Régression linéaire simple (continuation)

$$Y = \beta_0 + \beta_1 X + \mathcal{E} \quad , \quad \mathcal{E} \sim \mathcal{N}(0,\sigma^2)$$

Echantillon aléatoire $\{(x_i, y_i),\ i=1\ldots n\}$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \hat{\rho}(x,y) \cdot \frac{s_y}{s_x}$$

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 = \frac{1}{n-2} \sum_{i=1}^{n} \mathcal{E}_i^2$$

<u>Proposition</u> : $\hat{\beta}_0$, $\hat{\beta}_1$ et $\hat{\sigma}^2$ sont des estimateurs sans biais de β_0, β_1 et σ^2. De plus

$$\mathbb{V}(\hat{\beta}_0) = \frac{\sigma^2 \sum x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2} \ ; \ \mathbb{V}(\hat{\beta}_1) = \sigma^2 \cdot \frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\mathrm{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) = \frac{\sigma^2 \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

<u>Théorème</u> [Gauss-Markov] : $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}^2$ sont des variance minimale parmi tous les estimateurs linéaires en $\{y_i\}$.

Les résidus :

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

Propriété : $\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$

Décomposition de la variance de Y :
ANOVA de la régression

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\text{variance totale}} = \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{\text{résiduelle}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{\text{expliquée par le modèle}}$$

De plus

$$\frac{\text{variance expliquée par le modèle}}{\text{variance totale}} = r^2$$

r^2 = coefficient de détermination
(parfois noté avec R^2).

- Inférence dans le modèle linéaire simple

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\mathbb{V}(\hat{\beta}_1)}} \sim T_{n-2} \quad , \quad \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\mathbb{V}(\hat{\beta}_0)}} \sim T_{n-2}$$

$\boxed{H_0 : \beta_1 = 0}$ $\iff$ $\boxed{H_0 : \exists \text{ lien linéaire entre } X \text{ et } Y \text{ ?}}$

$\iff$ $\boxed{H_0 : \rho(X, Y) = 0}$

Pour tester la significativité du modèle linéaire on teste

TEST DE FISHER de la R.L.

$$\underline{H_0 : \rho^2(X,Y) = 0}$$

à l'aide de la statistique

$$F = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2} (n-2) = \frac{r^2}{1-r^2}(n-2) \sim F(1; n-2)$$

Diagnostic de la régression linéaire simple

1. Analyse des résidus : $\hat{\varepsilon}_i = y_i - \hat{y}_i$

- on fait deux hypothèses :
  - résidus normales ($\varepsilon \sim \mathcal{N}(0, \sigma^2)$)
  - indépendants et identiquement distribués (homoscédasticité)

On peut montrer que, même si l'hypothèse de homoscédasticité est vérifiée, on a :

$$\boxed{\mathbb{V}(\hat{\varepsilon}_i) = \sigma^2 (1 - h_{ii})}$$ où

h_{ij} sont les éléments de la matrice H telle que

$$\hat{y} = H \cdot y \quad \text{(vectoriel)}$$

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

Résidus standardisés (afin de les rendre comparables)

$$res_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma} \cdot \sqrt{1 - h_{ii}}}$$

Résidus studentisés

$$t_i = \frac{\hat{\varepsilon}_i}{s_{(i)}\sqrt{1-h_{ii}}}$$

où $s^2_{(i)}$ = estimation de la variance résiduelle par validation croisée

Remarque

Validation croisée et PRESS

Notons avec $\hat{y}_{i(i)}$ la prévision de y_i sans l'observation (x_i, y_i) dans l'échantillon.

