Modèle linéaire et extension
Régression linéaire multiple

M1 Math et Interactions – UEVE/ENSIIE

semestre d'automne 2016

`http://julien.cremeriefamily.info/teachings_M1MINT_Reg.html`

Laboratoire de Mathématiques et Modélisation d'Évry
LaMME

université evry val-d'essonne

Recommandations bibliographiques

- The Element of Statistical Learning: chapitre 2,
  T. Hastie, R. Tibshirani, J. Friedman.
  http://statweb.stanford.edu/~tibs/ElemStatLearn/
- Résumé du cours de modèle de régression, Y. Tillé
  https://www2.unine.ch/files/content/sites/statistics/files/shared/documents/cours_modeles_regression.pdf
- Bases du modèle linéaire, J.-J. Daudin, S. Robin, C. Vuillet
  http://moulon.inra.fr/~mag/modelstat/ModLin_2007.pdf
- Exemples d'applications du modèle linéaire, É. Lebarbier, S. Robin
  https://www.agroparistech.fr/IMG/pdf/ExemplesModeleLineaire-AgroParisTech.pdf

Plan

Modèle

Prérequis (rappels!)

Estimation

Analyse de la variance

Diagnostic

Un exemple: les processionaires de pins

Sélection de variables

Plan

Régression multiple
Objectif général I

Idée/Principe

Expliquer les variations

- d'une variable **quantitative** Y,
- par **plusieurs** variables quantitatives $x = (x_1, x_2, \ldots, x_p)$

Vocabulaire

- Y est la variable **réponse**, **à expliquer**
- Les x_j sont les variables **explicatives**, **covariables**, **régresseurs** ou **prédicteurs**

Régression multiple
Objectif général II

Exemples

- taux de DDT = f(age du brochet, age du brochet au carré)
- progression du diabète = f(age, indice masse corporelle, tension artérielle, concentration sanguine en diverses protéines)
- action au temps t = f(autres actions du marché à $t-1$)
- rendement d'une plante = f(expression de ses gènes)
- [VIH] à l'inclusion = f(variations du génotype)

$\rightsquigarrow$ potentiellement **beaucoup** de prédicteurs...

Régression linéaire multiple

Le modèle

On suppose que la vraie relation entre Y et x est linéaire:

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \varepsilon,$$

- β_0 est la **constante** (**intercept**)
- β_j sont les **coefficients de régression**
- ε est le **résidu** (variable aléatoire)
  - $\rightsquigarrow$ erreur de mesure, variabilité individuelle, facteur(s) non expliqué(s)

Hypothèses minimales sur le résidu

Centré et de variance finie:

- $\mathbb{E}(\varepsilon) = 0$,
- $\mathbb{V}(\varepsilon) = \sigma^2$.

Régression linéaire multiple

Échantillonnage et écriture matricielle

Collecte de données / échantillonnage aléatoire

Soit $\{(Y_i, x_i)\}_{i=1}^n$ un n-échantillon avec $Y_i \in \mathbb{R}$ et $x_i \in \mathbb{R}^p$. On a

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i,$$

avec $\{\varepsilon_i\}_{i=1}^n$ indépendants, identiquement distribués.

Notations

- $Y = (Y_1, \ldots, Y_n)^\intercal \in \mathbb{R}^n$ le vecteur des v.a. de réponse,
- $\mathbf{y} = (y_1, \ldots, y_n)^\intercal \in \mathbb{R}^n$ le vecteur d'observation de la réponse,
- $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^\intercal$ le vecteur d'observation du j^{e} prédicteur.
- $\boldsymbol{\varepsilon} = (\varepsilon_i, \ldots, \varepsilon_n)^\intercal$ le vecteur des résidus.

Régression linéaire multiple

Écriture matricielle

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i \quad i = 1, \ldots, n$$

$$Y = \mathbf{1}_n \beta_0 + \sum_{j=1}^{p} \beta_j \mathbf{x}_j + \boldsymbol{\varepsilon}$$

$$Y = (\mathbf{1}_n, \mathbf{x}_1, \ldots, \mathbf{x}_p) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \ldots \\ \beta_p \end{pmatrix} + \boldsymbol{\varepsilon} = \underbrace{\begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix}}_{\mathbf{X}, \text{ une matrice } n \times (p+1)} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

En résumé,

$$Y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Régression linéaire multiple

Écriture matricielle

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i \quad i = 1, \ldots, n$$

$$Y = \mathbf{1}_n \beta_0 + \sum_{j=1}^{p} \beta_j \mathbf{x}_j + \boldsymbol{\varepsilon}$$

$$Y = (\mathbf{1}_n, \mathbf{x}_1, \ldots, \mathbf{x}_p) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \ldots \\ \beta_p \end{pmatrix} + \boldsymbol{\varepsilon} = \underbrace{\begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix}}_{\mathbf{X}, \text{ une matrice } n \times (p+1)} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

En résumé,

$$Y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Régression linéaire multiple

Écriture matricielle

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i \quad i = 1, \ldots, n$$

$$Y = \mathbf{1}_n \beta_0 + \sum_{j=1}^{p} \beta_j \mathbf{x}_j + \boldsymbol{\varepsilon}$$

$$Y = (\mathbf{1}_n, \mathbf{x}_1, \ldots, \mathbf{x}_p) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \ldots \\ \beta_p \end{pmatrix} + \boldsymbol{\varepsilon} = \underbrace{\begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix}}_{\mathbf{X}, \text{ une matrice } n \times (p+1)} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

En résumé,

$$Y = \mathbf{X}\beta + \varepsilon.$$

Régression linéaire multiple

Écriture matricielle

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i \quad i = 1, \ldots, n$$

$$Y = \mathbf{1}_n \beta_0 + \sum_{j=1}^{p} \beta_j \mathbf{x}_j + \boldsymbol{\varepsilon}$$

$$Y = (\mathbf{1}_n, \mathbf{x}_1, \ldots, \mathbf{x}_p) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \ldots \\ \beta_p \end{pmatrix} + \boldsymbol{\varepsilon} = \underbrace{\begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix}}_{\mathbf{X}, \text{ une matrice } n \times (p+1)} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

En résumé,

$$Y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Régression linéaire multiple I

Linéarité en les paramètres

Le modèle est **linéaire en ses paramètres** (pas nécessairement en les x_j)

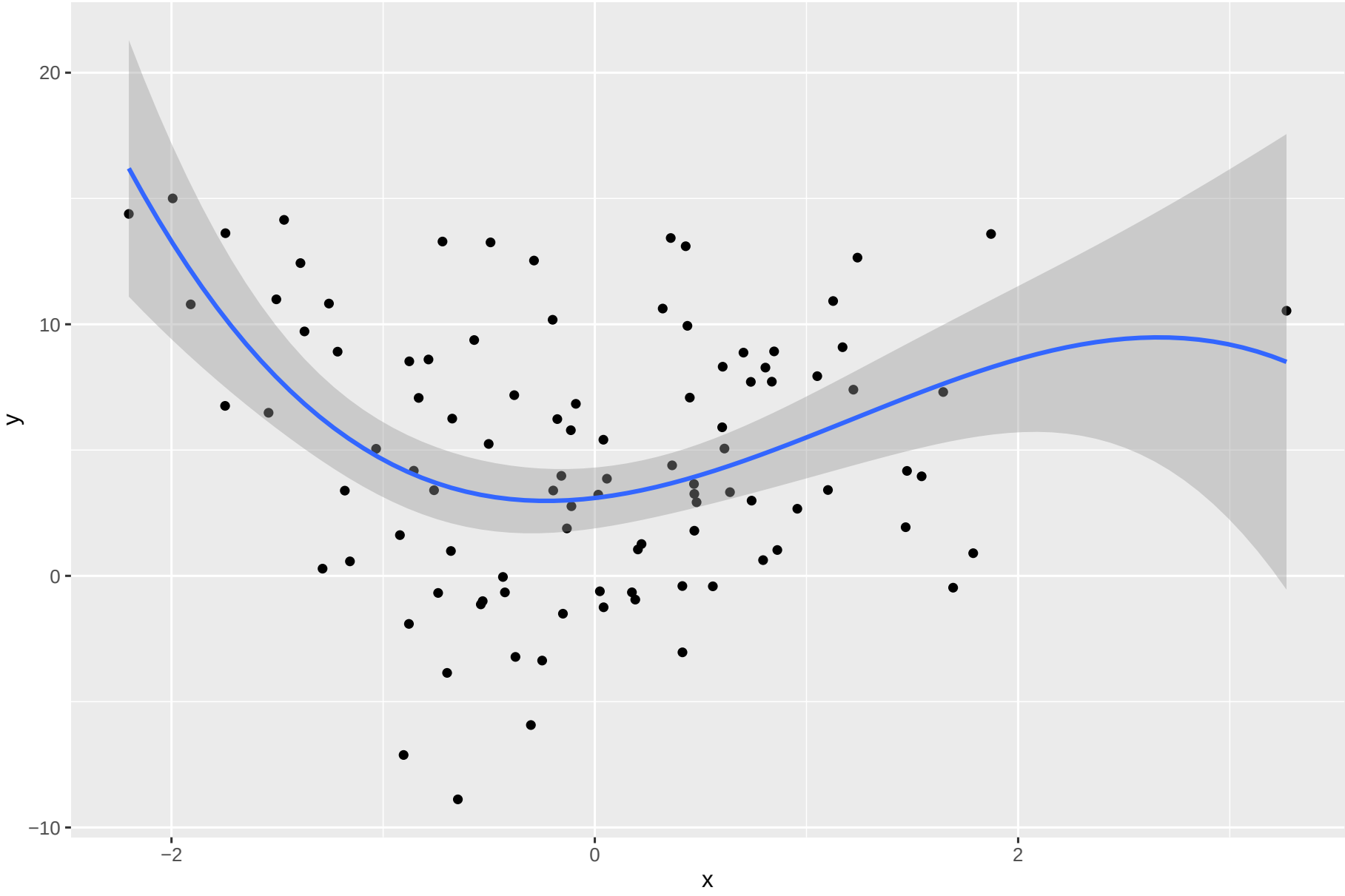
Exemple: régression sur base polynomiale

Un modèle de régression linéaire multiple représentable en 2D

```
## vrais paramètres: polynôme d'ordre 3
beta <- c(3, 1, 2, -1)
sigma <- 5
p <- length(beta)

## simulation des observations
n <- 100
x <- rnorm(n)
X <- cbind(1, x, x^2, x^3)
epsilon <- rnorm(n,0,sigma)
y <- X %*% beta + epsilon

ggplot(data.frame(x=x,y=y), aes(x,y)) + geom_point() +
    geom_smooth(method="lm", formula=y~poly(x,3))
```

Régression linéaire multiple II

Linéarité en les paramètres



Régression linéaire multiple
En résumé

Objectifs statistiques

1. Estimer les paramètres $\boldsymbol{\beta}$ et σ^2
2. Tester la nullité des paramètres $\{\beta_j\}_{j=1}^p$, i.e. l'influence de chacune des variables
3. Prédire Y_0 pour une nouvelle observation x_0
4. Tester la pertinence générale du modèle
5. Si p est grand, contrôler la complexité du modèle

Plan

Plan

Sous espaces orthogonaux

Définition (sous espaces vectoriels orthogonaux)

- *Les sous espaces V et W sont orthogonaux si tous les vecteurs de V sont orthogonaux à tous les vecteurs de W.*
- *L'ensemble de tous les vecteurs orthogonaux à V est appelé l'orthogonal de V et est noté $V^{\perp}$.*

Théorème

Soit V un sous-espace vectoriel de $\mathbb{R}^n$, alors tout vecteur de $\mathbb{R}^n$ se décompose de manière unique en une somme de vecteurs de V et de $V^{\perp}$.

