

# Régression linéaire

Bensalem .H

# objectif

- L'objectif de la régression linéaire est d'étudier l'effet de certains facteurs explicatifs appelés **régresseurs** sur un phénomène observé ou sur des résultats expérimentaux.

établir un lien entre une variable à expliquer Y et une variable explicative X



faire des **prévisions** sur Y lorsque X est mesurée.

# Le modèle

- Y: variable aléatoire
- X: variable **déterministe**  
(contrôlée = non aléatoire)
- Échantillon aléatoire  
 $(x_1, y_1), \dots, (x_n, y_n)$

- Y: variable aléatoire
- X: variable **aléatoire**
- Échantillon aléatoire  
 $(x_1, y_1), \dots, (x_n, y_n)$

Trouver une relation mathématique qui exprime Y en fonction de X

$Y = \beta_0 + \beta_1 X$ : modèle linéaire

$E[Y|X=x] = \beta_0 + \beta_1 X$ : fonction de régression  
/  $X \sim N(\mu_x, \delta_x)$   $Y \sim N(\mu_y, \delta_y)$

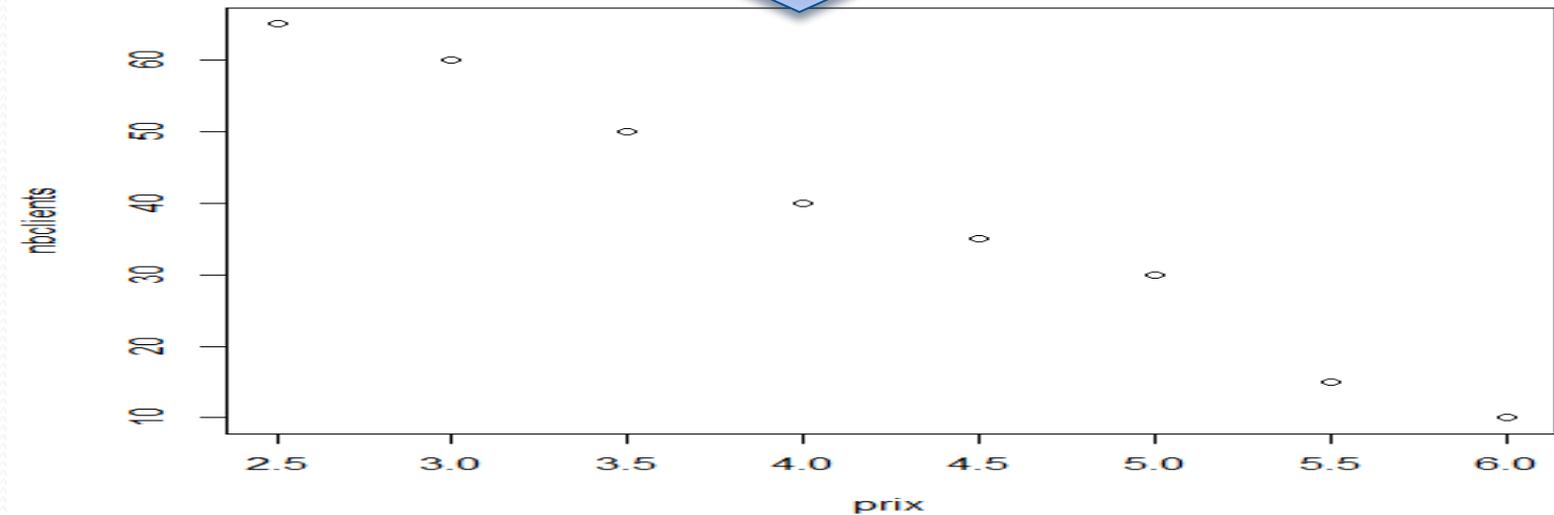
**Estimation des paramètres**: la méthode des moindres carrés les traite indifféremment

- Le modèle est dit :
- Simple, dû au fait qu'on explique Y avec une seule variable X,
- Linéaire, la relation linéaire est entre les coefficients
- Exemple:  $\log(y) = \beta_0 + \beta_1 X$ , est un modèle linéaire en les coefficients, en prenant  $y^* = \log(y) \Rightarrow y^* = \beta_0 + \beta_1 X$ .

# Exemple

Considérons une enquête effectuée par une direction commerciale auprès de 100 clients. Le tableau ci-dessous, indique pour chaque prix le nombre de clients s'étant déclarés potentiellement acheteur d'un produit

X: prix en milliers d'euros	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0
Y :nombre d'acheteurs potentiels	65	60	50	40	35	30	15	10



le nombre d'acheteurs potentiels en **fonction** des prix

- **Le coefficient de détermination** (Facteur de corrélation linéaire, pour le cas  $x$  aléatoire) et nuage de points donnent une idée sur la possibilité d'adopter un modèle de régression linéaire;