Alors on peut montrer que

$$\hat{\varepsilon}_{i(i)} = y_i - \hat{y}_{i(i)} = \frac{\hat{\varepsilon}_i}{1-h_{ii}}$$

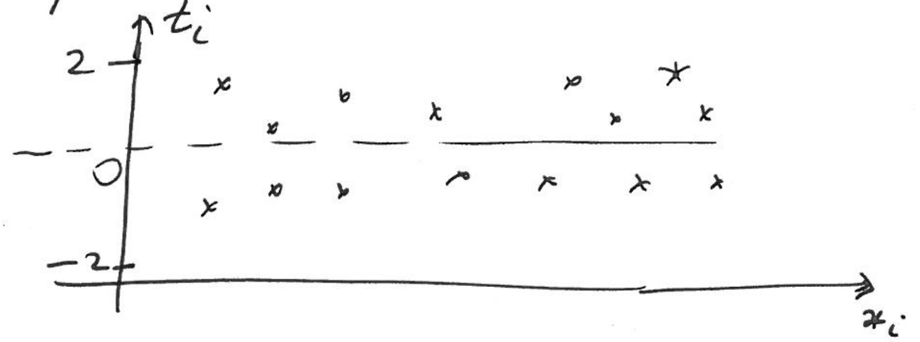
$$PRESS = \frac{1}{n}\sum \hat{\varepsilon}^2_{i(i)}$$

PRESS = predicted residual sum of squares
= leave-one-out cross-validation (validation croisée)
= mesure le caractère prédictif d'un modèle de prévision.

1.1 8 Normalité des résidus : [graphique]

Si les résidus sont $\mathcal{N}(0, \sigma^2)$ alors les résidus t_i (studentisés) sont distribués selon une loi de Student à $n-3$ ddl.

En pratique on vérifie si les $t_i \in [-2 ; 2]$.



On peut aussi tester la normalité des résidus à l'aide d'un test de

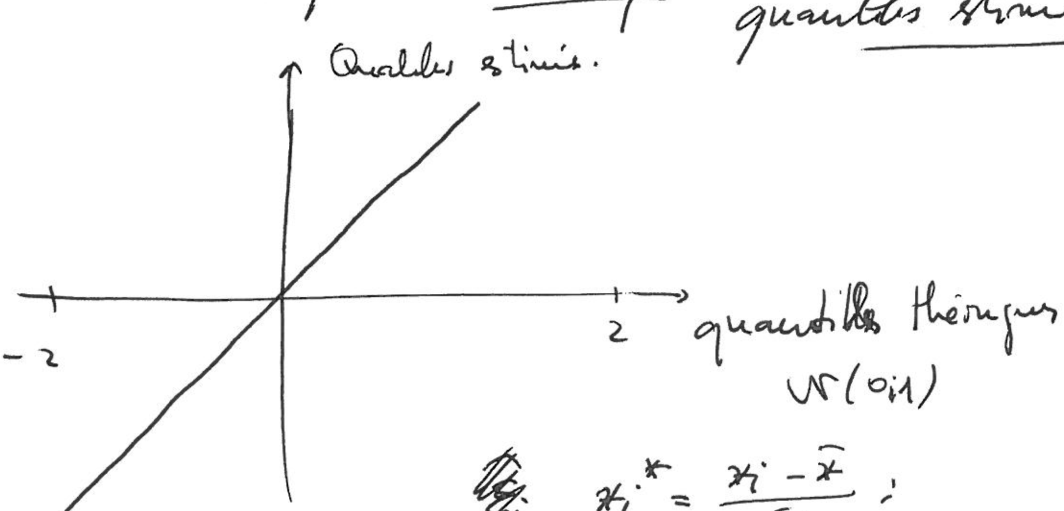
Kolmogorov-Smirnov : $\| F_n - \Phi \|_\infty$

ou

Shapiro-Wilk $W = \dfrac{\left[\sum_{i=1}^{[n/2]} a_i \left(x_{(n-i+1)} - x_{(i)} \right) \right]^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$

Droite d'Henry ou Q-Q plot

(quantiles théoriques versus quantiles estimés)