Projection orthogonale

Définition (Projection orthogonale)

Soit V un sous espace de $\mathbb{R}^n$, l'application linéaire qui à un vecteur $\mathbf{u} \in \mathbb{R}^n$ fait correspondre un vecteur $\mathbf{u}^\star \in V$ tel que $\mathbf{u} - \mathbf{u}^\star$ appartienne à $V^\perp$ est appelée projection orthogonale de $\mathbf{u}$ dans V.

Définition (Projecteur orthogonal et matrice)

Soit $\mathbf{X}$ une matrice $n \times p$ de plein rang telle que $n < p$.

La projection orthogonale de $\mathbf{u} \in \mathbb{R}^n$ dans l'image de $\mathbf{X}$ vaut

$$proj_{\mathbf{X}}(\mathbf{u}) = \underbrace{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'}_{\mathbf{P}_{\mathbf{X}}} \mathbf{u}$$

La projection orthogonale de $\mathbf{u} \in \mathbb{R}^n$ dans le noyau de $\mathbf{X}'$ vaut

$$proj_{\mathbf{X}}(\mathbf{u}) = \underbrace{\left(\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)}_{\mathbf{P}_{\mathbf{X}}^\perp} \mathbf{u}$$

Projection orthogonale

Définition (Projection orthogonale)

Soit V un sous espace de $\mathbb{R}^n$, l'application linéaire qui à un vecteur $\mathbf{u} \in \mathbb{R}^n$ fait correspondre un vecteur $\mathbf{u}^\star \in V$ tel que $\mathbf{u} - \mathbf{u}^\star$ appartienne à $V^\perp$ est appelée projection orthogonale de $\mathbf{u}$ dans V.

Définition (Projecteur orthogonal et matrice)

Soit $\mathbf{X}$ une matrice $n \times p$ de plein rang telle que $n < p$.

- *La projection orthogonale de $\mathbf{u} \in \mathbb{R}^n$ dans l'image de $\mathbf{X}$ vaut*

$$proj_{\mathbf{X}}(\mathbf{u}) = \underbrace{\mathbf{X}\left(\mathbf{X}^\intercal\mathbf{X}\right)^{-1}\mathbf{X}^\intercal}_{\mathbf{P_X}}\ \mathbf{u}.$$

- *La projection orthogonale de $\mathbf{u} \in \mathbb{R}^n$ dans le noyau de $\mathbf{X}$ vaut*

$$proj^{\perp}_{\mathbf{X}}(\mathbf{u}) = \underbrace{\left(\mathbf{I} - \mathbf{X}\left(\mathbf{X}^\intercal\mathbf{X}\right)^{-1}\mathbf{X}^\intercal\right)}_{\mathbf{I}-\mathbf{P_X}}\ \mathbf{u}.$$

Projection orthogonale

Définition (Projection orthogonale)

Soit V un sous espace de $\mathbb{R}^n$, l'application linéaire qui à un vecteur $\mathbf{u} \in \mathbb{R}^n$ fait correspondre un vecteur $\mathbf{u}^\star \in V$ tel que $\mathbf{u} - \mathbf{u}^\star$ appartienne à $V^\perp$ est appelée projection orthogonale de $\mathbf{u}$ dans V.

Définition (Projecteur orthogonal et matrice)

Soit $\mathbf{X}$ une matrice $n \times p$ de plein rang telle que $n < p$.

- *La projection orthogonale de $\mathbf{u} \in \mathbb{R}^n$ dans l'image de $\mathbf{X}$ vaut*

$$proj_{\mathbf{X}}(\mathbf{u}) = \underbrace{\mathbf{X}\left(\mathbf{X}^\intercal\mathbf{X}\right)^{-1}\mathbf{X}^\intercal}_{\mathbf{P_X}}\ \mathbf{u}.$$

- *La projection orthogonale de $\mathbf{u} \in \mathbb{R}^n$ dans le noyau de $\mathbf{X}$ vaut*

$$proj^{\perp}_{\mathbf{X}}(\mathbf{u}) = \underbrace{\left(\mathbf{I} - \mathbf{X}\left(\mathbf{X}^\intercal\mathbf{X}\right)^{-1}\mathbf{X}^\intercal\right)}_{\mathbf{I}-\mathbf{P_X}}\ \mathbf{u}.$$

Plan

Gradient

Définition (gradient)

Soit f une application de $\mathbb{R}^p$ dans $\mathbb{R}$. On appelle gradient de f le vecteur des dérivées partielles

$$\nabla f(\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}} f(\mathbf{x}) = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_p} \right)^\intercal .$$

De cette définition, on déduit en particulier la dérivée par rapport à un vecteur d'une forme linéaire, d'une application linéaire et d'une forme quadratique.

Dérivée par rapport à un vecteur

Proposition (dérivée par rapport à un vecteur)

Soit $\mathbf{u}, \mathbf{x} \in \mathbb{R}^p$ *et* $\mathbf{A} \in \mathcal{M}_{mp}$ *et* $\mathbf{S} \in \mathcal{M}_{pp}$.

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{u}^\intercal \mathbf{x} = \mathbf{u}$$
$$\frac{\partial}{\partial \mathbf{x}} \mathbf{A}\mathbf{x} = \mathbf{A}$$
$$\frac{\partial}{\partial \mathbf{x}} \mathbf{x}^\intercal \mathbf{S}\mathbf{x} = \mathbf{S}\mathbf{x} + \mathbf{S}^\intercal \mathbf{x}$$

Si de plus $\mathbf{S}$ *est symétrique, alors*

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{x}^\intercal \mathbf{S}\mathbf{x} = 2\mathbf{S}\mathbf{x}$$

Plan

Vecteur aléatoire, espérance et variance-covariance

Soit $X = (X_1, \ldots, X_p)^\intercal$ un vecteur de variables aléatoires dont la distribution est définie par la densité jointe $f(\mathbf{x}) = f(x_1, \ldots, x_p)$.

Définition (Espérance)

L'espérance de X est le vecteur d'espérance de chaque composant:

$$\mathbb{E}X = (\mathbb{E}(X_1), \ldots, \mathbb{E}(X_p))^\intercal .$$

Définition (Variance)

La variance de X est la matrice (de variance-covariance) définie par

$$\mathbb{V}(X) = \mathbb{E}\left[(X - \mathbb{E}X)(X - \mathbb{E}X)^\intercal\right]$$

Propriétés

Soit $\mathbf{A}$ une matrice $m \times p$ de constantes, alors

$$\mathbb{E}(\mathbf{A}X) = \mathbf{A}\mathbb{E}(X), \quad \mathbb{V}(\mathbf{A}X) = \mathbf{A}\mathbb{V}(X)\mathbf{A}^\intercal$$

Vecteur aléatoire, espérance et variance-covariance

Soit $X = (X_1, \ldots, X_p)^\intercal$ un vecteur de variables aléatoires dont la distribution est définie par la densité jointe $f(\mathbf{x}) = f(x_1, \ldots, x_p)$.

Définition (Espérance)

L'espérance de X est le vecteur d'espérance de chaque composant:

$$\mathbb{E}X = (\mathbb{E}(X_1), \ldots, \mathbb{E}(X_p))^\intercal .$$

Définition (Variance)

La variance de X est la matrice (de variance-covariance) définie par

$$\mathbb{V}(X) = \begin{pmatrix} \mathbb{V}(X_1) & \ldots & \mathrm{cov}(X_1, X_j) & \ldots & \mathrm{cov}(X_1, X_p) \\ \vdots & & \vdots & & \vdots \\ \mathrm{cov}(X_1, X_j) & \ldots & \mathbb{V}(X_j) & \ldots & \mathrm{cov}(X_j, X_p) \\ \vdots & & \vdots & & \vdots \\ \mathrm{cov}(X_1, X_p) & \ldots & \mathrm{cov}(X_j, X_p) & \ldots & \mathbb{V}(X_p) \end{pmatrix}$$

Vecteur aléatoire, espérance et variance-covariance

Soit $X = (X_1, \ldots, X_p)^\intercal$ un vecteur de variables aléatoires dont la distribution est définie par la densité jointe $f(\mathbf{x}) = f(x_1, \ldots, x_p)$.

Définition (Espérance)

L'espérance de X est le vecteur d'espérance de chaque composant:

$$\mathbb{E}X = (\mathbb{E}(X_1), \ldots, \mathbb{E}(X_p))^\intercal .$$

Définition (Variance)

La variance de X est la matrice (de variance-covariance) définie par

$$\mathbb{V}(X) = \mathbb{E}\left[(X - \mathbb{E}X)(X - \mathbb{E}X)^\intercal\right]$$

Propriétés

Soit $\mathbf{A}$ une matrice $m \times p$ de constantes, alors

$$\mathbb{E}(\mathbf{A}X) = \mathbf{A}\mathbb{E}(X), \quad \mathbb{V}(\mathbf{A}X) = \mathbf{A}\mathbb{V}(X)\mathbf{A}^\intercal$$

Vecteur gaussien

Définition

Le vecteur $X \in \mathbb{R}^p$ suit une distribution normale multivariée de moyenne $\boldsymbol{\mu}$ et de variance $\boldsymbol{\Sigma}$ si la fonction densité d'une réalisation de $\mathbf{x}$ est données par

$$f(\mathbf{x}) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}.$$

On note $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ un vecteur gaussien de $\mathbb{R}^p$.

