Statistiques multivariées

Rermerciements

Les transparents de ce chapitre sont une modification marginale des transparents d'Agathe Guilloux.

Chapitre 1 : régression linéaire

Introduction

Données “Vulnerability”, Patt et al., PNAS (2009)

Les pays les moins développés sont-ils plus vulnérables aux changements climatiques ?

Les auteurs ont voulu expliquer `ln_death_risk`, log du risque mortel dû aux évènement climatiques en fonction

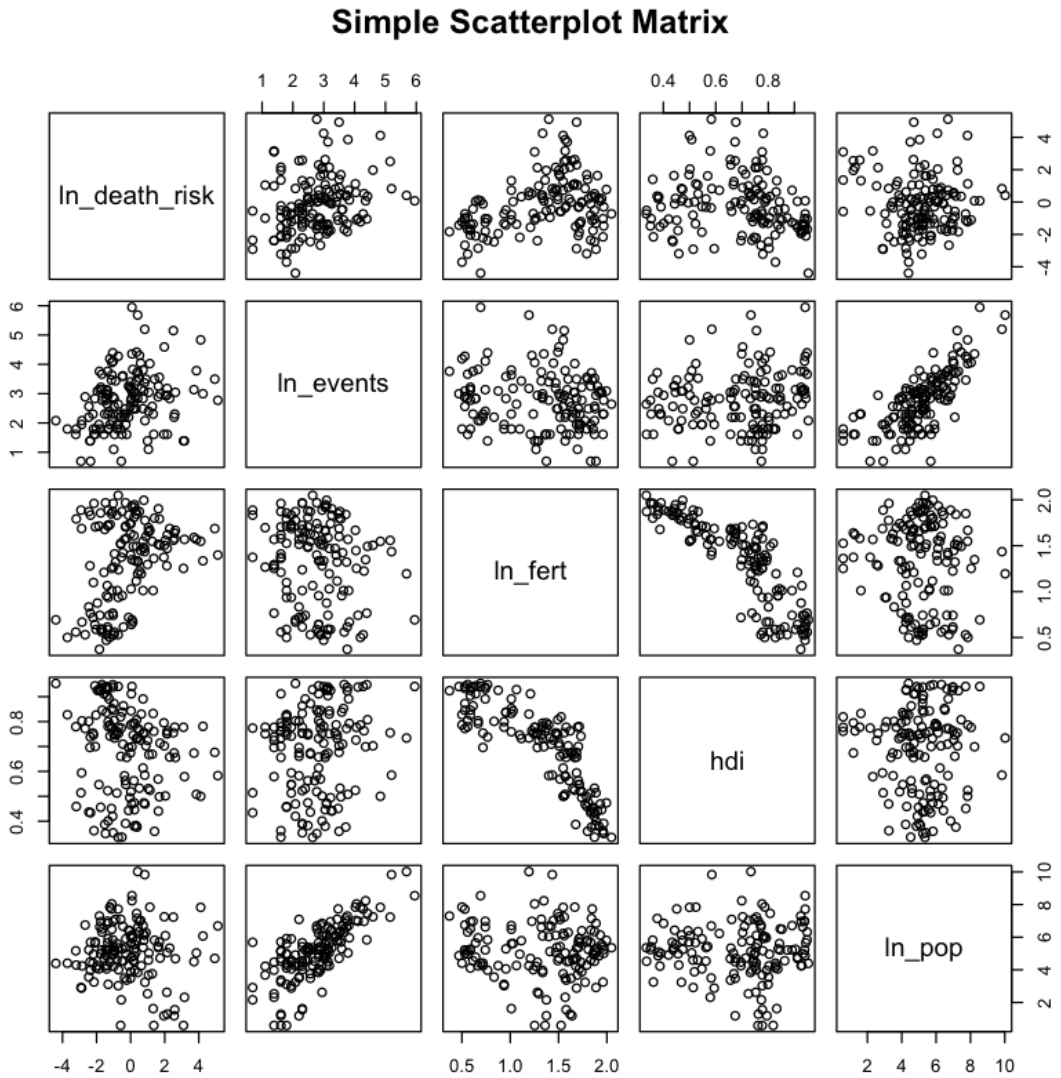
- du log du nombre d'évènements climatiques `ln_death_risk`
- du log de la fertilité `ln_fertility`
- de l'indice de développement humain `hdi` (United Nations)
- du log de la population `ln_pop`

Ils concluent que le développement socio-économique a un lien sur la fragilité aux événements climatiques, et ce lien pourrait se révéler dans le deuxième quart du 21ième siècle.

Visualisation des données "Vulnerability"

country_name	ln_events	ln_fert	hdi	ln_pop	ln_death_risk
Albania	2.3025850	1.2383740	0.7530000	4.0061200	-0.7102835002
Algeria	3.4965080	1.5993880	0.7025000	6.2838850	0.8961844999
Angola	3.0445230	1.9459100	0.4460000	5.5560560	0.2246879996
Argentina	3.6375860	1.0116010	0.8525001	6.4835150	-1.1036180004
Armenia	1.3862940	0.7654679	0.7380000	3.9765620	-2.3671239981
Australia	4.3944490	0.7654679	0.9480000	5.8379250	-1.0504329996
Austria	3.0910430	0.5306283	0.9330000	4.9908860	-1.4073670018
Azerbaijan	1.7917590	1.0986120	0.7460000	4.9572340	-2.1846459984
Bahamas	2.3025850	1.0116010	0.8325000	1.6226830	1.3217560000
Bangladesh	4.8362820	1.5475620	0.5000000	7.8287280	4.1112999999
Belarus	1.6094380	0.5596158	0.7795000	5.1651670	-3.2192569975

Visualisation des données "Vulnerability"



Notations

On note

- $X^1, X^2, \ldots, X^p$ les variables explicatives
- Y la variable dépendante, ou la variable à expliquer

Dans notre exemple :

- X^1 est `ln_events`
- X^2 est `ln_fert`
- X^3 est `hdi`
- X^4 est `ln_pop` et
- Y est `ln_death_risk`

donc $p = 4$.

Le modèle de regression linéaire classique : le modèle

On suppose que Y est composé

- d'une moyenne dépendant linéairement des $X^1, \ldots, X^p$, supposées fixées (pas aléatoires)
- d'une erreur aléatoire notée ε qui correspond à
  - une erreur de mesure ou
  - l'effet d'autres variables oubliées

.

On a

$$Y = \beta_0 + \beta_1 X^1 + \beta_2 X^2 + \ldots + \beta_p X^p + \varepsilon$$

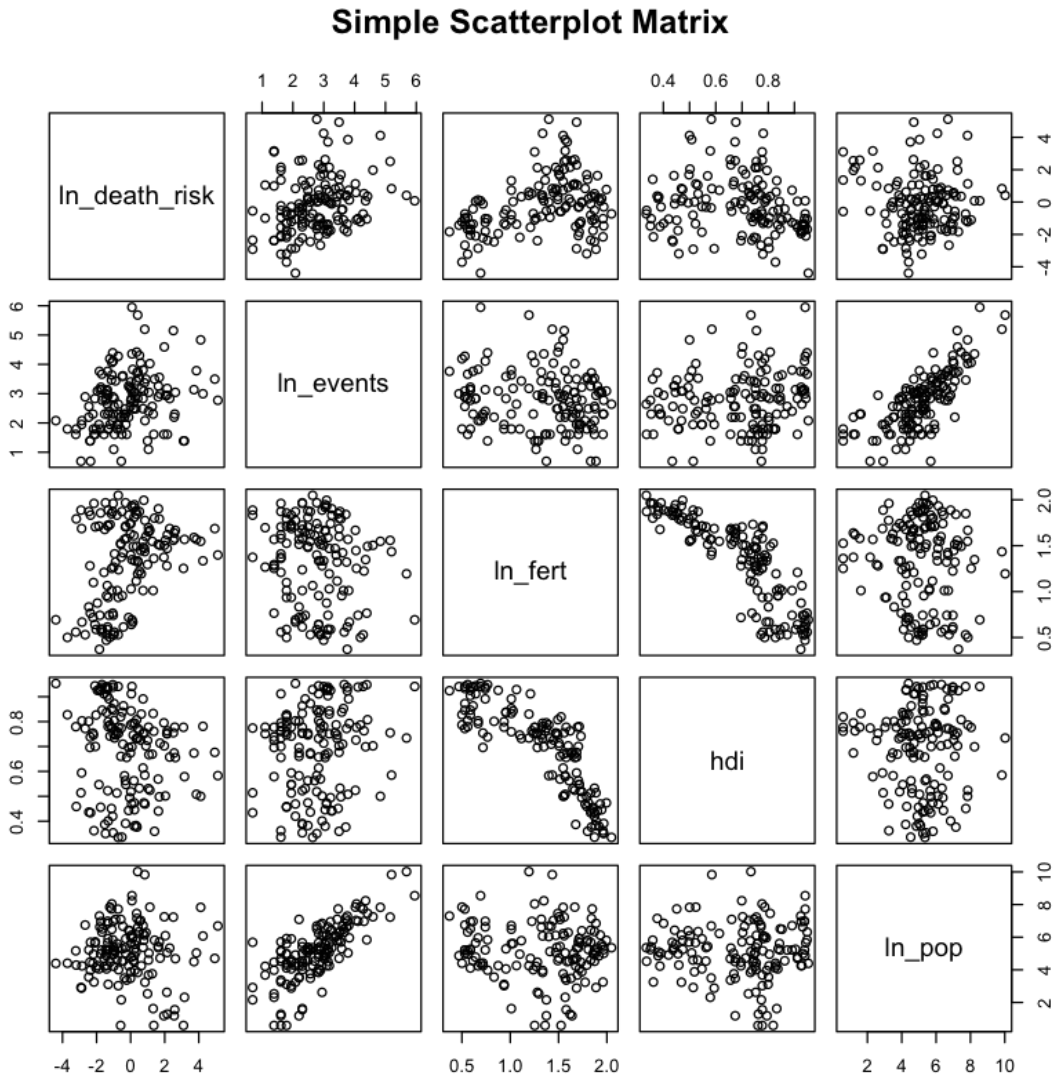
$$\text{variable dépendante} = \text{moyenne(dépendant de } X^1, \ldots, X^p) + \text{erreur}$$

Hypothèse de linéarité

Le modèle est dit **linéaire** car la moyenne de Y dépend **linéairement** de chaque $X^1, \ldots, X^p$

$$\mathbb{E}Y = \beta_0 + \beta_1 X^1 + \beta_2 X^2 + \ldots + \beta_p X^p.$$

Vérification de l'hypothèse de linéarité sur les données "Vulnerability"



Observations et données

On suppose qu'on observe Y et les $X^1, \ldots, X^p$ pour n individus **indépendants**, on obtient les données

$$\begin{pmatrix} 1 & X_1^1 & X_1^2 & \ldots & X_1^p & Y_1 \\ 1 & X_2^1 & X_2^2 & \ldots & X_2^p & Y_2 \\ \ldots & & & & & \\ 1 & X_n^1 & X_n^2 & \ldots & X_n^p & Y_n \end{pmatrix} \text{ ou } \begin{pmatrix} X_1^1 & X_1^2 & \ldots & X_1^p & Y_1 \\ X_2^1 & X_2^2 & \ldots & X_2^p & Y_2 \\ \ldots & & & & \\ X_n^1 & X_n^2 & \ldots & X_n^p & Y_n \end{pmatrix}$$

country_name	ln_events	ln_fert	hdi	ln_pop	ln_death_risk
Albania	2.3025850	1.2383740	0.7530000	4.0061200	-0.7102835002
Algeria	3.4965080	1.5993880	0.7025000	6.2838850	0.8961844999
Angola	3.0445230	1.9459100	0.4460000	5.5560560	0.2246879996
Argentina	3.6375860	1.0116010	0.8525001	6.4835150	-1.1036180004
Armenia	1.3862940	0.7654679	0.7380000	3.9765620	-2.3671239981
Australia	4.3944490	0.7654679	0.9480000	5.8379250	-1.0504329996
Austria	3.0910430	0.5306283	0.9330000	4.9908860	-1.4073670018
Azerbaijan	1.7917590	1.0986120	0.7460000	4.9572340	-2.1846459984
Bahamas	2.3025850	1.0116010	0.8325000	1.6226830	1.3217560000
Bangladesh	4.8362820	1.5475620	0.5000000	7.8287280	4.1112999999
Belarus	1.6094380	0.5596158	0.7795000	5.1651670	-3.2192569975

Le modèle est vérifiée pour chaque individu, on a donc

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_1^1 + \beta_2 X_1^2 + \ldots + \beta_p X_1^p + \varepsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_2^1 + \beta_2 X_2^2 + \ldots + \beta_p X_2^p + \varepsilon_2 \\ \ldots & \\ Y_n &= \beta_0 + \beta_1 X_n^1 + \beta_2 X_n^2 + \ldots + \beta_p X_n^p + \varepsilon_n \end{aligned}$$

Hypothèses sur les erreurs

- Pour tout individu i : $\mathbb{E}(\varepsilon_i) = 0$ **les erreurs sont centrées**
- Pour tout individu i : $\mathbb{V}(\varepsilon_i) = \sigma^2$ **les erreurs sont de variance constante**
- Pour tous individus i et j : $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ **les erreurs sont decorrélées.**

Ecriture matricielle

Pour un individu i, on a

$$Y_i = \beta_0 + \beta_1 X_i^1 + \beta_2 X_i^2 + \ldots + \beta_p X_i^p + \varepsilon_i.$$

On peut récrire

$$Y_i = (1, X_i^1, \ldots, X_i^p) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \ldots \\ \beta_p \end{pmatrix} + \varepsilon$$

ou bien, pour tous les individus

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \ldots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_1^1 & X_1^2 & \ldots & X_1^p \\ 1 & X_2^1 & X_2^2 & \ldots & X_2^p \\ \ldots & & & & \\ 1 & X_n^1 & X_n^2 & \ldots & X_n^p \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \ldots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \ldots \\ \varepsilon_n \end{pmatrix}$$

$$\begin{array}{ccccccc} Y & = & X & & \beta & + & \varepsilon. \\ n \times 1 & & n \times (p+1) & & (p+1) \times 1 & + & n \times 1 \end{array}$$

Modèle linéaire : définition et hypothèses

$$Y = X\beta + \epsilon$$

où

- Y est un vecteur $n \times 1$ **observé**
- X est une matrice $n \times (p+1)$ **observée** de **rang** $p+1$
- β est un vecteur $(p+1) \times 1$ de paramètres **inconnus**
- ϵ est un vecteur $n \times 1$ de v.a. **non-observées** supposées décorrélées avec

$$\mathbb{E}(\epsilon_i) = 0 \text{ et } \mathbb{V}(\epsilon_i) = \sigma^2$$

où σ^2 est un paramètre **inconnu**.

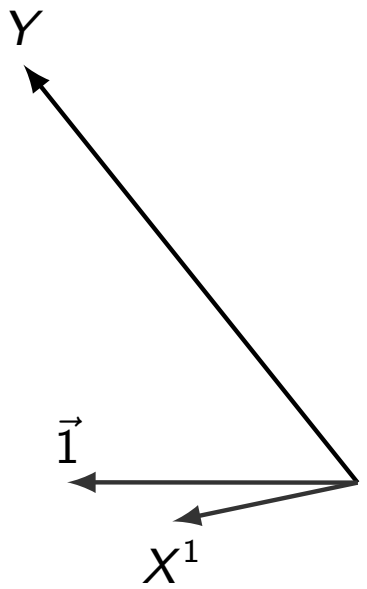
Quelques rappels d'algèbre linéaire

Les données comme des points de $\mathbb{R}^n$

On a écrit

$$\underset{n\times 1}{Y} = \underset{n\times(p+1)}{X}\ \underset{(p+1)\times 1}{\beta} + \underset{n\times 1}{\varepsilon}.$$

Chaque vecteur en jeu : Y, $\vec{1}$, X^1, $X^2, \ldots, X^p$ est un vecteur de $\mathbb{R}^n$.



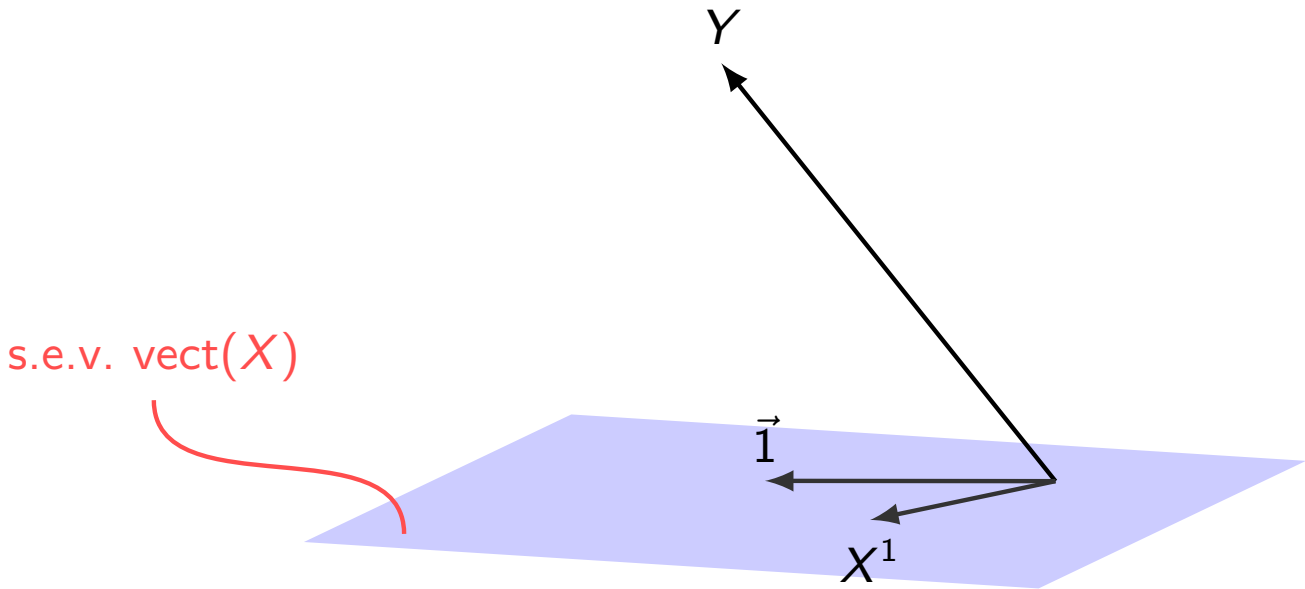
Dans $\mathbb{R}^3$

Sous-espace vectoriel, sous-espace vectoriel engendré

Dans cet espace vectoriel $\mathbb{R}^n$, on s'intéresse au **sous-espace vectoriel (s.e.v.)** engendré par les colonnes de X, c'est-à-dire à l'ensemble des vecteurs qui s'écrivent

$$\alpha_0\vec{1} + \alpha_1 X^1 + \alpha_2 X^2 + \ldots + \alpha_p X^p$$

qu'on note vect$(\vec{1}, X^1, X^2, \ldots, X^p)$ ou vect(X) pour faire court !



