

# 1 Régression linéaire simple

## Postulats

- H<sub>1</sub>** Linéarité :  $E[\varepsilon_i] = 0$   
**H<sub>2</sub>** Homoscédasticité :  $Var(\varepsilon_i) = \sigma^2$   
**H<sub>3</sub>** Indépendance :  $Cov(\varepsilon_i, \varepsilon_j) = 0$   
**H<sub>4</sub>** Normalité :  $\varepsilon_i \sim N(0, \sigma^2)$

## Modèle

$$E[Y_i|x_i] = \beta_0 + \beta_1 x_i$$

$$Var(Y_i|x_i) = \sigma^2$$

$$Y_i|x_i \stackrel{H_4}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$$

## 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

$$E[\hat{\beta}_1] = \beta_1, \quad Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{XX}}$$

$$\hat{\beta}_1 \stackrel{H_4}{\sim} N(\beta_1, \frac{\sigma^2}{S_{XX}})$$

$$E[\hat{\beta}_0] = \beta_0, \quad Var(\hat{\beta}_0) = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{XX}} \right)$$

$$\hat{\beta}_0 \stackrel{H_4}{\sim} N(\beta_0, \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{XX}} \right))$$

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x} \sigma^2}{S_{XX}}$$

## Tests d'hypothèse sur les paramètres

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

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

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	SS	MS	F
Model	$p$	$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$ (SSR)	$SSR/dl_1$ (MSR)	$\frac{MSR}{MSE}$
Residual error	$n - p'$	$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ (SSE)	$SSE/dl_2$ (MSE = $s^2$ )	
Total	$n - 1$	$\sum_{i=1}^n (Y_i - \bar{Y})^2$ (SST)		

## Coefficient de détermination

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

On a aussi la relation suivante avec  $F_{obs}$  :

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - p'}{p}$$

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

On rejette  $H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$  si

$$F_{obs} = \frac{MSR}{MSE} \geq F_{n, n-p'}(1 - \alpha)$$

où  $p$  est le nombre de variables explicatives dans le modèle (régression linéaire simple,  $p = 1$  et  $p' = p + 1$ ).

## 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

## Le modèle et ses propriétés

$$\mathbf{Y}_{n \times 1} = \mathbf{X}_{n \times p'} \boldsymbol{\beta}_{p' \times 1} + \boldsymbol{\varepsilon}_{n \times 1}$$

$$E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}, \quad Var(\mathbf{Y}) = \sigma^2 \mathbf{I}_{n \times n}$$

$$\mathbf{Y} \stackrel{H_4}{\sim} N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_{n \times n})$$

## Paramètres du modèle

### Estimation et propriétés des paramètres

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

$$E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}, \quad Var(\mathbf{Y}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$$

$$\hat{\boldsymbol{\beta}} \stackrel{H_4}{\sim} N_p(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1})$$

### Intervalle de confiance sur les paramètres

$$\left[ \hat{\beta} \pm t_{n-p'} \left( 1 - \frac{\alpha}{2} \right) \sqrt{s^2 v_{jj}} \right]$$

où  $v_{jj}$  est l'élément  $(i, i)$  de la matrice  $(\mathbf{X}^\top \mathbf{X})^{-1}$ .

**Estimation de  $\sigma^2$** 

$$\hat{\sigma}^2 = s^2 = \frac{\hat{\varepsilon}^\top \hat{\varepsilon}}{n - p'}$$

**Test d'hypothèse sur un paramètre du modèle**

On rejette  $H_0 : \beta_j = 0$  si

$$|t_{obs,j}| = \frac{\beta_j \sqrt{n - p'}}{\sqrt{v_{jj} \hat{\varepsilon}^\top \hat{\varepsilon}}} > t_{n-p'} \left(1 - \frac{\alpha}{2}\right)$$

**Propriétés de la droite de régression**

$$\begin{aligned}\hat{\mathbf{Y}} &= \mathbf{X}\beta \\ &= \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \\ &= \mathbf{H}\mathbf{Y}\end{aligned}$$

où  $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$  est la *hat matrix*.

On a aussi que

$$\begin{aligned}E[\hat{\mathbf{Y}}] &= \mathbf{X}\beta, \quad \text{Var}(\hat{\mathbf{Y}}) = \sigma^2 \mathbf{H} \\ \hat{\mathbf{Y}} &\stackrel{H_4}{\sim} N_n(\mathbf{X}\beta, \sigma^2 \mathbf{H})\end{aligned}$$

Pour les résidus de la droite de régression, on a

$$\begin{aligned}E[\hat{\varepsilon}] &\stackrel{H_1}{=} 0, \quad \text{Var}(\hat{\varepsilon}) = \sigma^2 (\mathbf{I}_{n \times n} - \mathbf{H}) \\ \hat{\varepsilon} &\stackrel{H_4}{\sim} N_n(0, \sigma^2 (\mathbf{I}_{n \times n} - \mathbf{H}))\end{aligned}$$

**Intervalle de confiance pour la prévision****Théorème de Gauss-Markov**

Selon les postulats  $H_1$  à  $H_4$ , l'estimateur

$$\mathbf{a}^\top \hat{\beta} = \mathbf{a}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

est le meilleur estimateur pour  $\mathbf{a}^\top \beta$

(BLUE : *Best linear unbiased estimator*).

**I.C. pour la prévision de la valeur moyenne  $E[\mathbf{Y}|\mathbf{X}^*]$** 

$$\left[ \mathbf{X}^{*\top} \hat{\beta} \pm t_{n-p'} \left(1 - \frac{\alpha}{2}\right) \sqrt{s^2 \mathbf{X}^{*\top} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^{*\top}} \right]$$

**I.C. pour la valeur prédite  $\hat{\mathbf{Y}}|\mathbf{X}^*$** 

$$\left[ \mathbf{X}^{*\top} \hat{\beta} \pm t_{n-p'} \left(1 - \frac{\alpha}{2}\right) \sqrt{s^2 \left(1 + \mathbf{X}^{*\top} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^{*\top}\right)} \right]$$

**Analyse de la variance****Tableau ANOVA**

- On utilise le même tableau ANOVA qu'en régression linéaire simple.
- $SSR_{\text{régression}} = \sum_{i=1}^p SSR_i$ , où  $SSR_i$  représente le SSR individuel de la variable explicative  $i$  calculé par R. On peut ensuite trouver  $MSR$  et la statistique  $F_{obs}$ .

**Test F pour la validité globale de la régression**

Même test qu'en régression linéaire simple.

**Test F partiel pour la réduction du modèle**

Avec  $k < p$ , on va rejeter

$$H_0 : Y_i = \beta_0 + \beta_1 x_{i1} + \dots \beta_{ik} \quad (\text{modèle réduit})$$

Pour

$$H_1 : Y_i = \beta_0 + \beta_1 x_{i1} + \dots \beta_{ip} \quad (\text{modèle complet})$$

Si

$$F_{obs} = \frac{(SSE^{(0)} - SSE^{(1)}) / \Delta dl}{SSE^{(1)} / (n - p')} \geq F_{p-k, n-p'}(1 - \alpha)$$

où  $\Delta dl = p - k$ ,  $SSE^{(0)}$  pour le modèle réduit ( $H_0$ ) et  $SSE^{(1)}$  pour le modèle complet ( $H_1$ ).