Log-vraisemblance

Soit $\mathbf{X}$ la matrice $n \times p$ dont les lignes, notées $\mathbf{x}_i$, sont des réalisation indépendantes de X.

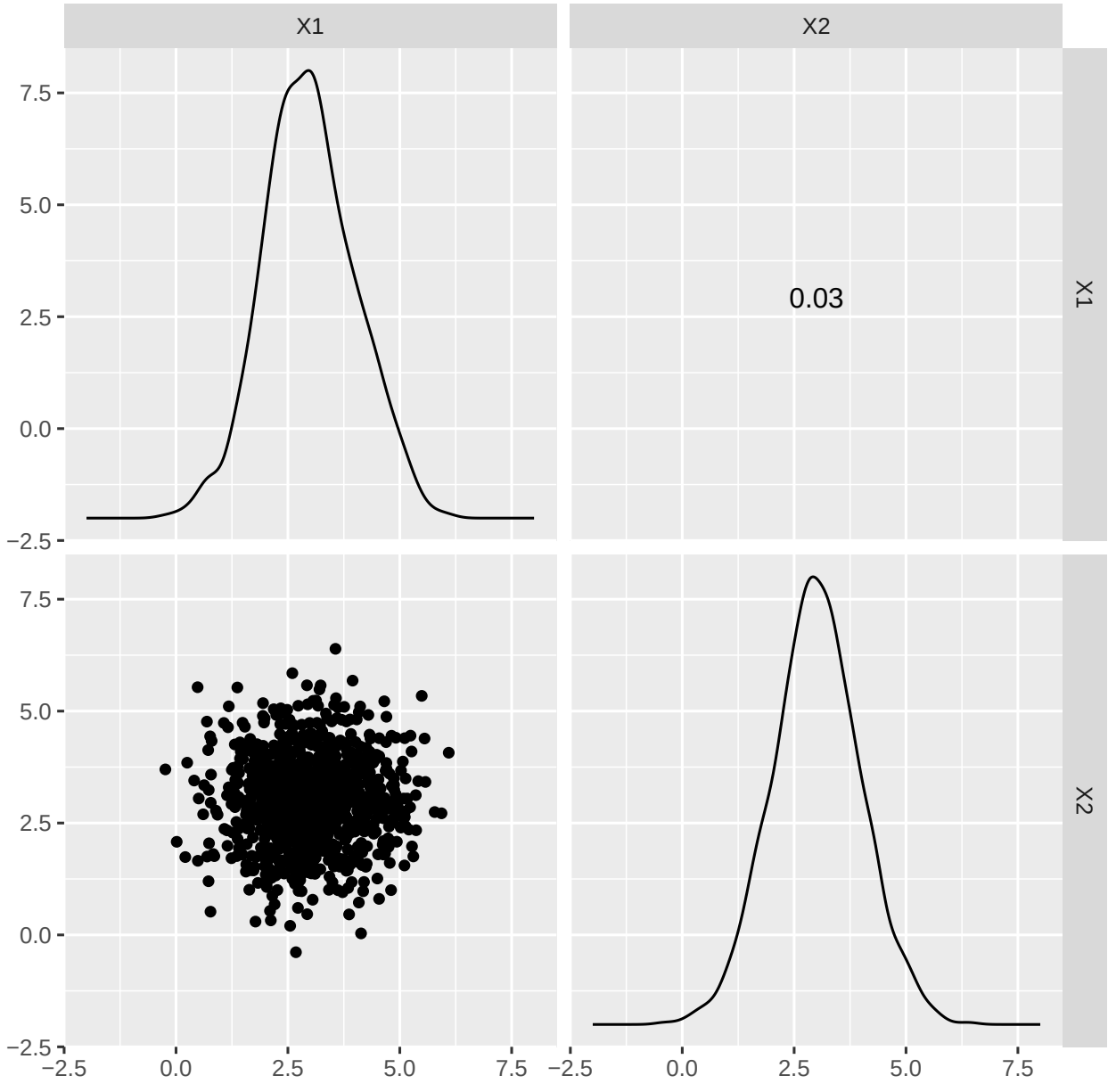
$$\log L(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X}) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})$$

Vecteur gaussien
Exemples bivariés

```
library(mvtnorm)
mu <- c(3,3)
Sigma.id   <- matrix(c(1,0,0,1), 2, 2)
Sigma.diag <- matrix(c(.5,0,0,5), 2, 2)
Sigma.cov1 <- matrix(c(1,0.5,0.5,1), 2, 2)
Sigma.cov2 <- matrix(c(.5,-0.75,-0.75,3), 2, 2)

X.id    <- rmvnorm(1000,mu,Sigma.id)
X.diag  <- rmvnorm(1000,mu,Sigma.diag)
X.cov1 <- rmvnorm(1000,mu,Sigma.cov1)
X.cov2 <- rmvnorm(1000,mu,Sigma.cov2)
```

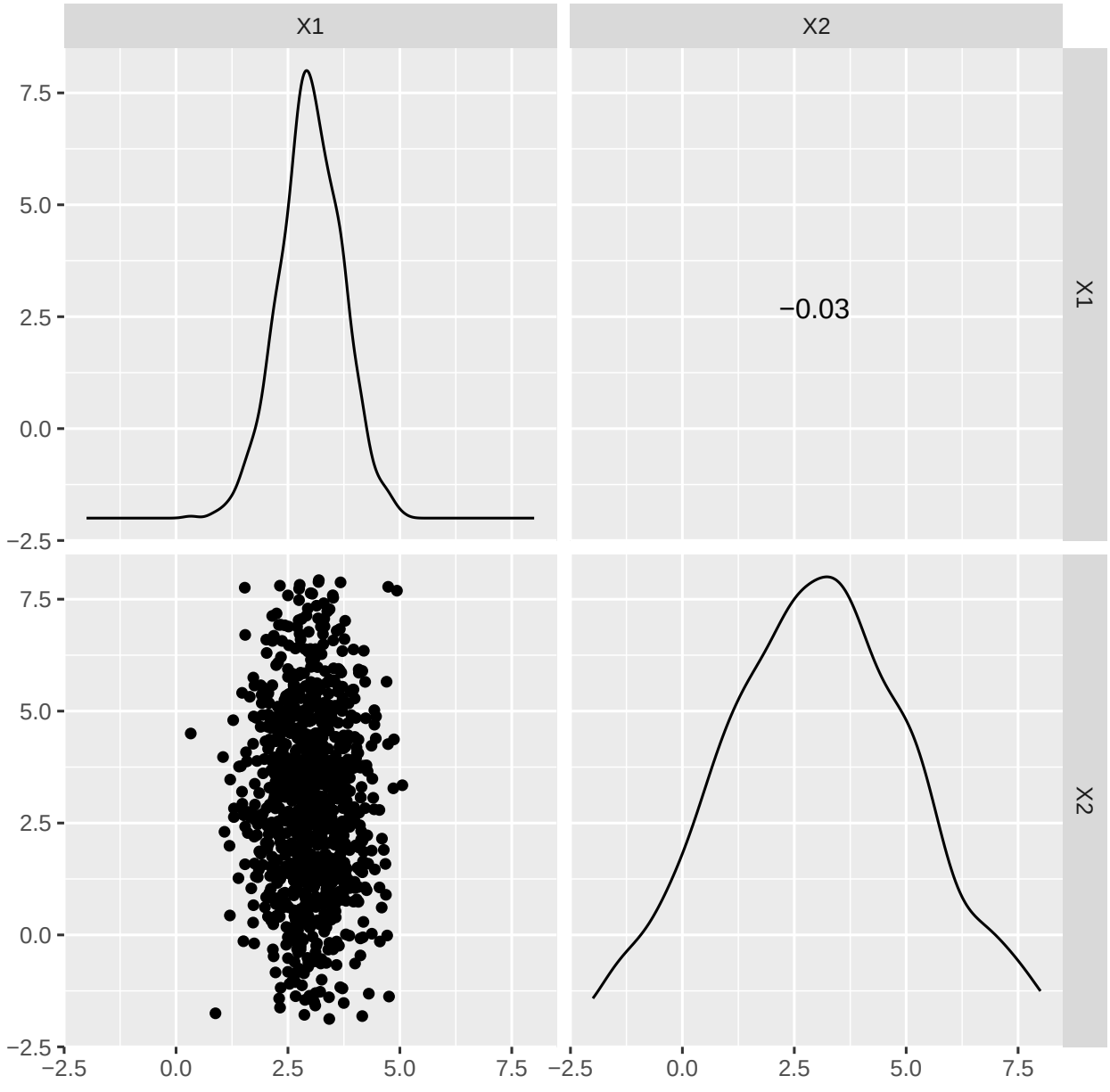
Vecteur gaussien

Exemples bivariés (I)