Dans $\mathbb{R}^3$

Question cruciale : Quelle est la dimension de vect(X) ?

Exercice

Les vecteurs Soient

- la matrice

$$A = \begin{pmatrix} 1 & -3 & -4 \\ -4 & 6 & -2 \\ -3 & 7 & 6 \end{pmatrix}$$

- le vecteur

$$b = \begin{pmatrix} 3 \\ 3 \\ -4 \end{pmatrix}.$$

Est ce que b est dans vect(A) ?

Vecteurs linéairement indépendant

On dit que des vecteurs $\vec{x_1}, \vec{x_2}, \ldots, \vec{x_p}$ sont **linéairement dépendant** s'il existe des réels $a_1, a_2, \ldots, a_p$ non tous nuls tels que

$$a_1\vec{x_1} + a_2\vec{x_2} + \ldots + a_p\vec{x_p} = 0.$$

Exercice

Les vecteurs

$$\vec{x_0} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \vec{x_1} = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} \text{ et } \vec{x_2} = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}$$

sont-ils linéairement indépendants ? Et

$$\vec{z_0} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \vec{z_1} = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} \text{ et } \vec{z_2} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} ?$$

,

Exercice

Est ce que 3 vecteurs de $\mathbb{R}^2$ peuvent être linéairement indépendants ?

Rang d'une matrice et dimension du s.e.v engendré par ses colonnes

On dit qu'une matrice M de taille $n \times p$ avec $n \geq p$ est de **rang** p

- si ses colonnes sont des vecteurs linéairement indépendants
- ou bien si ses colonnes engendrent un s.e.v. de dimension p.

Si ses colonnes sont des vecteurs linéairement dépendants, elle est de rang $p' < p$ où p' est taille de la plus grande sous-famille linéairement indépendante.

Exercice

Quel est le rang de la matrice

$$\begin{pmatrix} 1 & 2 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} ?$$

Et celui de

$$\begin{pmatrix} 1 & 2 & 3 \\ 1 & 1 & 2 \\ 1 & 0 & 1 \end{pmatrix} ?$$

Exemple sur des données

Données auction

On s'intéresse aux données "auction" dans lesquelles sont enregistrés pour 19 foires aux bestiaux

- `markedid` : l'identifiant de la foire
- `cattle` : le volume des boeufs achetés
- `calves` : le volume des veaux achetés
- `hogs` : le volume des porcs achetés
- `sheep` : le volume des moutons achetés
- `cost` : le coût total des transactions
- `volume` : le volume total des transactions

On veut expliquer la variable `cost` à partir des autres (sauf l'identifiant).

- Charger les données
- Quelle est la dimension de l'espace linéaire engendré par les vecteurs `cattle`, `calves`, `hogs`, `sheep`, `volume` ? Pourquoi ?
- Quelle hypothèse du modèle linéaire n'est pas vérifiée ? Que faire ?

Produit scalaire (1)

On définit le **produit scalaire** entre 2 vecteurs $\mathbf{u}$ et $\mathbf{v}$ de $\mathbb{R}^n$ comme

$$\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^\top \mathbf{v} = \begin{pmatrix} u_1, u_2, \ldots, u_n \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \ldots \\ v_n \end{pmatrix} = u_1 v_1 + u_2 v_2 + \ldots + u_n v_n$$

Exercice

Calculer les produits scalaires $\langle \mathbf{u}, \mathbf{v} \rangle$ et $\langle \mathbf{v}, \mathbf{u} \rangle$ pour

$$\mathbf{u} = \begin{pmatrix} 2 \\ -5 \\ 1 \end{pmatrix} \text{ et } \mathbf{v} = \begin{pmatrix} 3 \\ 2 \\ -3 \end{pmatrix}$$

Produit scalaire (2)

Propriétés du produit scalaire

Soient 3 vecteurs $\mathbf{u}$, $\mathbf{v}$ et $\mathbf{w}$ de $\mathbb{R}^n$ et c un nombre réel alors

- $\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle$
- $\langle (\mathbf{u} + \mathbf{v}), \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{w} \rangle + \langle \mathbf{v}, \mathbf{w} \rangle$
- $\langle (c\mathbf{u}), \mathbf{v} \rangle = c\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{u}, (c\mathbf{v}) \rangle$
- $\langle \mathbf{u}, \mathbf{u} \rangle \geq 0$,et $\langle \mathbf{u}, \mathbf{u} \rangle = 0$ si et seulement si $\mathbf{u} = \mathbf{0}$.

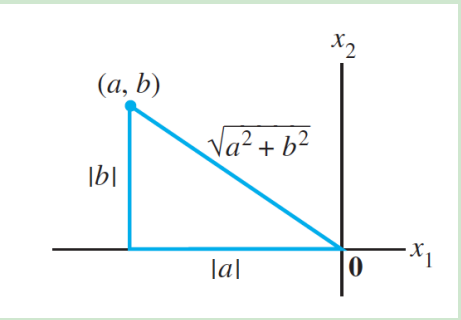
Longueur d'un vecteur

Pour $\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}$, la **longueur** ou la **norme** de $\mathbf{v}$ est le réel positiv ou nul $\|\mathbf{v}\|$ défini par

$$\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle} = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2} \quad \text{et} \quad \|\mathbf{v}\|^2 = \langle \mathbf{v}, \mathbf{v} \rangle.$$

Exercice

- Montrer que, si c est un réel positif et $\mathbf{v}$ est un vecteur de $\mathbb{R}^n$, $\|c\mathbf{v}\| = c\,\|\mathbf{v}\|$.
- Pour $\mathbf{v} = \begin{pmatrix} a \\ b \end{pmatrix}$
  - calculer $\|\mathbf{v}\|$
  - et trouver un vecteur $\mathbf{u}$ colinéaire à $\mathbf{v}$ qui a pour norme 1.



Distance dans $\mathbb{R}^n$

La **distance** entre $\mathbf{u}$ et $\mathbf{v}$ dans $\mathbf{R}^n$ est définie par

$$\text{dist}\,(\mathbf{u},\mathbf{v}) = \|\mathbf{u}-\mathbf{v}\|\,.$$

Soient $\mathbf{u} = (u_1, u_2)$ et $\mathbf{v} = (v_1, v_2)$ alots $\mathbf{u}-\mathbf{v} = (u_1 - v_1, u_2 - v_2)$ et

$$\text{dist}\,(\mathbf{u},\mathbf{v}) = \|\mathbf{u}-\mathbf{v}\| = \|(u_1 - v_1, u_2 - v_2)\| = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2}$$

Exercice

Calculer la distance entre $\mathbf{u} = \begin{pmatrix} 7 & 1 \end{pmatrix}^\top$ et $\mathbf{v} = \begin{pmatrix} 3 & 2 \end{pmatrix}^\top$.

Exercice

Soient $\mathbf{u}$ et $\mathbf{v}$ deux vecteurs de $\mathbb{R}^n$:
On peut écrire

$$[\text{dist}\,(\mathbf{u},\mathbf{v})]^2 = \|\mathbf{u}-\mathbf{v}\|^2 = \langle \mathbf{u}-\mathbf{v}, \mathbf{u}-\mathbf{v}\rangle = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - 2\langle \mathbf{u},\mathbf{v}\rangle.$$

Dans l'autre sens, on obtient

$$[\text{dist}\,(\mathbf{u},-\mathbf{v})]^2 = \|\mathbf{u}+\mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 + 2\langle \mathbf{u},\mathbf{v}\rangle.$$

Montrer les égalités précédentes.

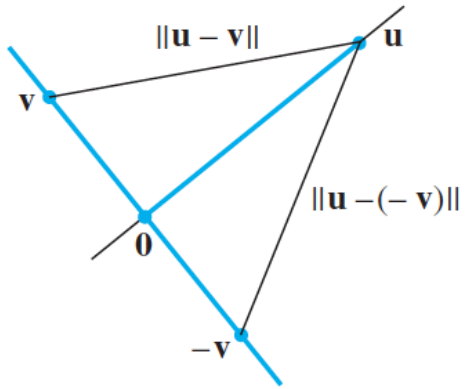
Exercice

Vérifier la loi du parallélogramme pour $\mathbf{u}$ et $\mathbf{v}$ dans $\mathbb{R}^n$

$$\|\mathbf{u}+\mathbf{v}\|^2 + \|\mathbf{u}-\mathbf{v}\|^2 = 2\|\mathbf{u}\|^2 + 2\|\mathbf{v}\|^2.$$

Vecteurs orthogonaux et théorème de Pythagore

Si $\mathbf{u}$ et $\mathbf{v}$ ont des directions perpendiculaires



alors $[\text{dist}\,(\mathbf{u}, \mathbf{v})]^2 = [\text{dist}\,(\mathbf{u}, -\mathbf{v})]^2$, on a donc

$$\langle \mathbf{u}, \mathbf{v} \rangle = 0$$

et on dit que $\mathbf{u}$ et $\mathbf{v}$ sont **orthogonaux**.

Théorème de Pythagore

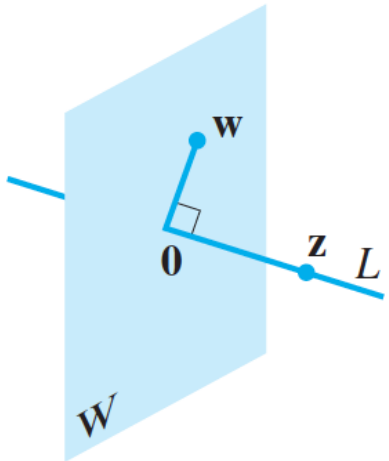
Deux vecteurs $\mathbf{u}$ and $\mathbf{v}$ sont orthogonaux si et seulement si

$$\|\mathbf{u} + \mathbf{v}\| = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 .$$

Compléments orthogonaux

Quand un vecteur **z** est orthogonal à tous les vecteurs d'un sous-espace W de $\mathbf{R}^n$, alors **z** est **orthogonal à** W.

L'ensemble des vecteurs orthogonaux à W est appelé **le complément orthogonal** de W et est noté $W^{\perp}$ (on dit "W orthogonal").



Exercice

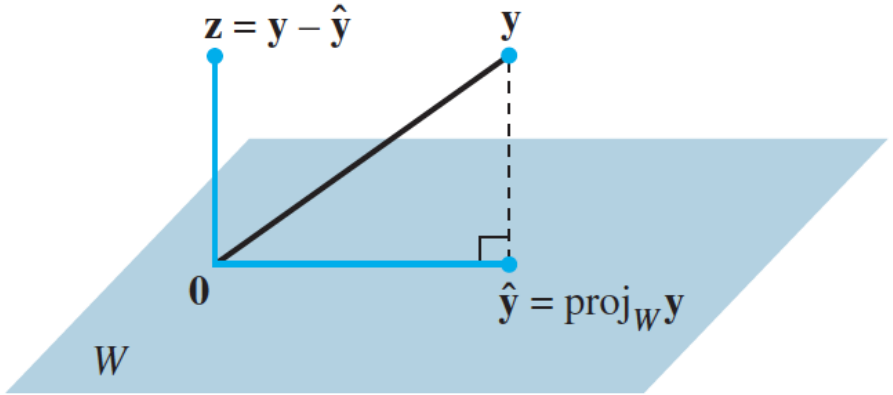
Soit $\mathbf{v} = \begin{pmatrix} a \\ b \end{pmatrix}$. Décrire l'ensemble $(\text{vect}(\mathbf{v}))^{\perp}$ des vecteurs orthogonaux **v**

Théorème de la projection orthogonale

Soit W un s.e.v de $\mathbb{R}^n$. Chaque vecteur $\mathbf{y}$ de $\mathbb{R}^n$ s'écrit de **manière unique** comme

$$\mathbf{y} = \text{proj}^W(\mathbf{y}) + \mathbf{z}$$

avec $\text{proj}^W(y) \in W$ et $\mathbf{z} \in W^\perp$. On appelle l'unique $\text{proj}^W(\mathbf{y})$ **la projection orthogonale** de y sur W.



Exercices sur la projection orthogonale

Exercice

Soient

$$\mathbf{u}_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \mathbf{u}_2 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} \text{ et } \mathbf{y} = \begin{pmatrix} -1 \\ 4 \\ 3 \end{pmatrix}.$$

Quelle est la projection orthogonale de $\mathbf{y}$ sur $W = \text{vect}(\mathbf{u}_1, \mathbf{u}_2)$?

Exercice

Soient $\mathbf{u}_1$ et $\mathbf{u}_2$ deux vecteurs orthogonaux de $\mathbb{R}^3$ et $W = \text{vect}(\mathbf{u}_1, \mathbf{u}_2)$. Soit $\mathbf{v}$ un vecteur de $\mathbb{R}^3$.

- Quel est le projeté orthogonal de $\mathbf{v}$ sur W ?
- Vérifier avec l'exercice précédent.

Exercice

Soient $\mathbf{u}_1$, $\mathbf{u}_2$ et $\mathbf{v}$ trois vecteurs de $\mathbb{R}^3$ et $W = \text{vect}(\mathbf{u}_1, \mathbf{u}_2)$. Quel est le vecteur $w \in W$ le plus proche de $\mathbf{v}$?

Propriétés de la projection orthogonale

Idempotence

- W un s.e.v. de $\mathbb{R}^n$
- $\mathbf{y}$ un vecteur de $\mathbb{R}^n$

Si $\mathbf{y} \in W$ alors $\text{proj}^W(\mathbf{y}) = y$.

Meilleure approximation

Soient

- W un s.e.v. de $\mathbb{R}^n$
- $\mathbf{y}$ un vecteur de $\mathbb{R}^n$
- et $\text{proj}^W(\mathbf{y})$ la projection orthogonale de $\mathbf{y}$ sur W

alors $\text{proj}^W(\mathbf{y})$ est le point de W le plus proche de $\mathbf{y}$.

L'estimateur des moindres carrés

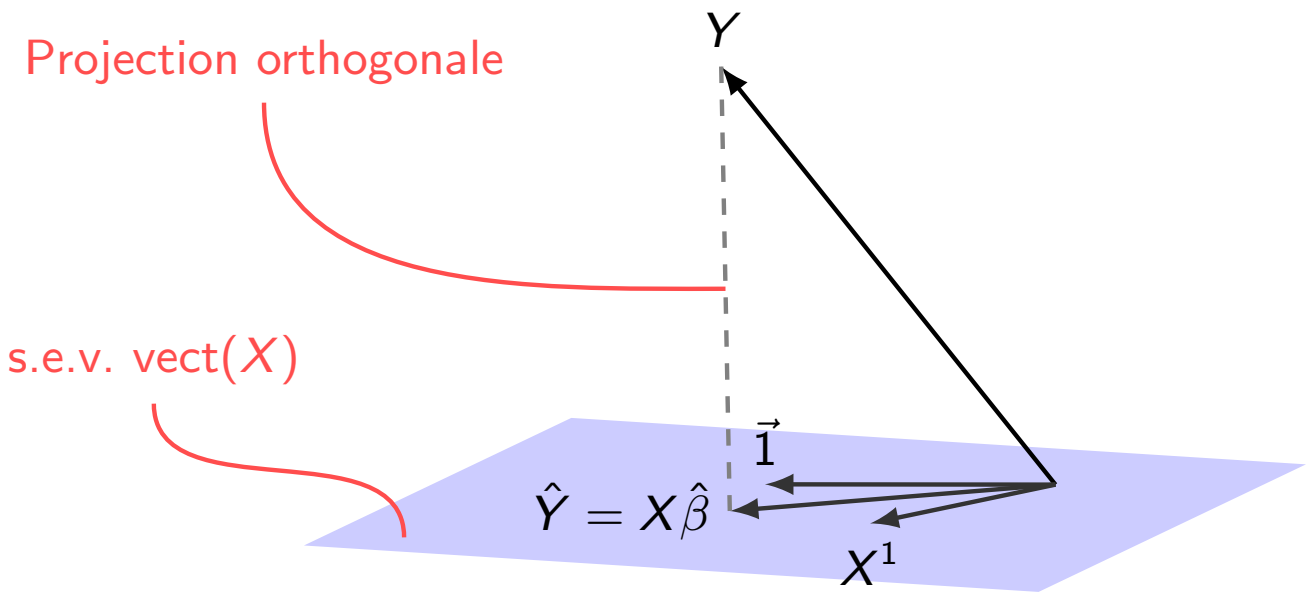
L'estimateur des moindres carrés

Dans le modèle linéaire

$$Y = X\beta + \varepsilon,$$

on définit $\hat{Y} = X\hat{\beta}$ comme le **projeté orthogonal de Y sur** vect(X), c'est le point de vect(X) le plus proche de Y.

$$\|Y - X\hat{\beta}\|^2 = \min_{\gamma \in \mathbb{R}^{p+1}} \|Y - X\gamma\|^2.$$



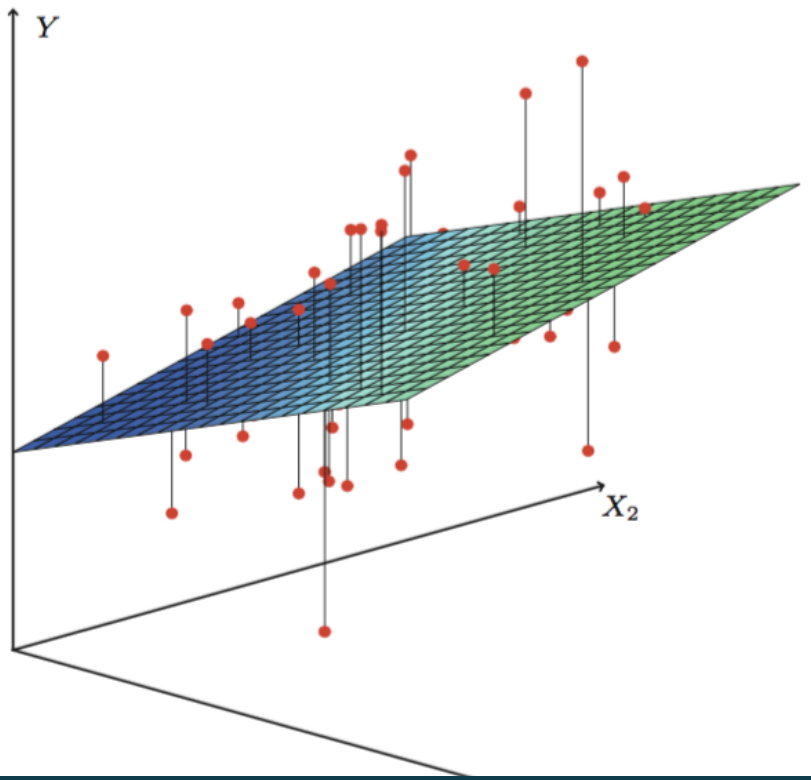
Dans $\mathbb{R}^3$

Meilleure approximation (1)

