Le modèle linéaire

$$Y_i = \alpha + \beta x_i + E_i$$
$$E_i \quad i.i.d., \quad V(E_i) = \sigma^2$$

- Y_i Variable dépendante pour le i^{eme} individu de l'échantillon
  $=>$ Alétaoire
- x_i Variable indépendante pour le i^{eme} inidividu de l'échantillon.
  $=>$ Non aléatoire
- E_i terme d'erreur pour le i^{eme} individu de l'échantillon.
  $=>$ Alétaoire
  - Les erreurs sont supposées **indépendantes**.
  - Les erreurs sont supposées **homoscédastiques**, la variance est identique quelque soit l'individu : σ^2.

$$Y_i = \alpha + \beta x_i + E_i$$
$$E_i \quad i.i.d., \quad V(E_i) = \sigma^2$$

- α, β sont les paramètres de régression à estimer.
  $=>$ Non aléatoire
  - α est l'ordonnée à l'origine, il a la même unité que Y_i
  - β est la pente, il s'exprime en unité de Y_i par unité de x_i. β représente l'augmentation ou la diminution de Y par unité d'augmentation de X.
- σ^2 est le paramètre de variance à estimer.
  $=>$ Non aléatoire

Remarque :

- La variable indépendante est considérée comme fixe (non aléatoire). C'est en pratique le cas, dans les expériementations, lorsqu'on fixe les conditions initiales. Les résultats des moindres carrés sont les mêmes que l'on considère X comme aléatoire ou non. C'est ce qui distingue le modèle de **regréssion** du modèle **linéaire**.

Solutions : les équations du premier ordre

On pose la fonction de coût suivante

$$F(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha + \beta x_i)^2$$

Selon la méthode des moindres carrés ordinaires on estime α et β par

$$(a, b) = \underset{(\alpha,\beta)\in\mathbb{R}^2}{\text{argmin}}\, F(\alpha, \beta)$$

(a, b) est la solution du système **d'équations normales**

$$\left\{ \begin{array}{l} \frac{\partial F}{\partial \alpha}(\alpha, \beta) = 0 \\ \frac{\partial F}{\partial \beta}(\alpha, \beta) = 0 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{ll} \overline{y} = \alpha + \beta \overline{x} & (1) \\ \sum_{i=1}^{n} x_i (y_i - (\alpha + \beta x_i)) = 0 & (2) \end{array} \right.$$

avec $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ et $\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$.

Solutions

$$b = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$$
$$a = \overline{y} - \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\overline{x}$$

Remarques :

- b est la version empirique de $\frac{Cov(X,Y)}{V(X)}$ qui est la pente de la droite d'équation $E(Y|X) = \alpha + \beta X$.
- a et b dépendent des y_i qui sont des réalisations des variables aléatoires Y_i, a et b sont donc les réalisations de variables aléatoires.
- On pose $y_i^* = a + bx_i = \overline{y} + b(x_i - \overline{x})$, c'est l'estimation de $E(Y_i|X = x_i)$
- La droite estimée passe par le point de gravité du nuage de points, de coordonnée $(\overline{x}, \overline{y})$.

Et σ^2 ?

- Dans le modèle, σ^2 est la variance des erreurs,
- puisque Y_i^* est un estimateur de $\alpha + \beta x_i$, on va utiliser la variance des résidus $E_i^* = Y_i - Y_i^*$ pour estimer σ^2.

On montrera (plus tard) que

$$S^2 = \frac{\sum_{i=1}^{n} (Y_i - Y_i^*)^2}{n-2}$$

est un estimateur sans biais de σ^2.

