ÉCOLE D'INGÉNIEURS

Régression linéaire Interpolation polynomiale

Lotfi HOCINI

Régression linéaire

Descriptive Statistics

- Collection of data;
- Organization of these data into tables;
- Interpretation by means of graphs and numerical parameters.

- **Population** Set being studied;
- **Individual** Each element of the population from which data is collected;
- **Sample** The part of the population from which data is collected;
- **Statistical Variable** some aspect of the individual in the sample.

One-Dimensional Statistics

We deal with just a single statistical variable.

- Statistical graphs;
- Statistical parameters: Mean, Median, Mod, Std-deviation, Variance, ...

Two-Dimensional Statistics and Linear Regression

- N individuals
- Each individual is represented by a couple (x_i, y_i)

The population can be presented by two vectors X and Y

- $X - [x_1, x_2, ..., x_N]$
- $Y - [y_1, y_2, ..., y_N]$

Example

Study of height and weight of N students.
- Response to questions like: is there a correlation between their height and weight?

Two-Dimensional Statistics and Linear Regression

The statistical variables can be studied separately (1D Statistics):

- $X \rightarrow \overline{X}, V_x, \sigma_x \ldots$
- $Y \rightarrow \overline{Y}, V_y, \sigma_y \ldots$

In 2D-Statistics we study the two statistical variables simultaneously:

- Representing the data cloud;
- Studying relationship between the two variables;
- Predicting one variable when knowing the other ...

Two-Dimensional Statistics and Linear Regression

Example

Considering the height (inch) and weight (pound) of a population of students (see .CSV file).

	Gender	Height	Weight
0	Male	73.847017	241.893563
1	Male	68.781904	162.310473
2	Male	74.110105	212.740856
3	Male	71.730978	220.042470
4	Male	69.881796	206.349801

Covariance

$X, Y \longrightarrow (x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$ N individuals

x_i	y_i	x_i^2	y_i^2	$x_i \cdot y_i$
x_1	y_1	x_1^2	y_1^2	$x_1 \cdot y_1$
x_2	y_2	x_2^2	y_2^2	$x_2 \cdot y_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_N	y_N	x_N^2	y_N^2	$x_N \cdot y_N$
$\sum x_i$	$\sum y_i$	$\sum x_i^2$	$\sum y_i^2$	$\sum x_i \cdot y_i$

$$\bar{X} = \frac{\sum x_i}{N} \quad \sigma_X^2 = \frac{\sum x_i^2}{N} - \bar{X}^2 \quad \sigma_X$$

$$\bar{Y} = \frac{\sum y_i}{N} \quad \sigma_Y^2 = \frac{\sum y_i^2}{N} - \bar{Y}^2 \quad \sigma_Y$$

- $\sigma_{XY} = \dfrac{\sum x_i \cdot y_i}{N} - \bar{X} \cdot \bar{Y}$ Covariance

Covariance

- $\sigma_{XY} = \frac{\sum x_i \cdot y_i}{N} - \bar{X} \cdot \bar{Y}$ Covariance

Interpretation:

i) Positive and large values of σ_{XY} indicate the tendency that if one variable increases, so does the other.

ii) Negative and large values of σ_{XY} indicate a tendency that when the value of one of the variables increases, the other decreases.

iii) Values of σ_{XY} close to 0 indicate that there little relationship between the variables.

σ_{XY} is sensitive to changes in scale.

Correlation coefficient

Correlation coefficient:

$$r = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y} \qquad -1 \leq r \leq 1$$

