Régression linéaire

- **La régression simple**

$$Y,\ X \ : \Omega \longrightarrow \mathbb{R}$$

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- La régression **multiple**

$$Y,\ \underbrace{X_1\ , X_2, \ldots X_p}$$

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

But :

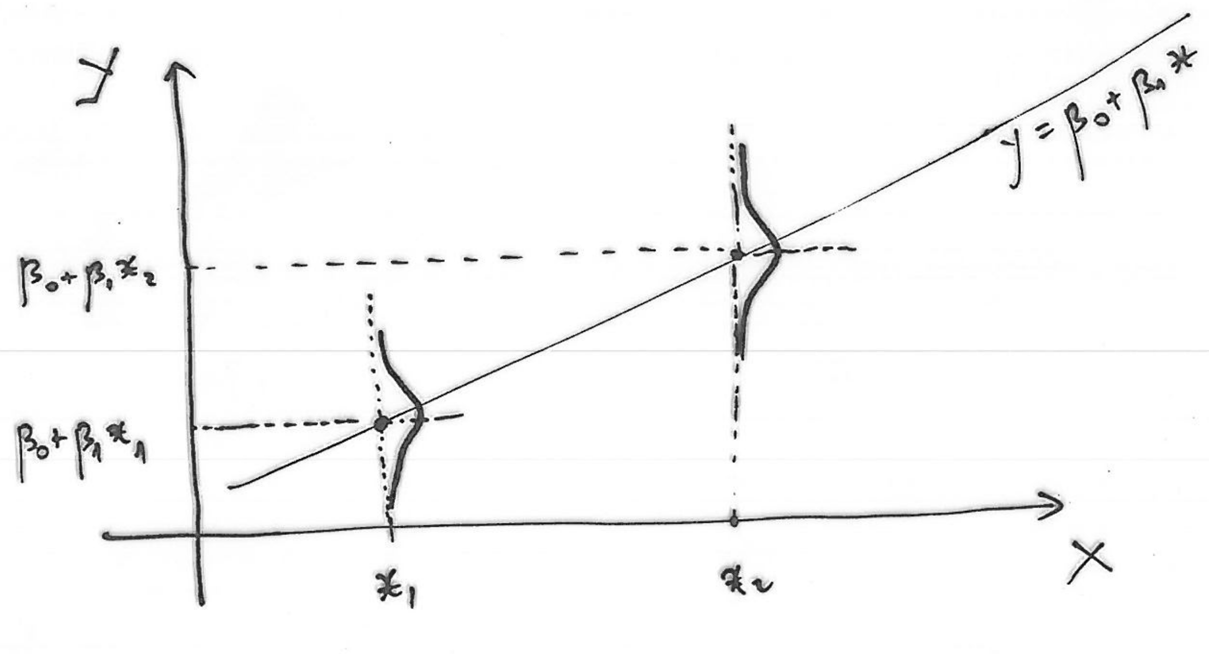
- estimation des $\{\beta_i\}_{i \geq 0}$
- qualité du modèle

La régression simple

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Pour $X = x$ fixe

$$\begin{cases} Y|_{X=x} = \beta_0 + \beta_1 x + \underbrace{\varepsilon} \\ \qquad \begin{cases} E(\varepsilon) = 0 \\ Var(\varepsilon) = \sigma^2 \end{cases} \\ E(Y|X=x) = \beta_0 + \beta_1 x \end{cases}$$



Notons par $\underline{\hat{y} = \beta_0 + \beta_1 X}$

Alors $y = \underbrace{\hat{y}}_{\text{modèle}} + \underbrace{\varepsilon}_{\text{aléa}}$

<u>Estimation de $\beta_0, \beta_1, \sigma^2$</u>

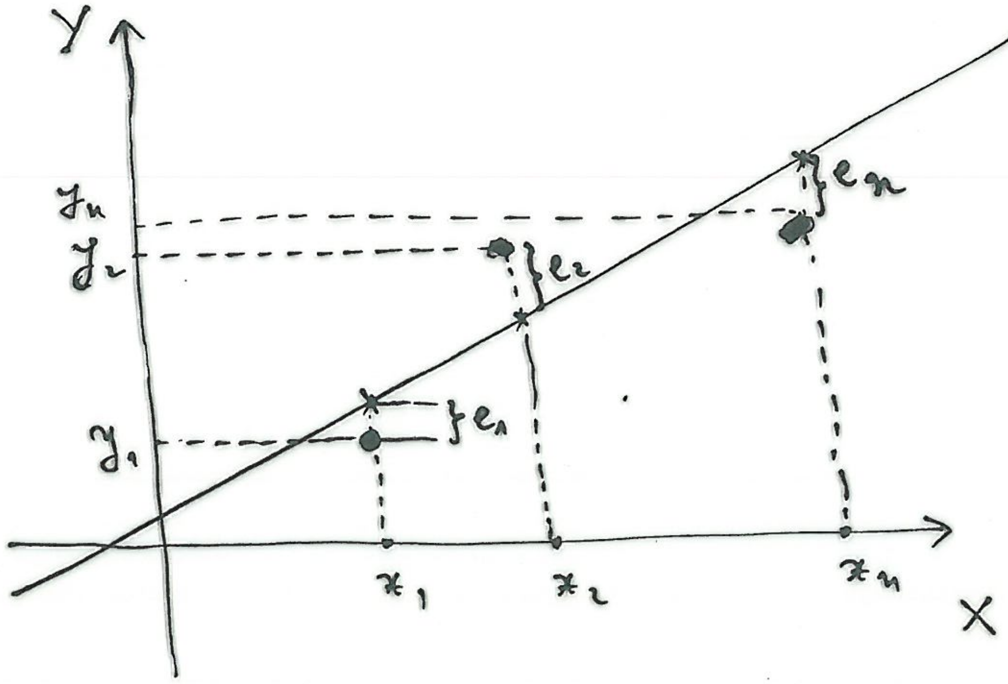
$(y_1, x_1), \ \ldots \ (y_n, x_n) \ , \ n \geq 1$