Propriétés des résidus

- Les résidus $e_i = y_i - y_i^*$ sont de moyenne nulle.
- Les résidus ne sont donc pas des réalisations de variables aléatoires indépendantes (comme sur le modèle).
- On note $s_n^2 = \frac{1}{n}\sum_{i=1}^n e_i^2$ la variance empirique des résidus.
- On a $s_n^2 = (1 - r^2)s_y^2$ avec r le coefficient de corrélation empirique, défini par

$$r = \frac{\sum_{i=1}^n (y_i - \overline{y})(x_i - \overline{x})}{\sqrt{\sum_{i=1}^n (y_i - \overline{y})^2 \sum_{i=1}^n (x_i - \overline{x})^2}}$$

Le modèle linéaire

On a les résultats suivants :

- $Y_i|X_i = x_i \sim \mathcal{N}(\alpha + \beta x, \sigma^2)$ i.i.d.
- B, A et Y_i^* suivent des lois normales, comme combinaisons linéaires de variables aléatoires normales indépendantes.

$$B \sim \mathcal{N}(\beta, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2})$$

$$A \sim \mathcal{N}(\alpha, \sigma^2(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}))$$

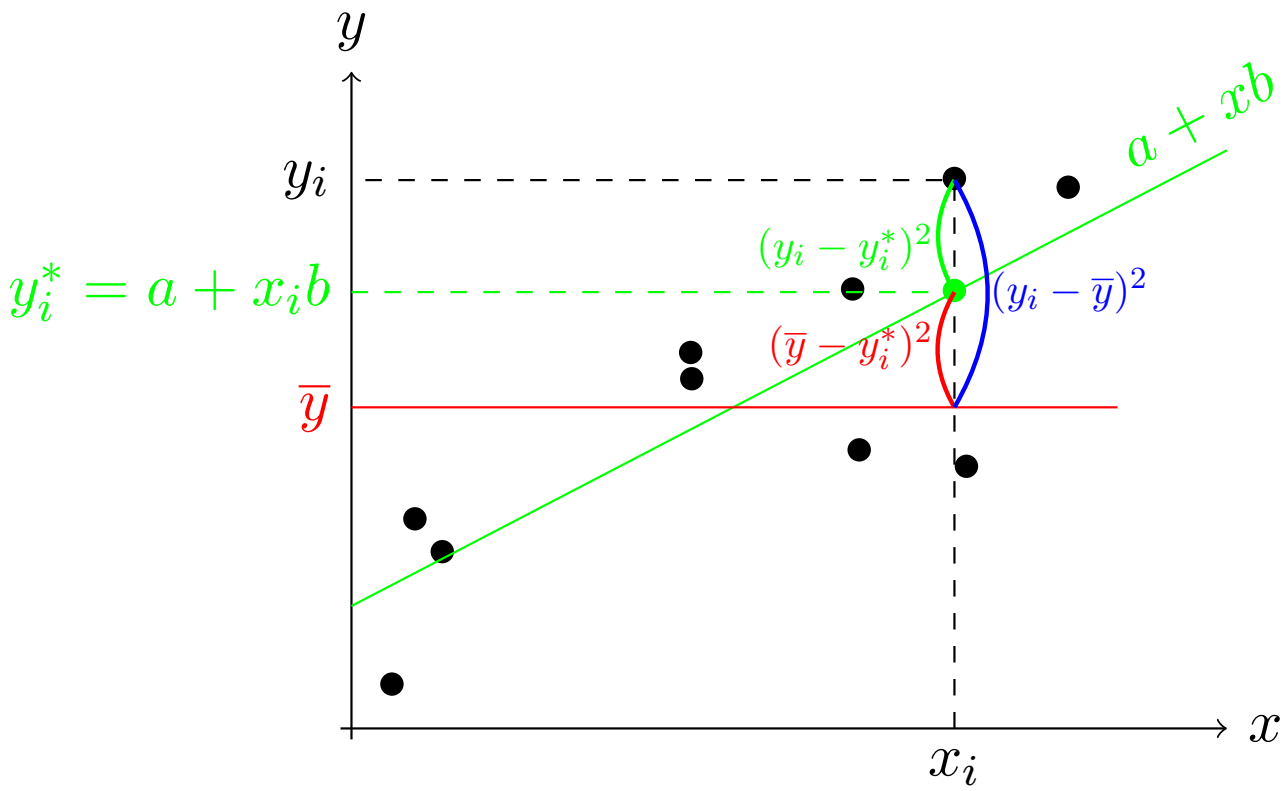
$$Y_i^* \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2(\frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}))$$

- Selon le théorème de Cochran

$$\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$$

- S^2 est indépendant de $\overline{Y}$, B et A.