- $Y = \beta_0 + \beta_1 X$

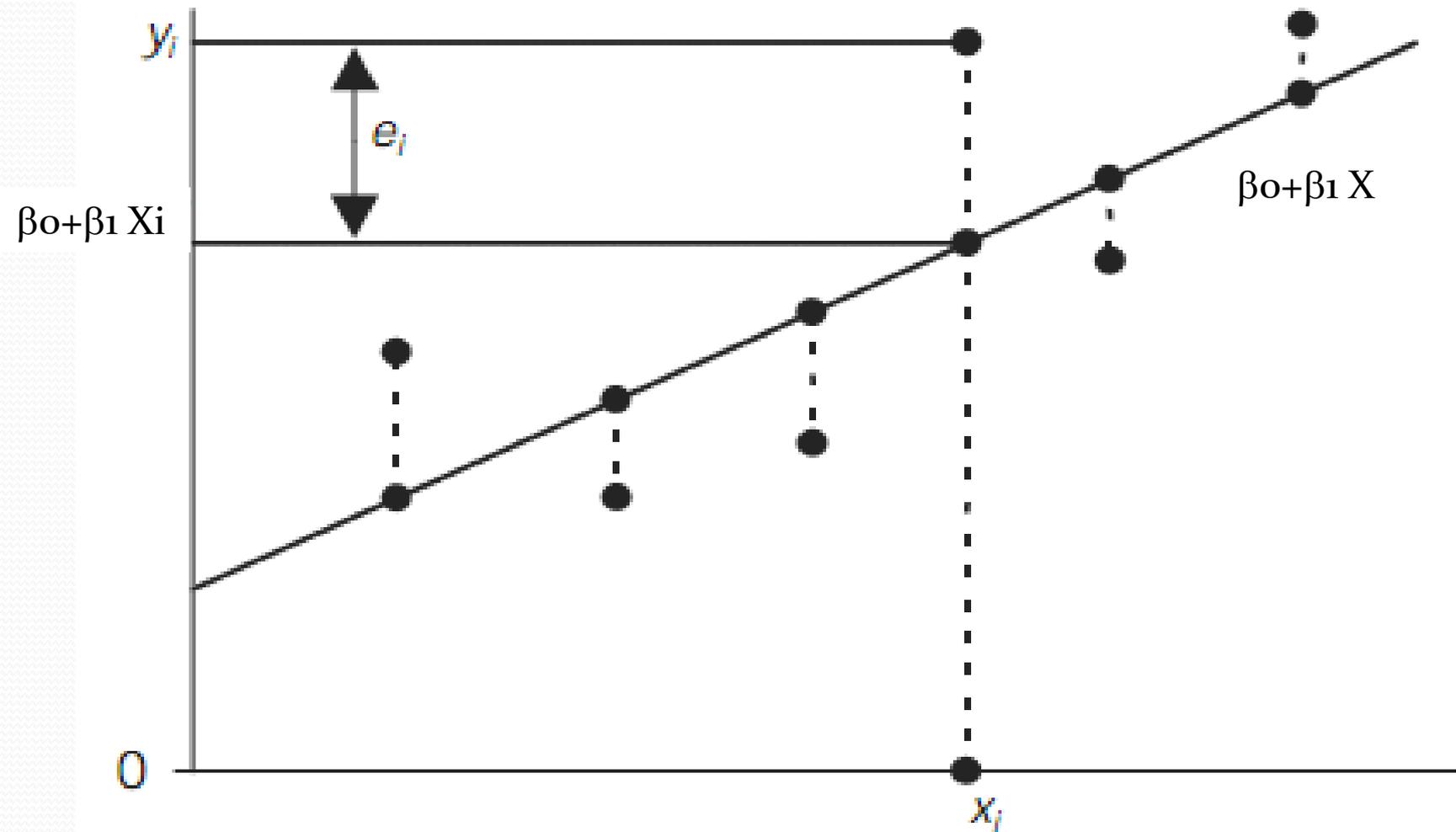
On cherche une droite par laquelle passent le **maximum** de points du nuage



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

les **résidus**, pour les points non alignés

- $\varepsilon_i = y_i - \beta_0 - \beta_1 X_i$  : est l'erreur résiduelle.



# Régression linéaire simple

## hypothèses

- (1)  $E[\varepsilon_i] = 0, \forall i = 1, \dots, n \Rightarrow E[y_i] = \beta_0 + \beta_1 x_i, \forall i = 1, \dots, n$
- les erreurs sont **centrées** (moyenne nulle) ce qui implique que  $y_i$  dépend seulement de  $x_i$  et que les autres sources de variations de  $y_i$  sont aléatoires.
- (2)  $V[\varepsilon_i] = \sigma^2, \forall i = 1, \dots, n \Rightarrow V[y_i] = \sigma^2, \forall i = 1, \dots, n$
- hypothèse d'homoscédasticité (homogénéité des variances, plusieurs échantillons), supposée constante cette variance est indépendante de  $x_i$  et à **estimer**.
- (3)  $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j \Rightarrow \text{Cov}(y_i, y_j) = 0, \forall i \neq j$ , c.a.d les termes d'erreur  $\varepsilon_i$  sont **non corrélés**.
- Lorsque l'on rajoutera une hypothèse de normalité sur les  $\varepsilon_i$ , les erreurs  $\varepsilon_i$  seront alors **indépendantes**.