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_1 + e_1 \\ y_2 = \beta_0 + \beta_1 x_2 + e_2 \\ \vdots \\ y_n = \beta_0 + \beta_1 x_n + e_n \end{cases}$$

Moindres carrés :

Chercher β_0, β_1 t.q :

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_i) \right]^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$



$$\frac{\partial L}{\partial \beta_0} = 0 \qquad \frac{\partial L}{\partial \beta_1} = 0$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \text{avec} \begin{cases} \bar{y} = \frac{1}{n} \sum y_i \\ \bar{x} = \frac{1}{n} \sum x_i \end{cases}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2$$

Th : $\underline{\hat{\beta}_0, \hat{\beta}_1 \text{ et } \hat{\sigma}^2}$ sont estimateurs sans biais de $\underline{\beta_0, \beta_1, \sigma^2}$.

$$\begin{cases} E(\hat{\beta}_0) = \beta_0 \\ Var(\hat{\beta}_0) = \sigma^2 \cdot \left(\dfrac{1}{n} + \dfrac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right) \end{cases}$$

$$\begin{cases} E(\hat{\beta}_1) = \beta_1 \\ Var(\hat{\beta}_1) = \sigma^2 \left(\dfrac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right) \end{cases}$$

Remarque

Soit $s_y^2 = \dfrac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2$

$$s_x^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

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

$$r = \frac{1/n \sum (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n} \sqrt{\sum (x_i - \bar{x}_i)^2 \cdot \sum (y_i - \bar{y})^2}} = \frac{cov(x,y)}{s_x \cdot s_y}$$

$$s_e^2 = (1 - r^2)\, s_y^2$$

ou

$$\underbrace{s_y^2}_{\text{var totale}} = \underbrace{s_e^2}_{\text{var residuelle}} + \underbrace{r^2 s_y^2}_{\text{variance expliquée par le modèle}}$$

On a également :

$$\underbrace{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} + \underbrace{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}$$

Inférence

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{V}ar(\hat{\beta}_1)}} \dot{\sim} T_{n-2}$$

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{Var(\hat{\beta}_0)}} \sim T_{n-2}$$

Rem : Si $\rho = \frac{Cov(x,y)}{\sqrt{V(x)\cdot V(y)}} = 0$ alors $\beta_1 = 0$

$$\frac{\hat{\beta}_1}{\sqrt{Var(\hat{\beta}_1)}} \sim T_{n-2}$$

$$\Downarrow$$

$$\boxed{\frac{r}{\sqrt{1-r^2}}\sqrt{n-2} \sim T_{n-2}}$$

Test pour $r=0$!

Tests dans le modèle linéaire

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

- à l'aide de Student
- utilisant la décomposition de la variance.

$$\frac{1}{\sigma^2} \times \left| \quad \underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\chi^2_{n-1}} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\substack{\text{si } \beta_1 = 0 \\ \chi^2}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\chi^2_{n-2}} \right.$$

$$\boxed{\frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}(n-2) \sim \mathcal{F}(1, n-2) \quad \beta_1 = 0}$$

$$\| $$

$$\frac{r^2}{1 - r^2}(n-2)$$

Etude des résidus

$e_1, e_2, \dots e_n$

$e_1 + e_2 + \dots + e_n = 0$ (pas d'indépendance)

On teste généralement une tendance ou une dependance entre $\underline{e_i}$ et $\underline{e_{i+1}}$

Test de <u>Durbin-Watson</u> :

$$\begin{cases} H_0 : \text{il n'y a pas corrélations entre } \varepsilon_i \text{ et } \varepsilon_{i+1} \\ H_1 : \varepsilon_{i+1} = \rho\, \varepsilon_i + u_i \end{cases}$$

<u>Statistique de test</u> :

$$d = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \underset{H_0}{\approx} \text{près de } \underline{\underline{2}}$$

$\downarrow$

On vérifie que $\underline{0 \le d \le 4}$ | table statistique.

Prévision

Soit $X = x_0$.

Alors

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

$$\mathrm{Var}(\hat{y}_0) = \sigma^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_i (x_i-\bar{x})^2}\right)$$ (Démo).

Mais y_0 et $\hat{y}_0$ sont indépendantes !

$\Rightarrow$ $y_0 - \hat{y}_0$:

$$\mathrm{Var}\underbrace{(y_0-\hat{y}_0)}_{\varepsilon} = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_i (x_i-\bar{x})^2}\right)$$

$$\boxed{\frac{y_0-\hat{y}_0}{\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum (x_i-\bar{x})^2}}} \sim T_{n-2}}$$

$IC^{**}(y_0)$: $\nearrow$ grand si $|x_0 - \bar{x}|$ grand.