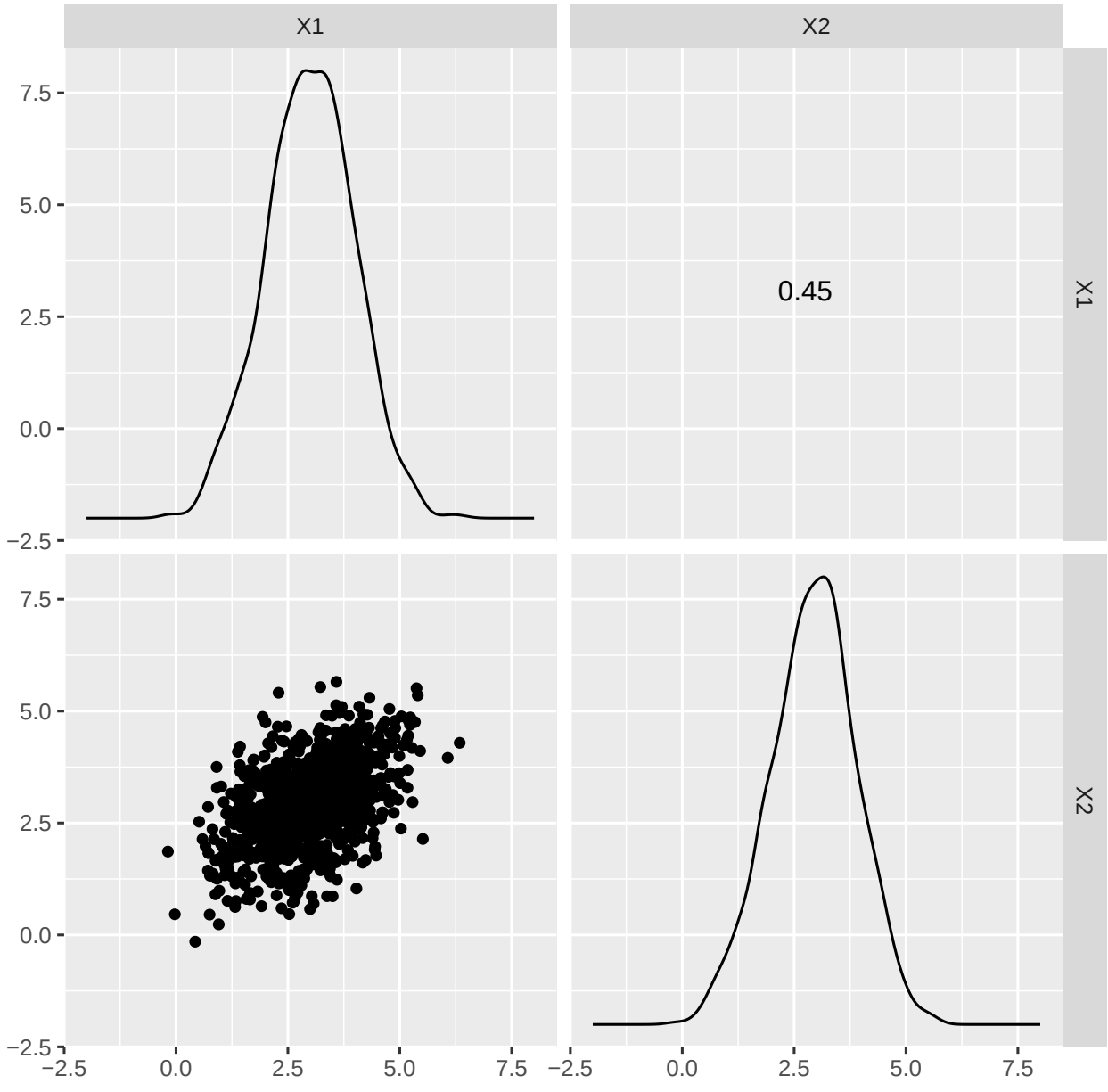
Vecteur gaussien

Exemples bivariés (II)



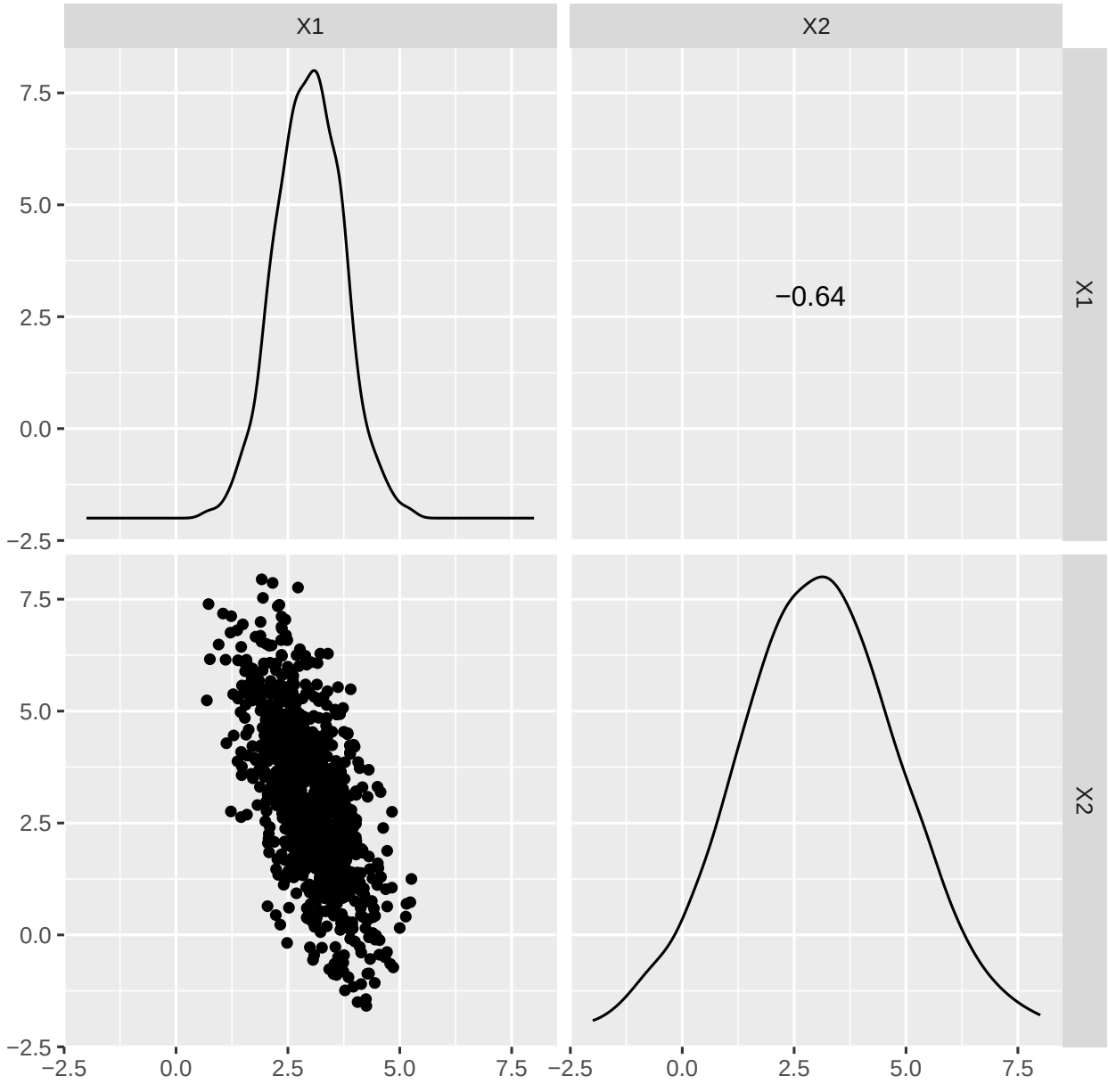
Vecteur gaussien

Exemples bivariés (III)



Vecteur gaussien

Exemples bivariés (IV)



Plan

Plan

Moindres carrés ordinaires
Intuition

- La **"vraie"** "droite" de $\mathbb{R}^{p+1}$ (un hyperplan) passe au plus près des points de la **population**.
- On cherche l'hyperplan passant **au plus près** des point de l'**échantillon**

Moindres carrés ordinaires

Intuition (II)

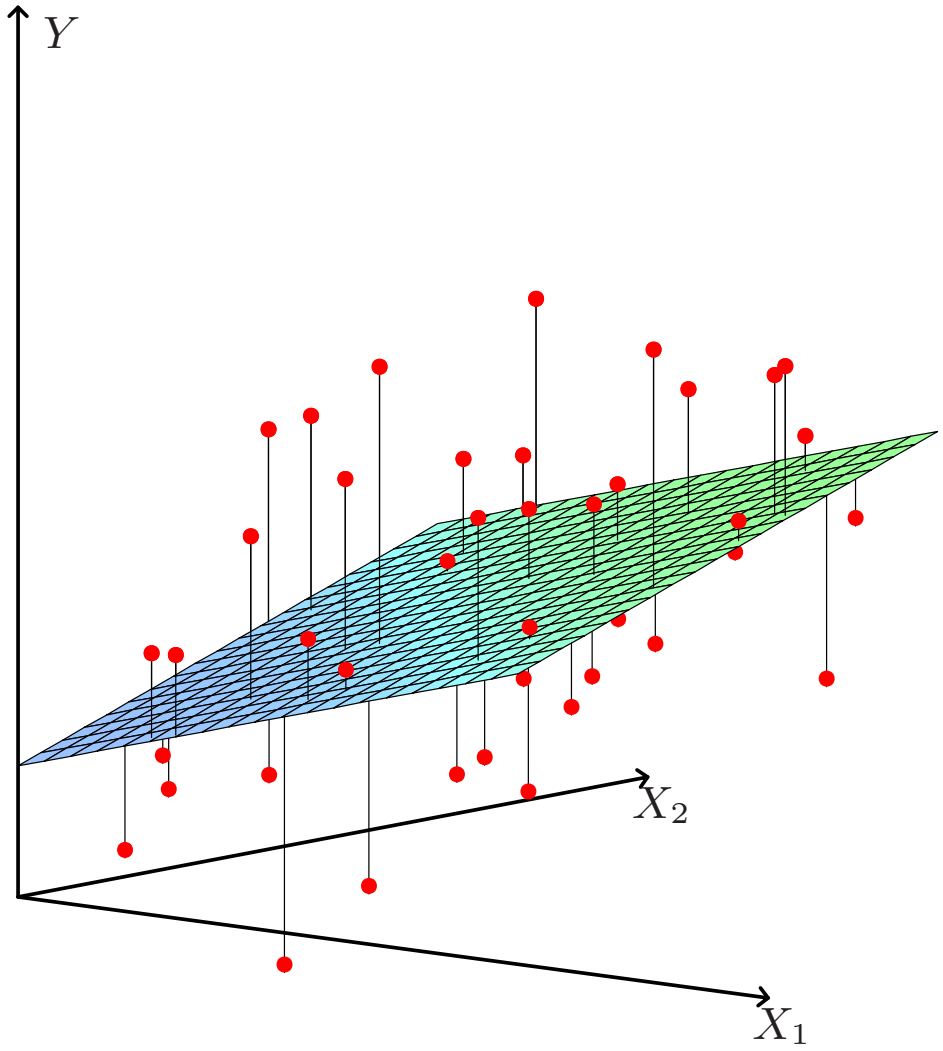


Figure: OLS: géométrie dans l'espace des variables $\mathbb{R}^{p+1}$

Moindres carrés ordinaires

Le critère

Formalisation

Trouver le plan de $\mathbb{R}^{p+1}$ de la forme

$$\beta_1 x_1 + \cdots + \beta_p x_p - y_i + \beta_0 = 0$$

telle que la distance à l'ensemble des points soit la plus petite possible.