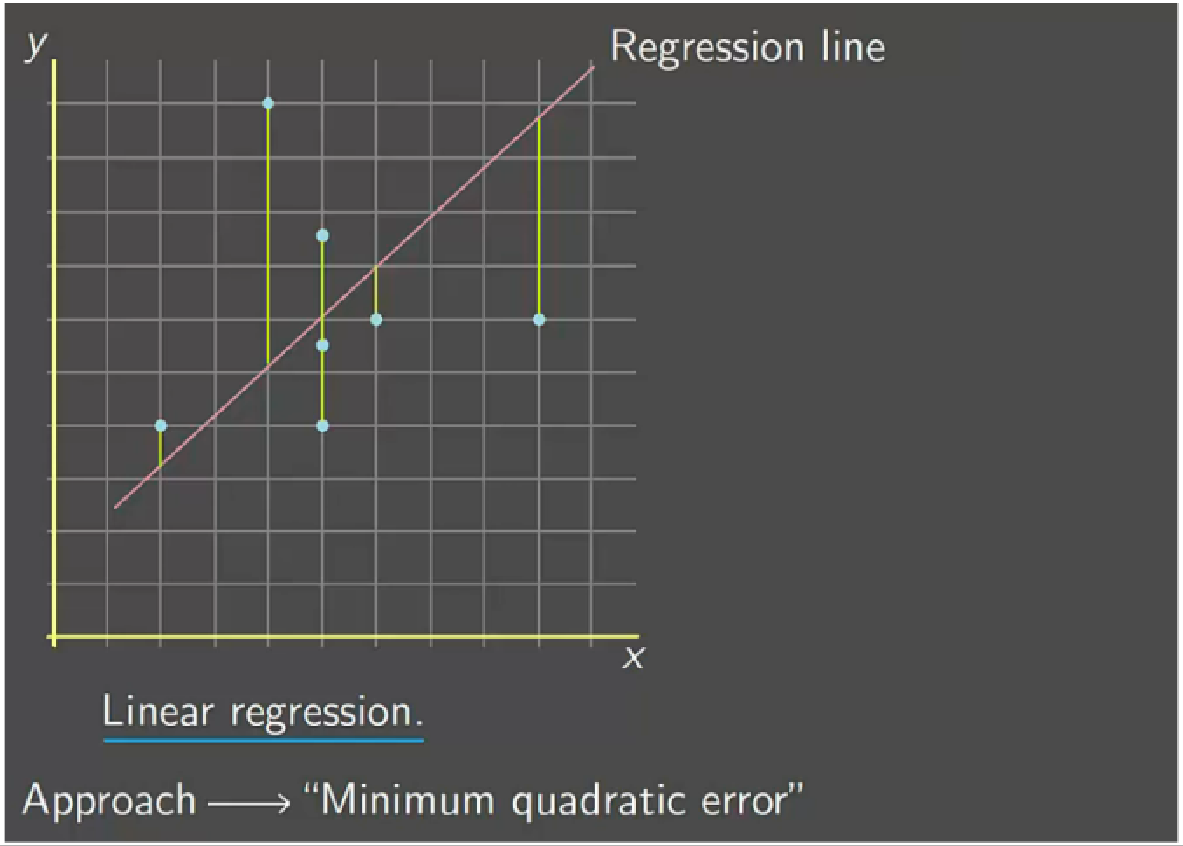
Interpretation:

i) If r is close to 1 or -1 this indicates that the regression line approximates well to the point cloud.

- r close to 1 $\Rightarrow$ Direct correlation.
- r close to -1 $\Rightarrow$ Inverse correlation.

ii) If r is close to 0, then the variables X and Y are essentially independent.

Linear Regression



Linear Regression

Linear regression.

- Regression line of Y on X:

$$Y - \bar{Y} = \frac{\sigma_{XY}}{\sigma_X^2}(X - \bar{X})$$

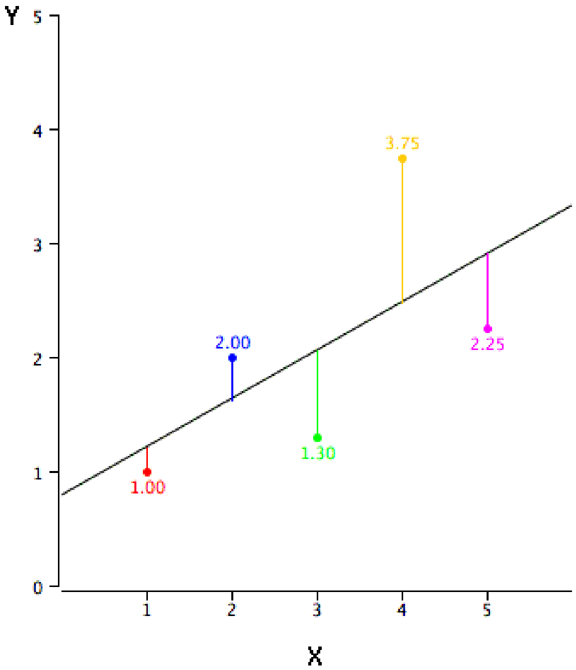
Allows us to obtain predictions of Y when we have values for X.

- Regression line of X on Y:

$$X - \bar{X} = \frac{\sigma_{XY}}{\sigma_Y^2}(Y - \bar{Y})$$

Allows us to obtain predictions of X when we have values for Y.

