Q 2.1.1

From dataset the dependent variable of the samples y_n is the total in flow for each station

$$Y = \begin{bmatrix} 3459623 & 3914019 & 8100630 & 13460142 & 2535732 \end{bmatrix}$$

the independent variable of the samples x_n consists of 4-dimensional data& which extends its value in the axis of Total Population near each station (x_{n1}) number of households that own 0 vehicles (x_{n2}) total employment (x_{n3}) and total road network density (x_{n4}). In each axis& the x_{nj} is a column vector containing the data value of each sample in the dataset.

$$X = \begin{bmatrix} x_{n1}^T & x_{n2}^T & x_{n3}^T & x_{n4}^T \end{bmatrix}$$
$$= \begin{bmatrix} 50383 & 4784 & 28318 & 28.7 \\ 11084 & 1664 & 33120 & 42.23 \\ 51122 & 16059 & 61815 & 36.3 \\ 25970 & 5383 & 181995 & 40.15 \\ 29222 & 2891 & 23981 & 31.3 \end{bmatrix}$$

Adding a column vector of 1 to the independent variable matrix to introduct an intercept to the model

$$X = \begin{bmatrix} 1 & 50383 & 4784 & 28318 & 28.7 \\ 1 & 11084 & 1664 & 33120 & 42.23 \\ 1 & 51122 & 16059 & 61815 & 36.3 \\ 1 & 25970 & 5383 & 181995 & 40.15 \\ 1 & 29222 & 2891 & 23981 & 31.3 \end{bmatrix}$$

Since the model is built as a linear regression model, there is the coefficient matrix w, which by taking dot product to the independent variable will give a close proximation to the dependent variable

$$y_n \approx w^T x_n$$

We obtains the close proximation of the model by maximum likeihood estimate, i.e. minimizing the error function

$$E(w) = \frac{1}{2} \sum_{n=1}^{N} [y_n - w^T x_n]^2$$

Using the close form solution formula (25) for lecture 2, obtained by taking derivitive to the error function and setting it equal 0

$$w = (X^T X)^{-1} X^T Y$$
$$= [5288434.545, 39.269, 127.973, 59.180, 156149.881]^T$$

Q2.1.2

Declare the objective matrix and the sample matrix

In [1]:
```
import numpy as np
np.set_printoptions(suppress=True)

X1 = np.array([[1,50383,4784,28318,28.7],[1,11084,1664,33120,42.23],
               [1,51122,16059,61815,36.3],[1,25970,5383,181995,40.15],
               [1,29222,2891,23981,31.3]])
Y1 = np.array([3459623,3914019,8100630,13460142,2535732]).reshape(-1,1)
```

Solving the MLE with the matrix product

In [6]:
```
g1 = X1.T@X1
g2 = np.linalg.inv(g1)
g3 = X1.T @ Y1
w = np.linalg.inv(X1.T@X1)@X1.T @ Y1
print("The w matrix with 5 variables [w0, w1, w2, w3, w4] is " , w.flatten())
```

```
The w matrix with 5 variables [w0, w1, w2, w3, w4] is  [-5288434.54549981       39.26866197      127.97287239
59.18005265
   156149.88142737]
```

Solving the MLE with the sklearn regression model

In [7]:
```
from sklearn.linear_model import LinearRegression

X2 = np.array([[50383,4784,28318,28.7],[11084,1664,33120,42.23],
               [51122,16059,61815,36.3],[25970,5383,181995,40.15],
               [29222,2891,23981,31.3]])
Y2 = np.array([3459623,3914019,8100630,13460142,2535732])

reg = LinearRegression().fit(X2, Y2)

print('[w1, w2, w3, w4] is', reg.coef_)
print('w0 is', reg.intercept_)
```

```
[w1, w2, w3, w4] is [    39.26866197    127.97287239     59.18005265 156149.88142737]
w0 is -5288434.545499415
```

Q2.1.3

In [4]:
```
X3 = np.array([1, 34689, 9443, 148355, 38.9])
Y3 = w.T@X3
print('estimated rideship inflow at Montgomery is', Y3[0])
```

```
estimated rideship inflow at Montgomery is 12136091.00253777
```

Q2.2.1

True

Q2.2.2

True

Q2.2.3

from definition, $\hat{y} = Xw$

the difference vector of $\hat{y}$ and y = $y - \hat{y} = y - Xw$

Since $y - \hat{y}$ is orthogonal to X,

$$X^T(y - \hat{y}) = 0$$
$$X^T(y - Xw) = 0$$
$$X^Ty - X^TXw = 0$$
$$X^TXw = X^Ty$$
$$w = (X^TX)^{-1}X^Ty$$