- On a dit que $\hat{Y} = X\hat{\beta}$ est le point de vect(X) le plus proche de Y
- Par définition :
  - Les points de vect(X) s'écrivent tous $X\beta$, pour un certain β.
  - La distance entre Y et un $X\beta$ vaut $\|Y - X\beta\|^2$,

on a donc

$$\|Y - X\hat{\beta}\|^2 = \min_{\gamma \in \mathbb{R}^{p+1}} \|Y - X\gamma\|^2.$$



Meilleure approximation (2)

Conséquence

Puisque $\hat{Y} = X\hat{\beta}$ est la projection de Y sur vect(X), on a

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

Meilleure approximation (3)

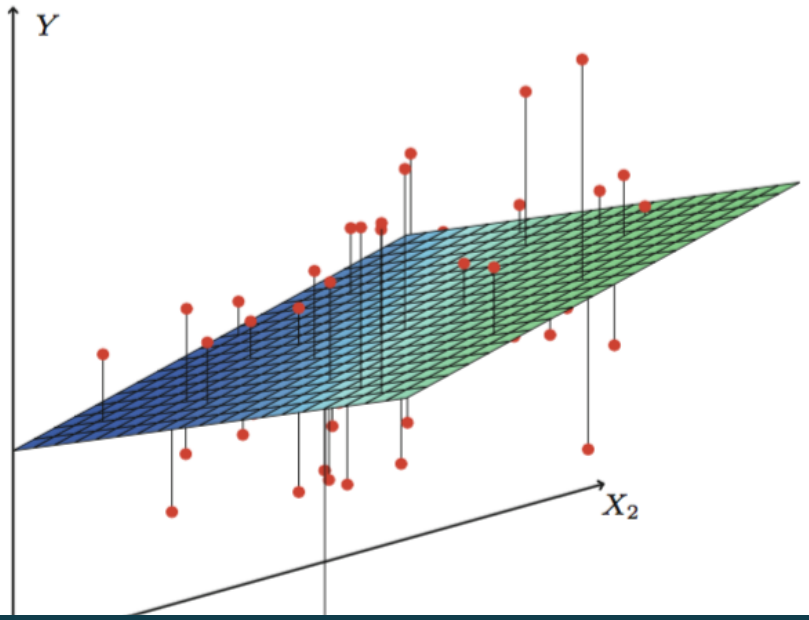
Définitions : erreurs résiduelles

On note le vecteur des erreurs résiduelles $e = Y - X\hat{\beta}$ (projection de Y sur vect$(X)^{\perp}$ on a

$$Y - X\hat{\beta} \perp X\hat{\beta}.$$

l'estimateur des moindres carrés de σ^2 est donné par

$$\hat{\sigma}^2 = \frac{\|e\|^2}{n - (p+1)}.$$



Conséquence géométrique : le R^2

On obtient des mesures de l'adéquation du modèle

- $Y - X\hat{\beta} \perp X\hat{\beta} - \bar{Y}\mathbf{1}$ si $\mathbf{1} \in \text{vect}(X)$ et donc

$$\underbrace{\|Y - \bar{Y}\mathbf{1}\|^2}_{\substack{\text{SC tot.} \\ \textit{SSTotal}}} = \underbrace{\|Y - X\hat{\beta}\|^2}_{\substack{\text{SC résiduelle} \\ \textit{SSError}}} + \underbrace{\|X\hat{\beta} - \bar{Y}\mathbf{1}\|^2}_{\substack{\text{SC expliquée} \\ \textit{SSModel}}}.$$

- On définit le R^2 par

$$0 \leq R^2 = \frac{\|X\hat{\beta} - \bar{Y}\mathbf{1}\|^2}{\|Y - \bar{Y}\mathbf{1}\|^2} = 1 - \frac{\|Y - X\hat{\beta}\|^2}{\|Y - \bar{Y}\mathbf{1}\|^2} \leq 1$$

et le R^2 ajusté du nombre de paramètres par

$$R^2_{\text{Adj}} = 1 - \frac{(n-1)(1-R^2)}{(n-(p+1)} \leq 1$$

Loi normale multivariée

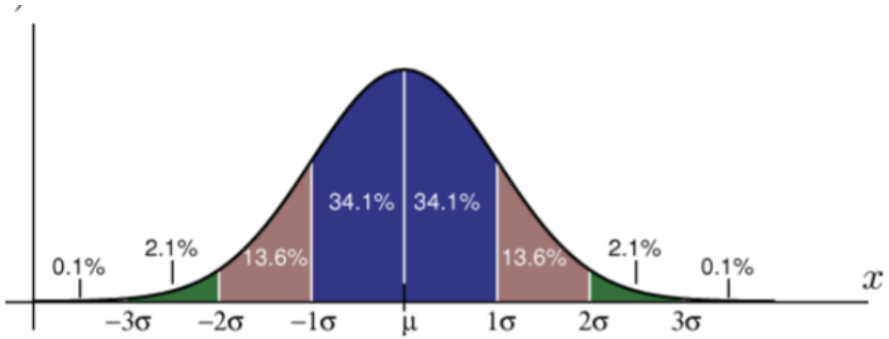
Définition/propriétés

On dit que ε est de loi normale $\mathcal{N}(\mu, \sigma^2)$ quand sa densité est donnée par

$$\frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}).$$

Alors

- $\mathbb{E}(\varepsilon) = \mu$
- $\mathbb{V}(\varepsilon) = \sigma^2$
- $(a\varepsilon + b) \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.



Définition/propriétés

Dans, $\mathbb{R}^n$, on dit que $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ est de loi normale multivariée standard quand

- tous les ε_i sont de loi $\mathcal{N}(0,1)$
- les $\varepsilon_1, \ldots, \varepsilon_n$ sont indépendants.

On note

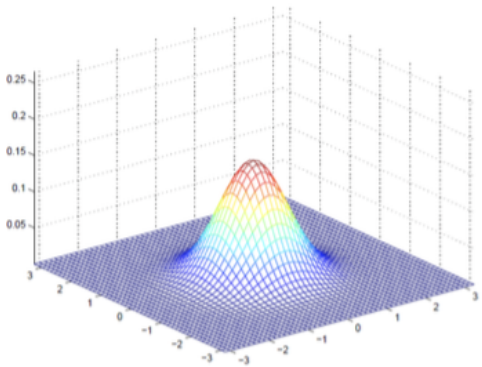
- $\mathbb{E}(\varepsilon) = \mathbf{0}$
- $\mathbb{V}(\varepsilon) = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \ldots & & & \\ 0 & 0 & \ldots & 1 \end{pmatrix}$
- sa densité vaut

$$\frac{1}{\sqrt{2\pi}^p} \exp\left(-\frac{1}{2}\|x\|^2\right).$$

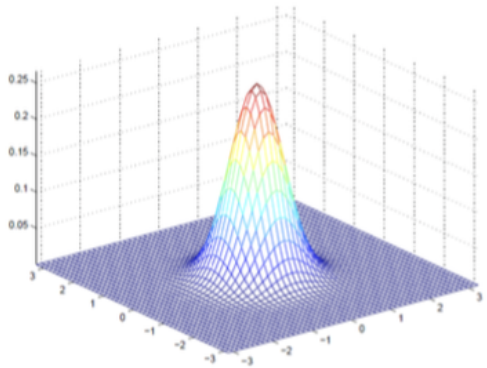
Propriété

Si A est une matrice de taille $k \times n$ alors le vecteur $\eta = A\varepsilon \sim \mathcal{N}(\mathbf{0}, \Sigma = AA^\top)$. En particulier

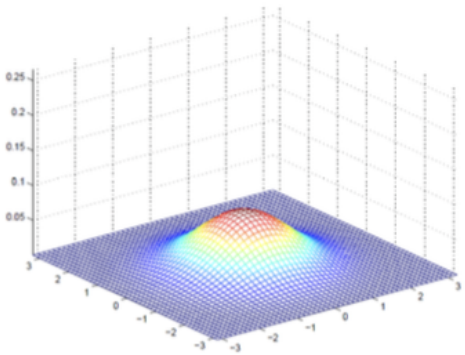
- chaque coordonnée η_i (entre 1 et k) de $\eta = A\varepsilon$ est de loi normale $\mathcal{N}(0, \Sigma_{ii})$
- $\mathrm{Cov}(\eta_i, \eta_j) = \Sigma_{ij}$
- $\mathrm{Cov}(\eta_i, \eta_j) = 0$ si et seulement si η_i et η_j sont indépendants.



- μ = [0; 0]
- Σ = [1 0 ; 0 1]

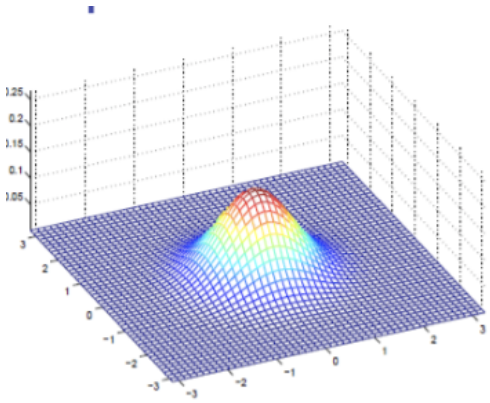


- μ = [0; 0]
- Σ = [.6 0 ; 0 .6]

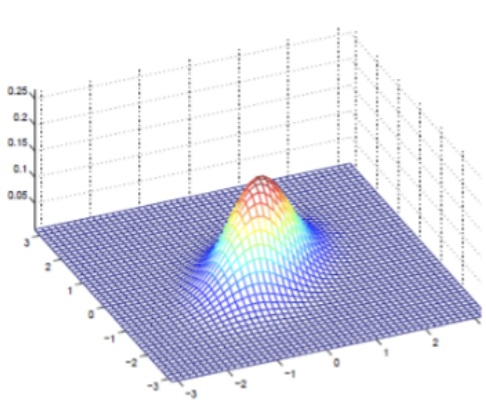


- μ = [0; 0]
- Σ = [2 0 ; 0 2]

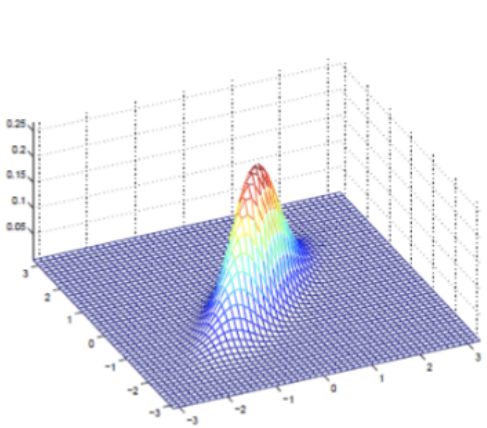
Graphique



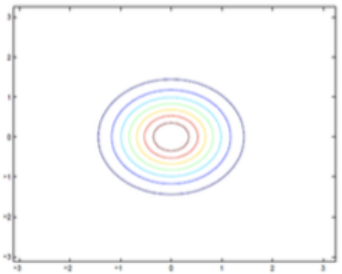
- $\mu = [0; 0]$
- $\Sigma = [1 \;\; 0; 0 \;\; 1]$

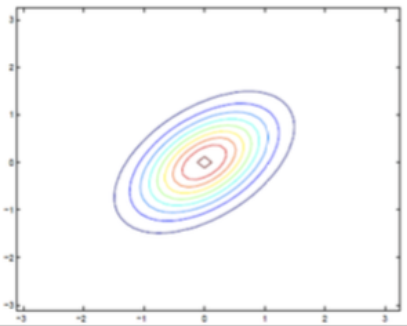


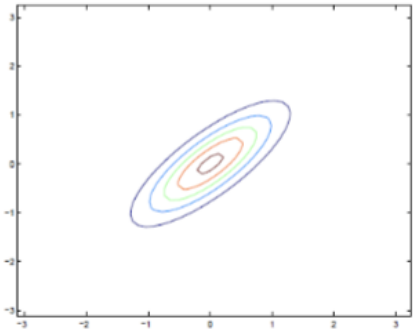
- $\mu = [0; 0]$
- $\Sigma = [1 \;\; 0.5; 0.5 \;\; 1]$



- $\mu = [0; 0]$
- $\Sigma = [1 \;\; 0.8; 0.8 \;\; 1]$



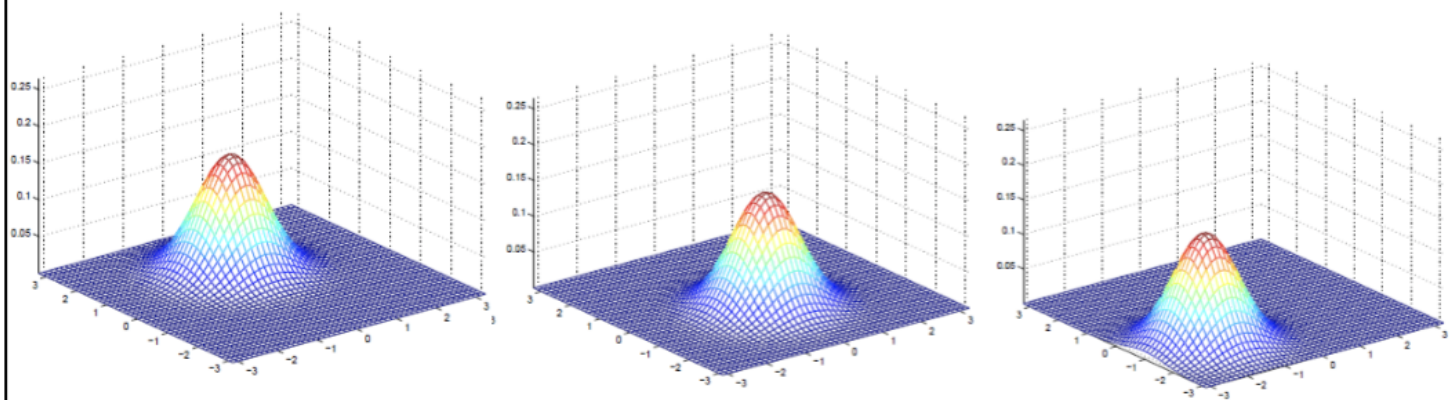




Propriété

Si μ est un vecteur de taille k et A une matrice de taille $k \times n$ alors le vecteur $\eta = A\varepsilon + \mu \sim \mathcal{N}(\mu, \Sigma = AA^\top)$. En particulier

- chaque coordonnée η_i (entre 1 et k) de $\eta = A\varepsilon$ est de loi normale $\mathcal{N}(\mu_i, \Sigma_{ii})$
- $\text{Cov}(\eta_i, \eta_j) = \Sigma_{ij}$
- $\text{Cov}(\eta_i, \eta_j) = 0$ si et seulement si η_i et η_j sont indépendants.



- μ = [1; 0]
- Σ = [1 0; 0 1]

- μ = [-.5; 0]
- Σ = [1 0; 0 1]

- μ = [-1; -1.5]
- Σ = [1 0; 0 1]

Exercice

Soient X et Y de loi normale $\mathcal{N}(0,1)$, et indépendants.

1. Que vaut $\mathrm{Cov}(X,Y)$?
2. Quelle est la loi de $\begin{pmatrix} X \\ Y \end{pmatrix}$?
3. On pose $Z = 2X + Y$. Que vaut $\mathrm{Cov}(X,Z)$?
4. Quelle est la loi de $\begin{pmatrix} X \\ Z \end{pmatrix}$?

Exercice

Soit $\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$. Chercher A telle que $AA^\top = \Sigma$

Propriétés

Si $\eta \sim \mathcal{N}(\mu, \Sigma)$ et si A est de plein rang

- $\Sigma = AA^\top$ est inversible et $\det(\Sigma) \neq 0$
- la densité de η est

$$\frac{1}{\sqrt{(2\pi)^p \det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

Lois dérivées de la gaussienne

Définition : loi du χ^2

Si ε est un vecteur gaussien standard dans $\mathbb{R}^k$, alors

$$\|\varepsilon\|^2 \sim \chi^2(k)$$

Définition : loi de student

Si U suit une loi du $\chi^2(k)$ et V suit une loi $\mathcal{N}(0,1)$ et U et V sont indépendants, alors

$$\frac{U}{\sqrt{V/k}} \sim \mathcal{T}(k)$$

Définition : loi de Fisher

Si R suit une loi du $\chi^2(k)$ et S suit une loi du $\chi^2(l)$ et R et S sont indépendants, alors

$$\frac{R/k}{S/l} \sim \mathcal{F}(k,l)$$

Théorème de Cochran

- Si V_1, V_2 sont des s.e.v. orthogonaux dans $\mathbb{R}^n$ de dimension n_1, n_2 et
- si Z_1, Z_2 sont les projections orthogonales d'un vecteur gaussien standard sur V_1, V_2

alors

- les v.a. Z_1, Z_2 sont gaussiens et deux à deux indépendants
- et, en particulier, $\|Z_1\|^2 \sim \chi^2(n_1)$ et $\|Z_2\|^2 \sim \chi^2(n_2)$.

Modèle linéaire gaussien

Modèle linéaire gaussien

$$Y = X\beta + \epsilon$$

où

- Y est un vecteur $n \times 1$ **observé**
- X est une matrice $n \times (p+1)$ **observée** de **rang** $p+1$
- β est un vecteur $p \times 1$ de paramètres **inconnus**

et

-
- ϵ est un vecteur $n \times 1$ de v.a. **non-observées** avec

$$\epsilon \sim \mathcal{N}((0, \ldots, 0)^\top, \sigma^2 I_n)$$

où σ^2 est un paramètre **inconnu**.