Linear Regression using linear algebra



Errors $\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$

To minimize $\sum \varepsilon_i{}^2$ is least squares

Linear Regression using linear algebra

In the diagram, errors are represented by red, blue, green, yellow, and the purple line correspondingly. To formulate this as a matrix solving problem, consider linear equation is given below, where Beta 0 is the intercept and Beta is the slope.

$$\beta_0 + X\,\vec{\beta} = \vec{y}$$

To simplify this notation, we will add Beta 0 to the Beta vector. This is done by adding an extra column with 1's in X matrix and adding an extra variable in the Beta vector. Consequently, the matrix form will be:

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ & . \\ & . \\ & . \\ 1 & x_n \end{bmatrix} x \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix}$$

Linear Regression using linear algebra

Then the least square matrix problem is:

$$\begin{bmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ . \\ . \\ . \\ \beta_0 + \beta_1 x_n \end{bmatrix} \textit{ is close to } \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix}$$

Linear Regression using linear algebra

Let us consider our initial equation:

$$X\,\vec{\beta} = \vec{y}$$

Multiplying both sides by X_transpose matrix:

$$X^T X\,\vec{\beta} = X^T\vec{y}$$

Where:

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ x_1 & x_2 & x_3 & \dots & x_n \end{bmatrix} x \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ & . \\ & . \\ & . \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} N & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}$$

$$X^T\vec{y} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ x_1 & x_2 & x_3 & \dots & x_n \end{bmatrix} x \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} \sum y \\ \sum X_i y_i \end{bmatrix}$$

Linear Regression using linear algebra

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

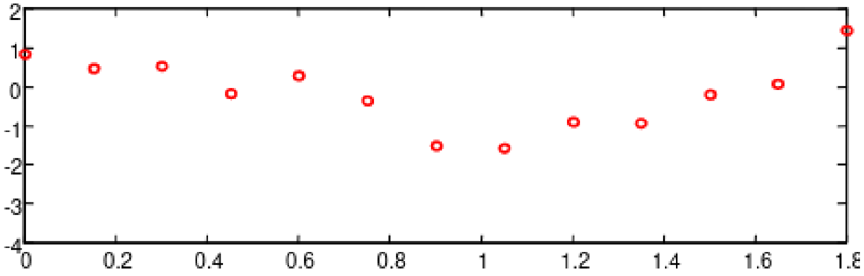
$$\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} N & \sum X_i \\ \sum X_i & \sum x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum y \\ \sum X_i y_i \end{bmatrix}$$

```
import numpy as np

X = np.matrix([[1, 1],
               [1, 2],
              [1, 3],
              [1, 4]])
print(X)

XT = np.matrix.transpose(X)
print(XT)

y = np.matrix([[1],
               [3],
              [3],
              [5]])
print(y)

XT_X = np.matmul(XT, X)
print(XT_X)

XT_y = np.matmul(XT, y)
print(XT_y)

betas = np.matmul(np.linalg.inv(XT_X), XT_y)
print(betas)
```

Approximation de fonctions (Interpolation)

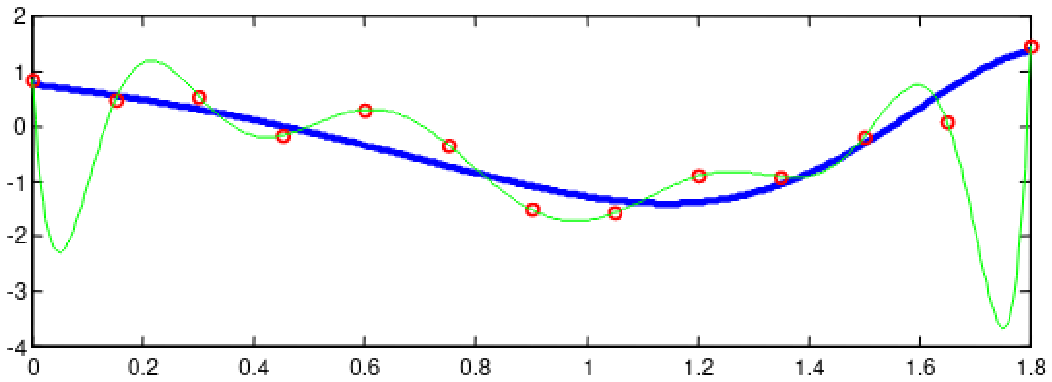
Approximation de fonctions

- Soit une fonction *f* (inconnue explicitement)
  - connue seulement en certains points $x_0, x_1 \ldots x_n$
  - ou évaluable par un calcul coûteux.
- Principe :
  - représenter *f* par une fonction simple, facile à évaluer
- Problème :
  - il existe une infinité de solutions !