Estimateurs OLS

Les valeurs estimées $\{\beta_j, j = 0, \ldots, p\}$ par OLS vérifient

$$(\hat{\beta}_0^{\text{ols}}, \hat{\beta}_j^{\text{ols}}) = \arg\min_{\beta_0, \beta_j \in \mathbb{R}} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij} \beta_j - \beta_0 \right)^2 \right\}.$$

Moindres carrés ordinaires

Le critère

Formalisation

Trouver le plan de $\mathbb{R}^{p+1}$ de la forme

$$\beta_1 x_1 + \cdots + \beta_p x_p - y_i + \beta_0 = 0$$

telle que la distance à l'ensemble des points soit la plus petite possible.

Estimateurs OLS

Les valeurs estimées $\{\beta_j, j = 0, \ldots, p\}$ par OLS vérifient

$$(\hat{\beta}_0^{\text{ols}}, \hat{\beta}_j^{\text{ols}}) = \underset{\beta_0, \beta_j \in \mathbb{R}}{\arg\min} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij} \beta_j - \beta_0 \right)^2 \right\}.$$

Moindres carrés ordinaires

Espace des observations (I)

Soit $\mathbf{X}_{i\cdot} = (\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_p)$ la i^{e} ligne de $\mathbf{X}$. Alors la valeur estimée est

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{\text{ols}} &= \underset{\beta_0, \beta_j \in \mathbb{R}}{\arg\min} \sum_{i=1}^{n} (y_i - \mathbf{X}_{i\cdot}\boldsymbol{\beta})^2 \\ &= \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{\arg\min} \left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right\|^2 .\end{aligned}$$

$\rightsquigarrow$ On cherche $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \in \text{vec}(\mathbf{x}_1, \ldots, \mathbf{x}_p)$ minimisant $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$.

Moindres carrés ordinaires

Espace des observations (II)

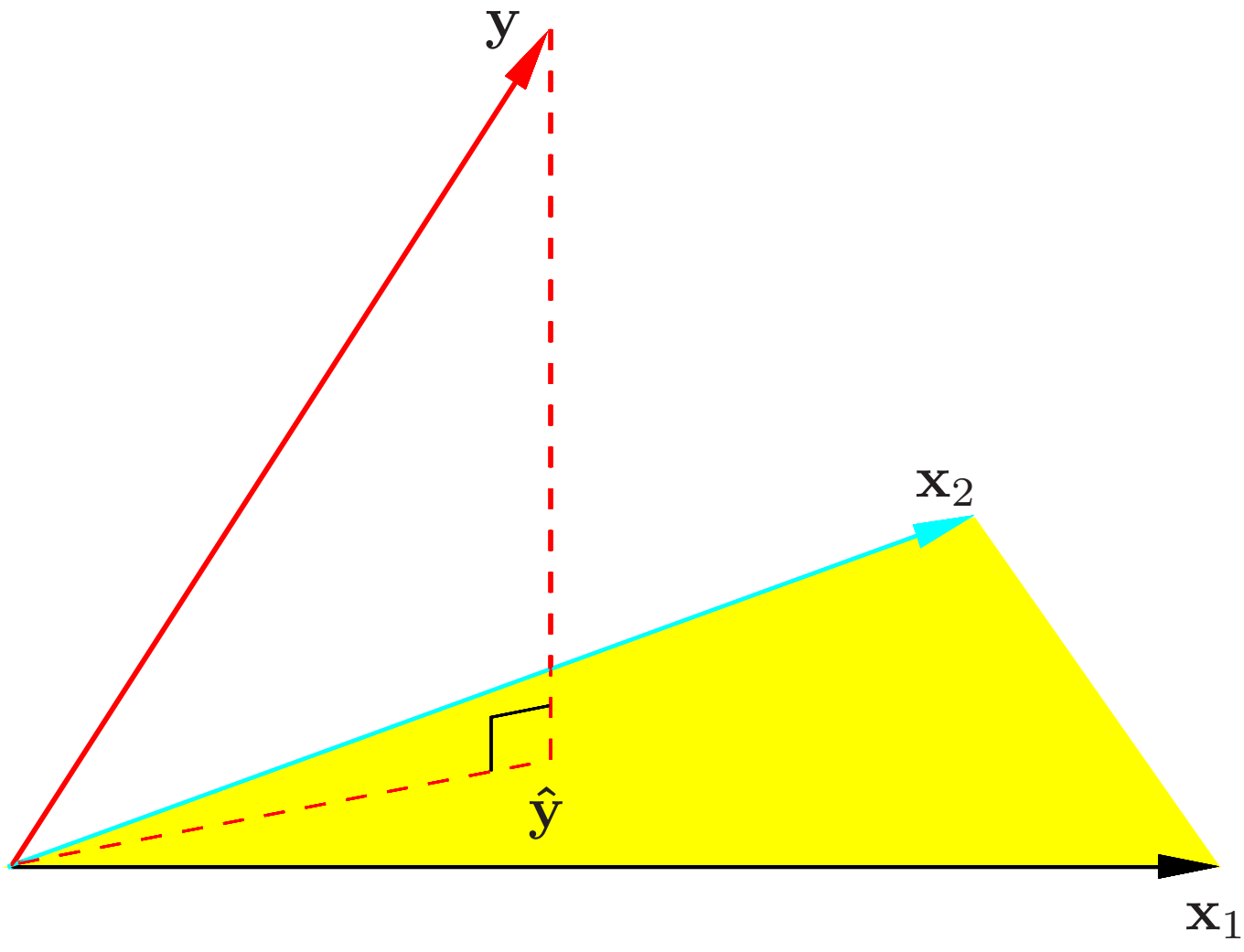


Figure: OLS: géométrie dans l'espace des observations $\mathbb{R}^n$

Moindres carrés ordinaires

Estimateurs

Théorème

L'estimateur des MCO vérifie les **équations normales** :

$$\mathbf{X}^\intercal \mathbf{X} \hat{\boldsymbol{\beta}}^{\text{ols}} = \mathbf{X}^\intercal Y$$

Si $\mathbf{X}^\intercal \mathbf{X}$ est inversible, alors

$$\hat{\boldsymbol{\beta}}^{\text{ols}} = (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal Y$$

Preuve

- montrer que $\hat{\boldsymbol{\beta}}^{\text{ols}}$ est tel que $\mathbf{X}\hat{\boldsymbol{\beta}}^{\text{ols}} = \text{proj}_{\mathbf{X}}(Y)$
- utiliser que $Y - \mathbf{X}\hat{\boldsymbol{\beta}}^{\text{ols}}$ est orthogonal à $\mathbf{x}_j$, pour tout $j = 1, \ldots, p$.

Moindres carrés ordinaires

Estimateurs

Théorème

L'estimateur des MCO vérifie les **équations normales** :

$$\mathbf{X}^\intercal \mathbf{X} \hat{\boldsymbol{\beta}}^{\text{ols}} = \mathbf{X}^\intercal Y$$

Si $\mathbf{X}^\intercal \mathbf{X}$ est inversible, alors

$$\hat{\boldsymbol{\beta}}^{\text{ols}} = (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal Y$$

Preuve

- montrer que $\hat{\boldsymbol{\beta}}^{\text{ols}}$ est tel que $\mathbf{X}\hat{\boldsymbol{\beta}}^{\text{ols}} = \text{proj}_{\mathbf{X}}(Y)$
- utiliser que $Y - \mathbf{X}\hat{\boldsymbol{\beta}}^{\text{ols}}$ est orthogonal à $\mathbf{x}_j$, pour tout $j = 1, \ldots, p$.

Projection orthogonale et matrice chapeau

Projection orthogonale dans l'image de $\mathbf{X}$

Si $\mathbf{X}^\intercal\mathbf{X}$ est inversible, la valeur prédite s'écrit

$$\hat{Y} = \mathbf{X}\hat{\boldsymbol{\beta}}^{\text{ols}} = \mathbf{X}\left(\mathbf{X}^\intercal\mathbf{X}\right)^{-1}\mathbf{X}^\intercal Y = \mathbf{P}_{\mathbf{X}}\, Y.$$