Conséquences du théorème de Cochran (1)

1. sur $\hat{\beta}$ (projection de Y sur vect(X))

$$\hat{\beta} - \beta \sim \mathcal{N}((0, \ldots, 0), \sigma^2 (X^\top X)^{-1})$$

2. sur $\hat{\sigma}$ (projection de Y sur vect$(X)^\perp$)

$$\frac{\|Y - X\hat{\beta}\|^2}{\sigma^2} = \frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p-1)$$

Conséquences du théorème de Cochran (2)

3 $\hat{\beta}$ et $\hat{\sigma}^2$ sont indépendants donc

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_j} \sim \mathcal{T}(n-p-1) \text{ où } \hat{\sigma}_j^2 = \hat{\sigma}^2 (X^\top X)^{-1}_{jj}.$$

Test de nullité d'un coefficient

On rejette $\mathcal{H}_0 : \beta_j = 0$ au niveau α quand

$$\|T_j\| = \left| \frac{\hat{\beta}_j}{\hat{\sigma}_j} \right| > t_{n-p-1,1-\alpha/2}.$$

En pratique, on utilise la p-value.

4 si $\mathbf{1} \in \text{vect}(X)$, on peut écrire

$$\mathbb{R}^n = \text{vect}(X)^{\perp} \overset{\perp}{\bigoplus} \left(\text{vect}(\mathbf{1})^{\perp_{\text{vect}(X)}} \right) \overset{\perp}{\bigoplus} \text{vect}(\mathbf{1})$$

Alors

$$\underbrace{Y - X\hat{\beta}}_{\substack{\in \text{vect}(X)^{\perp} \\ \text{de dim.} n-p}} \quad \text{et} \quad \underbrace{X\hat{\beta} - \bar{Y}\mathbf{1}}_{\substack{\in \text{vect}(\mathbf{1})^{\perp_{\text{vect}(X)}} \\ \text{de dim.} p-1}}$$

sont indépendants
donc :

$$\frac{\|X\hat{\beta} - \bar{Y}\mathbf{1}\|^2/p}{\|Y - X\hat{\beta}\|^2/(n-p-1)} \sim \mathcal{F}(p, n-p-1).$$

Test de nullité de tous les coefficients

On rejette $\mathcal{H}_0 : \beta_1 = \ldots = \beta_p = 0$ au niveau α quand

$$\frac{\|X\hat{\beta} - \bar{Y}\mathbf{1}\|^2/p}{\|Y - X\hat{\beta}\|^2/(n-p-1)} > f_{p,n-p-1,1-\alpha/2}.$$

En pratique, on utilise la p-value.

Erreur d'estimation de $X_i\beta$

Pour un individu i ($i = 1, \ldots, n$), la valeur Y_i (observée) est estimée par le modèle par

$$\hat{Y}_i = X_i\hat{\beta}.$$

On a

$$\mathbb{E}\hat{Y}_i = X_i\beta \text{ et}$$
$$\mathbb{V}(\hat{Y}_i) = X_i\mathbb{V}(\hat{\beta})X_i^\top = \sigma^2 X_i(X^\top X)^{-1}X_i^\top.$$

Intervalle de confiance pour $X_i\beta$

On a

$$\frac{\hat{Y}_i - X_i\beta}{\sqrt{\hat{\sigma}^2 X_i(X^\top X)^{-1}X_i^\top}} \sim \mathcal{T}(n-p-1).$$

donc

Intervalle de confiance pour l'estimation

On sait donc que, avec probabilité $1-\alpha$

$$X_i\beta \in [\hat{Y} \pm t_{n-p-1,1-\alpha/2}\sqrt{\hat{\sigma}^2 X_i(X^\top X)^{-1}X_i^\top}]$$

Erreur de prévision de Y

Si on considère un nouvel individu indépendant de $1, \ldots, n$ pour lequel on connaît X_+ (mais pas Y_+), on peut prédire la valeur de $Y_+ = X_+\beta + \epsilon_+$ par

$$Y_+^p = X_+\hat{\beta},$$

l'erreur commise est alors donnée par :

$$Y_+^p - Y_+ = X_+\hat{\beta} - (X_+\beta + \epsilon_+) = X_+(X^\top X)^{-1}X\epsilon - \epsilon_+.$$

Intervalle de prévision pour Y_k

$$\frac{Y_+^p - Y_+}{\sqrt{\hat{\sigma}^2(X_+(X^\top X)^{-1}X_+^\top + 1)}} \sim \mathcal{T}(n-p-1).$$

Intervalle de confiance pour l'estimation

On sait donc que, avec probabilité $1-\alpha$

$$Y_+ \in [Y_+^p \pm t_{n-p-1,1-\alpha/2}\sqrt{\hat{\sigma}^2 X_+(X^\top X)^{-1}X_+^\top + 1}].$$

Modèle avec 1 covariable

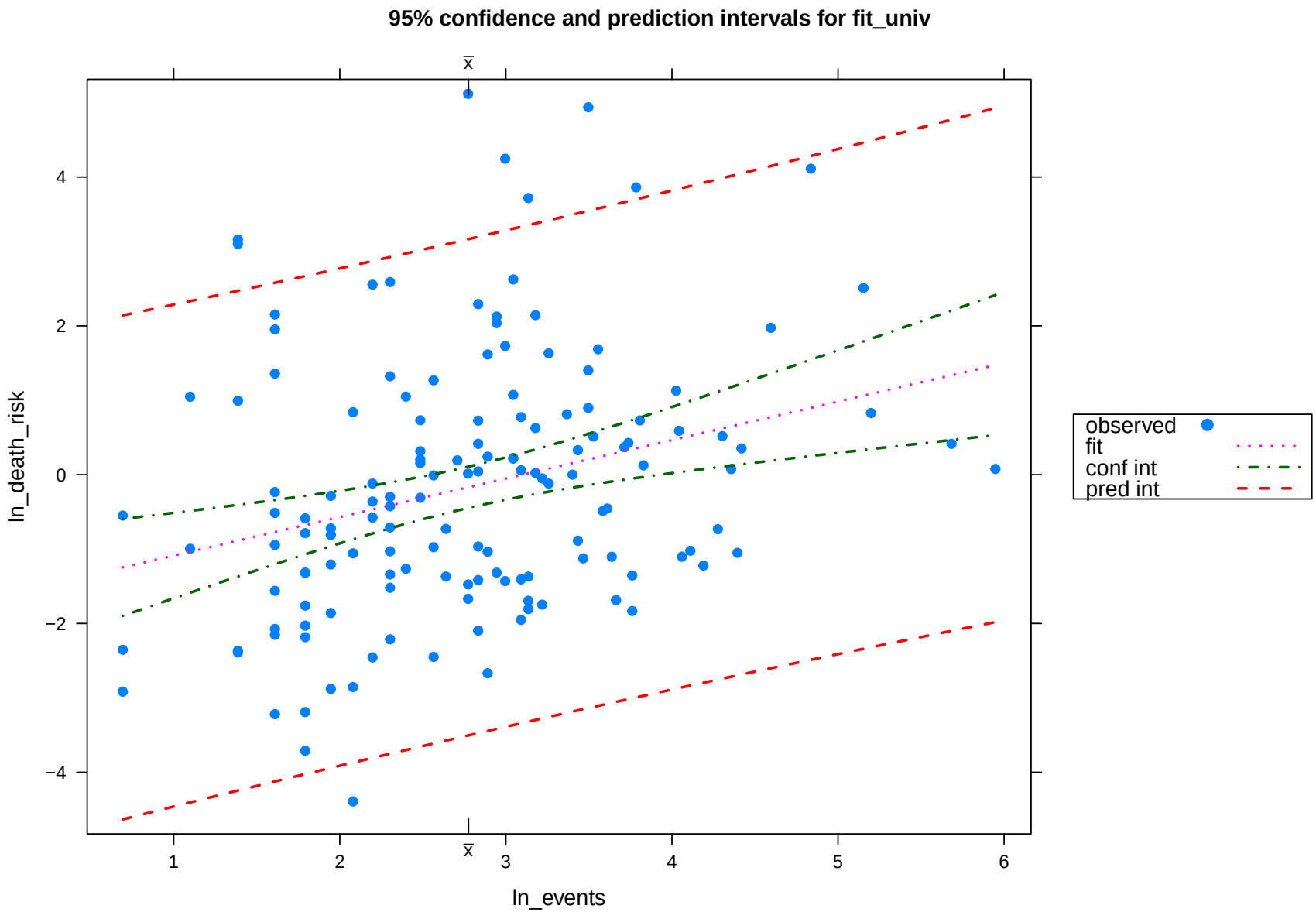
```
fit_univ = lm(ln_death_risk~ln_events)
newdata=data.frame(ln_events=3.4)
pred=predict(fit_univ,newdata,interval="predict")
ic=predict(fit_univ,interval="confidence")
print(pred)
```

```
##        fit       lwr      upr
## 1 0.1543123 -3.185642 3.494266
```

```
print(ic[1:5,])
```

```
##            fit          lwr         upr
## 1 -0.41346753 -0.72116718 -0.1057679
## 2  0.20424358 -0.14006908  0.5485563
## 3 -0.02960412 -0.31691647  0.2577082
## 4  0.27723443 -0.09224715  0.6467160
## 5 -0.88753758 -1.36903423 -0.4060409
```

```
ci.plot(fit_univ)
```



On doit vérifier les hypothèses du modèle, i.e.

- les hypothèses sur X (de plein rang)
- les hypothèses sur les erreurs
- la présence d'individus "influents"

Diagnostics sur X

Rang de la matrice X

- On veut vérifier l'hypothèse que X est de plein rang, i.e. que les $p+1$ colonnes de X engendrent un s.e.v. de $\mathbb{R}^n$ de dimension $p+1$.
- Si ce n'est pas le cas, la matrice $X^\top X$ n'est pas inversible, il n'y a donc pas de solution unique à l'équation
$$X^\top Y = X^\top X \hat{\beta}.$$
- On veut donc vérifier qu'il n'y pas de colinéarité entre les colonnes $\mathbf{1}, X^1, \ldots, X^p$ de X.

Valeurs propres de la matrice de corrélation

On définit la matrice R des corrélations empiriques entre les variables X^j, $j = 1, \ldots, p$:

$$R_{jj'} = \frac{\sum_{i=1}^{n}(X_i^j - \bar{X}^j)(X_i^{j'} - \bar{X}^{j'})}{\sqrt{\sum_{i=1}^{n}(X_i^j - \bar{X}^j)^2 \sum_{i=1}^{n}(X_i^{j'} - \bar{X}^{j'})^2}} = \text{cor}(X^j, X^{j'}).$$

- C'est une matrice symétrique positive de rang $= \dim(\text{vect}(X)) \leq p$ ($< p$ si il y a colinéarité).
- On calcule les p valeurs propres $\lambda_1 \geq \ldots \geq \lambda_p$ de cette matrice.
  - S'il y a une relation linéaire parfaite entre des X^j, une des valeurs propres vaut 0.
  - On définit l'indice de conditionnement $\kappa = \lambda_1/\lambda_p$ et la règle $\kappa > 500$ ou$1000 \Longrightarrow$ colinéarité trop forte
  - Si on veut une étude plus fine, il faut étudier les vecteurs propres associées aux trop petites valeurs propres.

Matrice de correlations

Definition de la matrice

```
X = vul[,c(3:6)]
cor_mat = cor(X)
```

Calcul des valeurs propres et vecteurs propres

```
propres = eigen(cor_mat)
1/ propres$values
```

```
## [1] 0.5009212 0.6255898 3.6824367 7.4835135
```

Variance inflation factor (VIF) et tolérance

Considérons la régression de la variable X^j sur les autres variables explicatives $X^1, \ldots, X^{j-1}, X^{j+1}, \ldots, X^p$, on note R_j^2 le R^2 associé à cette régression.

- Si $R_j^2 = 0$, X^j n'est pas fonction linéaire des autres variables
- Si $R_j^2 = 1$, X^j est fonction linéaire des autres variables $\Longrightarrow$ colinéarité

On définit les coefficients de "variance inflation factor" (VIF) pour $j = 1, \ldots, p$ par :

$$VIF_j = \frac{1}{1 - R_j^2}.$$

Règle

si $VIF > 10$ ou $100 \Longrightarrow$ colinéarité

Variance inflation factors

```
vif(fit)
```

```
## ln_events   ln_fert       hdi    ln_pop
##  2.421759  3.642415  3.767663  2.460624
```

Règle à suivre pour les problèmes de colinéarité

- Si on détecte un problème de colinéarité, il faut enlever les variables posant problème **une à une**.
- Le choix des variables devrait se faire avec ceux qui ont fourni le jeu de données.

Résidus

Analyse des résidus

On veut vérifier les hypothèses sur les erreurs ϵ, i.e.

- indépendantes (ou décorrélées) avec

$$\mathbb{E}(\epsilon_i) = 0 \text{ et } \mathbb{V}(\epsilon_i) = \sigma^2$$

- voire gaussiennes

Test de normalité

On suppose

$$\epsilon \sim \mathcal{N}((0,\dots,0)^\top, \sigma^2 I_n).$$

- Si les ϵ_i $(i = 1,\dots,n)$ étaient observables, on pourrait tracer un QQ-plot des ϵ_i/σ contre les quantiles de la $\mathcal{N}(0,1)$.
- On n'observe que les erreurs résiduelles e_i $(i = 1,\dots,n)$, et on définit **les résidus studentisés**

$$e_i^* = \frac{e_i}{\sqrt{\hat{\sigma}^2_{(i)}(1 - H_{ii})}},$$

  qu'on prend comme "estimateurs" des ϵ_i.

On conseille de faire le QQ-plot sur ces résidus (si $n - p - 1$ est grand, on peut le faire avec les quantiles gaussiens)

Valeurs ajustées $\hat{y}$

```
yhat = fit$fitted.values
```

Résidus $e = y - \hat{y}$

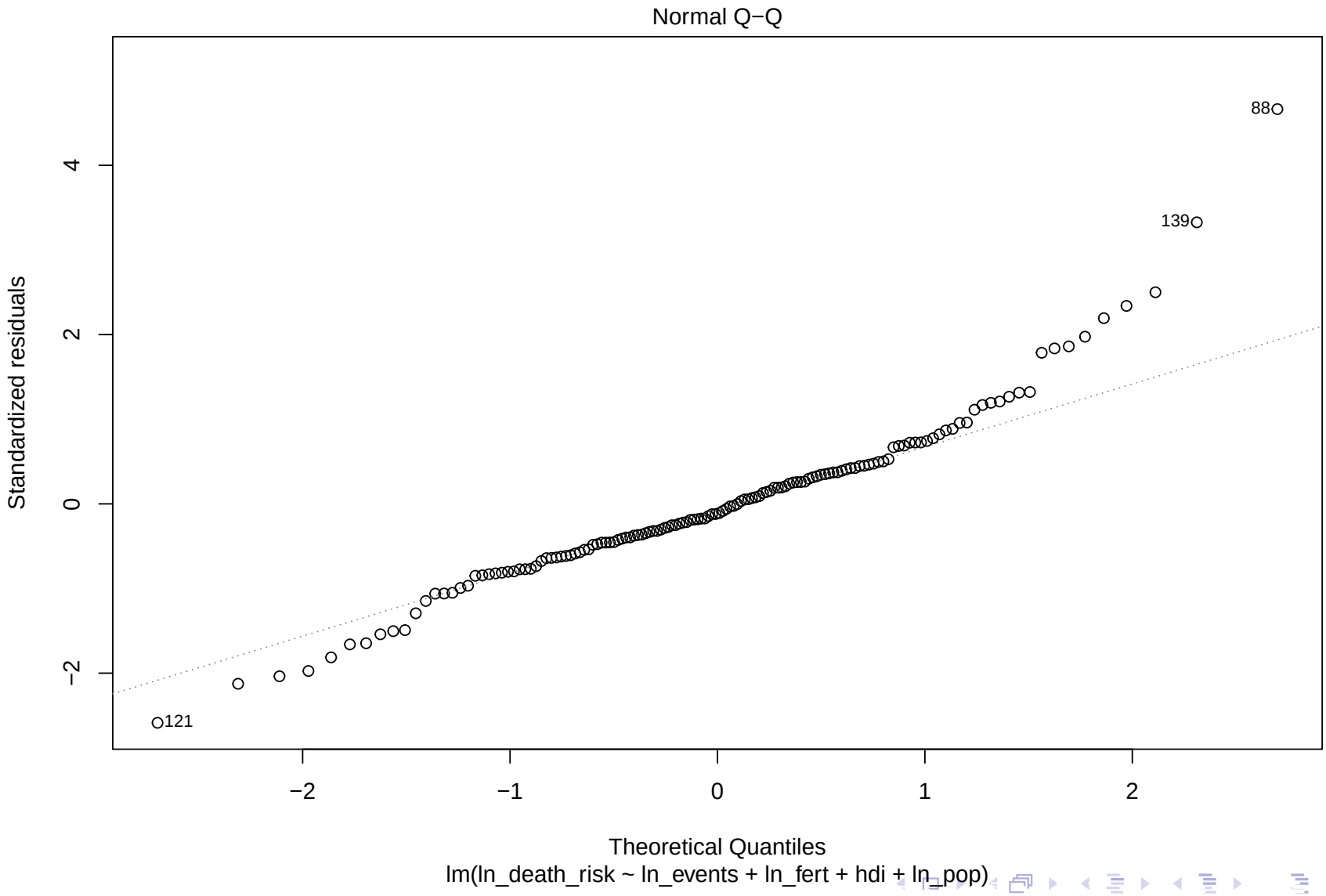
```
e = fit$residuals
```

Residus studentisés $e^{\star}$

```
e_star = rstudent(fit)
```

Normalité des erreurs

```
plot(fit,which=2)
```



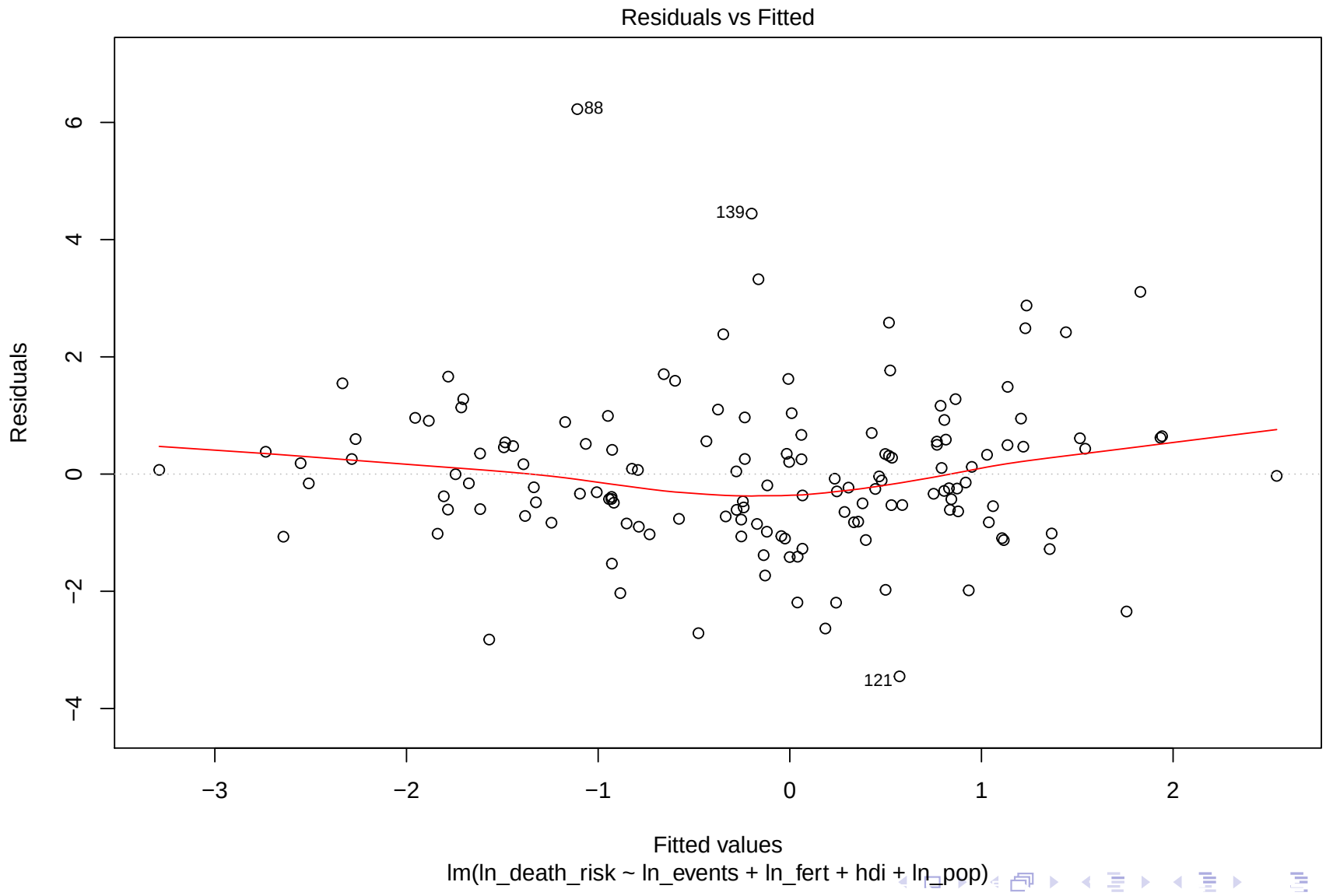
Dans le modèle linéaire

On veut vérifier les autres hypothèses sur les erreurs ϵ :

1. pour le centrage : c'est toujours vrai si on inclut l'intercept $(1, \ldots, 1)^\top$ dans la matrice X.
2. pour l'indépendance, il n'existe pas de test dans R. On conseille de représenter les résidus e contre les valeurs ajustées $X\hat{\beta}$ qui doivent décorrélées.