Décomposition de la variance

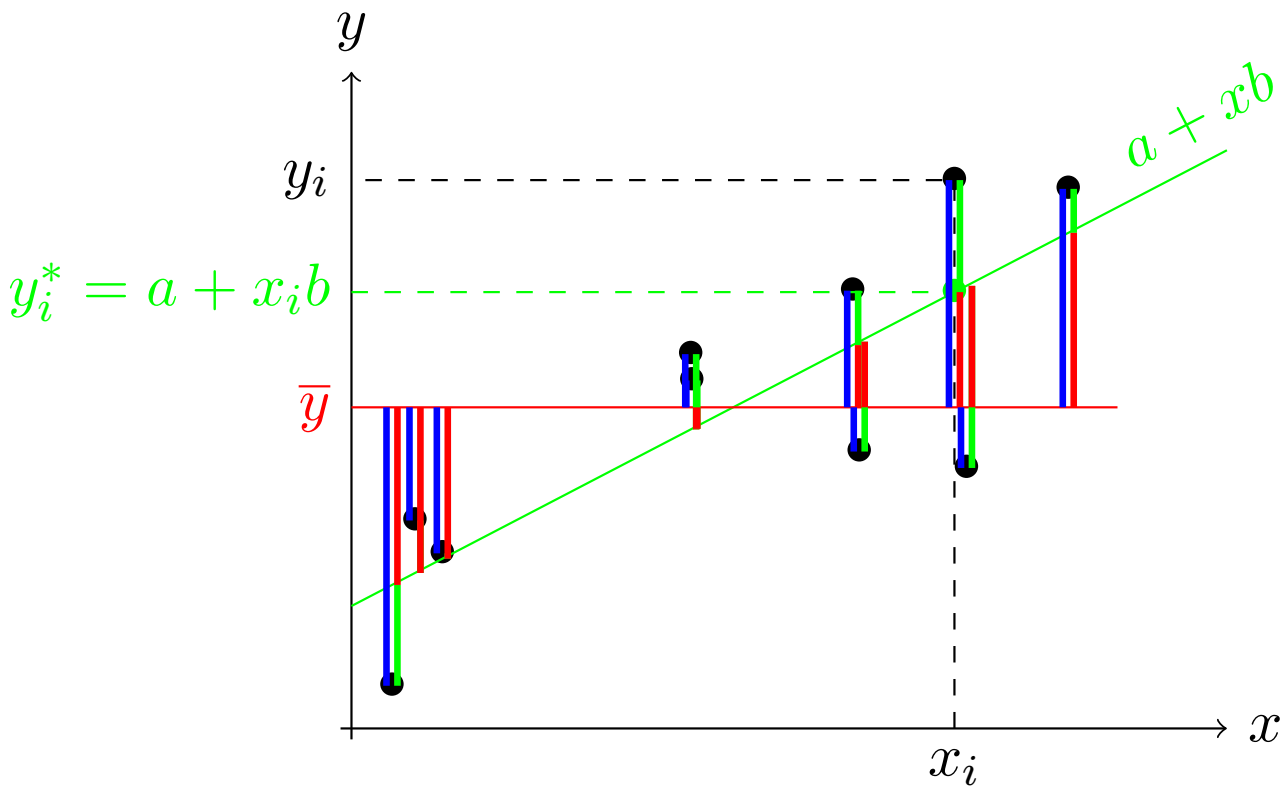


$SCM = \sum_{i=1}^{n}(y_i^* - \overline{y})^2$, somme des carrés du modèle

$SCR = \sum_{i=1}^{n}(y_i - y_i^*)^2$, somme des carrés résiduels

$SCT = \sum_{i=1}^{n}(y_i - \overline{y})^2$, somme des carrés totale

Décomposition de la variance



$SCM = \sum_{i=1}^{n}(y_i^* - \overline{y})^2$, somme des carrés du modèle

$SCR = \sum_{i=1}^{n}(y_i - y_i^*)^2$, somme des carrés résiduels

$SCT = \sum_{i=1}^{n}(y_i - \overline{y})^2$, somme des carrés totale

Décomposition de la variance

On peut montrer que

$$\underbrace{\sum_{i=1}^{n}(y_i - \overline{y})^2}_{SCT} = \underbrace{\sum_{i=1}^{n}(y_i^* - \overline{y})^2}_{SCM} + \underbrace{\sum_{i=1}^{n}(y_i - y_i^*)^2}_{SCR}$$

On a également les résultats suivants (Cf Th. de Cochran) :

- $\frac{SCM}{\sigma^2} = \frac{\sum_{i=1}^{n}(Y_i - Y_i^*)^2}{\sigma^2} \sim \chi^2_{n-2}$
- $\frac{SCR}{\sigma^2} = \frac{\sum_{i=1}^{n}(Y_i - Y_i^*)^2}{\sigma^2} \underset{\mathcal{H}_0}{\sim} \chi^2_1$
- $\frac{SCT}{\sigma^2} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{\sigma^2} \underset{\mathcal{H}_0}{\sim} \chi^2_{n-1}$
- SCR et SCM indépendants.

avec $\mathcal{H}_0$ l'hypothése $\{\beta = 0\}$.

Coefficient de détermination R^2

Définition

On appelle coefficient de détermination de la régression la proportion R^2 définie par

$$R^2 = \frac{\sum_{i=1}^{n}(y_i^* - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{SCM}{SCT}$$

- R^2 représente la part de la variabilité des données expliquée par le modèle.
- Plus le R^2 est grand, plus la variable explicative X explique une grande part de la variabilité de variable Y.