Estimation des paramètres du modèle:  
 $\beta_0$  ,  $\beta_1$  et  $\sigma^2$  (variance résiduelle)

- Pour estimer  $\beta_0$  et  $\beta_1$  , on peut utiliser la méthode des **moindres carrés** qui ne nécessite pas d'hypothèse supplémentaire sur la distribution de  $\varepsilon_i$  (ou  $y_i$ ),
- La méthode du maximum de vraisemblance est fondée sur la normalité de  $\varepsilon_i$  (ou  $y_i$ ).
- La méthode des moindres carrés ne fournit pas un estimateur de  $\sigma^2$  .

La méthode des moindres carrés est basée sur le principe de minimiser la somme des résidus

$\text{Cov}(x,y)/\text{var}(x)$

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

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

- Sous les trois hypothèses, les estimateurs de  $\beta_0$  et de  $\beta_1$  sont **sans biais** et de **variance minimale** parmi tous les estimateurs linéaires.

## Variance résiduelle

- Le paramètre  $\sigma^2$  est défini par  $\sigma^2 = V(\varepsilon_i) = V(y_i) = E[(y_i - E[y_i])^2]$
- En prenant  $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$  comme estimateur de  $E[y_i]$ , il paraît naturel d'estimer  $\sigma^2$  par:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2} = \frac{\sum_{i=1}^n (\hat{\varepsilon}_i)^2}{n - 2} = \frac{SCR}{n - 2}$$

- $\hat{\sigma}^2$  est un estimateur **sans biais** de  $\sigma^2$ , La perte de deux degrés de liberté dans l'expression  $n - 2$  est le "coût" de l'estimation de  $\beta_0$  **et** de  $\beta_1$ .
- SCR: somme des carrés des résidus(ou des erreurs)

# Exemple

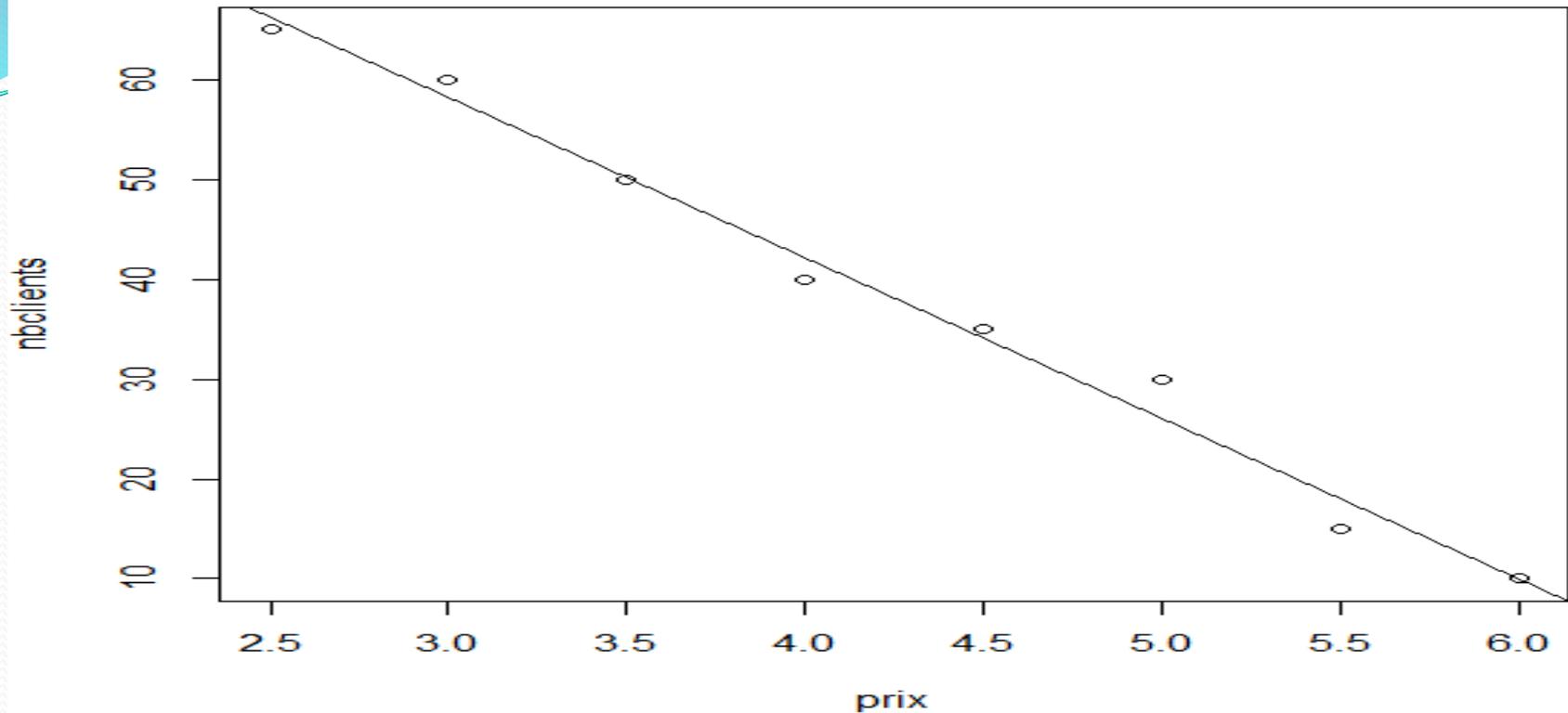
$x_i$	$y_i$	$x_i - \bar{x}$	$y_i - \bar{y}$	$(y_i - \bar{y}) * (x_i - \bar{x})$	$(x_i - \bar{x}) * (x_i - \bar{x})$
2.5	65	-1.75	26.875	-47.03125	3.0625
3.0	60	-1.25	21.875	-27.34375	1.5625
3.5	50	-0.75	11.875	-8.90625	0.5625
4.0	40	-0.25	1.875	-0.46875	0.0625
4.5	35	0.25	-3.125	-0.78125	0.0625
5.0	30	0.75	-8.125	-6.09375	0.5625
5.5	15	1.25	-23.125	-28.90625	1.5625
6.0	10	1.75	-28.125	-49.21875	3.0625
TOTAL	---	---	----	-168.75	10.5

