Validation et détection des écarts

Bruno Pelletier

Outline

Introduction

Hypothèses du modèle linéaire

Modèle

$$\mathbf{Y} = \mathbb{X}\beta + \epsilon.$$

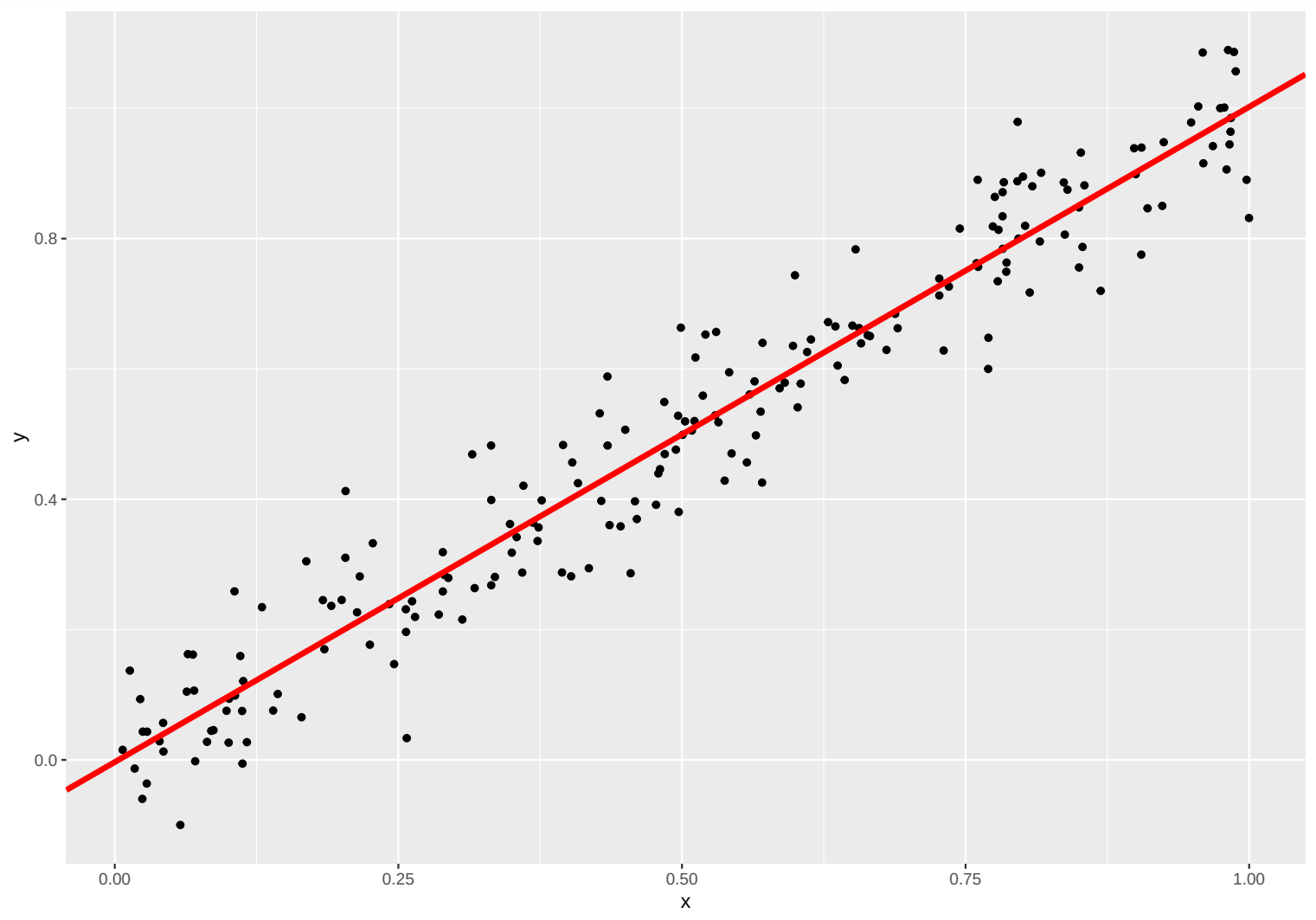
Hypothèses

1. Bruits centrés : $\mathbb{E}[\epsilon] = 0$
2. Homoscédasticité : $\mathbb{V}(\epsilon_i) = \sigma^2$, $\forall i = 1, \dots, n$.
3. Décorrélation : $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$, $\forall i \neq j$.
4. Hypothèse gaussienne : $\epsilon \sim \mathcal{N}_n(0, \sigma^2, \mathbf{I}_n)$.

Questions

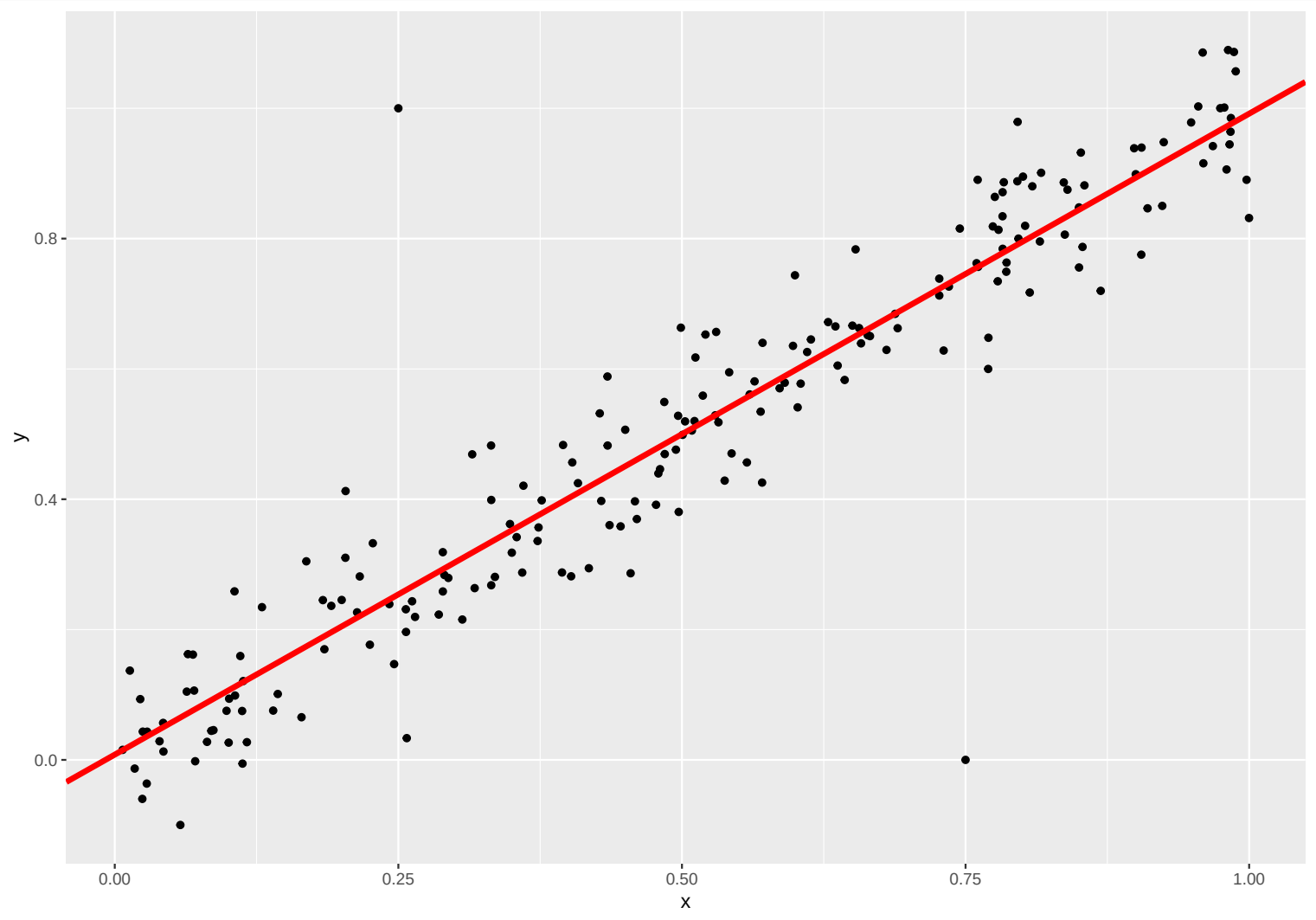
- Les hypothèses sont-elles satisfaites ?
- Certaines observations sont-elles ”atypiques”, susceptibles d’invalider le modèle ?