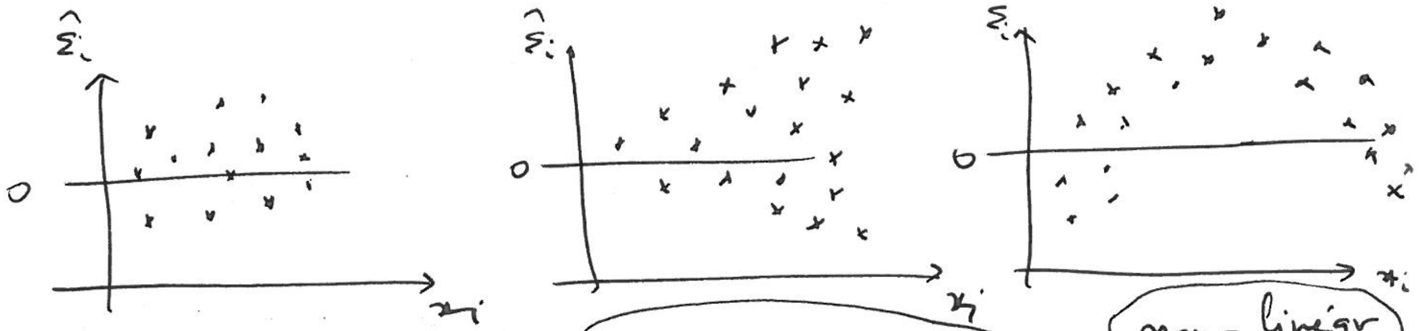


$x_i^* = \dfrac{x_i - \bar{x}}{s_x}$;

illustration en R : ks.test, shapiro.test, qqnorm + qqline

1.2 Homoscédasticité des résidus [analyse graphique]

(variance constante).



OK

variance non-constante

non-linéar

<u>Solution</u> :

transformer la variable y

$$y \to \sqrt{y}$$
$$y \to \log(y)$$

$$\Updownarrow$$

obtenir des modèles <u>non-linéaires</u>.

Pas de teste spécifique pour <u>homoscédasticité</u> à part le test de Breusch-Pagan : <u>bptest</u> dans le package "lmtest".

SAS : <u>PROC MODEL</u>

1.3 Indépendance des résidus

Tests d'indépendance des résidus :

Autocorrélation des résidus :

$$\varepsilon_i = \rho\, \varepsilon_{i-1} + \eta_i$$

H_0 : résidus non-corrélés $\Leftrightarrow \rho = 0$

Test de Durbin-Watson

$$DW = \frac{\sum_{i \geq 2}^{n} (\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1})^2}{\sum_{i=1}^{n} \hat{\varepsilon}_i^2} \sim \chi^2$$

en R : dwtest ("lmtest" package)

en SAS : proc Autoreg

possible tester autocorrélations d'ordre plus grands

option : dw

... Diagnostic de la régression linéaire simple
(continuation)

2 Influence des observations

$$\hat{y} = H \cdot y$$

$$\begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_i \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{bmatrix} & & \\ h_{i1} & \cdots \; h_{ii} \; \cdots & h_{in} \\ & & \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ \\ y_n \end{bmatrix}$$

avec $$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

h_{ii} mesure l'impact de l'observation y_i dans l'estimation de $\hat{y}_i$

si h_{ii} est grande alors y_i influe "beaucoup" sur $\hat{y}_i$

Un point est un point de levier si $h_{ii} > \frac{2}{n}$ (Hoaglin)

$> \frac{3}{n}$ (Welch)

> 0.5 (Huber)

Distance de Cook :

$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_{(i),j} - \hat{y}_j)^2}{2s^2} , \quad s^2 = \frac{1}{n-2} \sum (y_i - \hat{y}_i)^2$$

$D_i > 1$ – point influent.