$$\bar{x} = 1/n * \sum x_i = 1/8 * (2.5 + 3.0 + 3.5 + 4.0 + 4.5 + 5.0 + 5.5 + 6.0) = 4.25$$

$$\bar{Y} = 1/n * \sum Y_i = 1/8 * (65 + 60 + 50 + 40 + 35 + 30 + 15 + 10) = 38.125$$

$$\Rightarrow \hat{\beta}_1 = -168.75 / 10.5 = -16.071$$

$$\hat{\beta}_0 = 38.125 - (-16.071 * 4.25) = 106.4268 \Rightarrow Y = 106.43 - 16.07 X$$



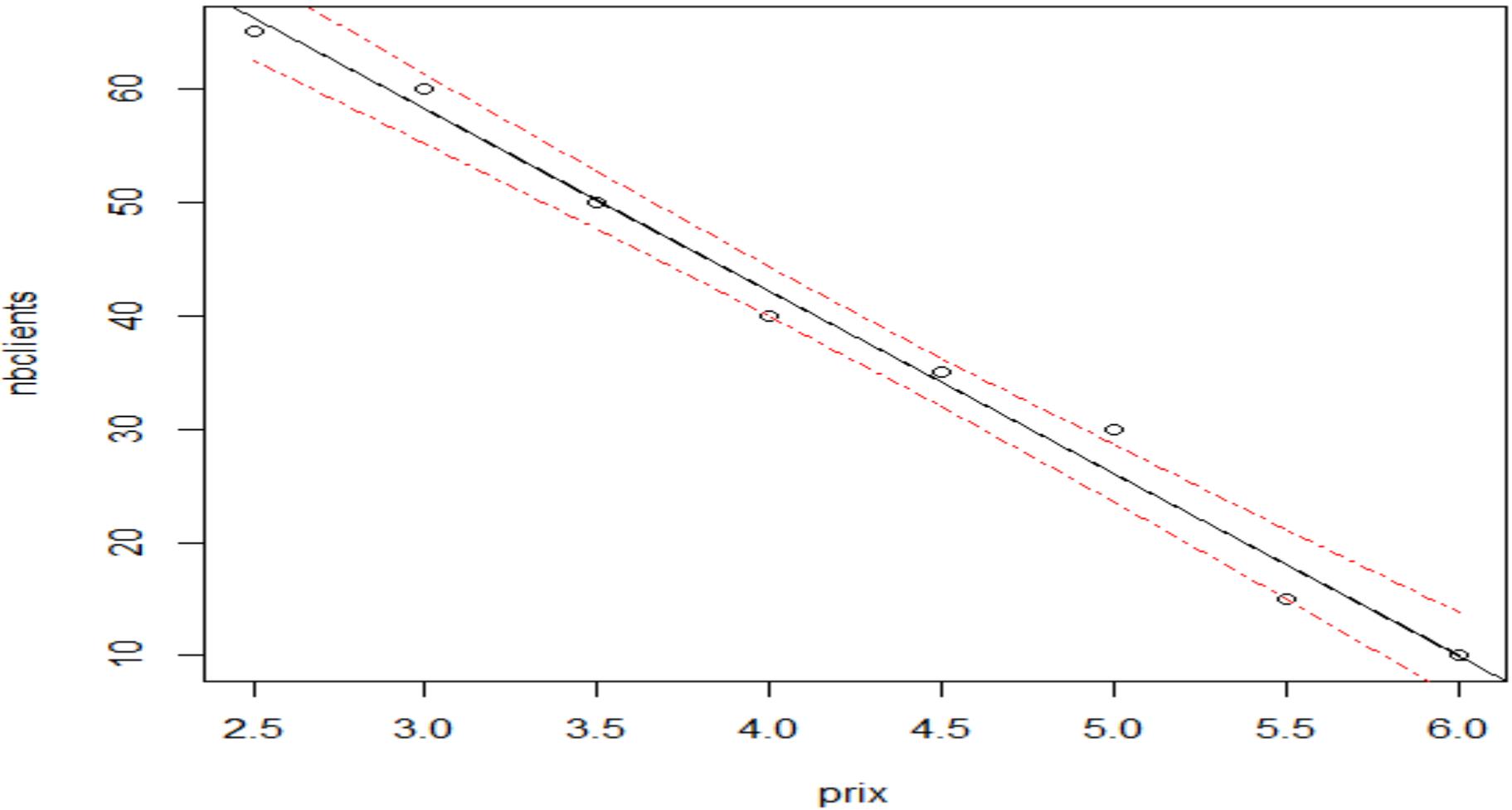
- Interprétation:
- si  $x=0$ ,  $y=106$  = estimation de  $\beta_0 > 100$  la taille de l'échantillon, c.à.d lorsque le produit est offert (gratuit), la totalité des clients se portent acquéreur.
- L'estimation de  $\beta_1$ : indique de combien varie Y lorsque X varie d'une unité.

