# 1 Régression linéaire simple

## Postulats

- $\mathbf{H}_1$ Linéarité : $E[\varepsilon_i] = 0$
- $\mathbf{H}_2$ Homoscédasticité : $Var(\varepsilon_i) = \sigma^2$
- $\mathbf{H}_3$ Indépendance : $Cov(\varepsilon_i, \varepsilon_j) = 0$
- $\mathbf{H}_4$ Normalité : $\varepsilon_i \sim N(0, \sigma^2)$

## Modèle

$$E[Y_i|x_i] = \beta_0 + \beta_1 x_i$$
$$Var(Y_i|x_i) = \sigma^2$$
$$Y_i|x_i \overset{\mathbf{H}_4}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$$

## Estimation des paramètres

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i Y_i - \bar{Y}\sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2 - \bar{x}\sum_{i=1}^n x_i} = \frac{S_{XY}}{S_{XX}}$$

## Estimation de $\sigma^2$

$$\hat{\sigma^2} = s^2 = \frac{\sum_{i=1}^n \hat{\varepsilon_i}^2}{n - p'} = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}$$

## Propriété des estimateurs

$$E\left[\hat{\beta}_1\right] = \beta_1 \quad , Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{XX}}$$
$$\hat{\beta}_1 \overset{H_4}{\sim} N(\beta_1, \frac{\sigma^2}{S_{XX}})$$
$$E\left[\hat{\beta}_0\right] = \beta_0 \quad , Var(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right)$$
$$\hat{\beta}_0 \overset{H_4}{\sim} N(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)$$
$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}\sigma^2}{S_{XX}}$$

## Tests d'hypothèse sur les paramètres

$H_0 : \hat{\beta} = \theta_0$ , $H_1 : \hat{\beta} \neq \theta_0$

$$t_{obs} = \frac{\hat{\beta} - \theta_0}{\sqrt{\hat{Var}(\hat{\beta}}} \sim T_{n-2}$$

On Rejette $H_0$ si $t_{obs} > |t_{n-2}(1 - \frac{\alpha}{2})|$

## Intervalle de confiance

### Pour la droite de régression ($E[Y_0|x_0]$)

Sachant que $E[Y_0|x_0] = \beta_0 + \beta_1 x_0$, on a l'IC suivant

$$\hat{Y_0} \pm t_{n-2}\left(\frac{\alpha}{2}\right)\sqrt{s^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}\right)}$$

### Pour la prévision de $Y_0$

Sachant que $Y_0 = \beta_0 + \beta_1 x_0 + \varepsilon$, on a l'IC suivant

$$\hat{Y_0} \pm t_{n-2}\left(\frac{\alpha}{2}\right)\sqrt{s^2\left(1 + \frac{1}{m} + \frac{(x_0 - \bar{x})^2}{S_{XX}}\right)}$$

## Analyse de la variance (ANOVA)

| Source | dl | Sum of squares | Mean squares | $f$ |
|---|---|---|---|---|
| Model | $p = 1$ | $SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$ | $MSR = \frac{SSR}{dl_1}$ | $F = \frac{MSR}{MSE}$ |
| Residual error | $n-2$ | $SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ | $MSE = \frac{SSE}{dl_2}$ | |
| Total | $n-2+p$ | $SST = \sum_{i=1}^n (Y_i - \bar{Y})i)^2$ | | |

## Test F de Fisher pour la validité globale de la régression

$H_0$ : La régression n'est pas pertinente (i.e. $\beta_1 = 0$)
$H_1$ : La régression est pertinente

$$F_{obs} = \frac{MSR}{MSE} \sim F_{1,n-2}$$

On rejette $H_0$ si $F_{obs} > F_{1,n-2}\left(\frac{\alpha}{2}\right)$

## Distribution d'un résidu $\varepsilon$

$$E[\hat{\varepsilon_i}] = 0 \quad , Var(\hat{\varepsilon_i}) = \sigma^2(1 - h_{ii})$$

où $h_{ii} = \frac{1}{n} + \frac{(\bar{x} - x_i)^2}{S_{XX}}$.

# 2 Régression linéaire multiple