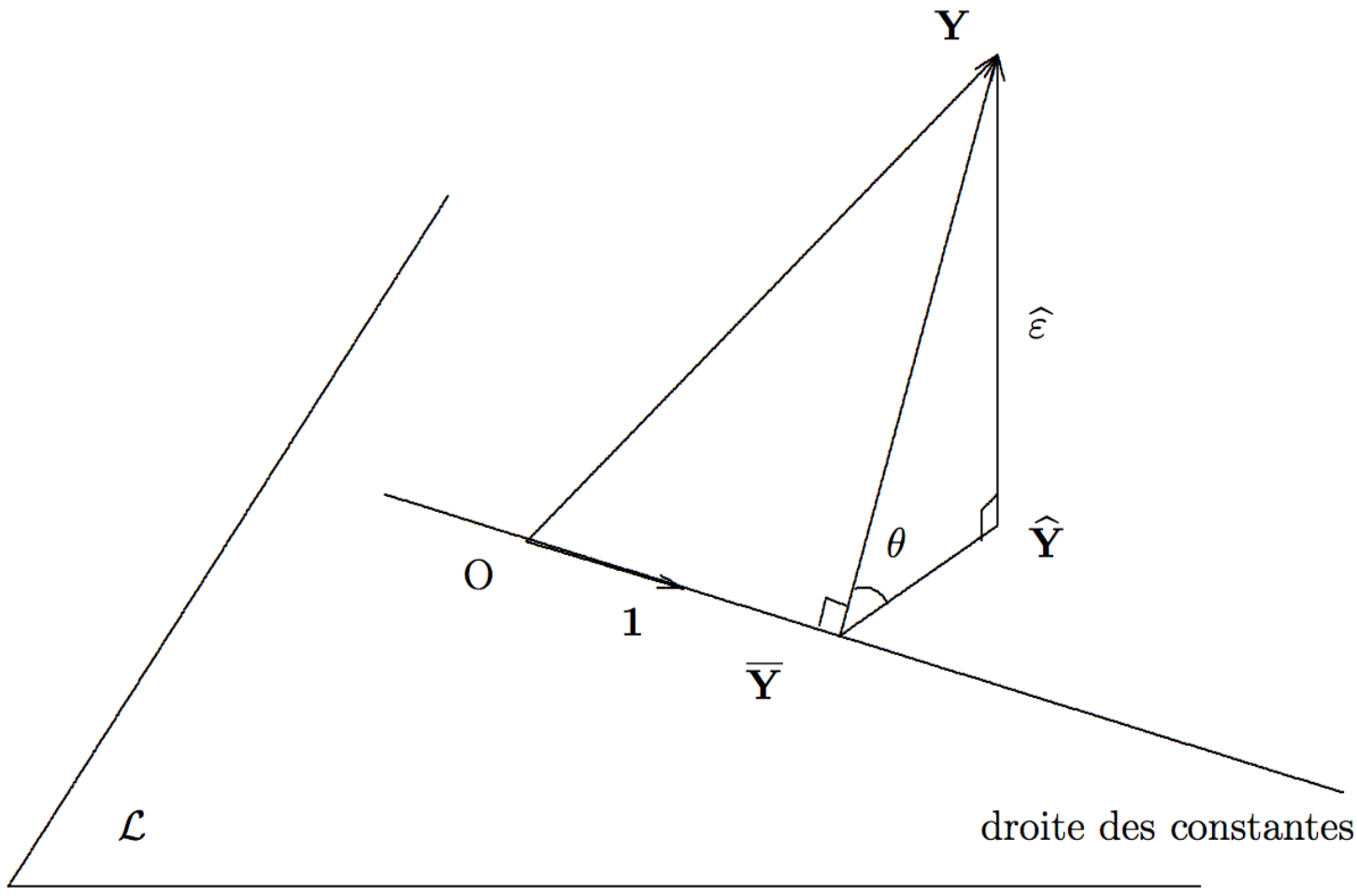
$\mathbf{P}_{\mathbf{X}}$ est parfois notée $\mathbf{H}$ et appelée "hat matrix" (*put a hat on* $\mathbf{y}$).

Projection orthogonale dans le noyau de $\mathbf{X}$

$$\hat{\varepsilon} = Y - \hat{Y} = \left(\mathbf{I} - \mathbf{P}_X\right) Y = \mathbf{P}_X^{\perp}\, Y.$$

$\rightsquigarrow$ les projecteurs $\mathbf{P}_X$ et $\mathbf{P}_X^{\perp}$ sont idempotents. Ils facilitent l'interprétation et les calculs !

Interprétation géométrique de l'OLS



Moindres carrés ordinaires

Propriétés découlant de l'interprétation géométrique

Proposition

Le vecteur des résidus estimés est orthogonal à la droite $\mathbf{1}_n$. On en déduit

$$\hat{\boldsymbol{\varepsilon}} \perp \bar{Y} \Rightarrow \sum_{i=1}^{n} \hat{\varepsilon}_i = 0$$

De plus, $\hat{Y} \perp \hat{\varepsilon}$.

Corollaire

- $\frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}$.
- La projection orthogonale de Y sur $\mathbf{1}_n$ a pour coordonnées $\bar{Y}$:

$$\text{proj}_{\mathbf{1}}(Y) = \mathbf{1}_n(\mathbf{1}_n^\intercal \mathbf{1}_n)^{-1}\mathbf{1}_n^\intercal Y = \mathbf{1}_n \bar{Y}.$$

Moindres carrés ordinaires
Remarques

Méthode purement géométrique

- ne repose pas sur l'hypothèse gaussienne des résidus
- ne dit **rien sur σ^2...**

Condition d'inversibilité de $\mathbf{X}^\intercal\mathbf{X}$

Il faut et il suffit que $\mathbf{X}$ soit de plein rang.

$\rightsquigarrow$ Aucune colonne n'est une combinaison linéaire des autres.

$\rightsquigarrow$ Chaque variable doit apporter "un peu d'information originale".

$\rightsquigarrow$ Les fortes corrélations induisent des instabilités numériques.

Plan

Maximum de vraisemblance

critère

Formalisation

Sous l'hypothèse où $\varepsilon_i \sim \mathcal{N}(0, \sigma)$,

- $Y \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma\mathbf{I}_n)$
- log-vraisemblance : $\log L(\mathbf{y}) = \log f(\mathbf{y})$

Estimateurs du MV

Les valeurs estimées (estimations) de β et σ vérifient

$$(\hat{\boldsymbol{\beta}}^{\text{mv}}, \hat{\sigma}^{\text{mv}}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}, \sigma>0}{\arg\min} \left\{ -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \right\}$$

Maximum de vraisemblance

critère

Formalisation

Sous l'hypothèse où $\varepsilon_i \sim \mathcal{N}(0, \sigma)$,

- $Y \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma\mathbf{I}_n)$
- log-vraisemblance : $\log L(\mathbf{y}) = \log f(\mathbf{y})$

Estimateurs du MV

Les valeurs estimées (estimations) de β et σ vérifient

$$(\hat{\boldsymbol{\beta}}^{\text{mv}}, \hat{\sigma}^{\text{mv}}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}, \sigma>0}{\arg\min} \left\{ -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \right\}$$

Maximum de vraisemblance

critère

Formalisation

Sous l'hypothèse où $\varepsilon_i \sim \mathcal{N}(0, \sigma)$,

- $Y \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma\mathbf{I}_n)$
- log-vraisemblance : $\log L(\mathbf{y}) = \log f(\mathbf{y})$

Estimateurs du MV

Les valeurs estimées (estimations) de $\boldsymbol{\beta}$ et σ vérifient

$$(\hat{\boldsymbol{\beta}}^{\mathsf{mv}}, \hat{\sigma}^{\mathsf{mv}}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}, \sigma>0}{\arg\min} \left\{ -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\left\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right\|^2 \right\}$$

Maximum de vraisemblance

Estimateurs

Théorème

Pour $n > p$, les estimateurs du maximum de vraisemblance sont

$$\hat{\boldsymbol{\beta}}^{\mathrm{mv}} = (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal Y$$

$$\hat{\sigma}^2 = \frac{1}{n} \| Y - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathrm{mv}} \|^2 = \frac{\hat{\varepsilon}^\intercal \hat{\varepsilon}}{n}$$

Preuve:

En annulant les dérivées de la fonction objectif, qui est concave.

Maximum de vraisemblance

Estimation pratique de la variance des résidus

Théorème

Soit $\hat{\varepsilon} = Y - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathrm{mv}} = \mathbf{P}_{\mathbf{X}}^{\perp} Y$, alors

$$\mathbb{E}[\hat{\varepsilon}^{\intercal}\hat{\varepsilon}] = (n - p - 1) \times \sigma^2.$$

Corollaire

Un estimateur non biaisé de la variance résiduelle est donné par

$$\hat{\sigma}^2 = \frac{1}{n - p - 1} \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathrm{mv}}\|^2$$

Vocabulaire

La quantité $n - p - 1$ est le **nombre de degrés de liberté des résidus**, égal au rang de $\mathbf{P}_{\mathbf{X}}^{\perp}$.

Maximum de vraisemblance

Estimation pratique de la variance des résidus

Théorème

Soit $\hat{\varepsilon} = Y - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathrm{mv}} = \mathbf{P}_{\mathbf{X}}^{\perp} Y$, alors

$$\mathbb{E}[\hat{\varepsilon}^{\intercal}\hat{\varepsilon}] = (n - p - 1) \times \sigma^2.$$

Corollaire

Un estimateur non biaisé de la variance résiduelle est donné par

$$\hat{\sigma}^2 = \frac{1}{n - p - 1} \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathrm{mv}}\|^2$$

Vocabulaire

La quantité $n - p - 1$ est le **nombre de degrés de liberté des résidus**, égal au rang de $\mathbf{P}_{\mathbf{X}}^{\perp}$.

Maximum de vraisemblance

Estimation pratique de la variance des résidus

Théorème

Soit $\hat{\varepsilon} = Y - \mathbf{X}\hat{\boldsymbol{\beta}}^{\text{mv}} = \mathbf{P}_{\mathbf{X}}^{\perp} Y$, alors

$$\mathbb{E}[\hat{\varepsilon}^{\intercal}\hat{\varepsilon}] = (n - p - 1) \times \sigma^2.$$

Corollaire

Un estimateur non biaisé de la variance résiduelle est donné par

$$\hat{\sigma}^2 = \frac{1}{n - p - 1} \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\text{mv}}\|^2$$

Vocabulaire

La quantité $n - p - 1$ est le **nombre de degrés de liberté des résidus**, égal au rang de $\mathbf{P}_{\mathbf{X}}^{\perp}$.

Plan

Estimation des paramètres

Propriétés des estimateurs

Cas général

$\hat{\boldsymbol{\beta}}$ sont des estimateurs sans biais de β de variance

$$\mathbb{V}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1}.$$

Cas gaussien

Si les résidus sont gaussien, i.e. $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, alors

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1}\right)$$

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\intercal \frac{\mathbf{X}^\intercal\mathbf{X}}{\sigma^2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \chi^2_{p+1}$$

$$(n-p-1)\hat{\sigma}^2 \sim \sigma^2 \sim \chi^2_{n-p-1}$$

Estimation des paramètres

Propriétés des estimateurs

Cas général

$\hat{\boldsymbol{\beta}}$ sont des estimateurs sans biais de β de variance

$$\mathbb{V}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1}.$$

Cas gaussien

Si les résidus sont gaussien, i.e. $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, alors

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1}\right)$$

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\intercal \frac{\mathbf{X}^\intercal\mathbf{X}}{\sigma^2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \chi^2_{p+1}$$

$$(n-p-1)\hat{\sigma}^2 \sim \sigma^2 \sim \chi^2_{n-p-1}$$

Estimation des paramètres
Propriétés des estimateurs (II)

Théorème de Gauss-Markov

- Cas gaussien: $\hat{\boldsymbol{\beta}}^{\text{ols}}$ est le meilleur estimateur sans biais (i.e. de variance minimale).
- Cas non gaussien: $\hat{\boldsymbol{\beta}}^{\text{ols}}$ est le meilleur estimateur **linéaire** sans biais (i.e. de variance minimale).