Exemple



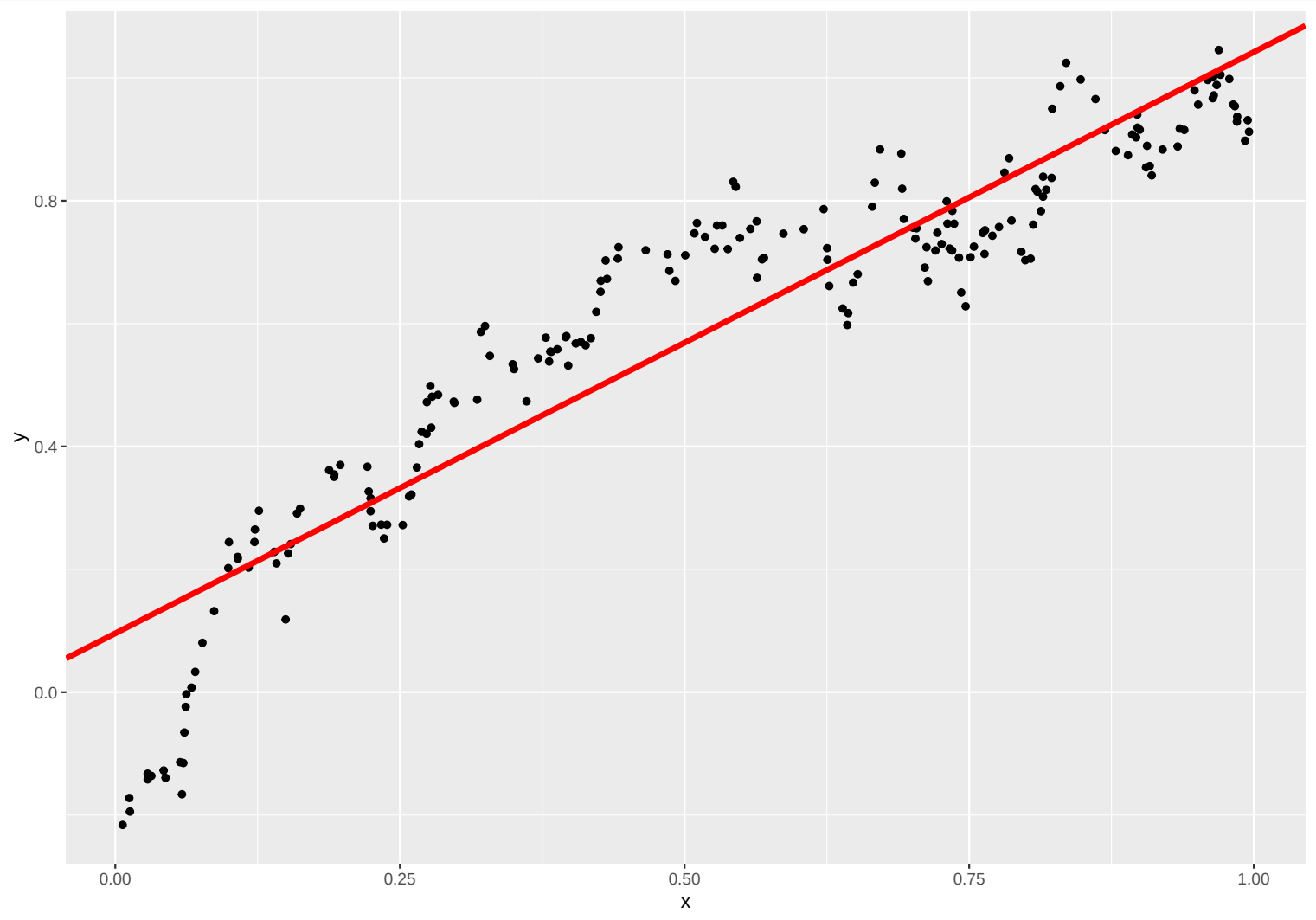
⇝ situation ”normale”.

Exemple



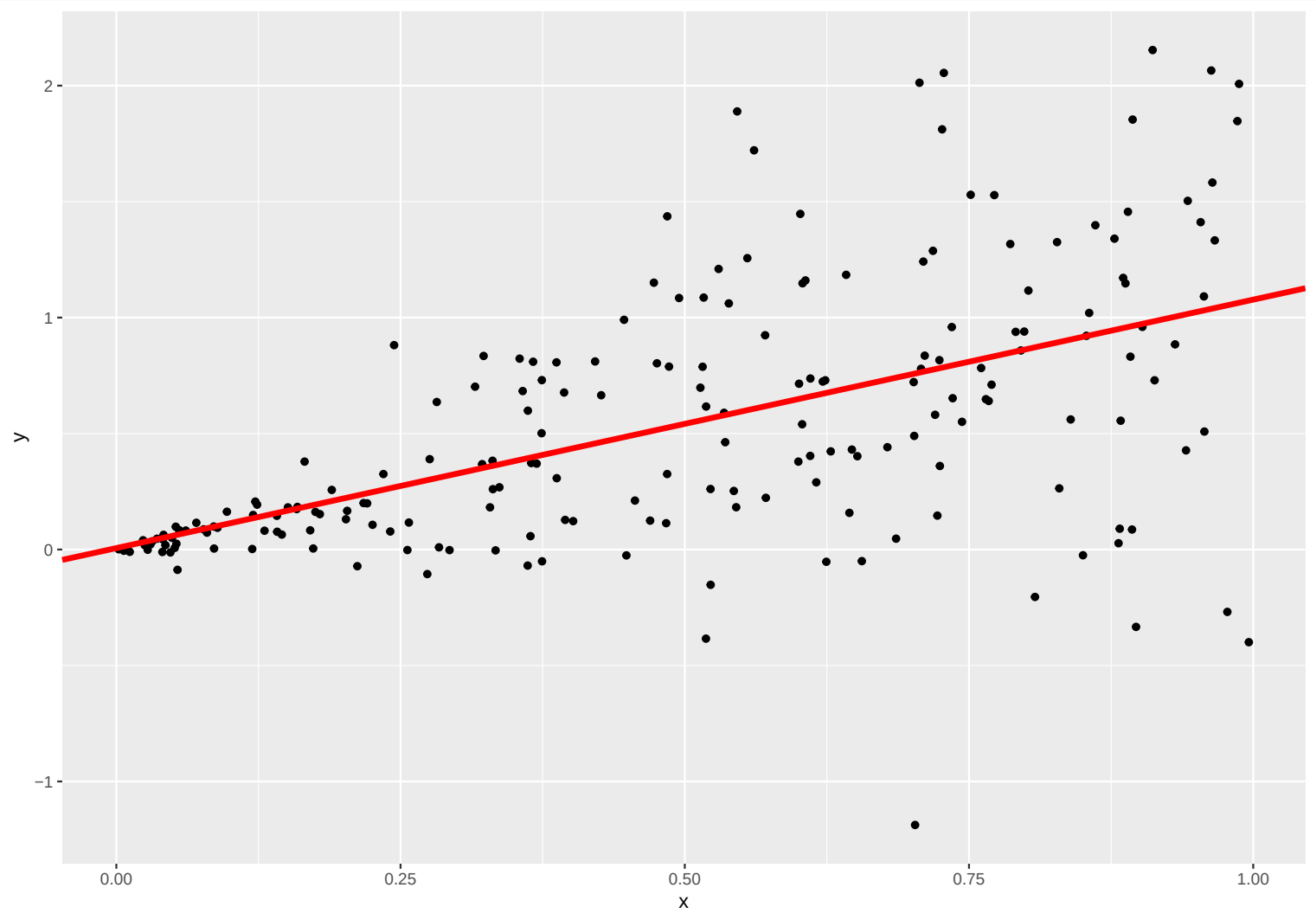
⇝ deux données aberrantes.

Exemple



⇝ défaut de décorrélation

Exemple



⇝ défaut d'homoscédasticité.

Etude des résidus

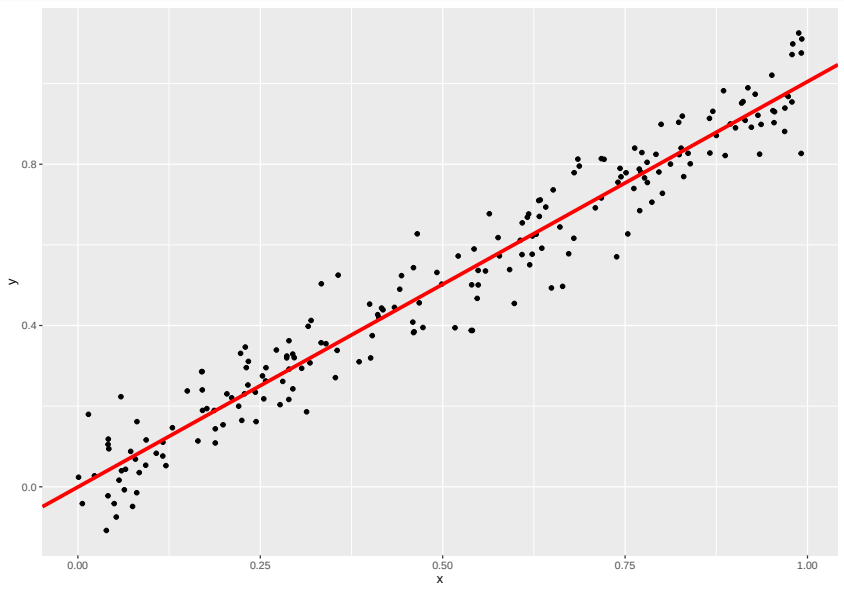
Etude des résidus

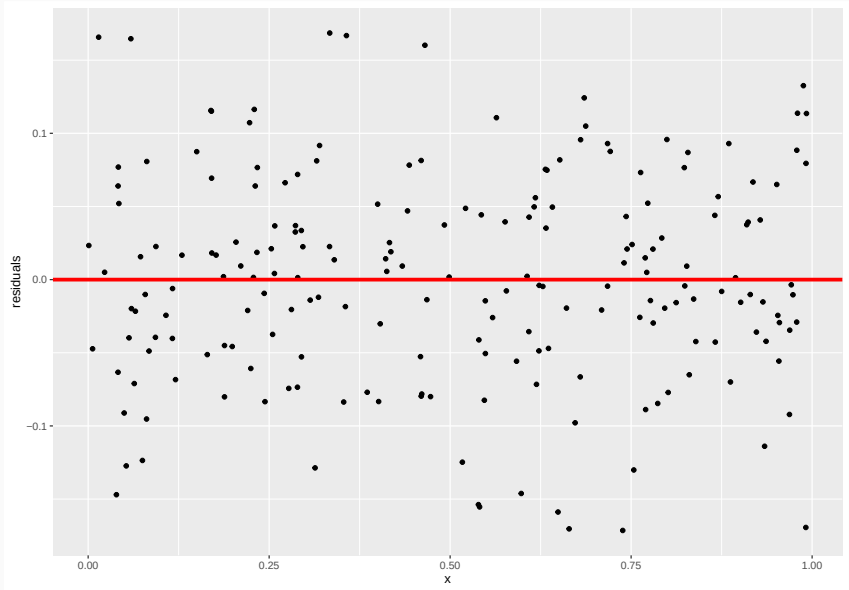
- L'examen des résidus constitue une étape primordiale de la régression linéaire.
- Les méthodes sont principalement graphiques. Il est difficile d'énoncer des règles strictes.