Régression linéaire multiple $(p \geq 2)$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

Hypothèses : $\varepsilon \sim \mathcal{N}(0, \sigma^2)$

$\varepsilon \perp \{X_1, X_2, \ldots, X_p\}$

Données

ind	X_1	X_2		X_p	Y
1	X_{11}	X_{12}		x_{1p}	Y_1
$\vdots$			$\cdots$		
i	x_{i1}	x_{i2}		x_{ip}	Y_i
$\vdots$	$\vdots$				
n	x_{n1}	x_{n2}		x_{np}	y_n

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & & x_{2p} \\ \vdots & \vdots & & & \\ \vdots & \vdots & & & \\ 1 & x_{n1} & x_{n2} & & x_{np} \end{pmatrix} \in \mathcal{M}_{n \times (p+1)}$$

$$\boxed{Y = X\beta + \varepsilon}$$ avec $\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$, $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$

La solution qui minimise les moindres carrés (MCO)

$$\sum_{i=1}^{n} \left(Y_i - (\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}\right)^2 = \sum_{i=1}^{n} \varepsilon_i^2$$

vaut

$$\hat{\beta} = \underbrace{(X^T X)^{-1} X^T}\, Y \quad \text{et}$$

$$\hat{Y} = \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_i \\ \vdots \\ \hat{y}_n \end{pmatrix} = X\hat{\beta} = \underbrace{X(X^T X)^{-1} X^T}_{H} Y .$$

$$\hat{Y} = HY \quad \text{avec}$$

$$H = X(X^T X)^{-1} X^T$$

("hat" matrix).

<u>Remarque</u> : Existance de $(X^T X)^{-1}$:

<u>cas</u> : $\begin{cases} - \ n < p \\ - \ \sum_{i=1}^{p} \alpha_i X_i = \alpha_0 \quad \text{(multicolinéarité)}. \end{cases}$

Proposition

1) $\hat{\beta}$ est un estimateur sans biais pour β

2) $\hat{\sigma}^2 = \dfrac{1}{n-p-1} \sum_{i=1}^{n} \hat{\varepsilon}_i^2$, $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$

est un estimateur sans biais pour σ^2

3) $\mathbb{V}(\hat{\beta}) = \sigma^2 \cdot (X^T X)^{-1}$

Théorème [Gauss-Markov]

$\hat{\beta}$ est optimal parmi tous les estimateurs linéaires en Y sans biais.

$$\widehat{\mathbb{V}(\hat{\beta})} = \hat{\sigma}^2 (X^T X)^{-1}$$

$$\widehat{\mathbb{V}}(\hat{\beta}_j) = \hat{\sigma}^2 \cdot (X^T X)^{-1}_{[j,j]}$$

$$T_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\widehat{\mathbb{V}}(\hat{\beta}_j)}} \sim \text{Student}(n-p-1)$$

Tests de $\boxed{H_0 : \beta_j = 0}$ vs $\beta_j \neq 0$

Qualité du modèle : R^2

ANOVA de la régression :

$$\underbrace{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}_{\text{var totale}} = \underbrace{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}_{\text{var residuelle}} + \underbrace{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}_{\text{var modèle}}$$

$$\boxed{R^2 = \frac{\sum(\hat{Y}_i-\bar{Y})^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2} = \rho^2(Y,\hat{Y})}$$

Est-ce que le modèle est significatif : $\beta \neq 0$?

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \\ H_1 : \exists \beta_j \neq 0 \quad, j = 1 \ldots p \end{cases}$$

La statistique de Fisher :

$$F = \frac{R^2}{1-R^2}(n-p-1) \underset{H_0}{\sim} F(p, n-p-1).$$

Diagnostics de la régression linéaire multiple

① Analyse des résidus

- homoscédasticité
- normalité
- indépendance

Pareil que dans le cas de la régression linéaire simple

② Influence des observations :

Distance de Cook

③ Influence des variables

- Est-ce que la variable X_j contribue à la qualité du modèle ?

$H_0 : \beta_j = 0$

- Test de Student
- Test de Fisher $F = \dfrac{\|\hat{Y}_{-j} - \hat{Y}\|}{\|Y - \hat{Y}\|^2 / n-p-1} \sim F(1, n-p-1)$

Choix des variables ≡ Choix du modèle

Critères ?   Objectifs
- description
- estimation
- prévision