- Dans le cadre de la régression simple

$$R^2 = r^2$$

avec r le coefficient de corrélation de Pearson.

Test d'analyse de la variance

- On considère l'hypothése suivante

$$\mathcal{H}_0 = \{\beta = 0\} = \left\{ \begin{array}{l} \text{le modèle de régression linéaire simple} \\ \text{est inutile par rapport à un modèle avec} \\ \text{une moyenne constante} \end{array} \right\}$$

- Sous l'hypothèse $\mathcal{H}_0$, $F = \frac{SCM/1}{SCR/(n-2)} \sim \mathcal{F}_{1,n-2}$.
- On rejette $\mathcal{H}_0$ si F est trop grande, c'est à dire si

$$F_{obs} \geq f_{1,n-2,1-\alpha}$$

- La probabilité critique associée à ce test est :

$$p_c = P(V > F_{obs})$$

avec $V \sim \mathcal{F}_{1,n-2}$.

Test sur le paramètre α

- On considère les hypothèses $\mathcal{H}_0 = \{\alpha = \alpha_0\}$ et $\mathcal{H}_1 = \{\alpha \neq \alpha_0\}$.
- On rappelle que
  - $A \sim \mathcal{N}(\alpha, \sigma^2(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}))$
  - $\frac{\sum_{i=1}^n (Y_i - Y_i^*)^2}{\sigma^2} = \frac{S^2(n-2)}{\sigma^2} \sim \chi^2_{n-2}$
  - Les deux estimateurs sont indépendants.
- Sous l'hypothèse $\mathcal{H}_0$, $\frac{(A-\alpha_0)}{S\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}} \sim \mathcal{T}_{n-2}$
- On rejette $\mathcal{H}_0$ au profit de $\mathcal{H}_1$, si $\frac{(A-\alpha_0)}{S\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}}$ s'écarte trop de 0.
- C'est à dire pour un risque de niveau α, si

$$|a - \alpha_0| \geq t_{n-2,1-\alpha/2} s \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}$$

avec $t_{n-2,1-\alpha/2}$ le quantile d'ordre $1 - \alpha/2$ de la loi de Student à $n - 2$ dl.