# Les intervalles de confiance

$$\hat{\beta}_0 + \hat{\beta}'x \mp U_{\alpha/2} * \hat{\sigma} \sqrt{1/n + (x - \bar{x})^2 / Scx}$$

- $U_{\alpha/2}$ : le fractile d'ordre  $\alpha/2$  de la loi de Student avec  $n-2$  degrés de liberté.
- $Scx = \text{var}(x)/n$
- Un intervalle de confiance bilatéral à  $(1 - \alpha)$  degrés de confiance, pour la valeur moyenne de  $Y$  lorsque  $x = x \in [x_1, x_n]$

# Exemple



# Prédiction

- Pour prédire la valeur de  $y$  lorsque  $x=x_i$  on utilise l'équation:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

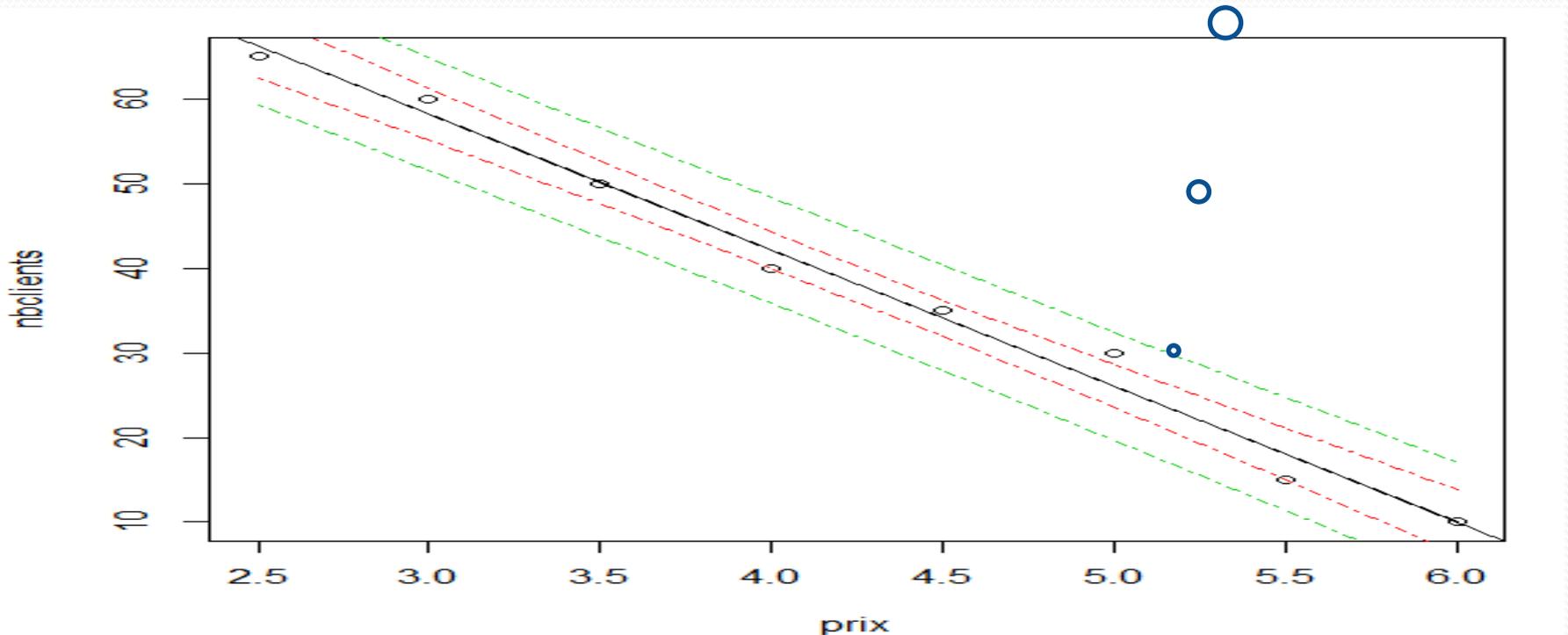
- La prédiction ne devrait être (théoriquement) utilisée que pour  $x \in [x_1, x_n]$
- Intervalle de confiance pour une nouvelle observation (prédiction) de  $y$  pour  $x=x$

$$\hat{\beta}_0 + \hat{\beta}'_1 x \pm U_{\alpha/2} * \hat{\sigma} \sqrt{1 + 1/n + (x - \bar{x})^2 / S_{cx}}$$