Rappels

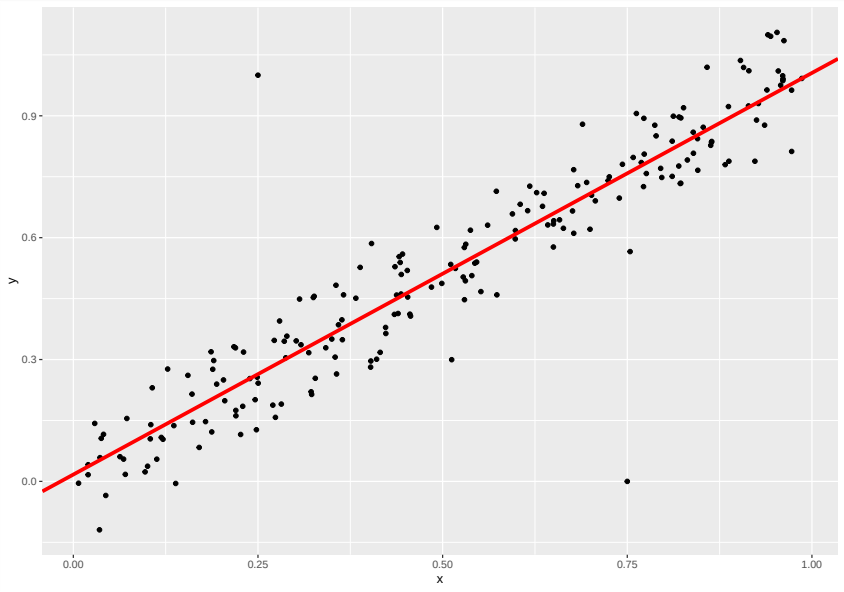
- Résidus : $\hat{\epsilon}_i = y_i - \hat{y}_i$ et $\hat{\epsilon} = \mathbf{Y} - \hat{\mathbf{Y}}$.

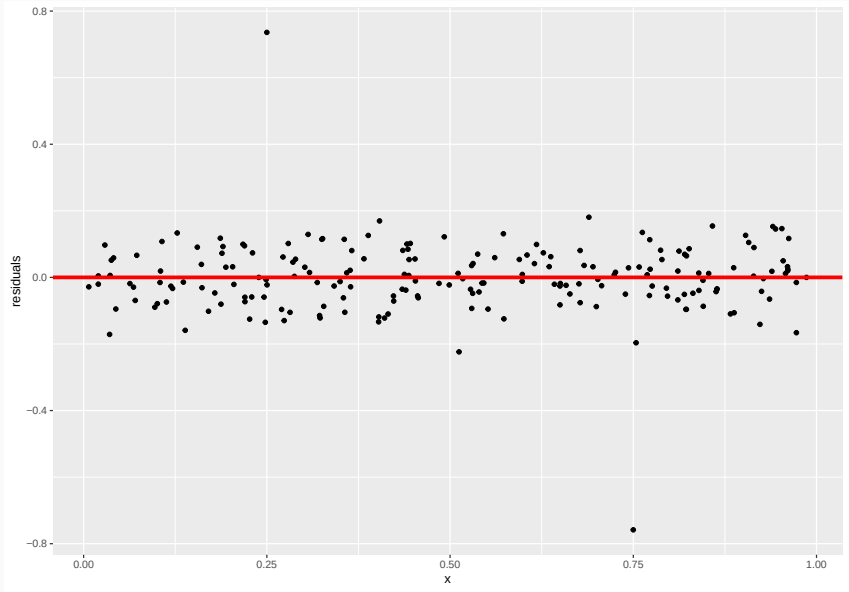
Exemple : situation ”normale”





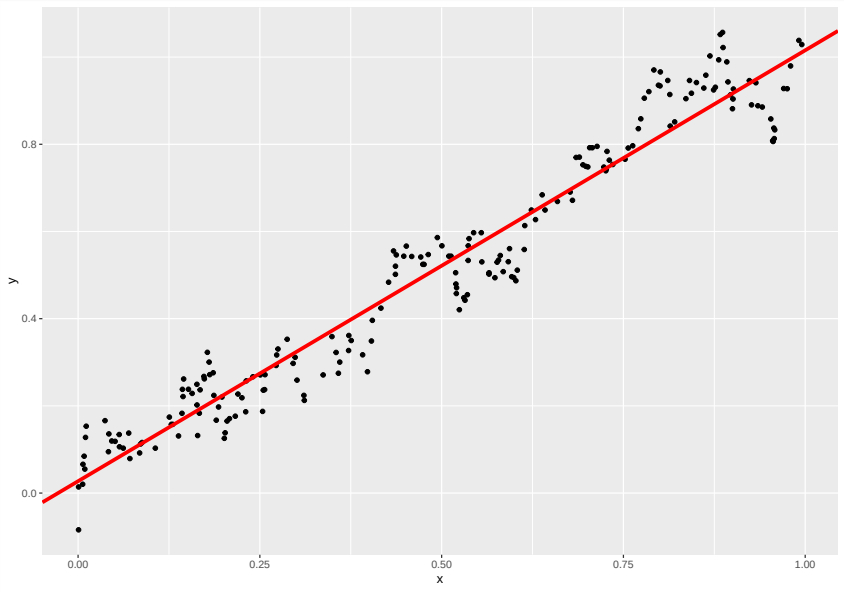
Exemple : données aberrantes/atypiques

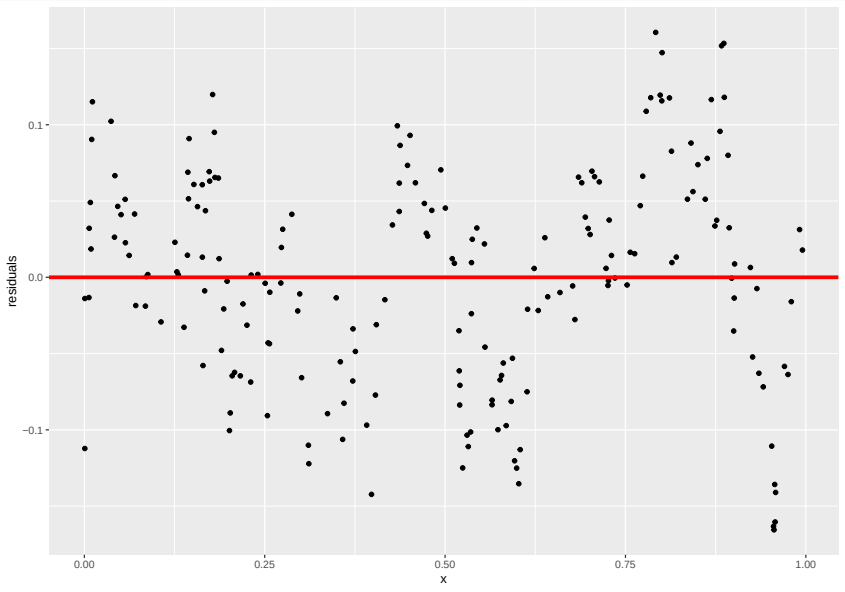




- Deux résidus ont des valeurs anormalement élevées, par rapport aux autres.
- Détection statistique : $\rightsquigarrow$ calibration, test.

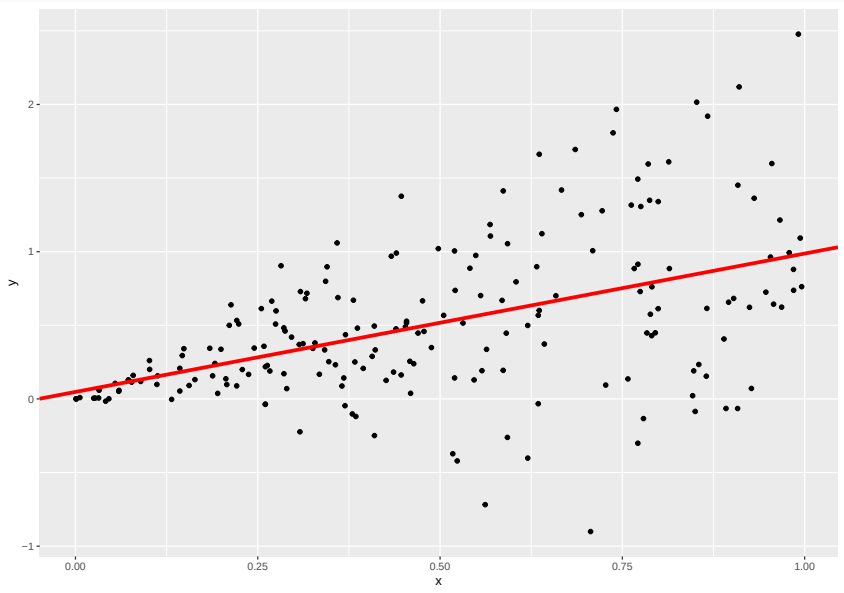
Exemple : défaut de décorrélation

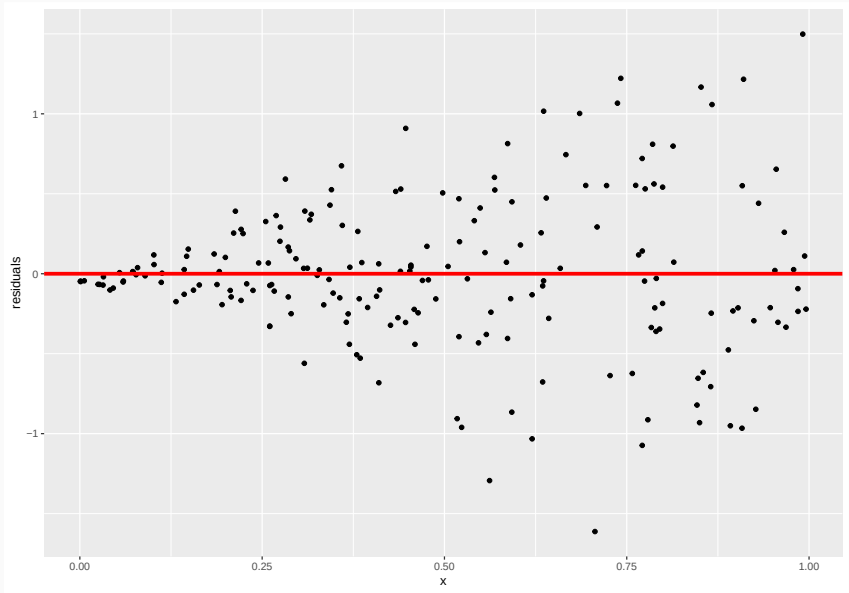




- Motif ”non linéaire” visible sur ces graphiques.
- Causes possibles : bruits autorégressifs, dépendance non linéaire en une variable explicative.