Approximation de fonctions

- Il faut se restreindre à une famille de fonctions
  - polynômes,
  - exponentielles,
  - fonctions trigonométriques…



Interpolation polynomiale - Position du problème

On se donne le tableau de données suivant

i	x_i	y_i
0	x_0	y_0
$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n

Définition

On cherche un polyôme P_n de degré au plus n ($P_n \in \mathcal{P}_n$)tel que

$$P_n(x_i) = y_i, \quad \text{pour } i = 0, \ldots, n.$$

POLYNÔMES DE LAGRANGE

Théorème 1 (POLYNÔMES DE LAGRANGE)

Pour tout choix de nœuds $x_0, x_1, ..., x_n$ dans $[a, b]$, il existe un unique polynôme P_n de degré inférieur ou égal à n qui coïncide avec f aux points $x_0, x_1, ..., x_n$ (*i. e.* $P(x_j) = f(x_j)$ pour tout $j = 0, ..., n$).
Ce polynôme s'écrit

$$P_n(x) = \sum_{j=0}^{n} f(x_j) L_j(x), \tag{1.1}$$

où

$$L_j(x) = \prod_{\substack{k=0, \\ k \neq j}}^{n} \frac{x - x_k}{x_j - x_k}.$$

pour tout $j = 0, ..., n$.

POLYNÔMES DE LAGRANGE
- Remarque

1. *Les polynômes de Lagrange sont tels que*

$$L_j(x_k) = \delta_{jk} = \begin{cases} 1, & si \;\; j = k, \\ 0, & sinon. \end{cases}$$

   on rappelle que δ_{jk} est appelé symbole de Kronecker.

2. *L'écriture (1.1) n'est pas utilisée en pratique. On ne peut pas calculer facilement le polynôme d'interpolation de f aux point $x_0, x_1, ..., x_n$ à partir du polynôme d'interpolation aux nœuds $x_0, x_1, ..., x_n$ étant donné que chacun des L_j dépend de tous les nœuds.*

Il existe une autre forme, plus pratique à utiliser : la forme de Newton.

MÉTHODE DE NEWTON

Théorème 2 (MÉTHODE DE NEWTON)

Pour tout $n \in \mathbb{N}^*$, pour tous nœuds $x_0, x_1, \ldots, x_n$ dans $[a, b]$, il existe un unique polynôme P_n de degré inférieur ou égal à n qui coïncide avec f aux points $x_0, x_1, \ldots, x_n$ (*i.e.* $P(x_j) = f(x_j)$ pour tout $j = 1, \ldots, n$).
Ce polynôme s'écrit

$$P_n(x) = f[x_0] + f[x_0, x_1](x - x_0) + \ldots + f[x_0, \ldots, x_n] \prod_{k=0}^{n-1} (x - x_k).$$

MÉTHODE DE NEWTON

Pour construire les coefficients de Newton nous procédons de la façon suivante :

$$a_0 = f(x_0)$$

$$a_1 = f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

$$a_2 = f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} = \frac{\dfrac{f(x_2) - f(x_1)}{x_2 - x_1} - \dfrac{f(x_1) - f(x_0)}{x_1 - x_0}}{x_2 - x_0}$$

$$\vdots$$

$$a_n = f[x_0, x_1, \cdots, x_n] = \frac{f[x_1, x_2, \cdots, x_n] - f[x_0, x_1, x_2, \cdots, x_{n-1}]}{x_n - x_0}$$

MÉTHODE DE NEWTON

$$
\begin{array}{llll}
x_1 & y_1 & & \\
 & & y_{2,1} = \dfrac{y_2 - y_1}{x_2 - x_1} & \\
x_2 & y_2 & & y_{3,2,1} = \dfrac{y_{3,2} - y_{2,1}}{x_3 - x_1} \\
 & & y_{3,2} = \dfrac{y_3 - y_2}{x_3 - x_2} & \\
x_3 & y_3 & & y_{4,3,2} = \dfrac{y_{4,3} - y_{3,2}}{x_4 - x_2} \\
 & & y_{4,3} = \dfrac{y_4 - y_3}{x_4 - x_3} & \\
x_4 & y_4 & &
\end{array}
\qquad y_{4,3,2,1} = \dfrac{y_{4,3,2} - y_{3,2,1}}{x_4 - x_1}
$$

$$y_{k,\cdots,j} = \frac{y_{k,\cdots,j+1} - y_{k-1,\cdots j}}{x_k - x_j}$$

$$p(x) = y_1 + y_{2,1}(x - x_1) + y_{3,2,1}(x - x_1)(x - x_2) + y_{4,3,2,1}(x - x_1)(x - x_2)(x - x_3)$$

MÉTHODE DE NEWTON

$n=2$ $(0,1)$, $(2,5)$ et $(4,17)$

MÉTHODE DE NEWTON

n=2 (0,1), (2,5) et (4,17)

	a_0	a_1	a_2
0	$f[x_0]=1$		
2	$f[x_1]=5$	$f[x_0, x_1] = (1-5)/(0-2)=2$	
4	$f[x_2]=17$	$f[x_1, x_2] = (5-17)/(2-4)=6$	$f[x_0, x_1, x_2] = (2-6)/(0-4)=1$

$p(x)=1 + 2x + x(x-2)$ (et on retombe sur $p(x) = 1 + x^2$)