# Exemple

Intervalle de confiance pour la prédiction en vert

- Pour  $X=3.3$ , et  $Y= 106.43 - 16.07 X \Rightarrow$
- $Y= 106.43 - 16.07 * 3.3=53.399 \approx 53.4$



c.a.d tester :variable  
x n'a pas d'influence  
sur Y

## Tests d'hypothèse

Peut être effectué  
par le test de Student

- Test de **signification** globale de la régression:
- $H_0: \beta_1=0$  contre  $H_1: \beta_1 \neq 0$
- Si l'on ne peut rejeter  $H_0$ , alors on doit conclure qu'il n'y a pas de relation linéaire entre x et y.

- Calcule du Coefficient de détermination du modèle,  $R^2$   
(coefficient de corrélation en cas de X aléatoire) /  $R^2=r^2$

Rejette ou accepte le  
modèle

- Le test de pertinence de Fisher : teste le pouvoir explicative du modèle.

- $R^2 = SC_E / SC_T$

- L'estimation de  $\delta = SC_R / n - 2$

- $SC_E$ : Somme des carrés expliqués (dû)

par la régression =  $\hat{\beta}_1^2 * SC_X$

- $SC_T$ : Somme des carrés total =  $\sum_{i=1}^n (y_i - \bar{y})^2$

- $SC_R$ : Somme des carrés résiduels =  $\sum_{i=1}^n (y_i - \hat{y})^2$

- $SC_T = SC_R + SC_E$

- $\Rightarrow$  si  $R^2 = 1 \Rightarrow SC_R = 0$ , notre système explique tout les points.

- Si  $R^2 = 0$ , la variation de  $y$  est dû aux résidus  $\Rightarrow$  la variation de  $x$  n'explique pas la variation de  $y$ ;

- Calcul :
- $SC_T=2746.875$
- $SC_E=2711.571$
- $R^2=2711.571/2746.875=0.987$ , notre système explique 98.7% des points de l'échantillon.

# Analyse des résidus (vérification des hypothèses)

- Adéquation du modèle et homoscedasticité(graphiquement)
- Test de normalité(test de shapiro)
- Test d'indépendance des résidus(Durbin Watson)
- Détection des points aberrants(distance de cooks)