Exemple : défaut d'homoscédasticité





- Tendance de variance croissante avec la variable explicative.
- Hétéroscédasticité des résidus.

Etude des résidus

Modèle gaussien

- Observations : $\mathbf{Y} = \mathbb{X}\beta + \epsilon$, où $\epsilon \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$.
- Valeurs ajustées : $\hat{\mathbf{Y}} = \mathbb{X}\hat{\beta} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y}$, où

$$\mathbf{H} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}',$$

est la matrice de projection sur $\mathrm{Im}\,(\mathbb{X})$ et s'appelle la matrice chapeau (hat matrix). Notons $\mathbf{H} = [h_{ij}]_{1 \leq i,j \leq n}$.

Résidus

- Résidus : $\hat{\epsilon} = \mathbf{Y} - \hat{\mathbf{Y}}$.
- Sous les hypothèses standard du modèle gaussien :
  - $\hat{\epsilon}$ suit une loi gaussiene
  - $\mathbb{E}\left[\hat{\epsilon}\right] = 0$ et $\mathbb{V}\left(\hat{\epsilon}\right) = \sigma^2(\mathbf{I} - \mathbf{H})$.

$$\rightsquigarrow \quad \hat{\epsilon}_i \sim \mathcal{N}(0, \sigma^2(1 - h_{ii})).$$

Résidus studentisés

- $\hat{\epsilon}_i \sim \mathcal{N}(0, \sigma^2(1 - h_{ii}))$.
- Donc

$$\frac{\hat{\epsilon}_i}{\sigma\sqrt{1 - h_{ii}}} \sim \mathcal{N}(0, 1).$$

Résidus studentisés

$$r_i^* = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}} \underset{\text{approx.}}{\sim} \mathcal{T}(n - p).$$

Remarques

- Les résidus studentisés ne suivent pas exactement une loi de Student.
- $\hat{\epsilon}$ et $\hat{\sigma}$ ne sont pas indépendants.
- Les r_i^* ont une variance approxmativement égale à 1.

Résidus studentisés par validation croisée

1. Ajuster un modèle linéaire sur les n observations pour former les résidus $\hat{\epsilon}_i$.
2. Pour chaque $i = 1, \dots, n$:
   2.1 Ajuster un modèle linéaire sur toutes les observations sauf la $i^{\text{ème}}$.
   2.2 Noter $\hat{\sigma}^2_{(i)}$ l'estimateur sans biais de la variance résultant.
   2.3 Définir

$$t_i^* = \frac{\hat{\epsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}}.$$

Théorème (Résidus studentisés par validation croisée)

$$t_i^* \sim \mathcal{T}(n - p - 1).$$

Remarque
On montre que

$$t_i^* = r_i^* \sqrt{\frac{n - p - 1}{n - p - (r_i^*)^2}}.$$

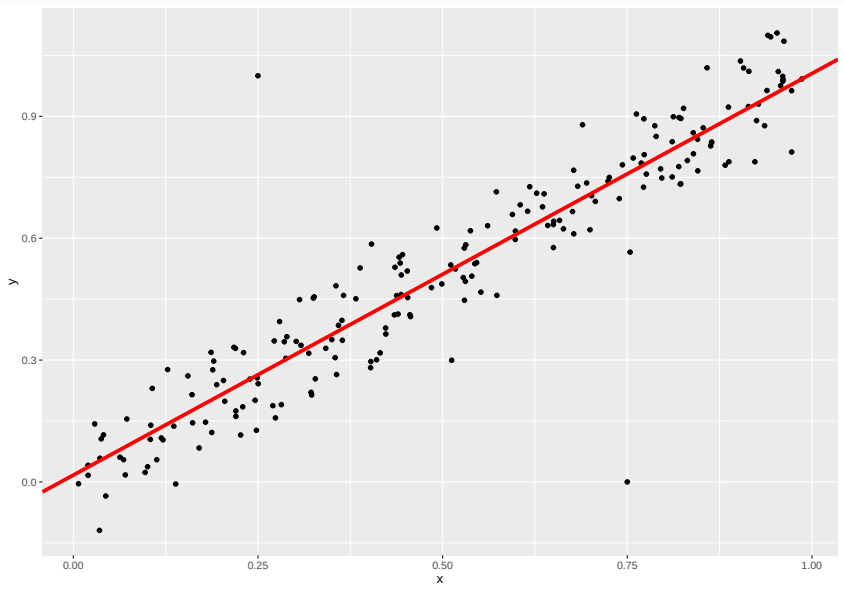
Utilisation des résidus studentisés.

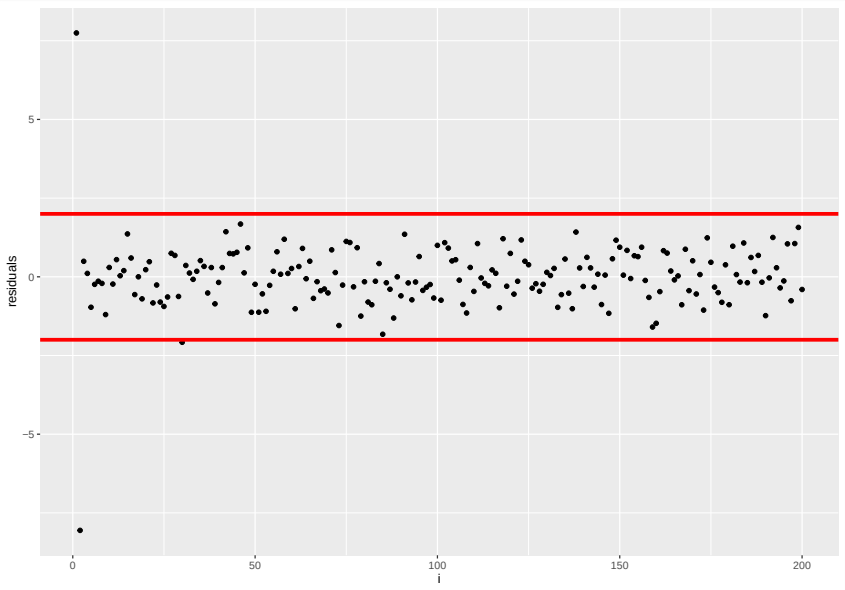
Donnée aberrante
Une donnée (x_i, y_i) est considérée comme aberrante si son résidu studentisé par validation croisée t_i^* excède un seuil relatif à la loi de Student $\mathcal{T}(n-p-1)$, par exemple si $|t_i^*| > t_{n-p-1}(1-\alpha/2)$.

Remarque :

- En pratique, si $n-p-1 \geq 30$, avec $\alpha = 5\%$, $t_{n-p-1}(1-\alpha/2) \approx 2$.
- Diagnostic graphique : tracés des résidus studentisés par VC en fonction de i, ou d'une variable explicative ou des valeurs ajustées $\hat{y}_i$.

Exemple : donnée aberrante





⇝ trois observations sont détectées comme aberrantes.

Analyse de la normalité

Outils graphiques

- Histogramme des résidus studentisés t_i^*.
- Tracés des t_i^* en fonction de l'espérance de ces qunatiles sour l'hypothèse de normalité (Q-Q plot).

Mesure diagnostique

- Test de normalité de Shapiro-Wilk.

Remarque

Les t_i^* suivent une loi de Student et ne sont pas indépendants. Néanmoins, si n est grand et n est grand devant p, alors cette loi est proche d'une loi $\mathcal{N}(0, 1)$.

Analyse de l'homoscédasticité et de la structure des résidus

Outils graphiques

- Tracés des résidus studentisés t_i^* en fonction des valeurs ajustées $\hat{y}_i$.
- Selon le problème, tracé des t_i^* en fonction d'autres variables.
- Détecter la présence d'une structure (tendance croissante, décroissante, oscillante, etc.)
- Estimer la courbe de régression de t_i^* sur $\hat{y}_i$ par polynômes locaux par exemple.

Mesure diagnostique

- Test de Breusch et Pagan : détection de l'hétéroscédasticité.
- Test de Durbin Watson : détection d'une auto-corrélation.

Etude de la matrice chapeau

Influence d'une observation

Question
Quelle est l'influence d'une observation x_i sur l'ajustement de Y_i dans le modèle ?

- y_i est prédit par $\hat{y}_i$ et le résidu associé est $\hat{\epsilon}_i \sim \mathcal{N}(0, \sigma^2(1 - h_{ii}))$.

$\rightsquigarrow$ Dépendance de la variance des résidus en les variables explicatives au travers des éléments h_{ii}.

- $\hat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j = h_{ii} Y_i + \sum_{j \neq i} h_{ij} Y_j$.

$\rightsquigarrow$ h_{ii} traduit l'influence de Y_i sur sa valeur prédite $\hat{Y}_i$.

Influence d'une observation

Ajustement sur les n observations

- La valeur ajustée est $\hat{Y}_i$
- L'erreur (résidu) est $Y_i - \hat{Y}_i$.

Ajustement sur toutes les observations sauf la $i^{\text{ème}}$.

- Ce modèle est utilisé pour prédire Y_i à partir de x_i.
- On note $\hat{Y}_i^{(p)}$ la valeur prédite.
- L'erreur en prévision est $Y_i - \hat{Y}_i^{(p)}$.

Propriété

$$Y_i - \hat{Y}_i = (1 - h_{ii})(Y_i - \hat{Y}_i^{(p)}).$$

$\rightsquigarrow$ variation dans l'erreur de prévision selon que la $i^{\text{ème}}$ observation est prise en compte ou non.

Propriétés de la matrice chapeau

Hat matrix (matrice de projection orthogonale sur $\mathrm{Im}\,(\mathbb{X})$)
Soit $\mathbf{H} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$.

Propriété 1

$\mathbf{H}$ est idempotente : $\mathbf{H}^2 = \mathbf{H}$.

En effet $\mathbf{H} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}' = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}' = \mathbf{H}$.