Graphique résidus/valeurs ajustées

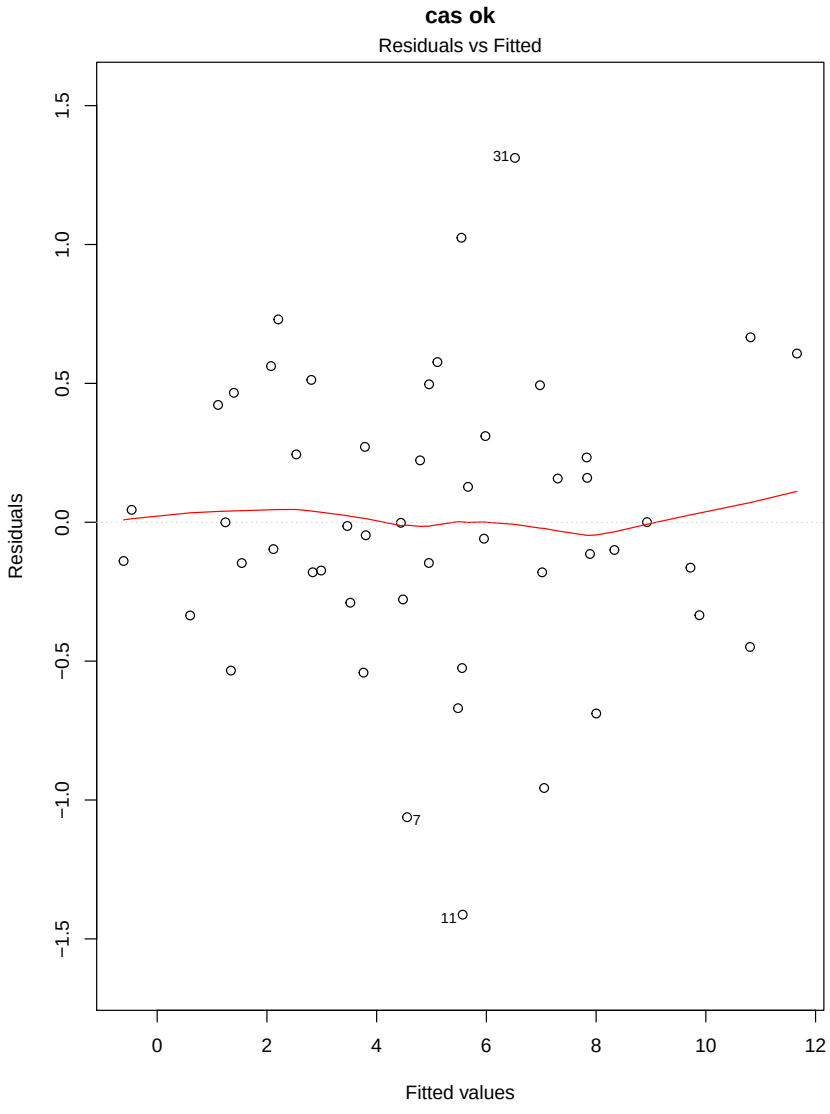
```
plot(fit,which=1)
```

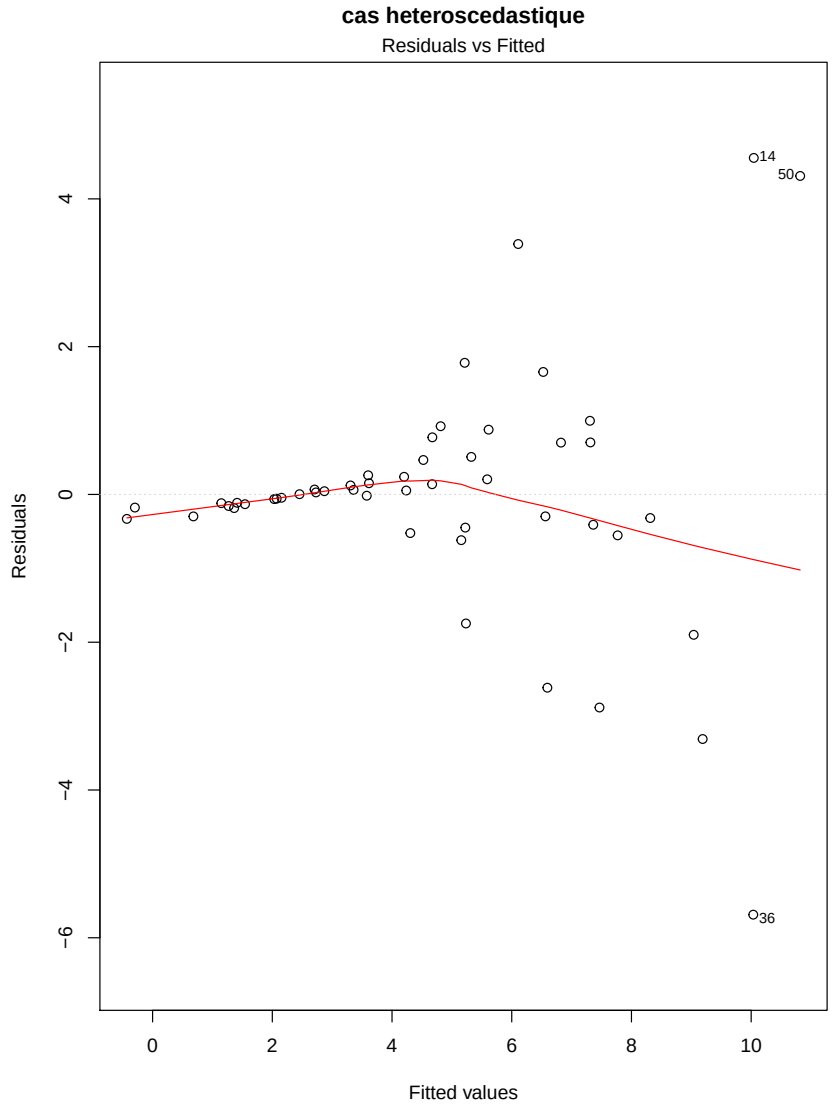


3 pour l'égalité des variances, on recommande de représenter $\sqrt{e^\star}$ contre les valeurs ajustées $X\hat{\beta}$. On ne doit pas voir de forme au nuage de points

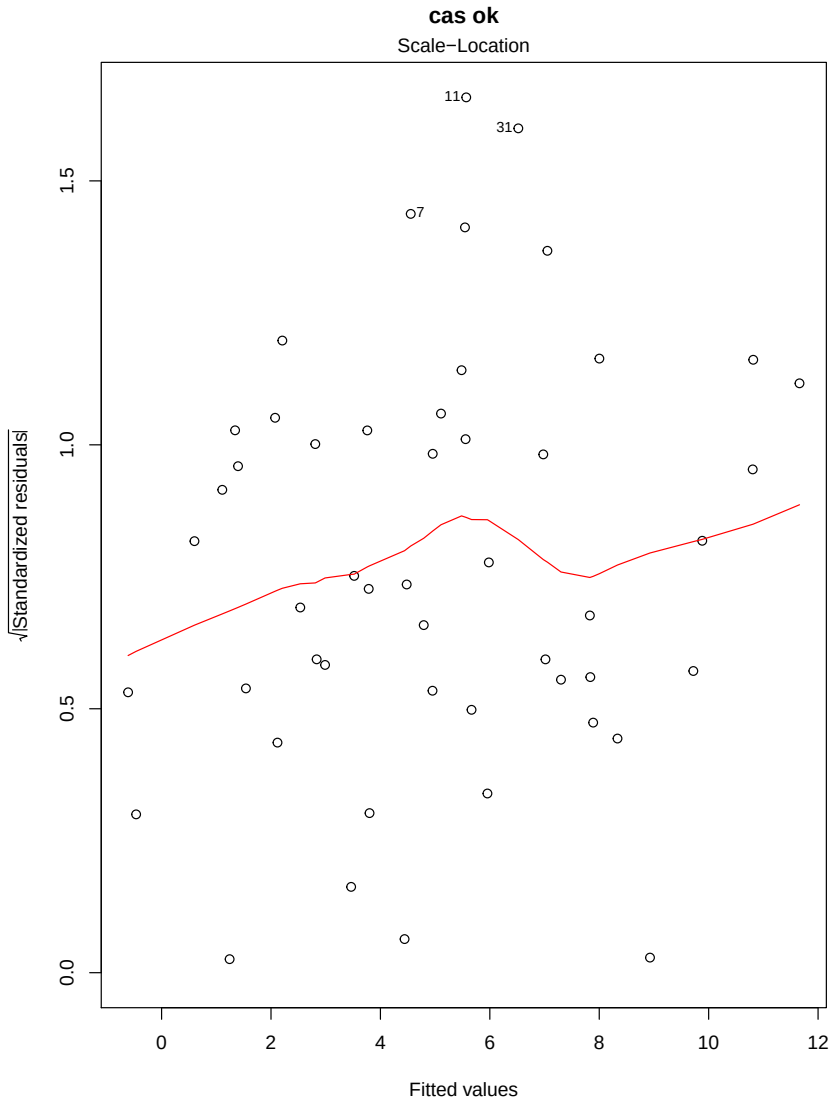
4 si on suspecte une autocorrélation, en particulier si on étudie une série chronologique, on utilise le test de Durbin-Watson (fonction dwtest du package lmtest).

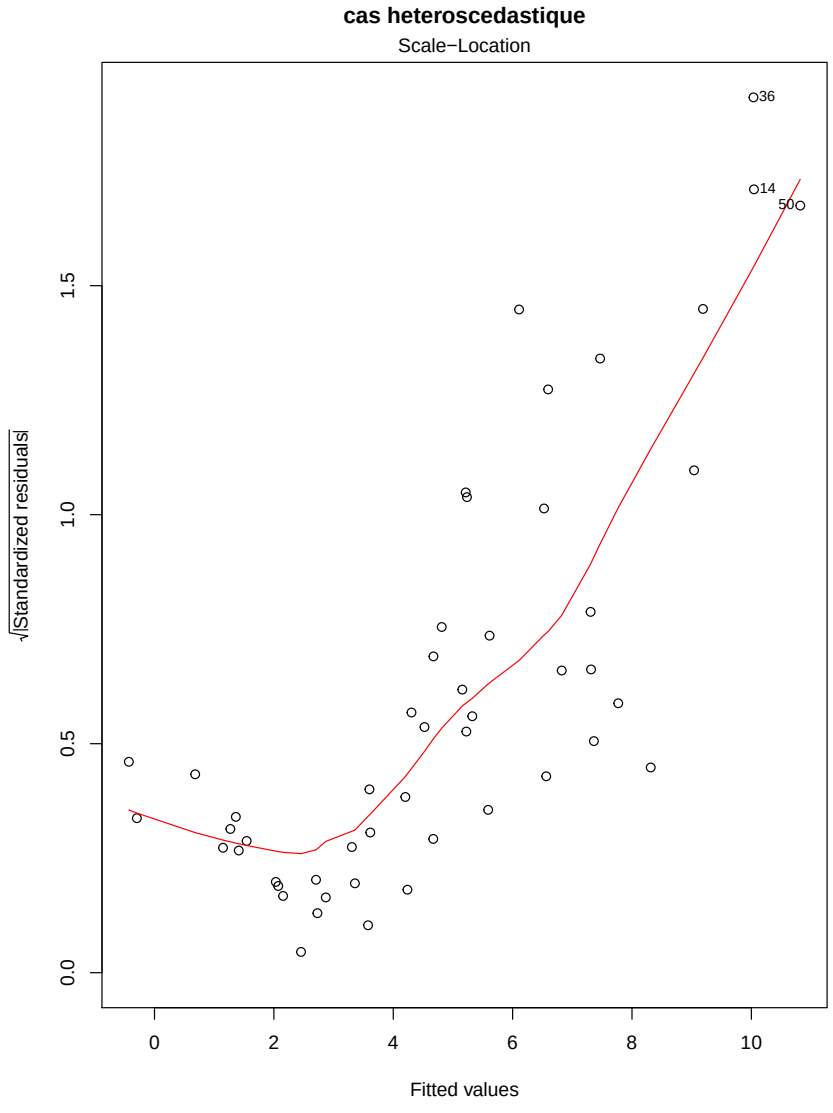
Graphiques résidus/ajustées





Graphiques scale/location





Observations

Différentes observations atypiques

On cherche maintenant des mesures de l'influence des observations dans l'estimation.

- Une "enquête" sur les observations/ les individus "trop influent(e)s" devra être faite, pour déterminer notamment s'il n'y a pas eu d'erreur de mesure, de relevé, etc.
- Le rôle du statisticien est de les détecter.

On peut distinguer deux types d'observations atypiques :

- celles qui ont un "trop" grand résidu (influence sur l'estimation de σ)
- celles qui sont trop isolées (influence sur l'estimation de β)

Observation aberrante

On connaît la loi des résidus studentisés e_i^*

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

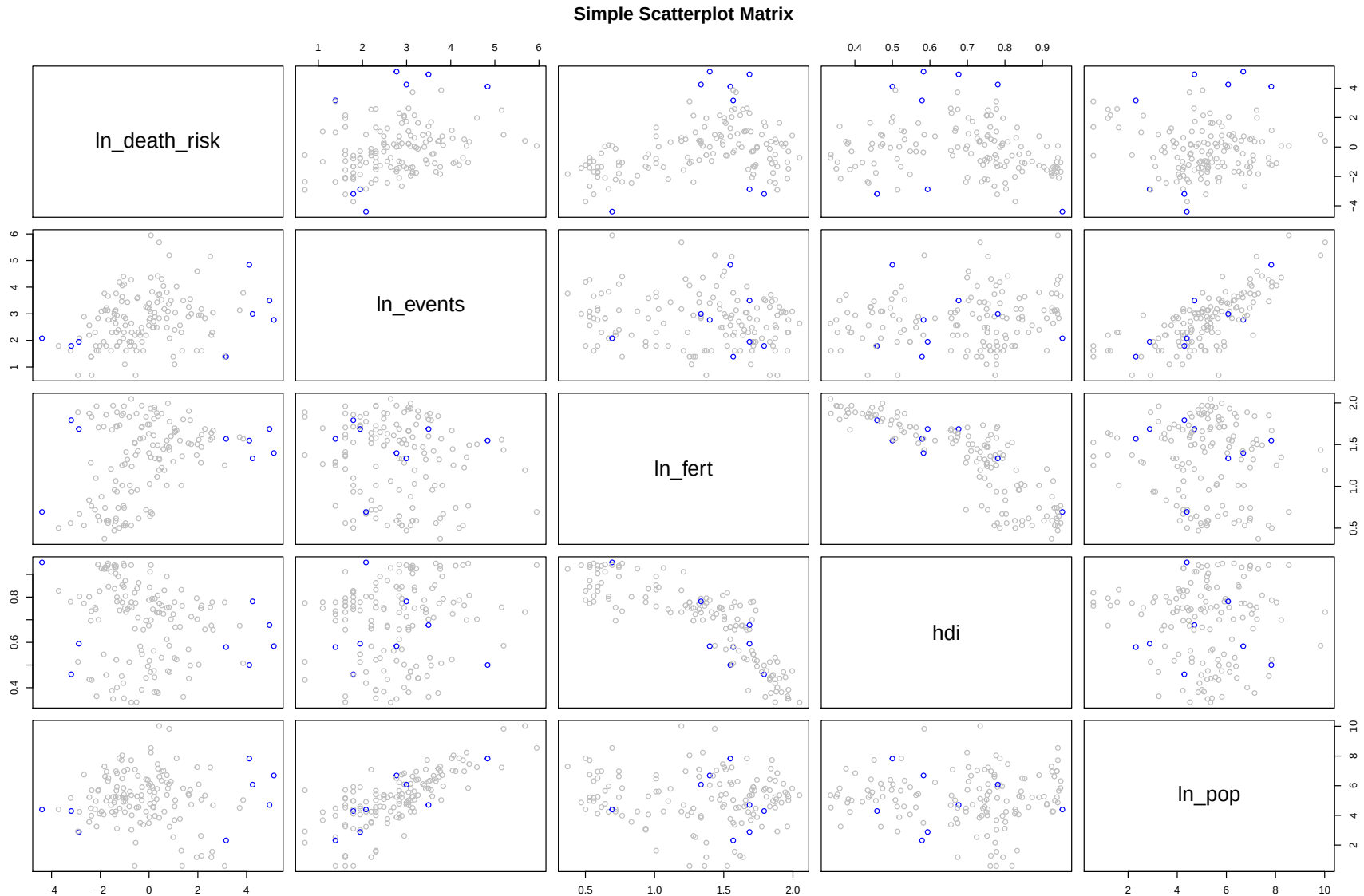
Règle

On dit qu'une observation est **aberrante** si

$$e_i^* > F^{-1}_{\mathcal{T}(n-p-1)}(1-\alpha).$$

On choisit souvent α de l'ordre de $1/n$ ou $F^{-1}_{\mathcal{T}(n-p-1)}(1-\alpha) = 2$.