# Exemple

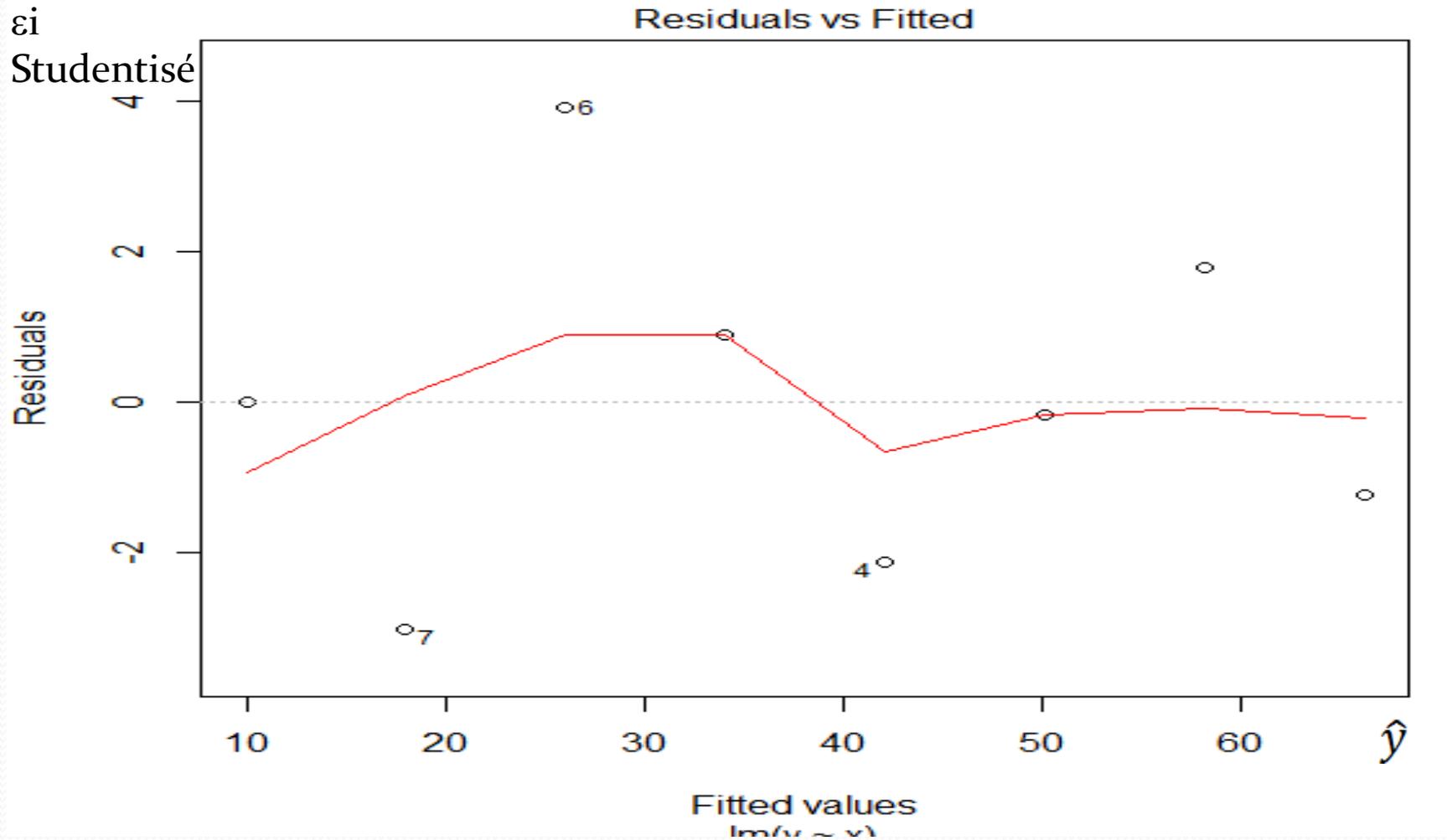
$$\hat{y} = 106.43 - 16.07 X$$

$x_i$	$y_i$	$\hat{y}_i$	$\varepsilon_i = (Y_i - \hat{y}_i)$	$(Y_i - \hat{y}_i)^* (Y_i - \hat{y}_i)$
2.5	65	66.25	-1.25	1.5625
3.0	60	58.21	1.79	3.204
3.5	50	50.18	-0.18	0.0324
4.0	40	42.14	-2.14	4.579
4.5	35	34.11	0.89	0.7921
5.0	30	26.07	3.93	15.445
5.5	15	18.03	-3.03	9.181
6.0	10	10	0	0
TOTAL	---		----	34.796

$$\hat{\sigma}^2 = 34.796 / 8 - 2 = 5.8$$

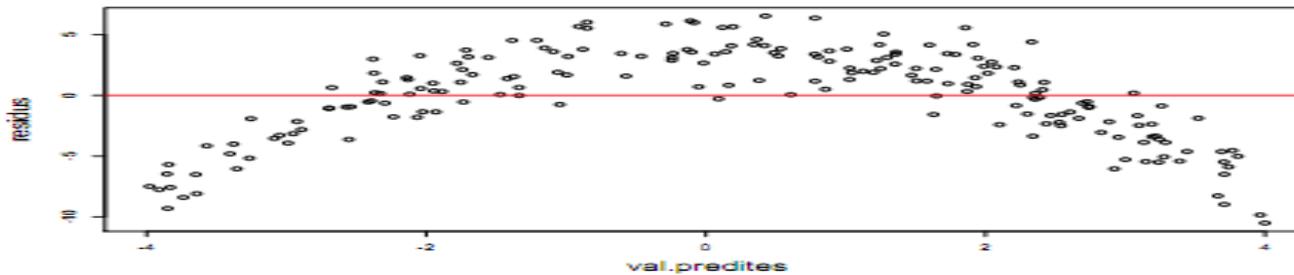
- Calcule de l'estimateur de la variance résiduel

# Aspect aléatoire et homoscedasticité

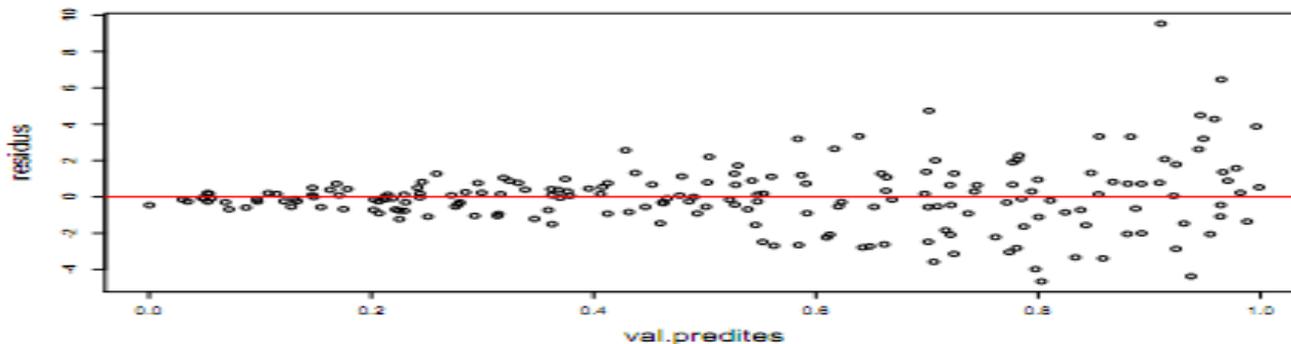


# Analyse des graphique des résidus (cas à changer le modèle)

- On observe un “structure évidente” dans les résidus (qui ne sont plus vraiment aléatoires).