Il en résulte que $\mathrm{tr}(\mathbf{H}^2) = \mathrm{tr}(\mathbf{H})$.

- D'une part : $\mathrm{tr}(\mathbf{H}^2) = \sum_{i=1}^n \sum_{j=1}^n h_{ij}h_{ji} = \sum_{i,j=1}^n h_{ij}^2$ car $\mathbf{H}$ est symétrique.
- D'autre part : $\mathrm{tr}(\mathbf{H}) = \sum_{i=1}^n h_{ii}$. En outre

$$\mathrm{tr}(\mathbf{H}) = \mathrm{tr}\left(\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\right) = \mathrm{tr}\left((\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{X}\right) = \mathrm{tr}(\mathbf{I}_p) = p.$$

$$\rightsquigarrow \quad \sum_{i,j=1}^n h_{ij}^2 = \sum_{i=1}^n h_{ii} = p.$$

Propriétés de la matrice chapeau

Propriété 2

$0 \leq h_{ii} \leq 1$

Soit e_i le $i^{\text{ème}}$ vecteur de la base canonique de $\mathbb{R}^n$. On a :

$$h_{ii} = e_i' \mathbf{H} e_i = e_i' \mathbf{H}^2 e_i = e_i' \mathbf{H}' \mathbf{H} e_i = \|\mathbf{H} e_i\|^2 \geq 0.$$

$\mathbf{H}$ étant une matrice de projection, $\|\mathbf{H}u\| \leq \|u\|$ pour tout $u \in \mathbb{R}^n$. Donc $\|\mathbf{H}e_i\| \leq \|e_i\| = 1$ et donc $h_{ii} \leq 1$.

$\rightsquigarrow$ cas extrémaux : $h_{ii} = 0$ et $h_{ii} = 1$.

Cas extrémaux

Cas $h_{ii} = 0$
On a $h_{ii} = \|\mathbf{H}e_i\|^2 = \sum_{j=1}^{n} h_{ji}^2$. Donc

$$h_{ij} = h_{ji} = 0 \qquad \text{pour tout } j = 1, \dots, n.$$

Conséquence

$$\hat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j = 0.$$

$\rightsquigarrow$ Y_i n'a pas d'influence sur $\hat{Y}_i$.

Cas extrémaux

Cas $h_{ii} = 1$
Comme $\mathbf{H}$ est idempotente, on a
$h_{ii} = \sum_{j=1}^{n} h_{ji}^2 = h_{ii}^2 + \sum_{j \neq i} h_{ji}^2$. Donc

$$h_{ij} = h_{ji} = 0 \quad \text{pour tout } j \neq i.$$

Conséquence

$$\hat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j = Y_i.$$

$\rightsquigarrow \hat{Y}_i$ est entièrement déterminée par Y_i.

Effet levier

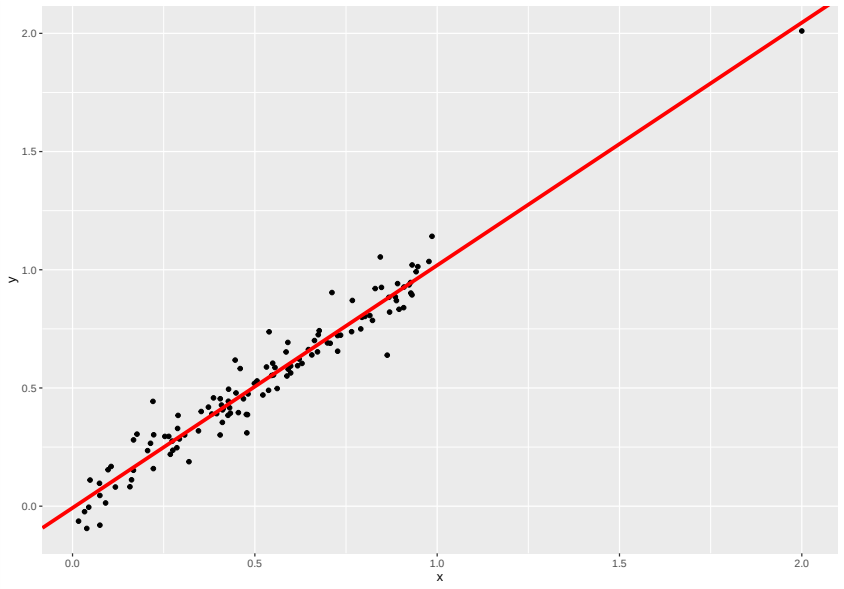
Ainsi $\hat{Y}_i$ est d'autant plus influencée par Y_i que h_{ii} est proche de 1. Seuil ?

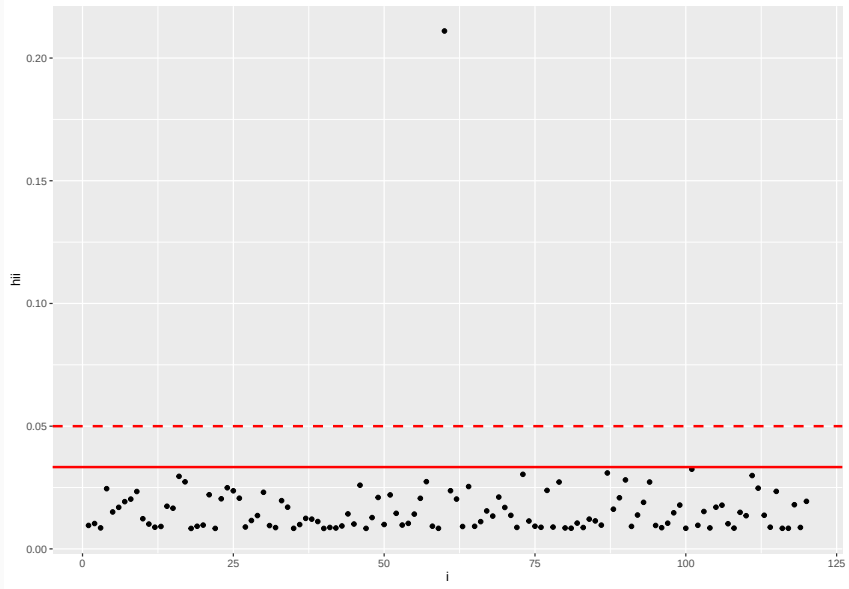
On a montré que $\sum_{i=1}^{n} h_{ii} = p$, donc la valeur moyenne des h_{ii} est $\frac{p}{n}$.

Définition (Effet levier)
On dit que la donnée (x_i, y_i) a un effet levier si :

- $h_{ii} \geq 2p/n$ (Hoaglin and Welsh, 1978).
- $h_{ii} \geq 3p/n$ pour $p > 6$ et $n - p > 12$ (Velleman and Welsh, 1981).
- $h_{ii} \geq 1/2$ (Huber, 1981).

Exemple : effet levier





⇝ Un point a un effet levier, mais n'est pas un point aberrant.

Résumé : donnée aberrante et effet levier

Donnée aberrante
Si $|t_i^*| > t_{n-p-1}(1-\alpha/2)$.

$\rightsquigarrow$ donnée atypique par rapport à la variable réponse.

Effet levier
Si $h_{ii} > cp/n$, pour $c = 2$ ou $c = 3$.

$\rightsquigarrow$ donnée atypique par rapport aux variables explicatives.

Mesures d'influence

Influence sur l'estimation de β

Quelle est l'influence des données aberrantes à effet levier sur l'estimation des paramètres du modèle ?

Approche

1. Estimer β à partir de toutes les observations : $\rightsquigarrow \hat{\beta}$.
2. Estimer β à partir de toutes les observations sauf la $i^{\text{ème}}$: $\rightsquigarrow \hat{\beta}^{(i)}$.
3. Former une mesure de distance entre $\hat{\beta}$ et $\hat{\beta}^{(i)}$, par exemple de la forme $(\hat{\beta} - \hat{\beta}^{(i)})' A (\hat{\beta} - \hat{\beta}^{(i)})$ avec $A \in \mathcal{M}_{p,p}(\mathbb{R})$ symétrique et définie positive, i.e. induite par un produit scalaire dans $\mathbb{R}^p$.

Rappel : Ellipsoïde de confiance de niveau $1-\alpha$ sur β.

$$I_{1-\alpha} = \left\{ \beta \in \mathbb{R}^p \ : \ \frac{1}{p\hat{\sigma^2}} (\hat{\beta} - \beta) \mathbb{X}' \mathbb{X} (\hat{\beta} - \beta) \leq f_{p,n-p}(1-\alpha) \right\}.$$

Distance de Cook

Distance de Cook

La distance de Cook pour la $i^{\text{ème}}$ observation est définie par :

$$C_i = \frac{1}{p\hat{\sigma}^2}(\hat{\beta}^{(i)} - \hat{\beta})'\mathbb{X}'\mathbb{X}(\hat{\beta}^{(i)} - \hat{\beta}).$$

Propriété

$$C_i = \frac{1}{p}\frac{h_{ii}}{(1-h_{ii})^2}\frac{\hat{\epsilon}_i^2}{\hat{\sigma}^2} = \frac{1}{p}\frac{h_{ii}}{1-h_{ii}}(r_i^*)^2.$$