Test sur le paramètre β

- On considère les hypothèses $\mathcal{H}_0 = \{\beta = \beta_0\}$ et $\mathcal{H}_1 = \{\beta \neq \beta_0\}$.
- On rappelle que
  - $B \sim \mathcal{N}(\beta, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2})$
  - $\frac{\sum_{i=1}^n (Y_i - Y_i^*)^2}{\sigma^2} = \frac{S^2(n-2)}{\sigma^2} \sim \chi^2_{n-2}$
  - Les deux estimateurs sont indépendants.
- Sous l'hypothèse $\mathcal{H}_0$, $\frac{(B-\beta_0)\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}}{S} \sim \mathcal{T}_{n-2}$
- On rejette $\mathcal{H}_0$ au profit de $\mathcal{H}_1$, si $\frac{(B-\beta_0)\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}}{S}$ s'écarte trop de 0.
- C'est à dire, pour un risque de niveau α, si

$$|b - \beta_0| \geq t_{n-2,1-\alpha/2} \frac{s}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}}$$

avec $t_{n-2,1-\alpha/2}$ le quantile d'ordre $1 - \alpha/2$ de la loi de Student à $n - 2$ dl.

Intervalle de confiance pour α et β

A partir des lois de A et B énoncées précédemment, on peut obtenir des statistiques pivotales et des intervalles de confiance pour α et β.

- Intervalle de confiance de niveau $1-\gamma$ pour α :

$$IC_{1-\gamma}(\alpha) = \left[a \pm t_{n-2,1-\gamma/2} s \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}} \right]$$

- Intervalle de confiance de niveau $1-\gamma$ pour β :

$$IC_{1-\gamma}(\beta) = \left[b \pm t_{n-2,1-\gamma/2} \frac{s}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}} \right]$$

Intervalle de confiance pour $E(Y_j|X = x_j)$

Comme

- $Y_j^* \sim \mathcal{N}(\alpha + \beta x_j, \sigma^2(\frac{1}{n} + \frac{(x_j-\overline{x})^2}{\sum_{i=1}^n (x_i-\overline{x})^2}))$
- $\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$
- S^2 est indépendant de $\overline{Y}$, B et A, donc de Y_j^*,

on en déduit un intervalle de confiance de niveau $1-\gamma$ pour $E(Y_j)$

$$IC_{1-\gamma}(E(Y_j)) = \left[y_j^* \pm t_{n-2,1-\gamma/2} s \sqrt{\frac{1}{n} + \frac{(x_j - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}} \right]$$

REMARQUE : Plus x_i est loin de $\overline{x}$ plus l'intervalle de confiance est grand. On obtient une hyperbole de confiance, lorsque x_i est proche de $\overline{x}$ c'est le terme $1/n$ qui domine, mais si x_i est éloigné de $\overline{x}$ c'est le terme quadratique $(x_i - \overline{x})^2$ qui domine.

Intervalle de prédiction pour une nouvelle donnée

On dispose d'un nouvel x_0 et l'on souhaite **prédire** le Y_0 correspondant avec notre modèle. C'est à dire $Y_0 = \alpha + \beta x_0 + E_0$, avec $E_0 \sim \mathcal{N}(0, \sigma^2)$. On peut prédire Y_0 avec le modèle estimé :

$$Y_0^p = A + Bx_0 + E_0^p \text{ avec } E_0^p \sim \mathcal{N}(0, S^2)$$

- La prédiction est $y_0^p = a + b_{x0}$
- La variance de la prédiction est donnée par