Où sont les points “aberrants” ?



Le levier

Une bonne mesure de l'isolement des observations est le **coefficient H_{ii} appelé "levier"** ("leverage") .
On sait (par propriété géométrique) que $0 \leq H_{ii} \leq 1$.

Régle pour les leviers

On sait aussi que $\sum_i H_{ii} = p$, on considère donc qu'une observation est **isolée** quand a un levier sup. à $2p/n$ (ou $(2p+2)/n$ ou $3p/n$).

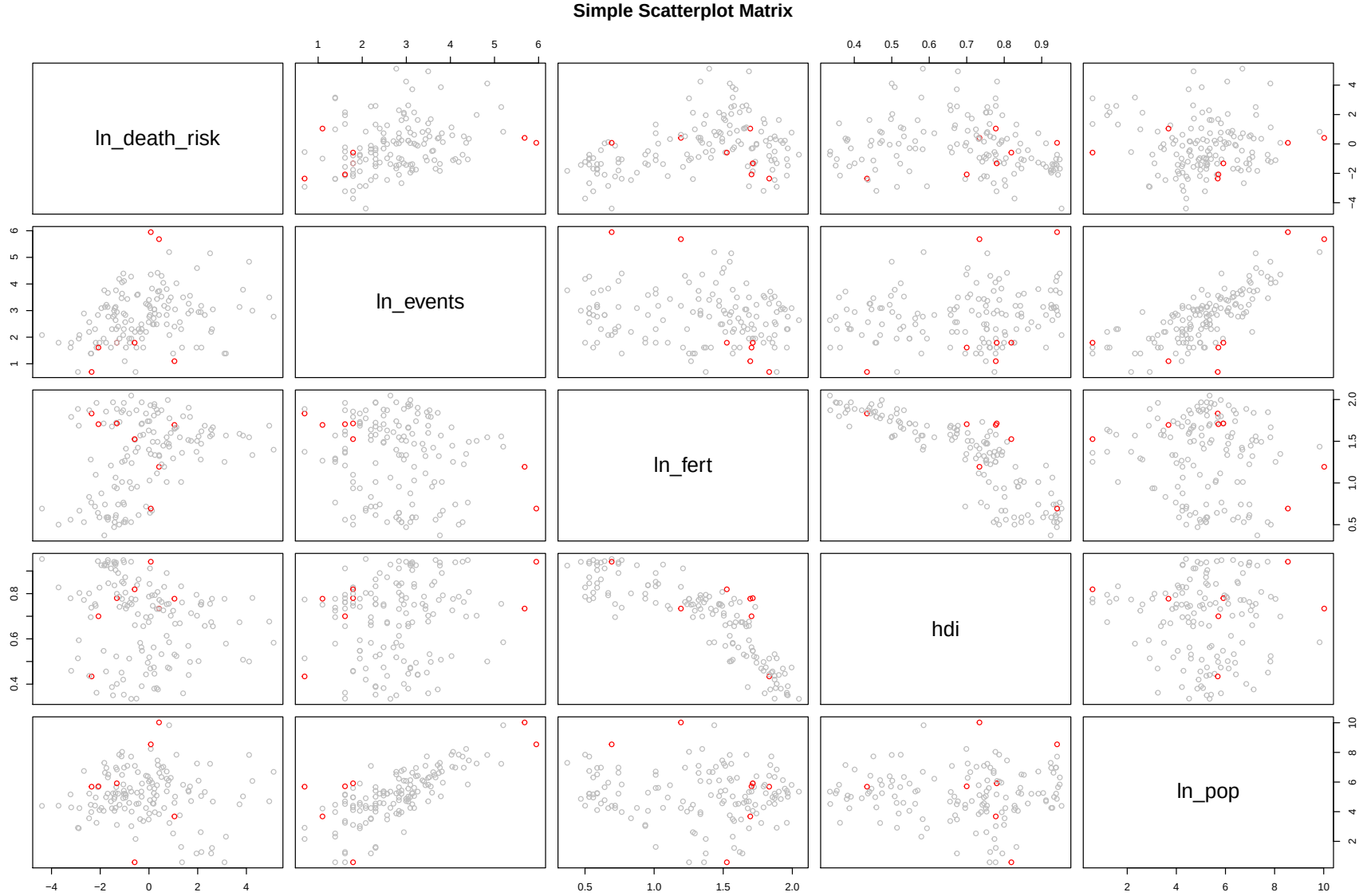
Distance de Cook

La **distance de Cook** est une mesure globale :

$$\text{DCOOK}_i = \frac{(e_i)^2 H_{ii}}{(1+p)\hat{\sigma}^2(1-H_{ii})^2} > 4/n \text{ ou } 1 \Longrightarrow \text{influence}$$

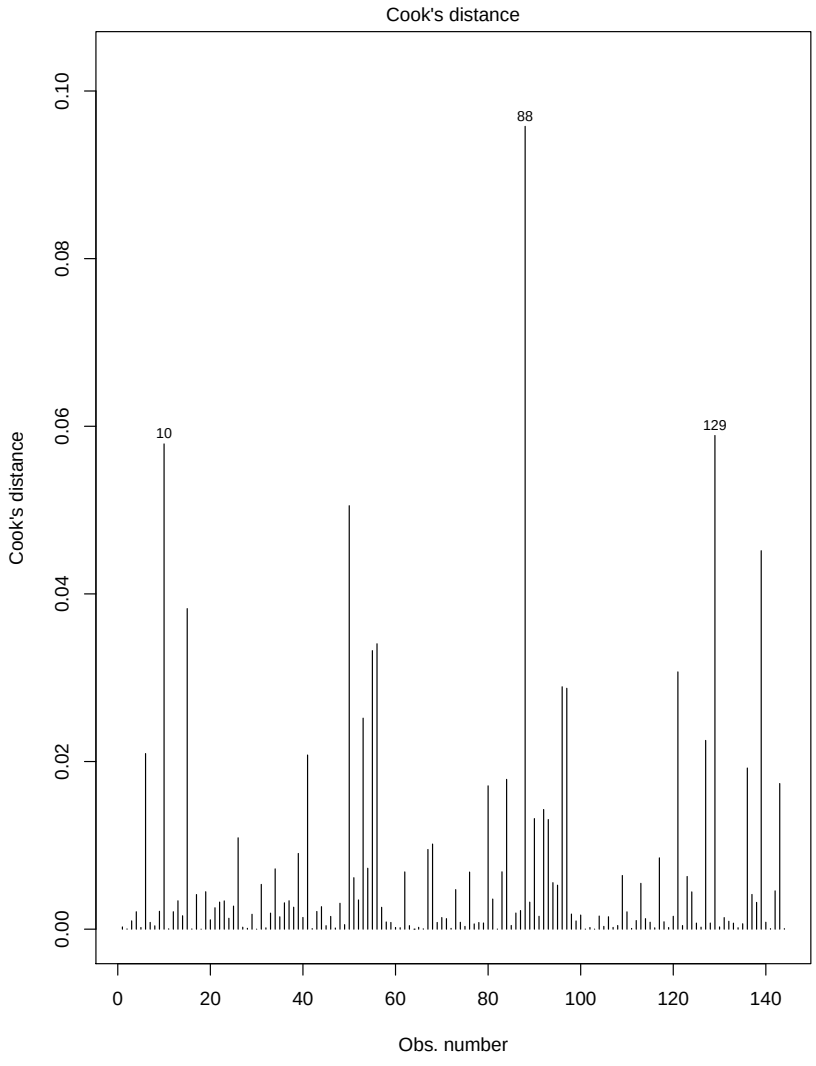
Leviers, observations influentes

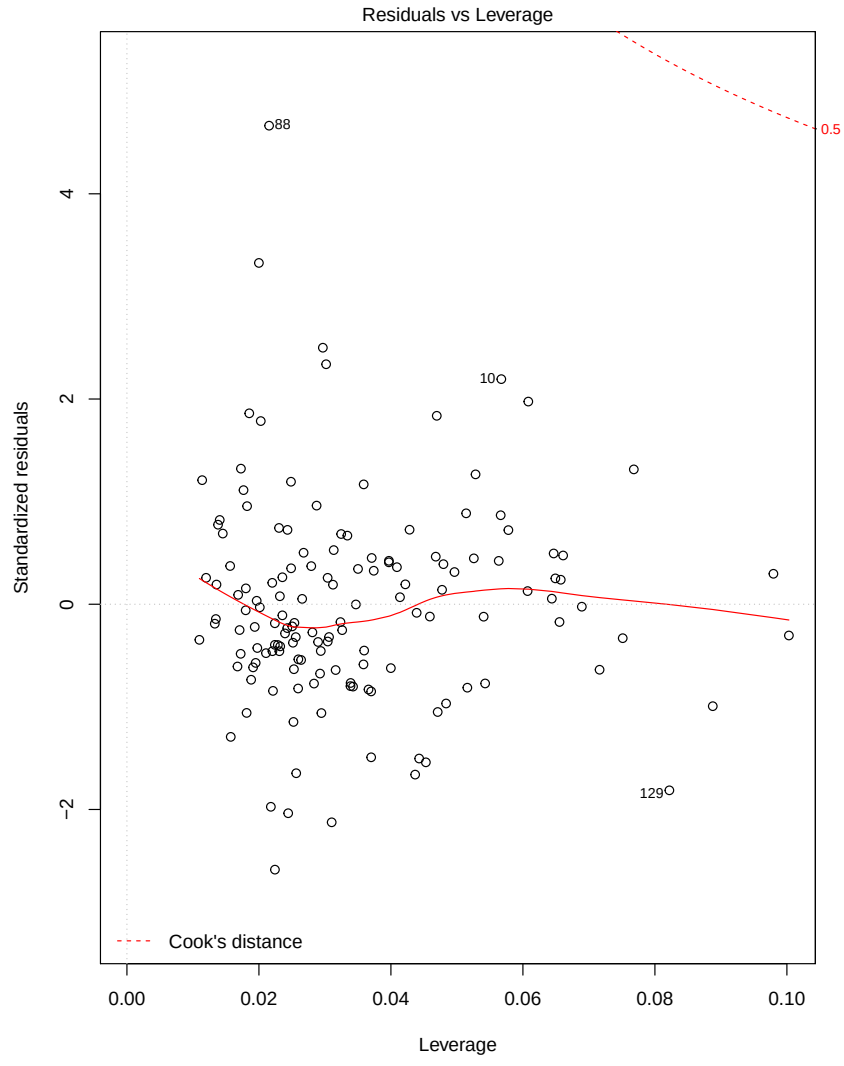
```
influences = lm.influence(fit)
hat = influences$hat
```


Simple Scatterplot Matrix
ln_death_risk
ln_events
ln_fert
hdi
ln_pop

```
detach(vul)
```

Graphique DCook





Interprétation

Définition : résidus partiels

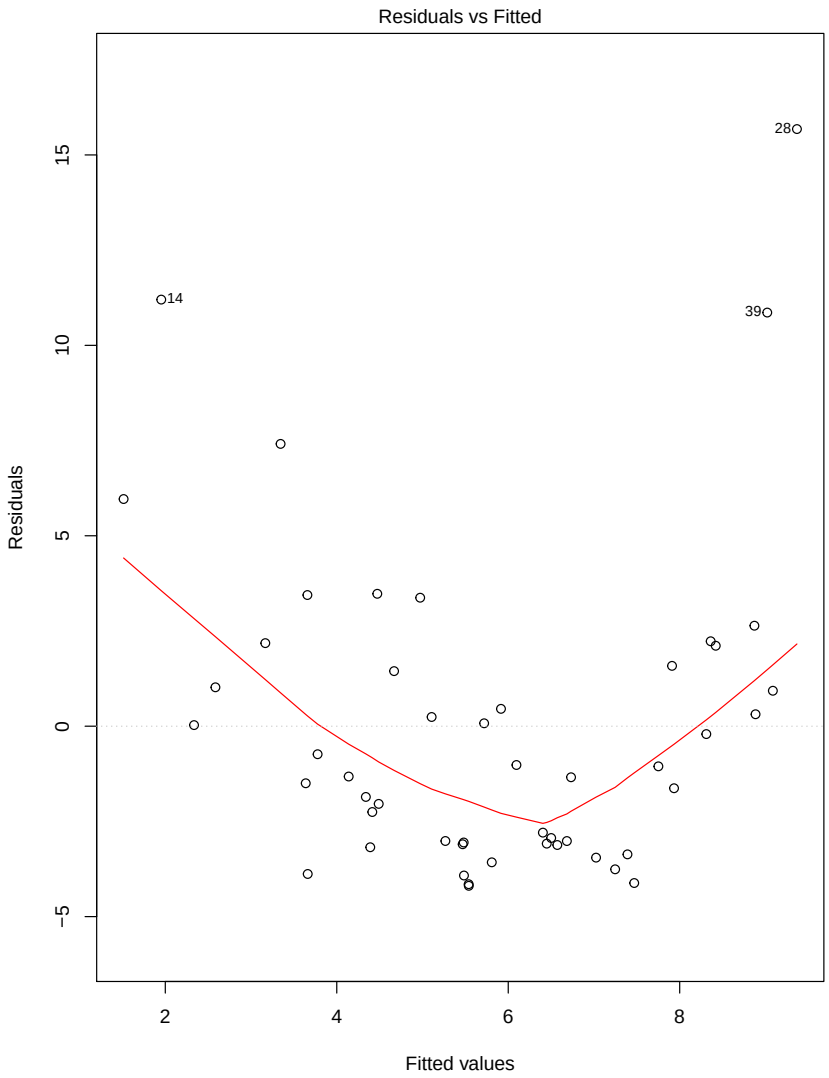
Quand on suspecte un problème de linéarité entre un X^j et Y, on représente le résidu partiel

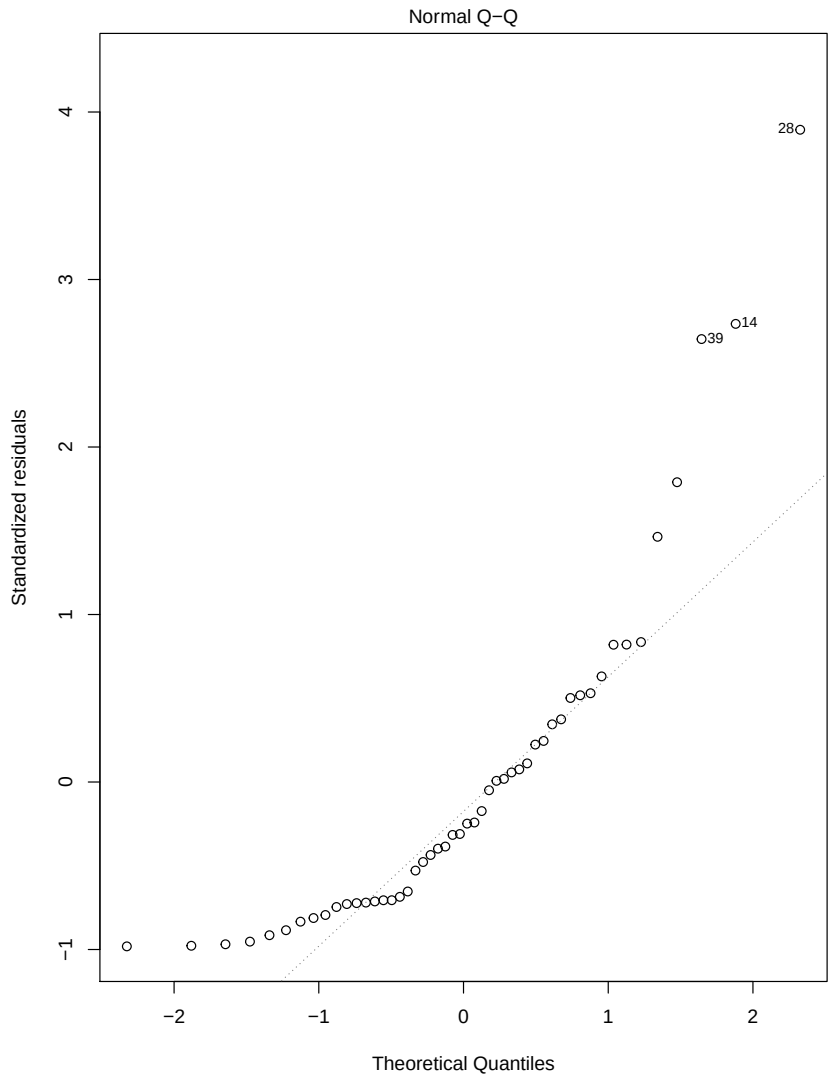
$$e_P^j = e + X^j \hat{\beta}_j$$

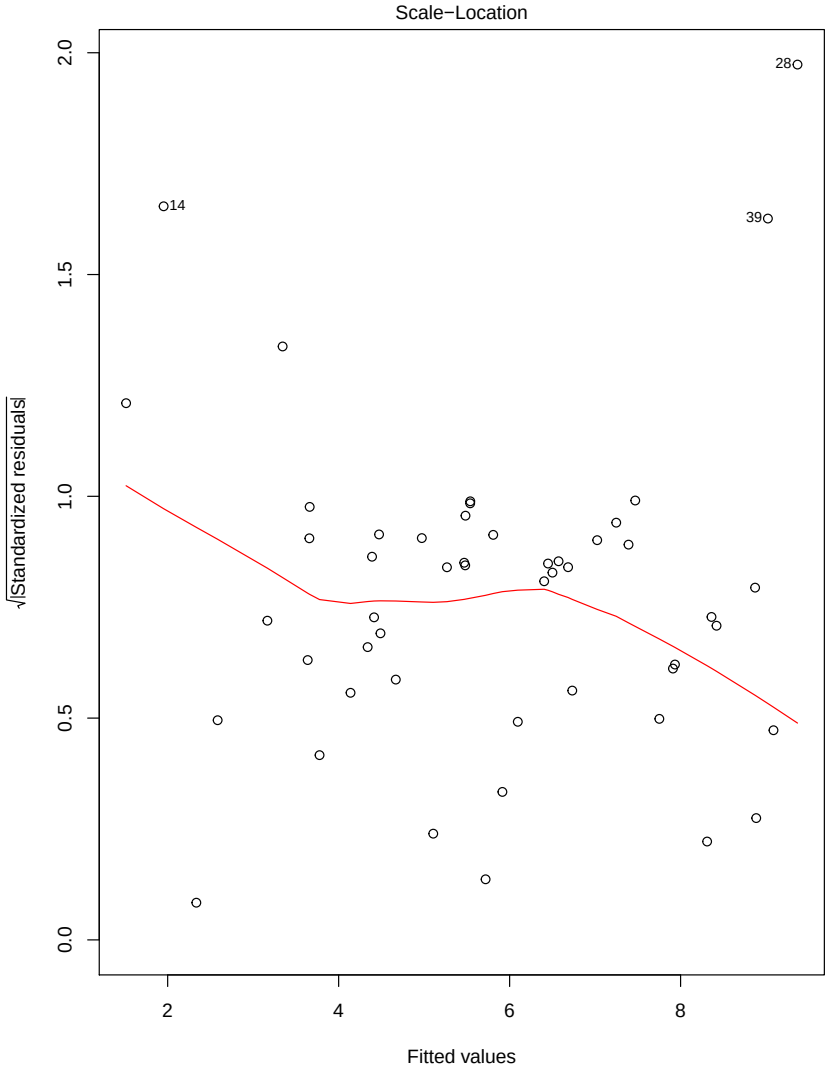
contre le régresseur X^j.