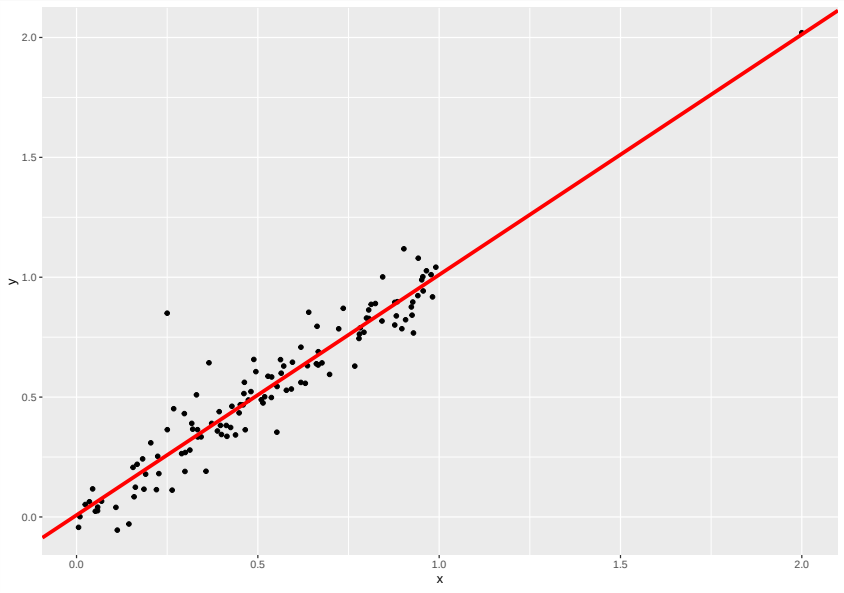
Interprétation

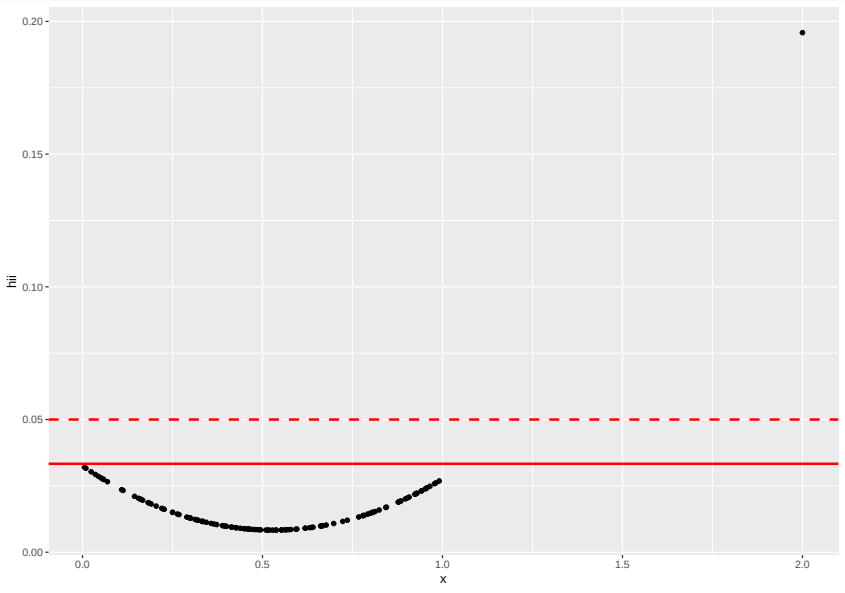
C_i est élevée si (x_i, y_i) influence fortement l'estimation de β.

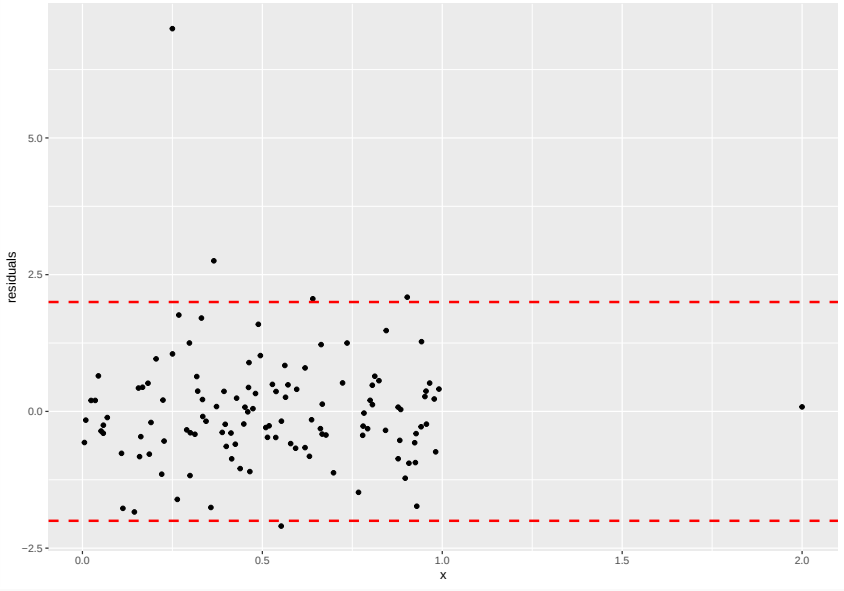
En outre, C_i est potentiellement élevée si :

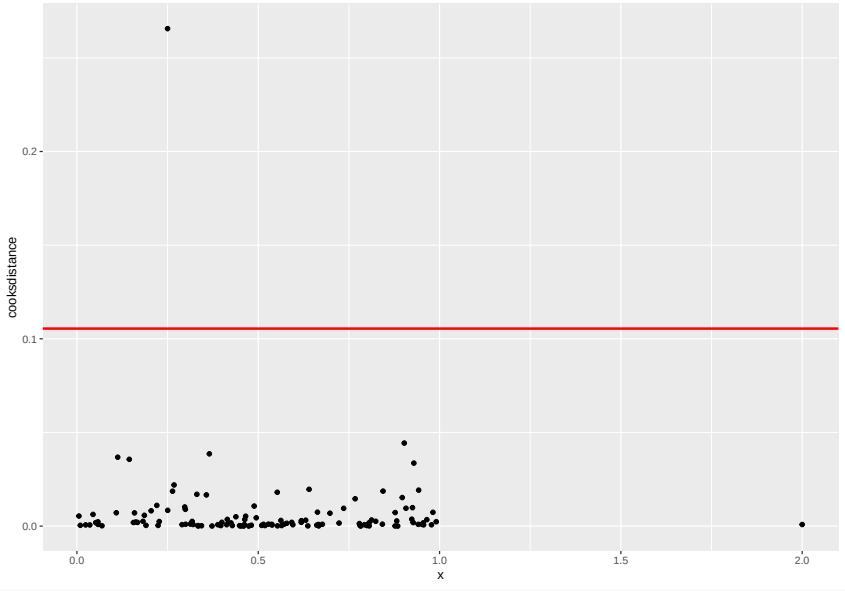
- $|r_i^*|$ est élevé $\rightsquigarrow$ donnée aberrante,
- h_{ii} est proche de 1 $\rightsquigarrow$ point levier.
- les deux.

