones_column(X_matrix)

     y_pred: np.ndarray = np.dot(X_matrix_ones, b_)
```

Graficamos el error cometido por cada casa.

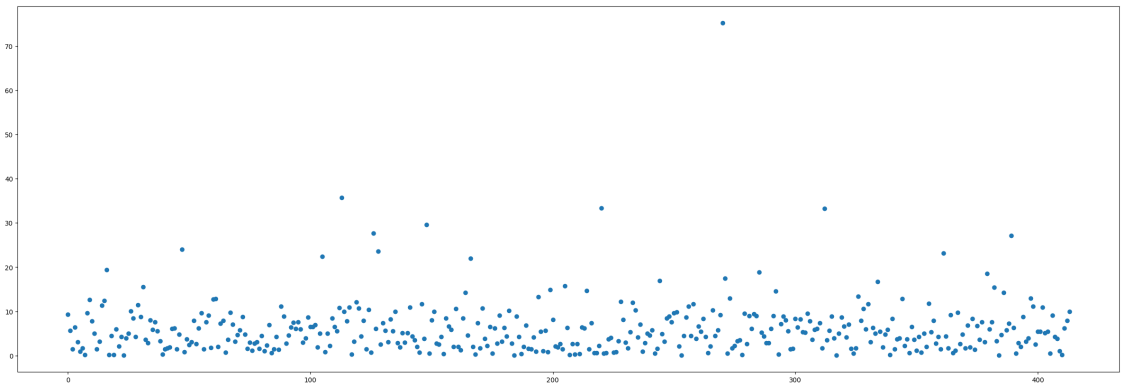
Error sobre precio de vivienda

```
[ ]: plot_regression_line(y_matrix, abs(y_matrix - y_pred), linear_regression_ones)
```



Error por cada vivienda

```
[ ]: plt.figure(figsize=(30, 10))
     plt.scatter(range(len(X_matrix)), abs(y_matrix - y_pred))
     plt.show()
```



4. Imaginemos que se agrega una nueva columna a los datos que informa el año en que la misma fue construida. ¿Disminuiría esto el ECM?

Como el año en el que la casa fue construida se puede obtener como combinación linea de la fecha de la transacción y la edad de la casa, se puede decir que son variables colineales. Por lo tanto, no creemos que aporte información adicional relevante para el modelo, por lo que no se espera que disminuya el ECM.

Podemos evaluar esto de forma empirica, entrenando un modelo con la nueva columna y comparando los ECM contra el modelo anterior.

$$\text{Año de construcción} = \text{Fecha de transacción} - \text{Edad de la casa}$$

```
[ ]: trans_date: np.ndarray = np.array(real_estate['X1 transaction date']).
      ↪reshape(-1, 1)
     age: np.ndarray = np.array(real_estate['X2 house age']).reshape(-1, 1)
     anio: np.ndarray = trans_date - age

     X_ext = np.concatenate((X_matrix, anio), axis=1)
```

```
[ ]: b_ext: np.ndarray = linear_regression_ones(X_ext, y_matrix)
     b_ext
```

```
[ ]: array([[-1.60413709e+04],
            [-4.39352583e+01],
            [ 4.95597342e+01],
            [-4.48750637e-03],
            [ 1.13332499e+00],
            [ 2.25470187e+02],
            [-1.24288921e+01],
```

```
       [ 4.98621799e+01]])
```

[ ]:
```
y_pred_ext: np.ndarray = np.dot(ones_column(X_ext), b_ext)
ECM(y_matrix, y_pred_ext)
```

[ ]:
```
1227.1544998065524
```

Podemos ver que el ECM para el conjunto total de datos original X es menor que el ECM para el conjunto de datos X' extendido que incluye la columna de año de construcción. Por lo tanto, la colinealidad de la variable año de construcción respecto al resto de las variables del modelo afecta negativamente al ECM.

0.4 Conclusiones

El objetivo del trabajo práctivo era el de implementar las funciones necesarias para realizar una regresión lineal y aplicarlas a un conjunto de datos reales. A su vez, se implementaron funciones para calcular el error cuadrático medio y graficar el error cometido por cada casa.

Se pusieron en práctica las funciones para predecir el precio de una vivienda en Taiwan. Para hacer esto se utilizaron los datos provistos por el Machine Learning Repository y se entrenó un modelo con los datos de entrenamiento y se evaluó con los datos de test.

Se obtuvieron los siguientes resultados:

Datos	$\hat{\beta}$	$\hat{y}$	ECM
Train	`lin_reg(X_train, y_train)`	X train $\cdot\, \hat{\beta}$	83.16321956992414
Test	`lin_reg(X_train, y_train)`	X test $\cdot\, \hat{\beta}$	58.66059708585313
Full	`lin_reg(X, y)`	$X \cdot \hat{\beta}$	57.38793541970973

Por ultimo se observó que el ECM para los datos de test es menor que el ECM para los datos de entrenamiento, lo que indica que el modelo es bueno para predecir el precio de una vivienda en Taiwan.

0.5 Bibliografía

[1] Machine Learning Repository

[2] Linear Regression

[3] Mean squared error