$\rightsquigarrow$ On dit que $\hat{\boldsymbol{\beta}}^{\text{ols}}$ est le **BLUE** (best linear unbiased estimator)

Plan

Tests et intervalle de confiance sur les paramètres β_j

Sous hypothèse de normalité des résidus

Hypothèse testée: nullité de β_j

Est-ce que la j^{e} variable apporte une information supplémentaire?

$$\begin{cases} H_0 : & \beta_j = 0 \\ H_1 : & \beta_j \neq 0 \end{cases}$$

Comme $\hat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1}\right)$, on a

Statistique de test et règle de décision

$$T_{\beta_j} = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{[(\mathbf{X}^\intercal\mathbf{X})^{-1}]_{jj}}} \underset{H_0}{\sim} \mathcal{T}_{n-p-1}, \text{ on rejette } H_0 \text{ si } |T_{\beta_j}| \geq t_{n-p-1,1-\frac{\alpha}{2}}$$

Intervalle de confiance sur les $\hat{\beta}_j$

$$IC_{1-\alpha}(\hat{\boldsymbol{\beta}}_j) = \left[\hat{\beta}_j \pm q_{t_{n-p-1},1-\frac{\alpha}{2}}\hat{\sigma}\sqrt{[(\mathbf{X}^\intercal\mathbf{X})^{-1}]_{jj}}\right]$$

Tests et intervalle de confiance sur les paramètres β_j

Sous hypothèse de normalité des résidus

Hypothèse testée: nullité de β_j

Est-ce que la j^{e} variable apporte une information supplémentaire?

$$\begin{cases} H_0 : & \beta_j = 0 \\ H_1 : & \beta_j \neq 0 \end{cases}$$

Comme $\hat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1}\right)$, on a

Statistique de test et règle de décision

$$T_{\beta_j} = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{\left[(\mathbf{X}^\intercal\mathbf{X})^{-1}\right]_{jj}}} \underset{H_0}{\sim} \mathcal{T}_{n-p-1}, \text{ on rejette } H_0 \text{ si } |T_{\beta_j}| \geq t_{n-p-1,1-\frac{\alpha}{2}}$$

Intervalle de confiance sur les $\hat{\beta}_j$

$$IC_{1-\alpha}(\hat{\boldsymbol{\beta}}_j) = \left[\hat{\beta}_j \pm q_{t_{n-p-1},1-\frac{\alpha}{2}}\hat{\sigma}\sqrt{\left[(\mathbf{X}^\intercal\mathbf{X})^{-1}\right]_{jj}}\right]$$

Tests et intervalle de confiance sur les paramètres β_j

Sous hypothèse de normalité des résidus

Hypothèse testée: nullité de β_j

Est-ce que la j^{e} variable apporte une information supplémentaire?

$$\begin{cases} H_0 : & \beta_j = 0 \\ H_1 : & \beta_j \neq 0 \end{cases}$$

Comme $\hat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1}\right)$, on a

Statistique de test et règle de décision

$$T_{\beta_j} = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{\left[(\mathbf{X}^\intercal\mathbf{X})^{-1}\right]_{jj}}} \underset{H_0}{\sim} \mathcal{T}_{n-p-1}, \text{ on rejette } H_0 \text{ si } |T_{\beta_j}| \geq t_{n-p-1,1-\frac{\alpha}{2}}$$

Intervalle de confiance sur les $\hat{\beta}_j$

$$IC_{1-\alpha}(\hat{\boldsymbol{\beta}}_j) = \left[\hat{\beta}_j \pm q_{t_{n-p-1},1-\frac{\alpha}{2}}\hat{\sigma}\sqrt{\left[(\mathbf{X}^\intercal\mathbf{X})^{-1}\right]_{jj}}\right]$$

Plan

Résidus et prédiction

Soit $\mathbf{x}_0 \in \mathbb{R}^p$ une nouvelle observation et $\hat{Y}_0 = \mathbf{x}_0 \hat{\boldsymbol{\beta}}$ le prédicteur associé.

Proposition

Soit $\hat{\varepsilon}_0 = Y_0 - \hat{Y}_0$ l'erreur de prévision au nouveau point. On a :

$$\mathbb{E}(\hat{\varepsilon}_0) = 0$$
$$\mathbb{V}(\hat{\varepsilon}_i) = \sigma^2 \left(1 + \mathbf{x}_0 \left(\mathbf{X}^\intercal \mathbf{X}\right)^{-1} \mathbf{x}_0\right)$$

Intervalle de confiance

$$IC_{1-\alpha}(\hat{Y}_0) = \left[\hat{Y}_0 \pm q_{t_{n-p-1}, 1-\frac{\alpha}{2}} \hat{\sigma} \sqrt{\mathbf{x}_0 \left(\mathbf{X}^\intercal \mathbf{X}\right)^{-1} \mathbf{x}_0}\right]$$

Résidus et prédiction

Soit $\mathbf{x}_0 \in \mathbb{R}^p$ une nouvelle observation et $\hat{Y}_0 = \mathbf{x}_0\hat{\boldsymbol{\beta}}$ le prédicteur associé.

Proposition

Soit $\hat{\varepsilon}_0 = Y_0 - \hat{Y}_0$ l'erreur de prévision au nouveau point. On a :

$$\mathbb{E}(\hat{\varepsilon}_0) = 0$$
$$\mathbb{V}(\hat{\varepsilon}_i) = \sigma^2 \left(1 + \mathbf{x}_0 \left(\mathbf{X}^\intercal \mathbf{X}\right)^{-1} \mathbf{x}_0\right)$$

Intervalle de prévision

$$IC_{1-\alpha}(Y_0) = \left[\hat{Y}_0 \pm q_{t_{n-p-1},1-\frac{\alpha}{2}} \hat{\sigma} \sqrt{1 + \mathbf{x}_0 \left(\mathbf{X}^\intercal \mathbf{X}\right)^{-1} \mathbf{x}_0}\right]$$

Plan

Décomposition de la variance

Théorème fondamental (Pythagore)

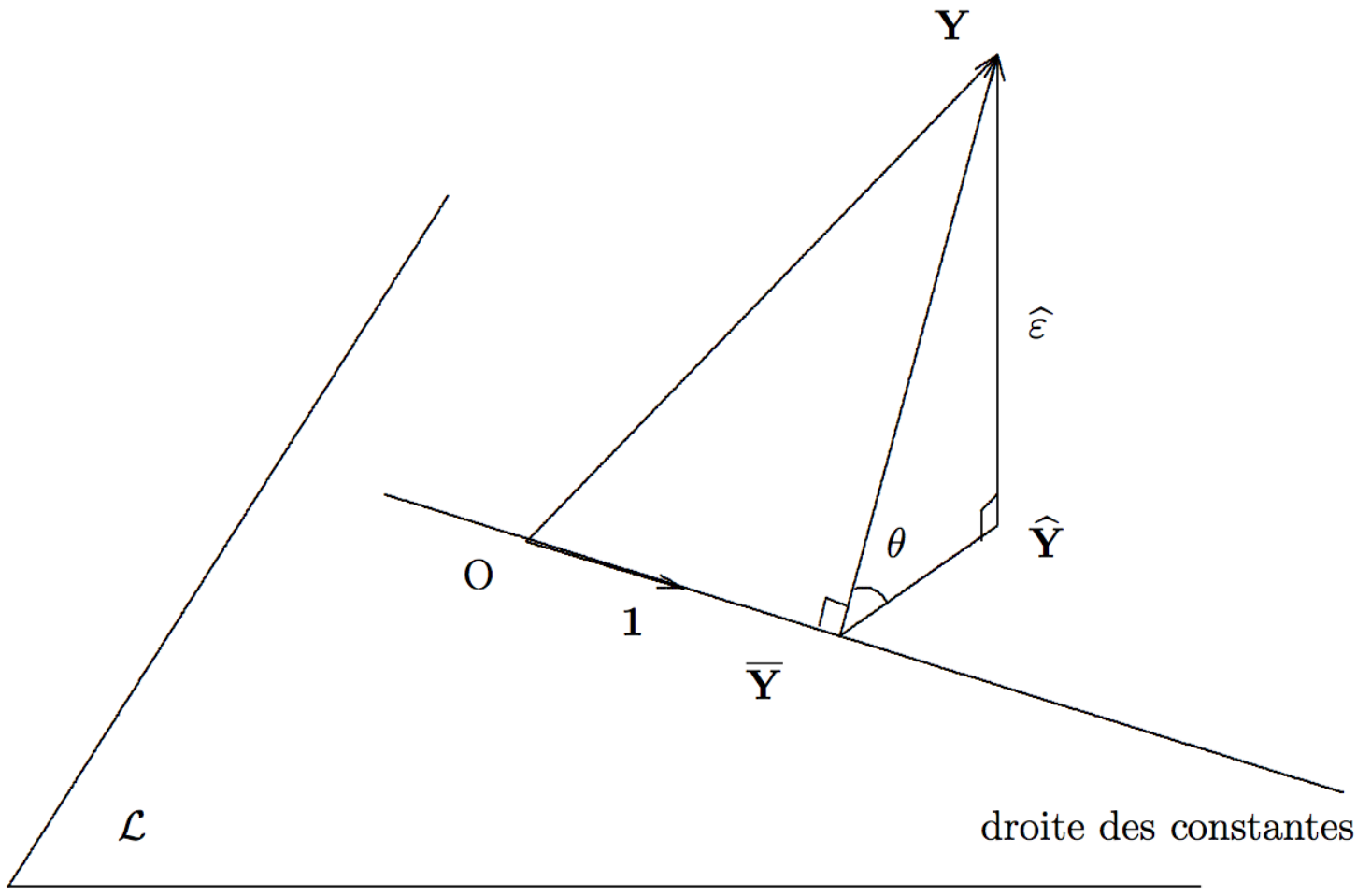
Comme $\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\mathbf{Y}}$ est orthogonal à $\hat{\mathbf{Y}} - \bar{\mathbf{Y}}$, on a

$$SCT = SCR + SCM$$
$$\|\mathbf{Y} - \bar{\mathbf{Y}}\|_2^2 = \|\mathbf{Y} - \hat{\mathbf{Y}}\|_2^2 + \|\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\|_2^2,$$

avec

- SCT = Somme des carrés totale
  $\rightsquigarrow$ variabilité totale à expliquer
- SCM = Somme des carrés due au modèle
  $\rightsquigarrow$ variabilité expliquée par le modèle
- SCR = Somme des carrés résiduelle
  $\rightsquigarrow$ variabilité non expliquée par le modèle

Rappel: Interprétation géométrique



Coefficient d'ajustement

Définition

R^2

Le coefficient de détermination est défini par :

$$R^2 = \frac{SCM}{SCT} = 1 - \frac{SCR}{SCT}$$

R^2 ajusté

Le coefficient de détermination ajusté est défini par :

$$\text{adjusted-}R^2 = 1 - \frac{SCR/(n-p-1)}{SCT/(n-1)}$$

Remarque

Le coefficient d'ajustement peut être interprété comme le pourcentage de variance expliquée par le modèle.