Exemple

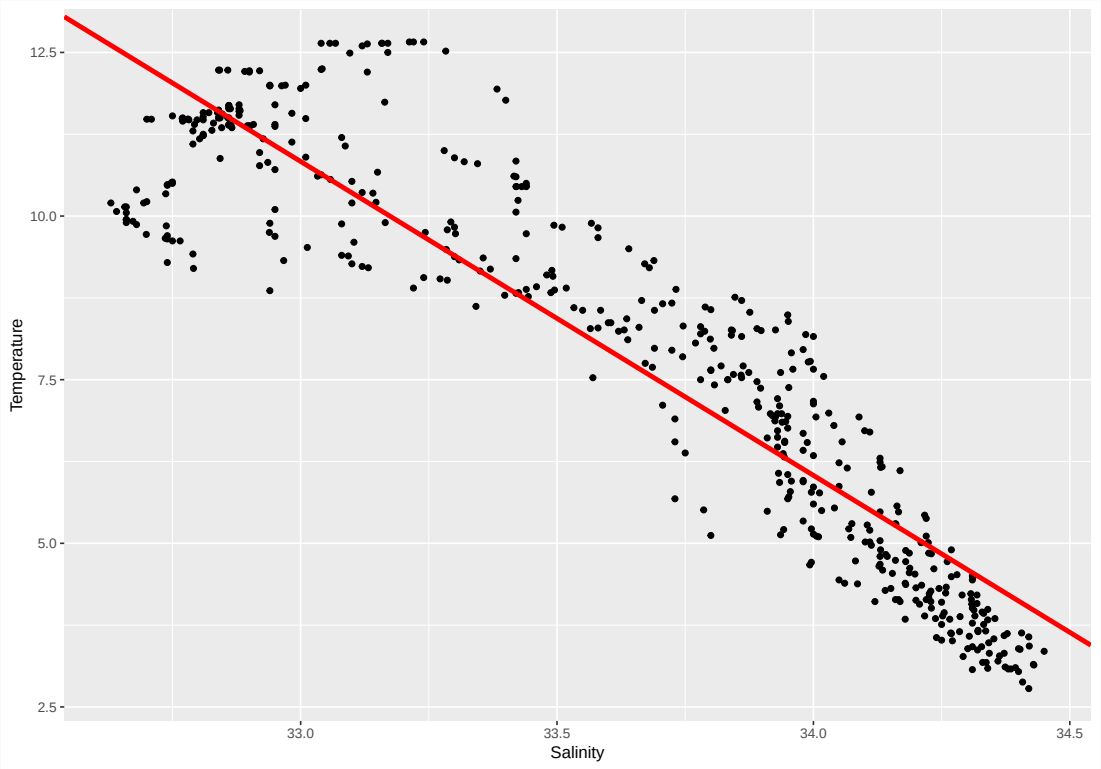

y
x
0.0
0.5
1.0
1.5
2.0


hii
x
0.00
0.05
0.10
0.15
0.20
0.0
0.5
1.0
1.5
2.0


residuals
x
−2.5
0.0
2.5
5.0
0.0
0.5
1.0
1.5
2.0

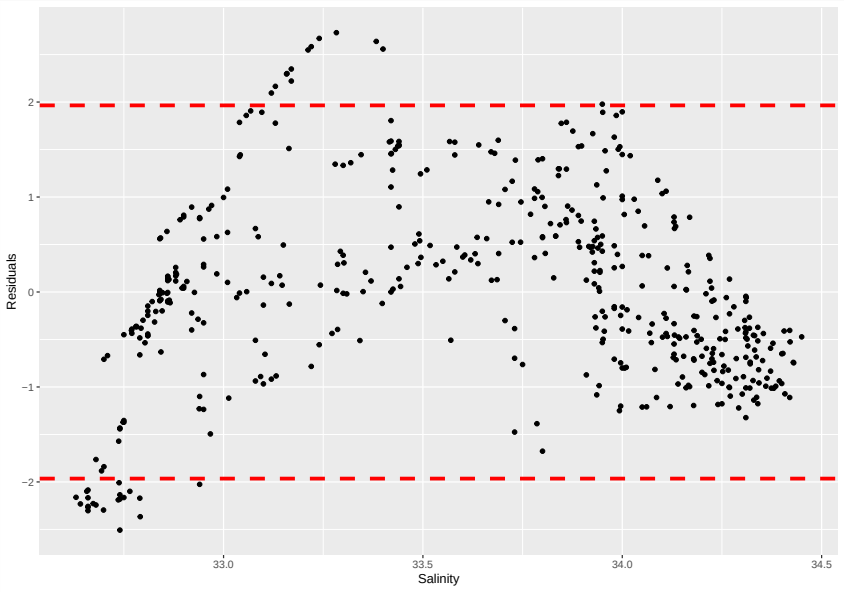

cooksdistance
x
0.0
0.1
0.2
0.0
0.5
1.0
1.5
2.0

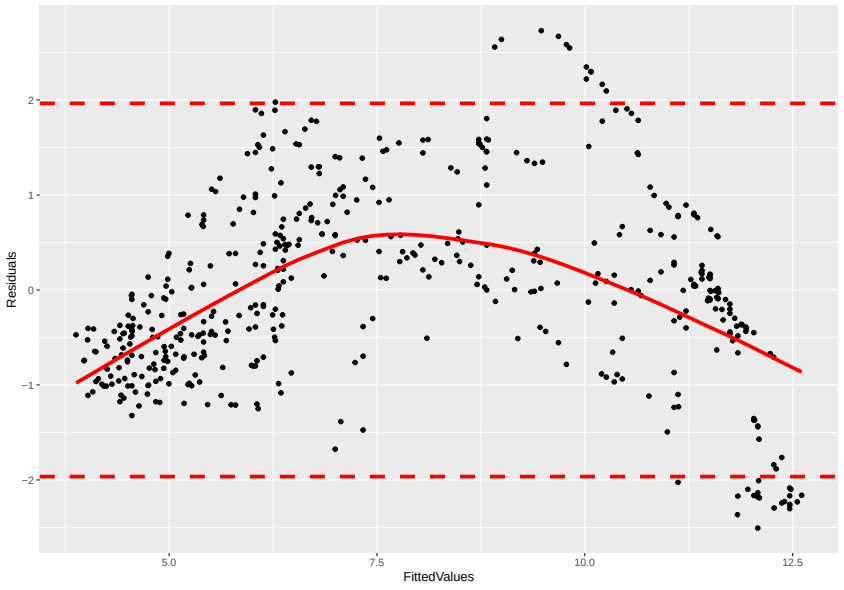
CalCOFI (1) : $y = \beta_0 + \beta_1 x + \epsilon_i$

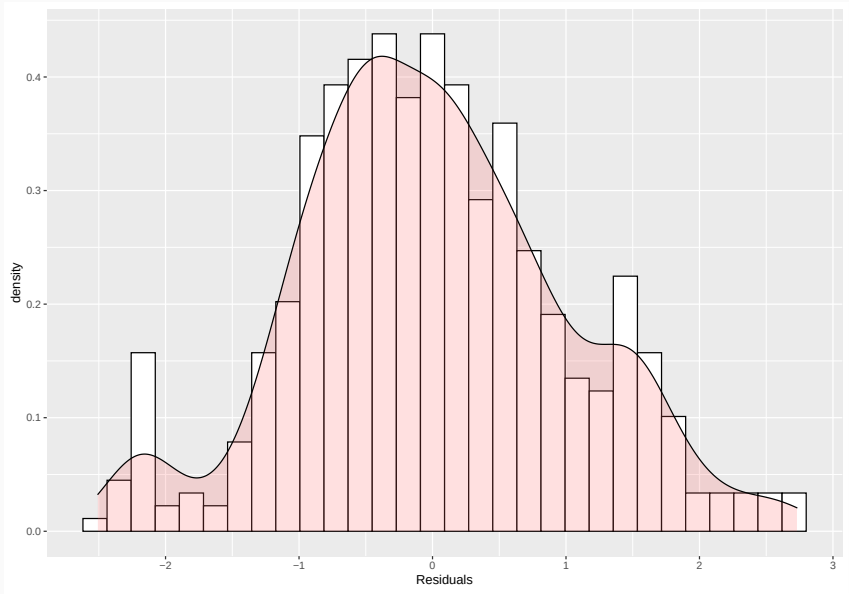


- Droite de régression : $y = 169.1178 - 4.79646x$.
- $R^2 = 0.8517$.

CalCOFI (1) : analyse des résidus



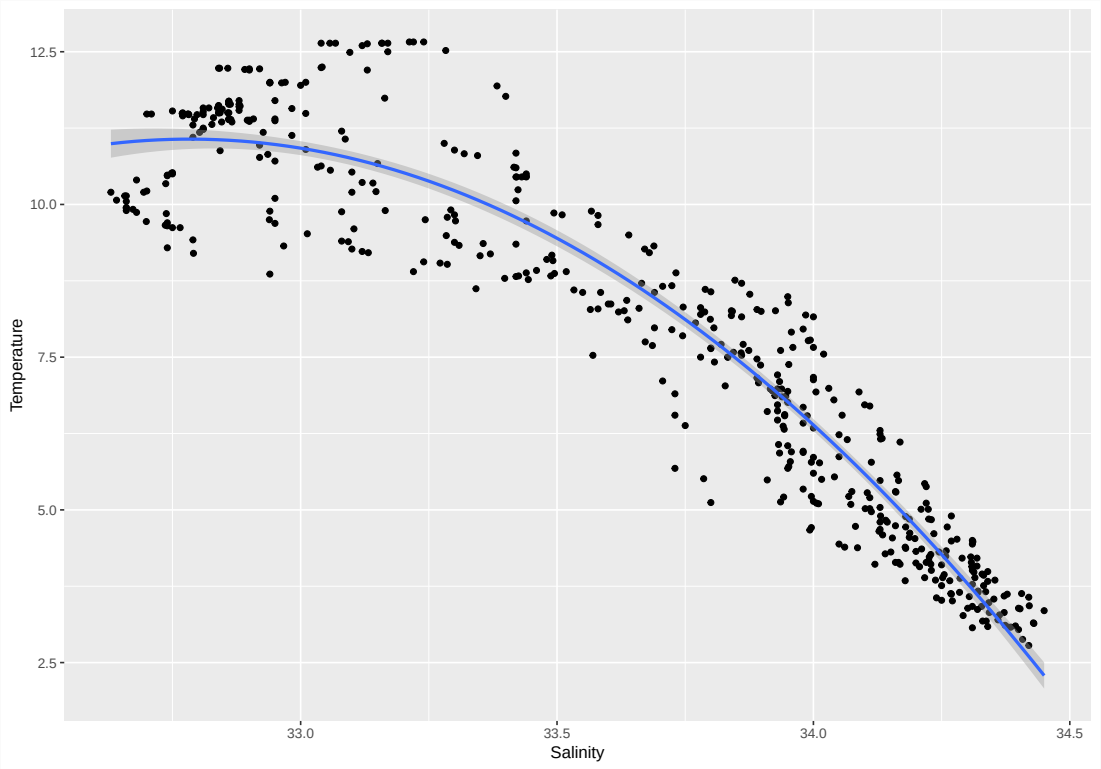




Sahpiro-Wilk's normality test

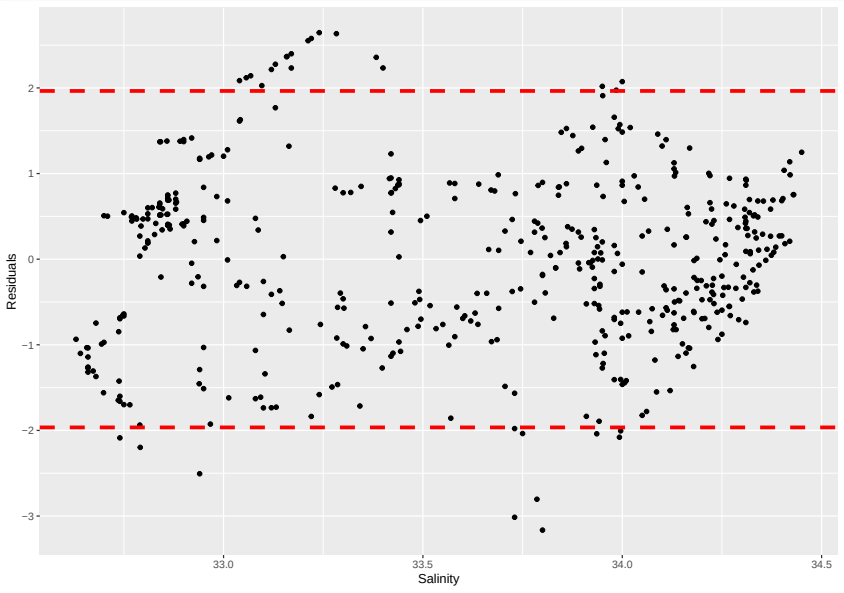
p-value : 0.001155

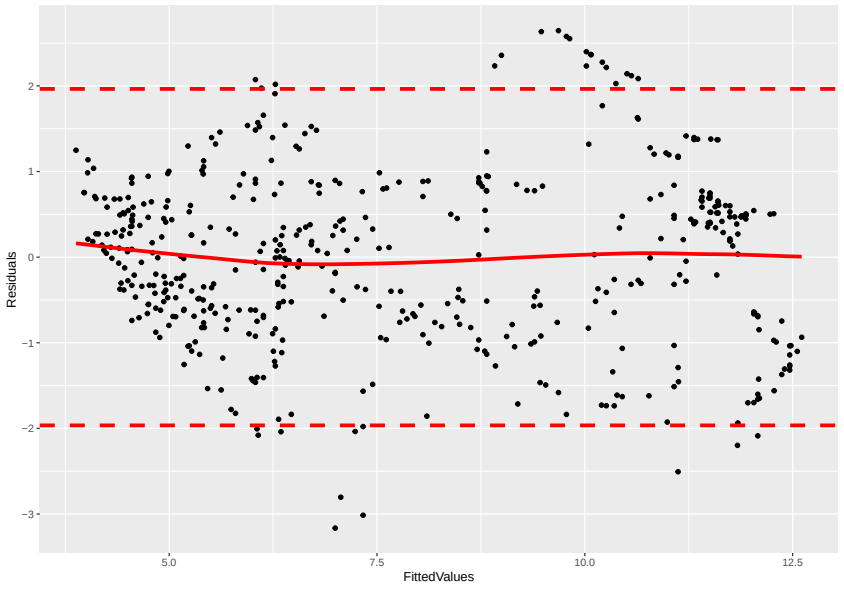
CalCOFI (2) : $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon_i$