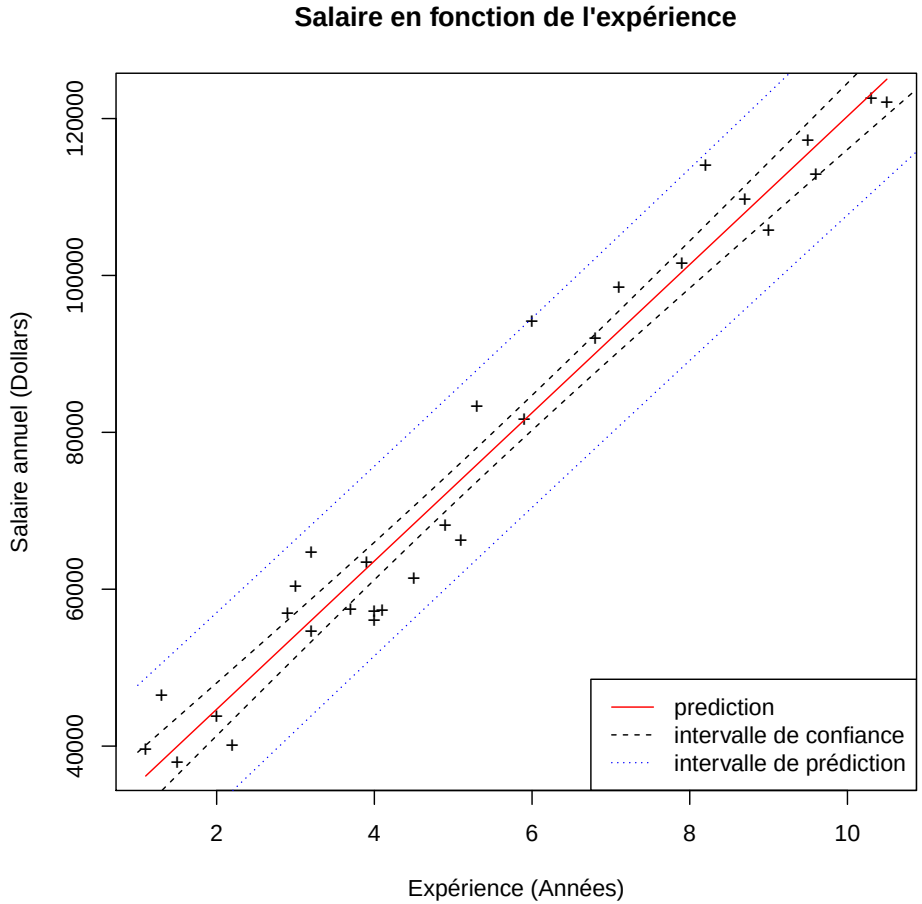
$$V(Y_0^p) = V(A + Bx_0) + V(E_0^p) = S^2 \left(1 + \frac{1}{n} + \frac{(x_j - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)$$

Remarques :

- la nouvelle donnée donnée n'intervient pas dans le calcul de A et B.
- Deux sources d'incertitudes sur Y_0^p : l'incertitude sur a, b et σ^2 et la variabilité de l'erreur.
- la variance augmente lorsque x_0 s'éloigne de $\overline{x}$.

Application : intervalle de confiance et de prédiction sur Y_0^*

```
> precon=predict(reslm,newdata=data.frame(YearsExperience=seq(1,11,0.01))
+                ,interval="confidence")
> prepre=predict(reslm,newdata=data.frame(YearsExperience=seq(1,11,0.01))
+                ,interval="prediction")
```



Validation des hypothèses du modèle

1. Hypothèse d'indépendance
2. Hypothèse d'homoscédasticité
3. Hypothèse de normalité
4. Hypothèse de linéarité

=> Cette étape est basée sur l'analyse des résidus.
=> On s'appuie sur des méthodes graphiques.
=>Il n'y a pas de règle fixe pour la prise de décision.