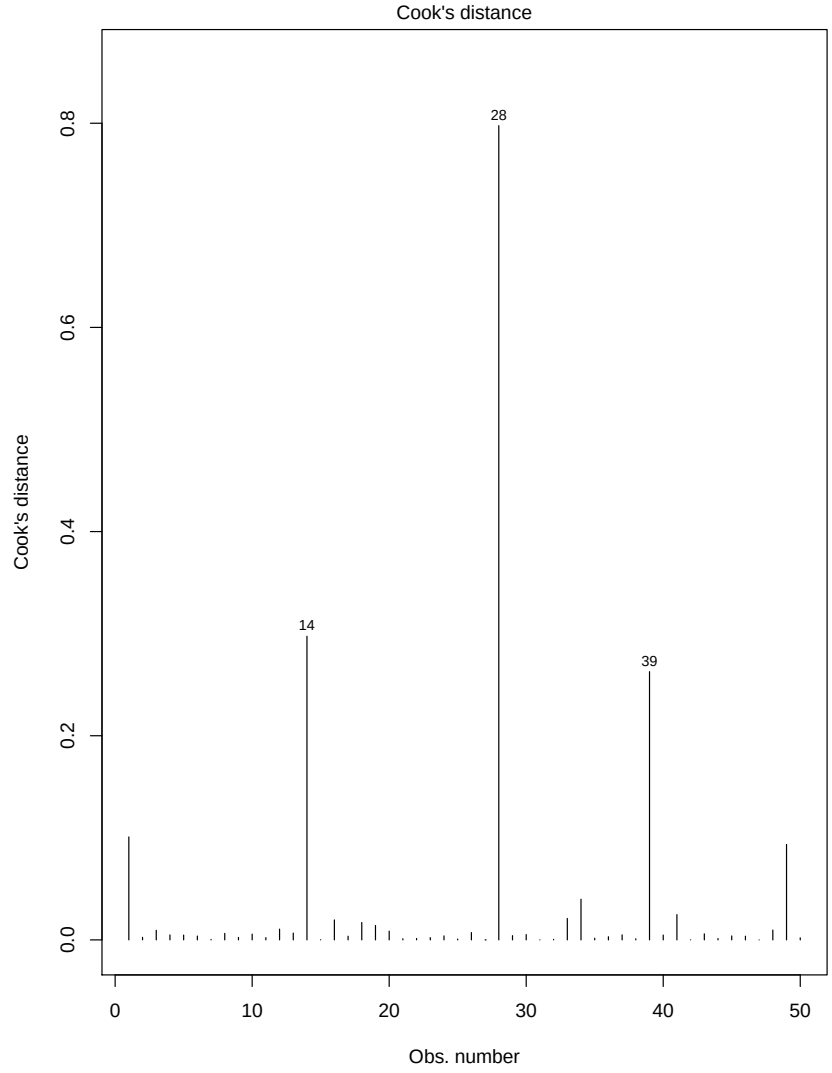
Relation non-linéaire due à une covariable

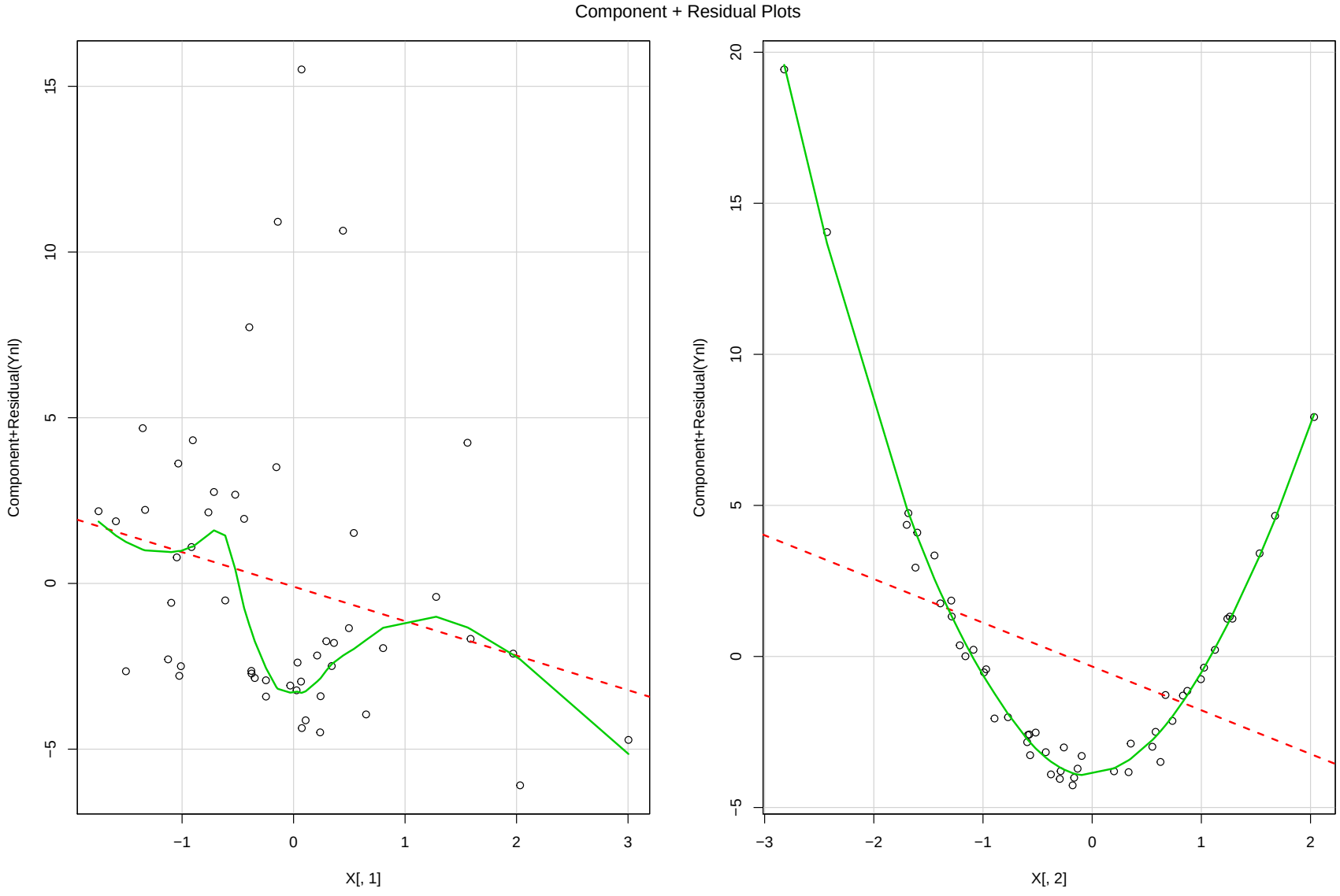
```
n =50
X=matrix(rnorm(n*2),ncol=2)
epsilon = rnorm(n,0,0.5)

Ynl = 2 - X[,1] + 3* X[,2]^2 + epsilon
lmnl = lm(Ynl~X[,1]+X[,2])
```


Residuals vs Fitted
Residuals
Fitted values
28
14
39
15
10
5
0
−5
2
4
6
8

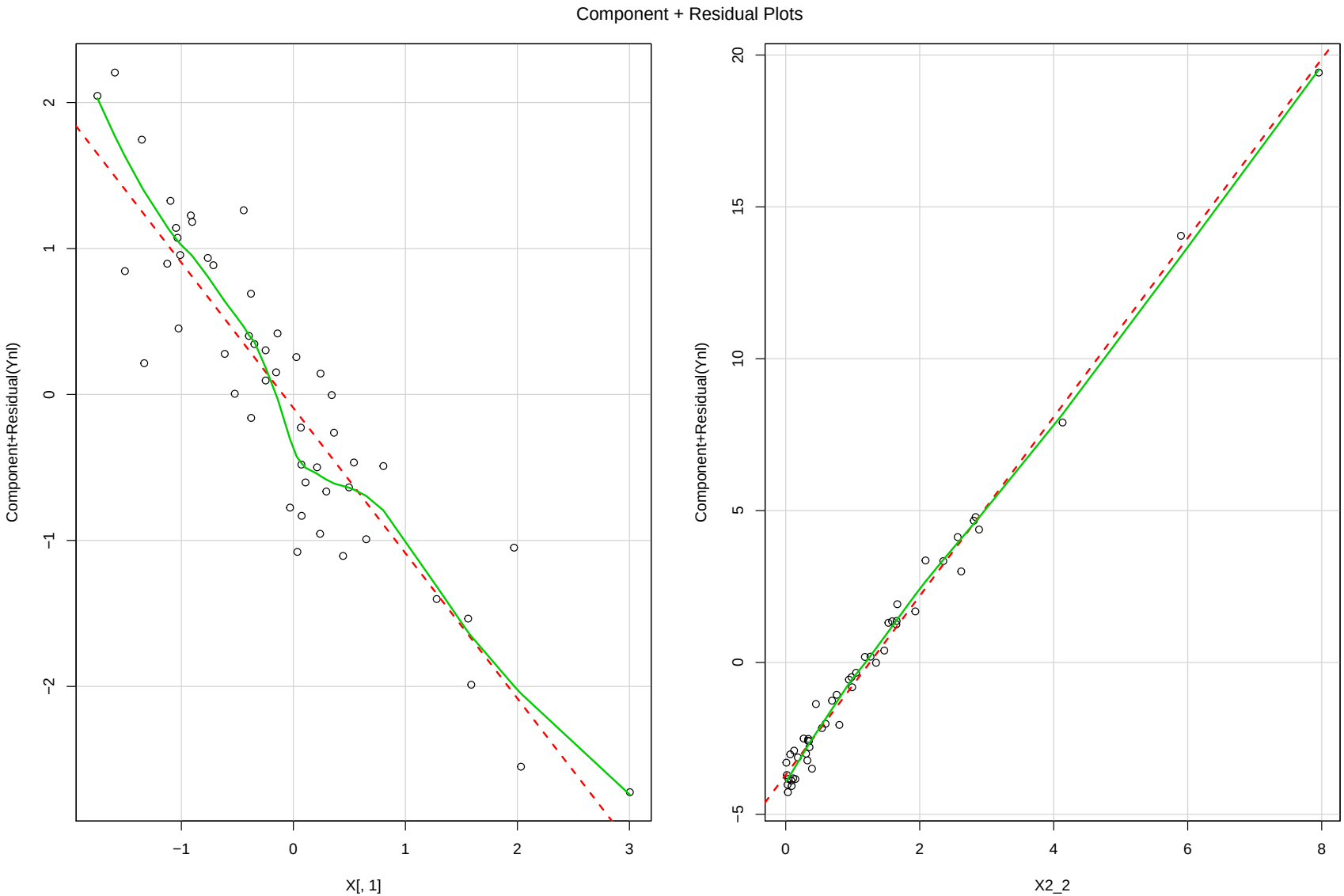

Normal Q−Q
Standardized residuals
Theoretical Quantiles
28
39
14
4
3
2
1
0
−1
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
Fitted values
0.0
0.5
1.0
1.5
2.0
2
4
6
8
14
28
39


Cook's distance
Cook's distance
Obs. number
0.0
0.2
0.4
0.6
0.8
0
10
20
30
40
50
14
28
39

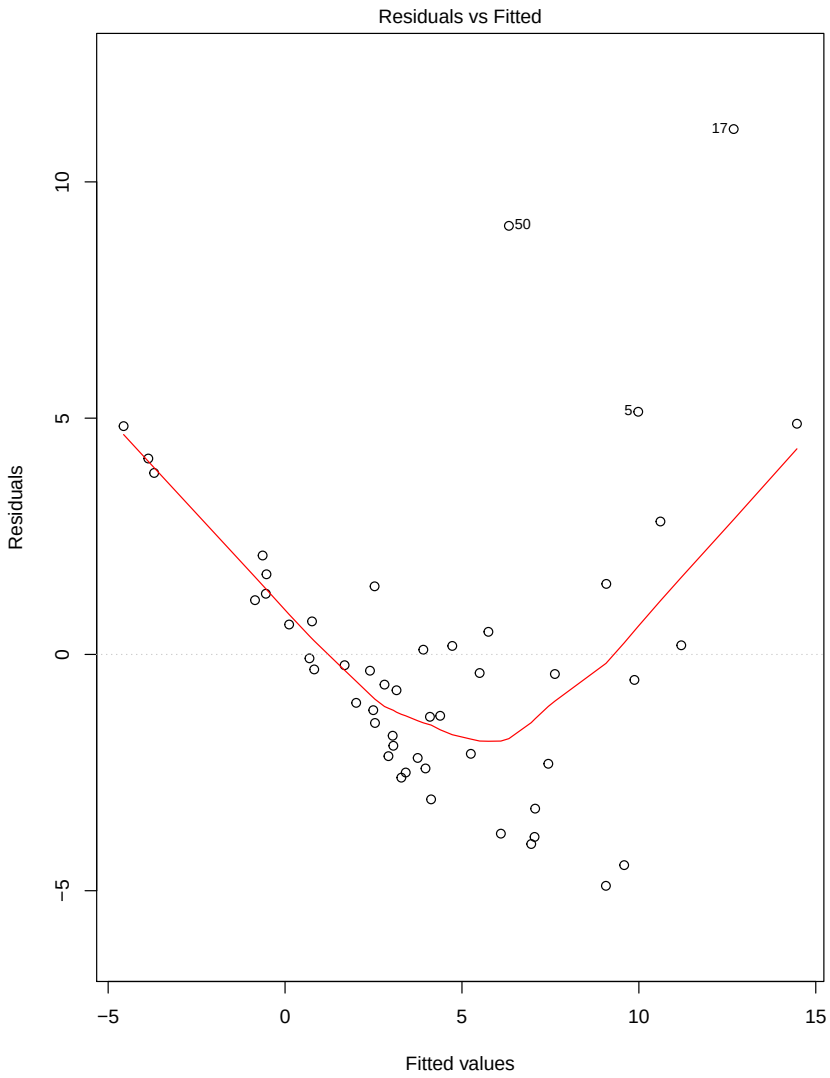

Component + Residual Plots
Component+Residual(Ynl)
X[, 1]
X[, 2]