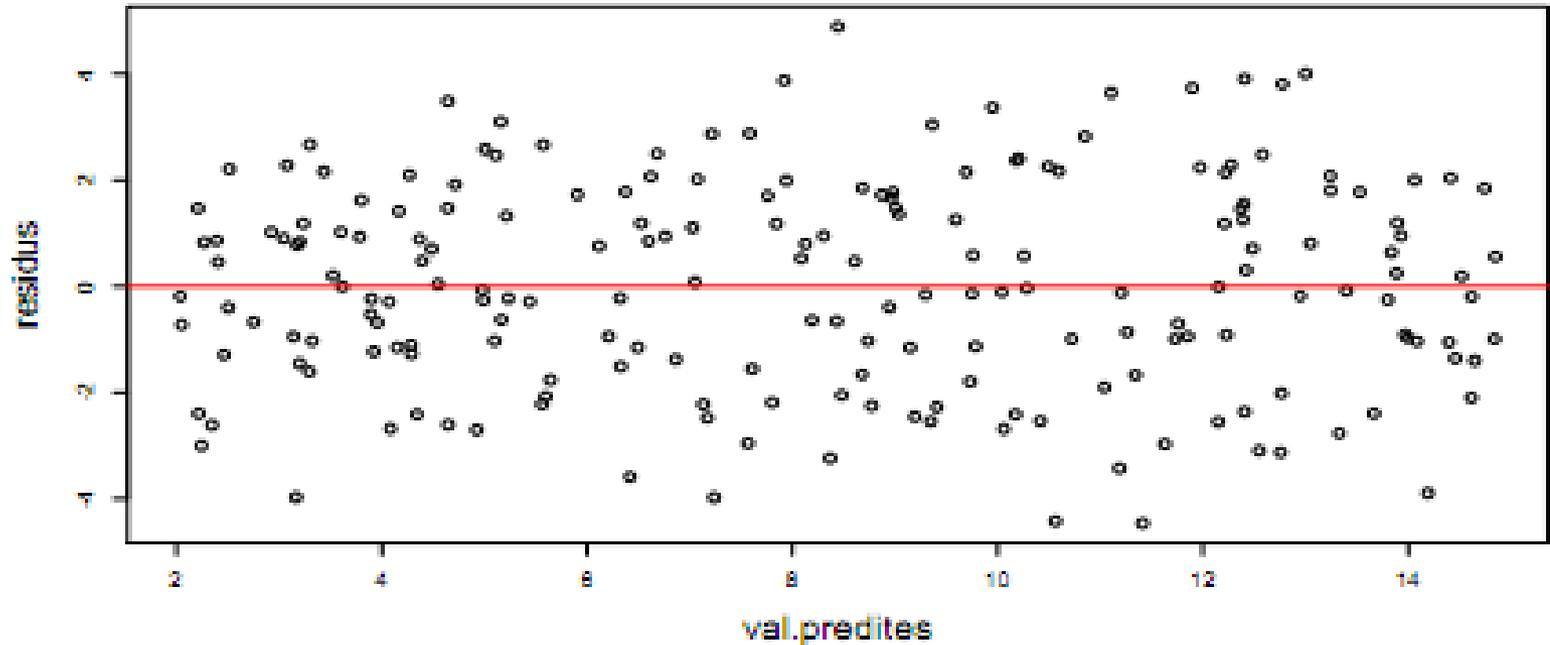


- La variance des résidus n'est pas constante (croît selon valeurs prédite) => pas d'homoscédascité.



# Analyse des graphique des résidus (cas vérifiant les hypothèses)

Comportement aléatoire des résidus et variance constante



# Régression linéaire multiple

- Principe:
- Déterminer la variable expliquée  $Y$ .
- Déterminer  $(p - 1)$  variables explicatives  $X_1, \dots, X_{p-1}$
- Il ne reste plus qu'à appliquer un modèle linéaire :
- $y = \beta_0 + \beta_1 x_1 + \dots + \beta_{p-1} x_{p-1} + \varepsilon$
- On mesure  $y_i$  pour  $x_{i,1}, \dots, x_{i,p-1}$  pour  $i = 1, \dots, n$ .

- Le problème posé est d'estimer les paramètres  $\beta_0, \dots, \beta_{p-1}$  du modèle de régression.
- Méthode des moindres carrés:

$$\min_{\beta_0, \dots, \beta_{p-1}} \sum_{i=1}^n \left( y_i - (\beta_0 + \beta_1 x_{i,1} + \dots + \beta_{p-1} x_{i,p-1}) \right)^2$$

- Sachant que :

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,p-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$y = X \beta + \varepsilon$$

- La méthode des moindres carrés consiste à trouver le vecteur  $\hat{\beta}$  qui minimise  $\|\varepsilon\|^2 = {}^t\varepsilon\varepsilon$
- Ce qui revient à résoudre :

$${}^tXy = {}^tXX\hat{\beta}$$

- D'où  $\hat{\beta} = ({}^tXX)^{-1} {}^tXy$
- t: la matrice transposée