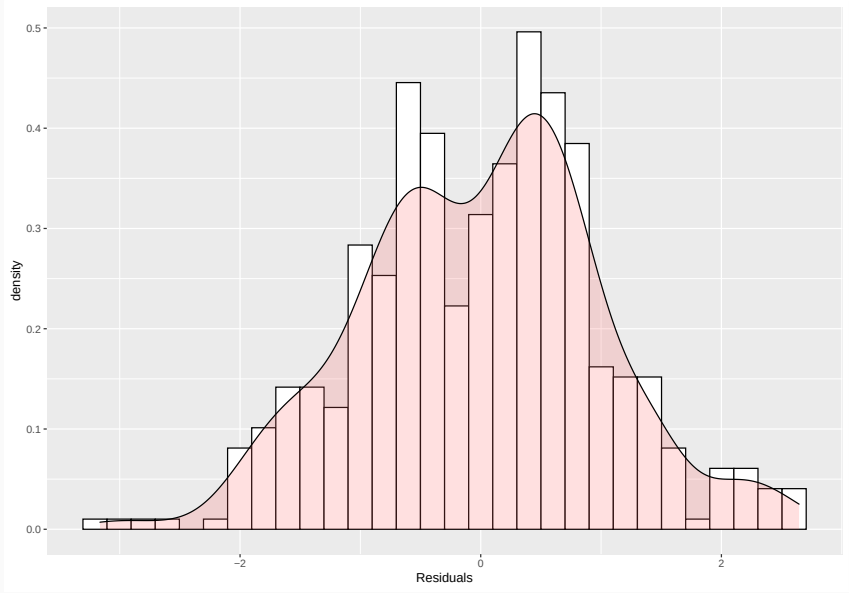


- Ajustement : $y = -3387.5941 + 207.3393x - 3.1622x^2$.
- $R^2 = 0.9138$.

CalCOFI (2) : analyse des résidus





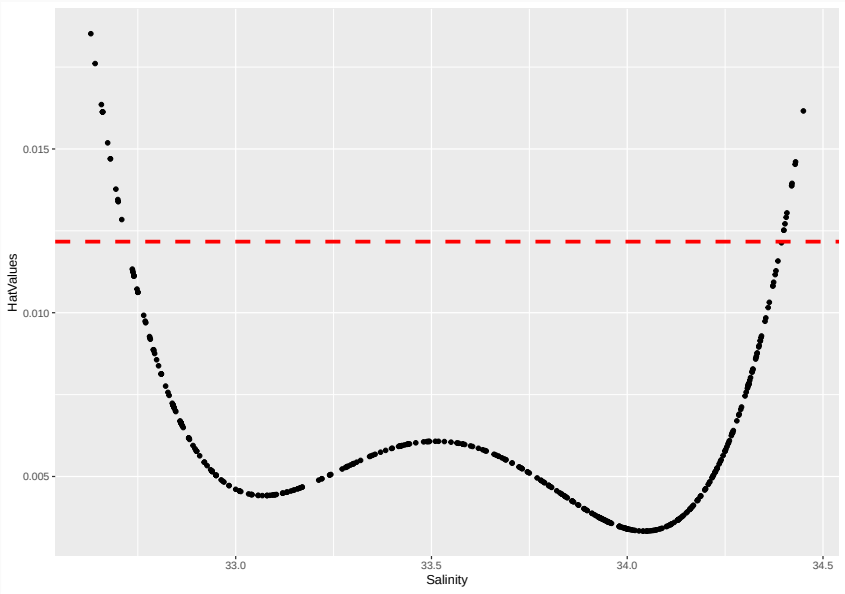


Sahpiro-Wilk's normality test

p-value : 0.08177

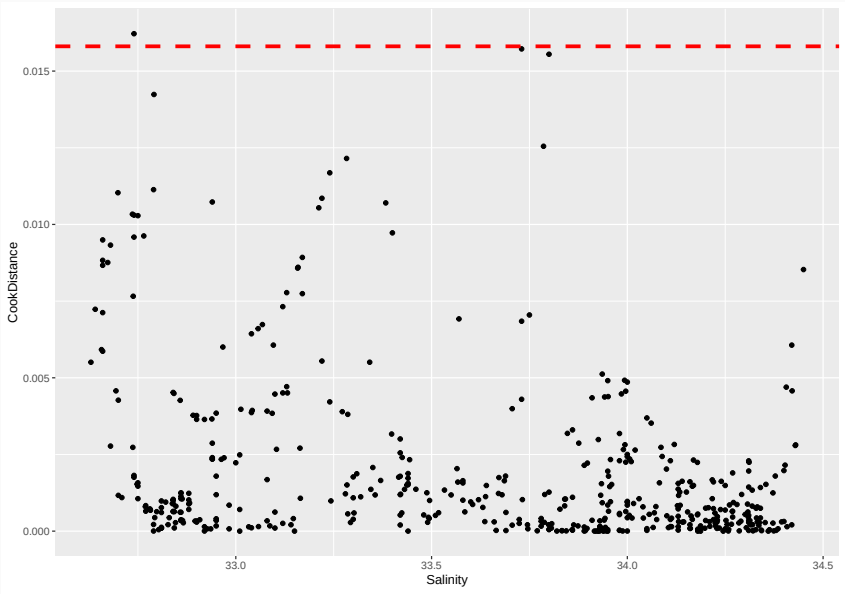
CalCOFI (2) : analyse des résidus

Fitted values



Seuil à $3p/n$

Cook's distances



Quantile d'ordre 10%

Autres diagnostics

Influence non linéaire : résidus partiels

Modèle linéaire

$$Y_i = \beta_0 + \beta_1 xi, 1 + \cdots + \beta_j x_{i,j} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i.$$

Résidus : $\hat{\epsilon}_i = Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_{p-1} x_{i,p-1}\right).$

Résidu partiel pour la $j^{\text{ème}}$ variable

$$\hat{e}_i^j = \hat{\epsilon}_i + \hat{\beta}_j x_{i,j}.$$

Diagnostic pour la $j^{\text{ème}}$ variable

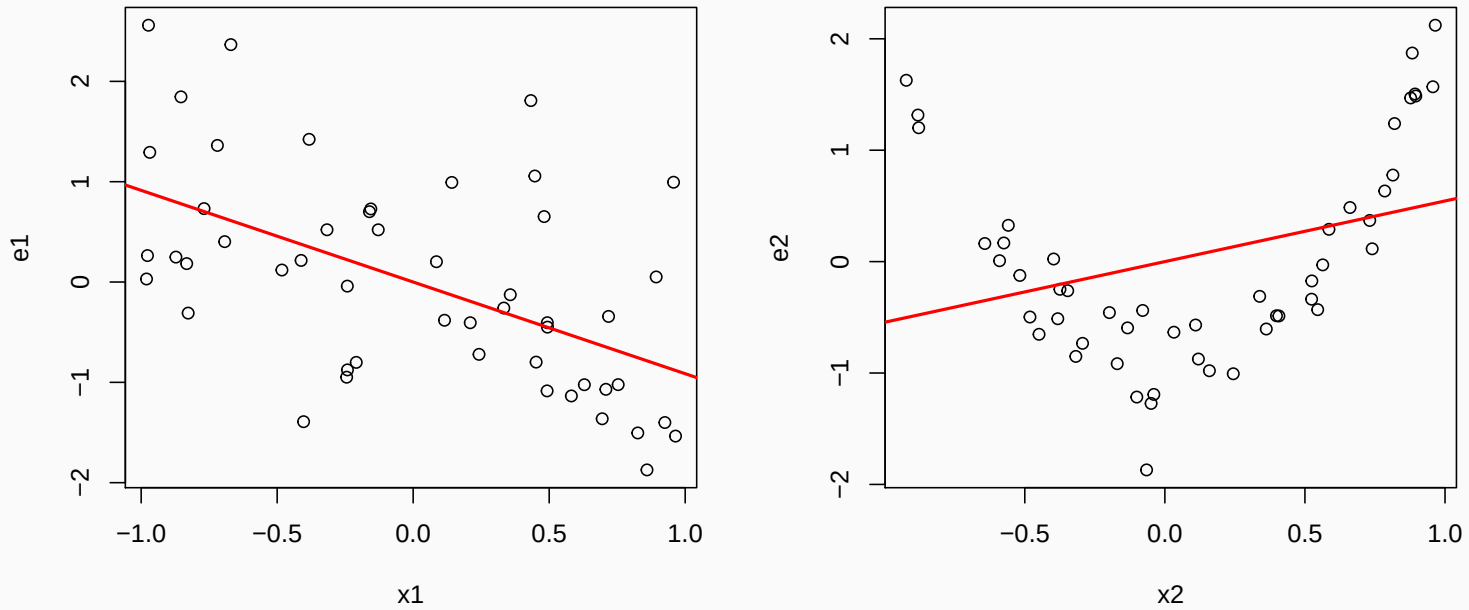
- Tracer les résidus partiels $\hat{e}_i^j$ en fonction des $x_{i,j}$.
- En effet, une régression linéaire sans constante pour expliquer $\hat{e}_i^j$ à partir des $x_{i,j}$ donne une pente égale à :

$$\frac{\langle \hat{e}^j, X^j \rangle}{\|X^j\|^2} = \hat{\beta}_j.$$

Exemple : tracés des résidus partiels

Soient des observations $(x_{i,1}, x_{i,2}, Y_i)$ issues du modèle

$$Y_i = 4 - x_{i,1} + 3x_{i,2}^2 + \epsilon_i.$$



$\rightsquigarrow$ Identification de l'influence non linéaire de X^2.