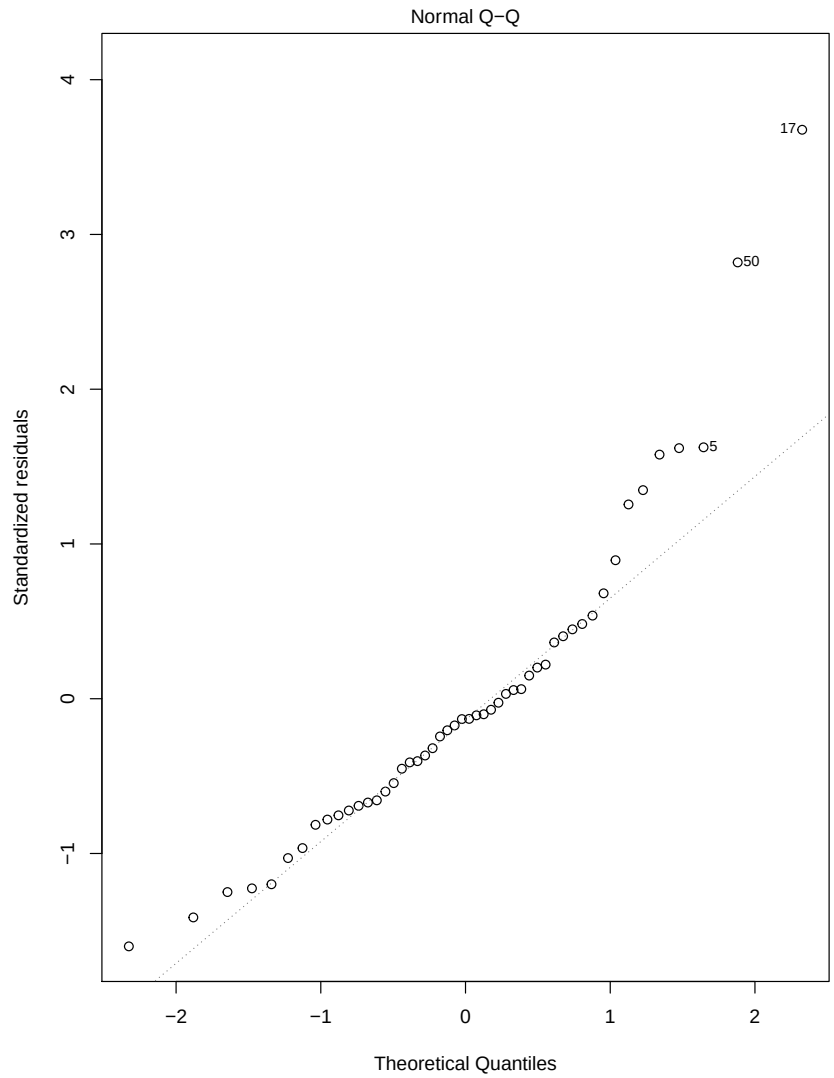
Transformation de la variable

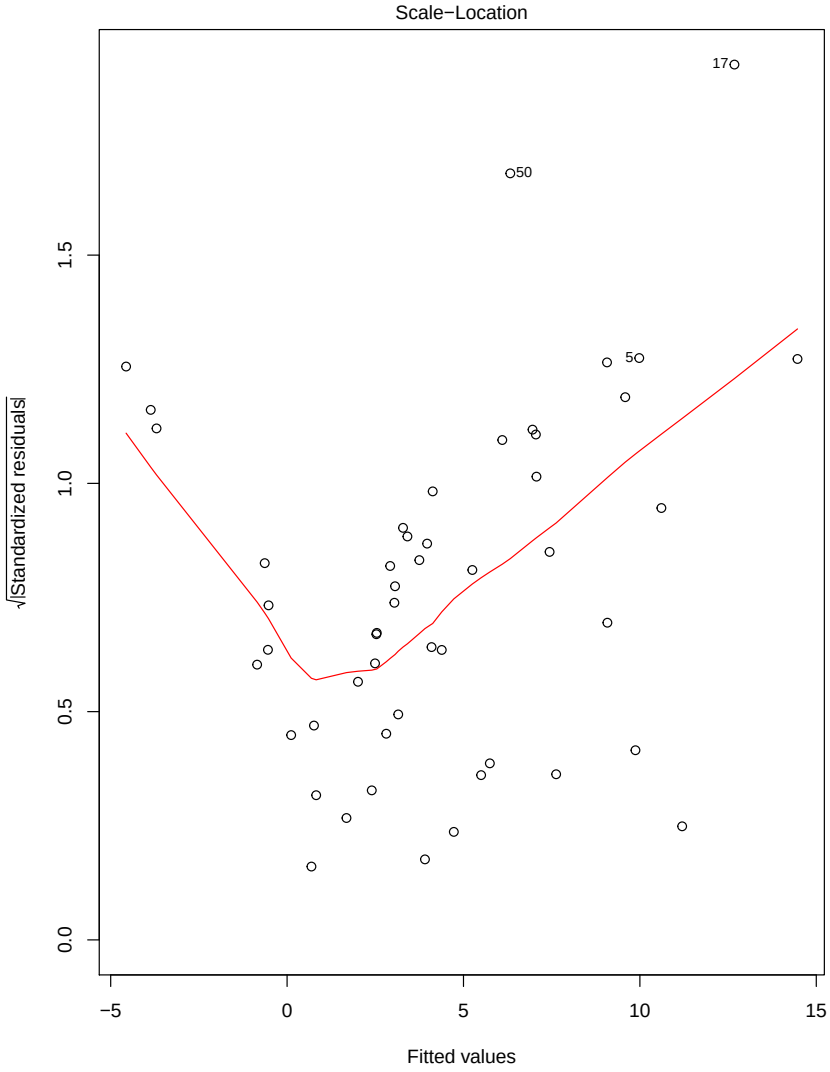


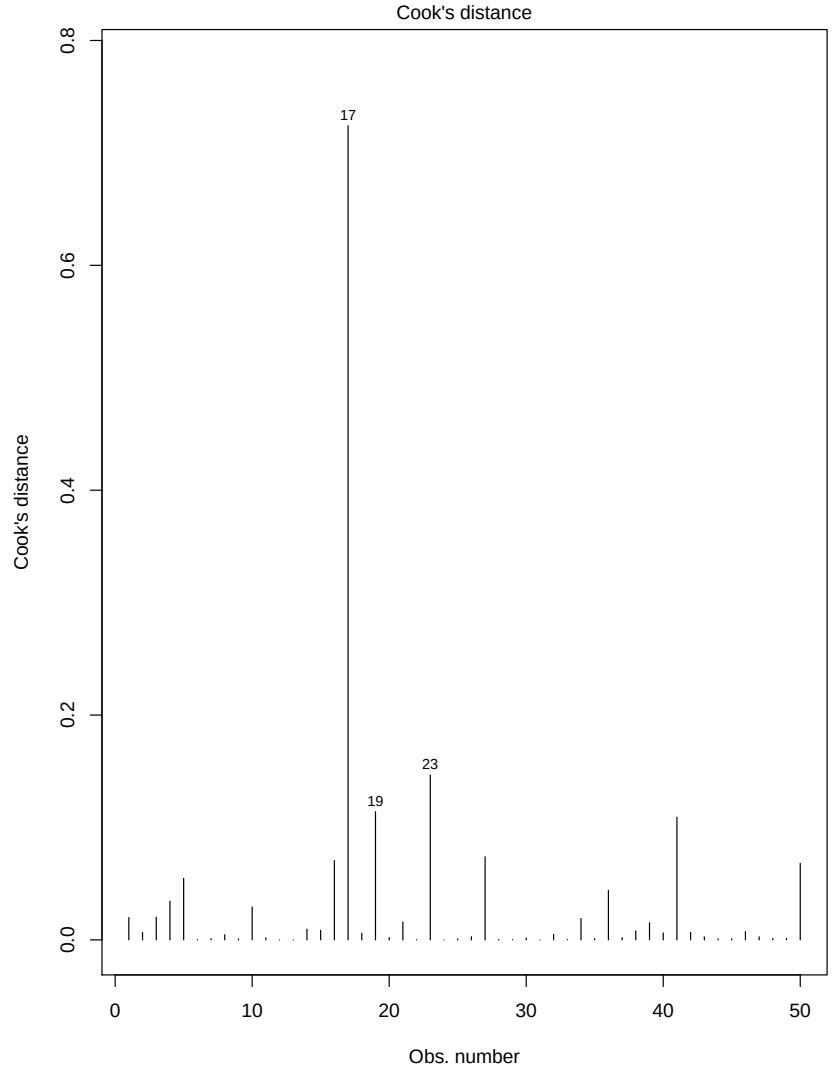
Relation non-linéaire due Y

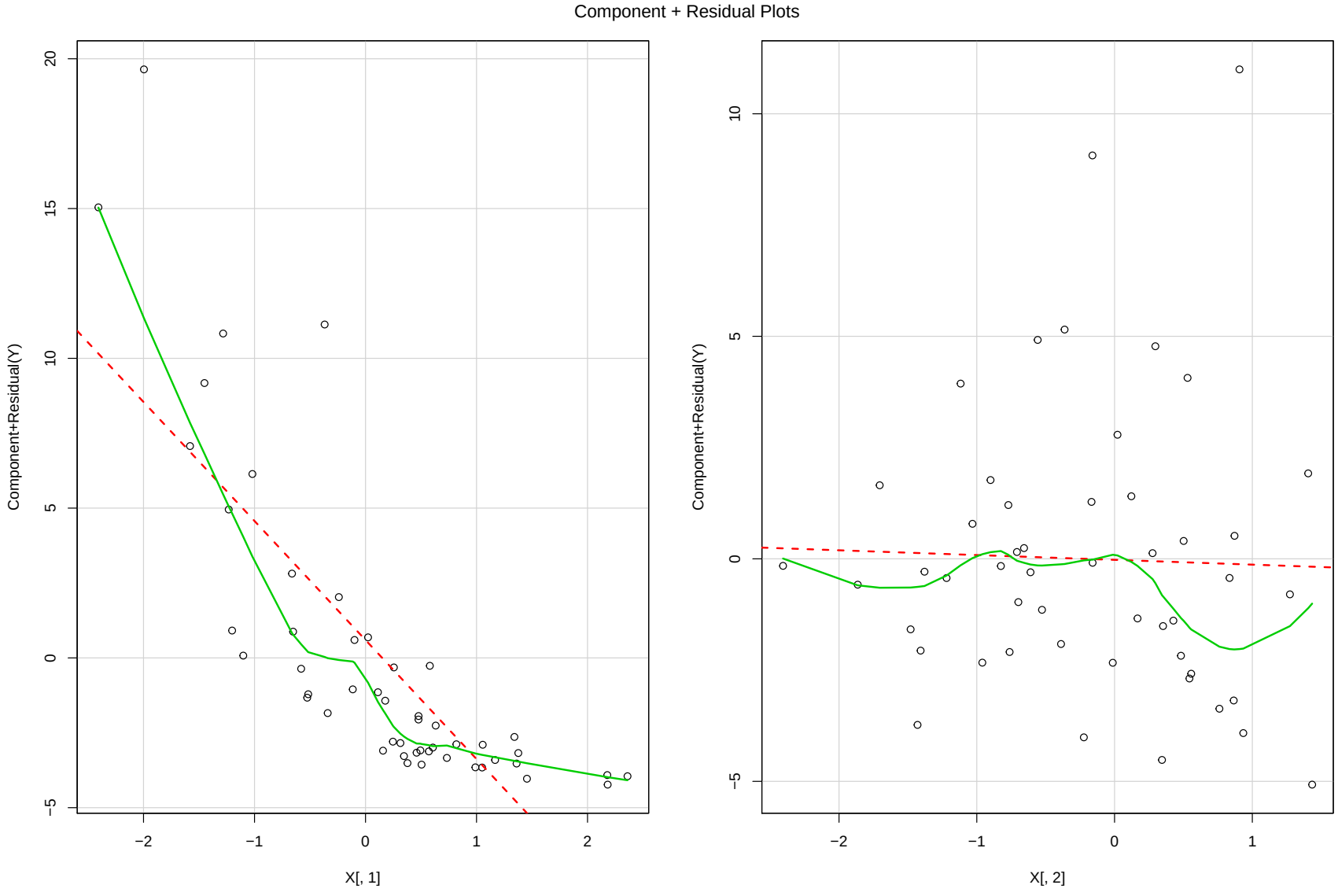
```
n =50
X=matrix(rnorm(n*2),ncol=2)
epsilon = rnorm(n,0,0.5)

ln_Y = 1 - X[,1] + 0.1* X[,2] + epsilon
Y = exp(ln_Y)
lm_ln = lm(Y~X[,1]+X[,2])
```


Residuals vs Fitted
Residuals
Fitted values
17
50
5
−5
0
5
10
15


Normal Q−Q
Standardized residuals
Theoretical Quantiles
17
50
5
−2
−1
0
1
2
3
4

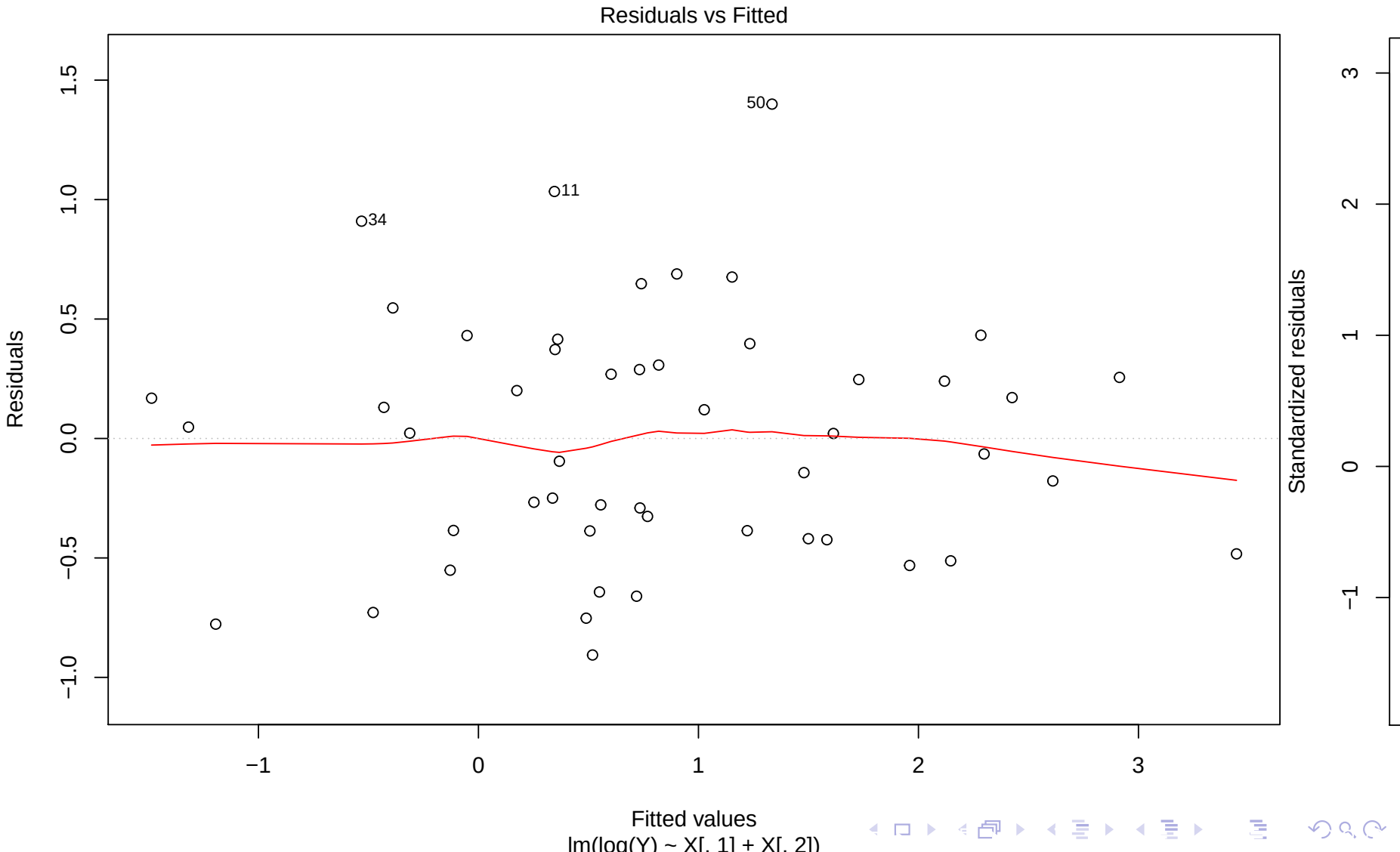

Scale−Location
√|Standardized residuals|
Fitted values
17
50
50
0.0
0.5
1.0
1.5
−5
0
5
10
15


Cook's distance
Cook's distance
Obs. number
17
23
19
0.0
0.2
0.4
0.6
0.8
0
10
20
30
40
50

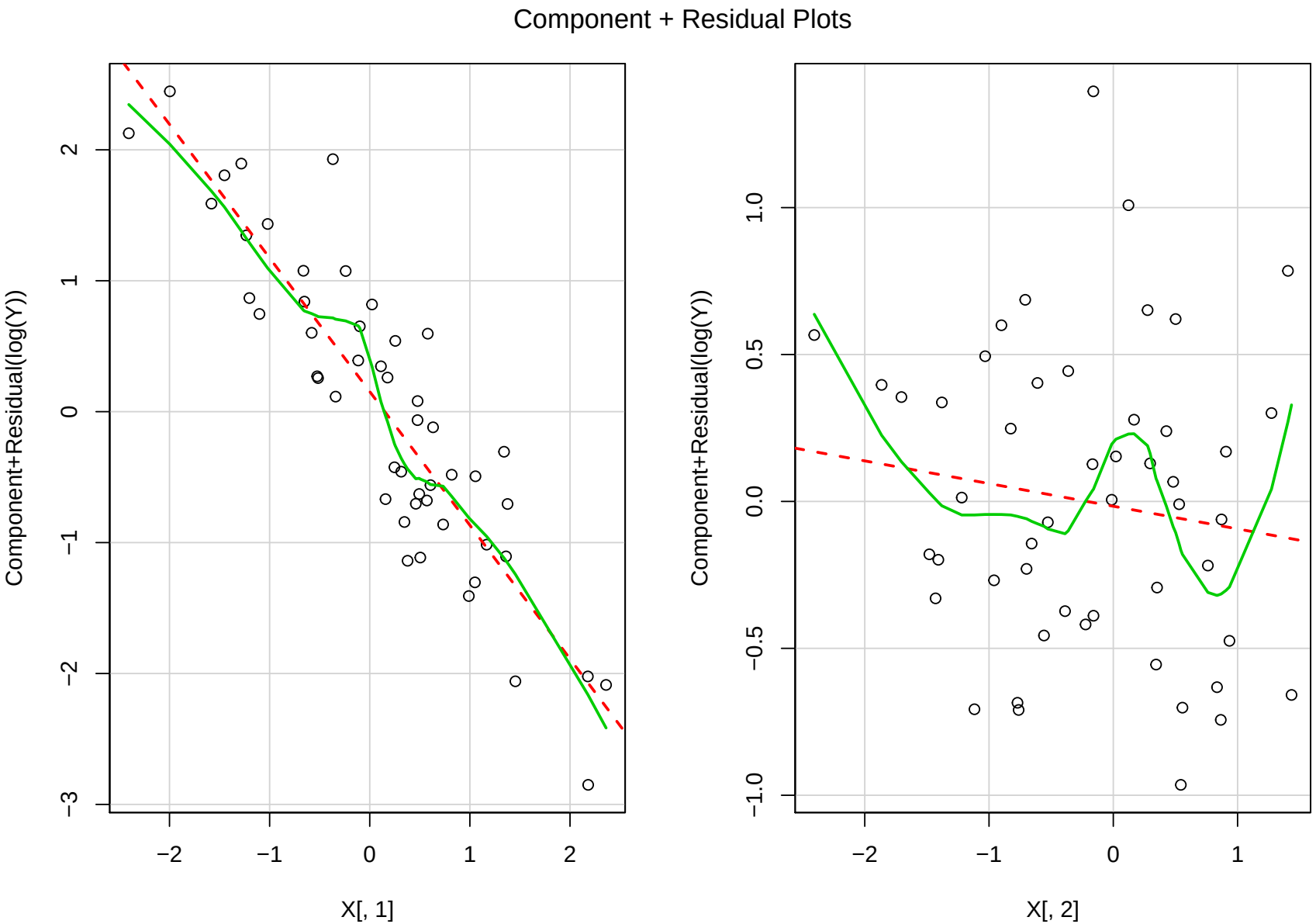

Component + Residual Plots
Component+Residual(Y)
X[, 1]
X[, 2]

Transformation de Y

```
lm_ln_trans = lm(log(Y)~X[,1]+X[,2])
plot(lm_ln_trans)
```



```
crPlots(lm_ln_trans)
```



##

A cette étape, on doit avoir un jeu de données propre pour le modèle linéaire :

- relations linéaires entre variables explicatives et variable à expliquer
- matrice X de plein rang
- résidus normaux
- pas d'observation aberrante ou trop influente

Il reste à sélectionner un modèle et à l'interpréter !