Test du modèle (I)

Hypothèse testée

$$\left\{ \begin{array}{ll} \mathcal{M}_0 : & \text{modèle le plus simple} \\ \mathcal{M}_1 : & \text{modèle le plus complexe} \end{array} \right. \Leftrightarrow \left\{ \begin{array}{ll} \mathcal{M}_0 : & Y_i = \beta_0 + \varepsilon_i \\ \mathcal{M}_1 : & Y_i = \mathbf{X}\boldsymbol{\beta} + \varepsilon_i \end{array} \right.$$

Loi des sommes de carrés sous H_0

- $SCR = \hat{\varepsilon}^\intercal \hat{\varepsilon}$, donc $SCR = (n-p-1)\hat{\sigma}^2 \sim \sigma^2 \chi^2_{n-p-1}$.
- $SCM = \|\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\|^2 = \|\text{proj}_{\mathbf{X}}(\mathbf{Y}) - \text{proj}_{\mathbf{1}}(\mathbf{Y})\|^2$, d'où $SCM \overset{H_0}{\sim} \sigma^2 \chi^2_p$.
- Comme $SCT = \|\mathbf{Y} - \bar{\mathbf{Y}}\|^2$, on a $SCT \overset{H_0}{\sim} \sigma^2 \chi^2_{n-1}$

Test du modèle (I)

Hypothèse testée

$$\left\{ \begin{array}{ll} \mathcal{M}_0 : & \text{modèle le plus simple} \\ \mathcal{M}_1 : & \text{modèle le plus complexe} \end{array} \right. \Leftrightarrow \left\{ \begin{array}{ll} \mathcal{M}_0 : & Y_i = \beta_0 + \varepsilon_i \\ \mathcal{M}_1 : & Y_i = \mathbf{X}\boldsymbol{\beta} + \varepsilon_i \end{array} \right.$$

Loi des sommes de carrés sous H_0

- $SCR = \hat{\varepsilon}^\intercal \hat{\varepsilon}$, donc $SCR = (n-p-1)\hat{\sigma}^2 \sim \sigma^2 \chi^2_{n-p-1}$.
- $SCM = \|\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\|^2 = \|\text{proj}_{\mathbf{X}}(\mathbf{Y}) - \text{proj}_{\mathbf{1}}(\mathbf{Y})\|^2$, d'où $SCM \overset{H_0}{\sim} \sigma^2 \chi^2_p$.
- Comme $SCT = \|\mathbf{Y} - \bar{\mathbf{Y}}\|^2$, on a $SCT \overset{H_0}{\sim} \sigma^2 \chi^2_{n-1}$

Test du modèle (II)

Statistique de test: Fisher

On rejette lorsque F, mesurant la part de variabilité expliquée par le modèle, est "grande":

$$F = \frac{SCM/\text{ddl}(SCM)}{SCR/\text{ddl}(SCR)} \underset{H_0}{\sim} \mathcal{F}_{p,n-p-1}.$$

Règle de décision

$$\text{On rejette } H_0 \text{ si } F \geq f_{p,n-p-1;1-\alpha}$$

$p-$valeur

$$p - \text{val} = \mathbb{P}_{H_0}\left(\mathcal{F}_{p,n-p-1} \geq f(\text{obs})\right)$$

Analyse de la variance

Tableau de synthèse

Source	Degrés de liberté	Sommes des carrés	Carrés moyens	F
Modèle	p	SCM	SCM/p	$F = \frac{(n-p-1)SCM}{SCR/p}$
Résiduelle	$n-p-1$	SCR	$\frac{SCR}{(n-p-1)}$	
Total	$n-1$	SCT		

Comparaison de modèles

Une question légitime lorsque l'on considère un modèle est

> *Est-ce que toutes les variables explicatives sont nécessaires pour expliquer la variable de sortie ?*

Une façon de poser la question consiste à considérer le test d'hypothèse suivant

$$\begin{cases} \mathcal{M}_\omega : & \text{modèle le plus simple} \\ \mathcal{M}_\Omega : & \text{modèle le plus complexe} \end{cases},$$

où $\mathcal{M}_\omega \subset \mathcal{M}_\Omega$: les modèles sont dits "emboîtés".

Comparaison de modèles

Aperçu géométrique

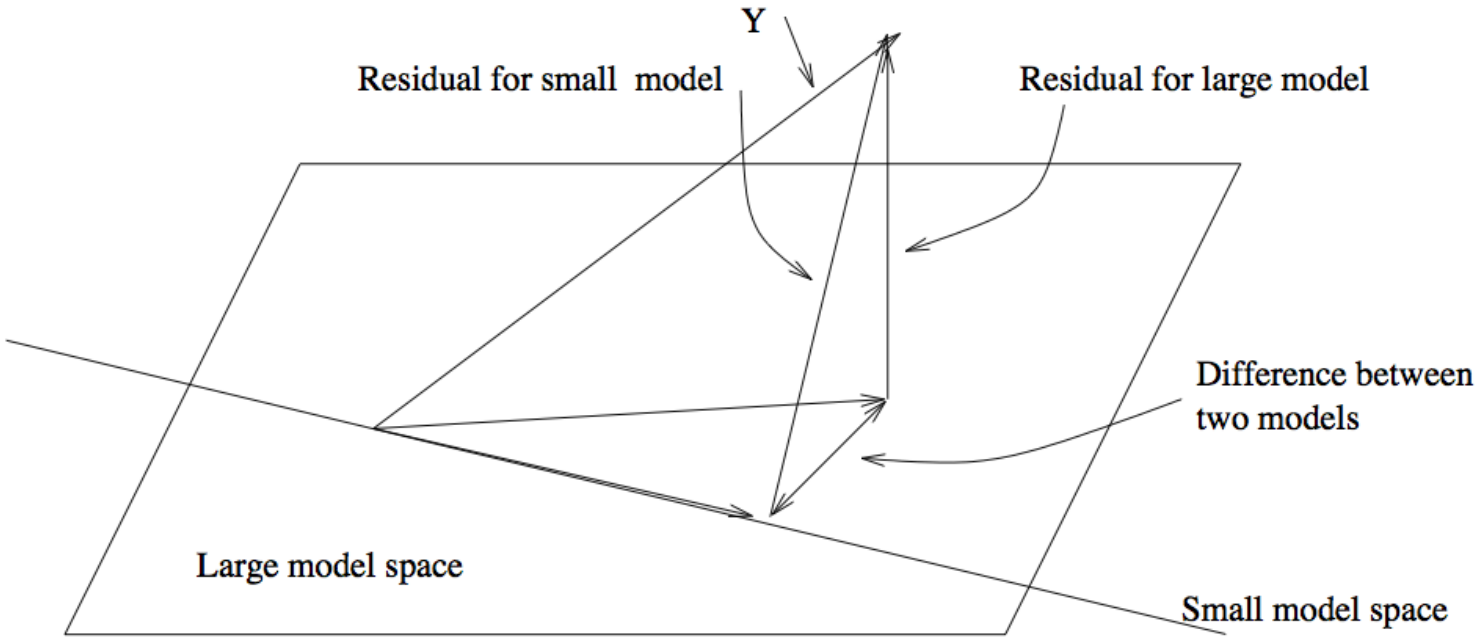


Figure: Source: *Pratical regression and anova using R*, J. Faraway

Comparaison de modèles emboîtés

Intuition

On choisira H_1 (le modèle le plus grand Ω) si les résidus de Ω sont vraiment petits comparés au modèle ω, *i.e.*,

$$SCR_\Omega < SCR_\omega \quad \text{ou} \quad \frac{SCR_\omega - SCR_\Omega}{SCR_\Omega} \gg 1$$

Sous H_0

- $SCR_\omega - SCM_\Omega \sim \sigma \chi^2_{\text{ddl}_\omega - \text{ddl}_\Omega}$
- $SCR_\Omega \sim \sigma \chi^2_{n - \text{ddl}_\Omega}$

Statistique de test

$$F = \frac{(SCR_\omega - SCR_\Omega)}{SCR_\Omega} \times \frac{(n - \text{ddl}_\Omega)}{(\text{ddl}_\omega - \text{ddl}_\Omega)} \underset{H_0}{\sim} \mathcal{F}_{n - \text{ddl}_\Omega, \text{ddl}_\omega - \text{ddl}_\Omega}.$$

Comparaison de modèles emboîtés

Intuition

On choisira H_1 (le modèle le plus grand Ω) si les résidus de Ω sont vraiment petits comparés au modèle ω, *i.e.*,

$$SCR_\Omega < SCR_\omega \quad \text{ou} \quad \frac{SCR_\omega - SCR_\Omega}{SCR_\Omega} \gg 1$$

Sous H_0

- $SCR_\omega - SCM_\Omega \sim \sigma \chi^2_{\text{ddl}_\omega - \text{ddl}_\Omega}$
- $SCR_\Omega \sim \sigma \chi^2_{n - \text{ddl}_\Omega}$

Statistique de test

$$F = \frac{(SCR_\omega - SCR_\Omega)}{SCR_\Omega} \times \frac{(n - \text{ddl}_\Omega)}{(\text{ddl}_\omega - \text{ddl}_\Omega)} \underset{H_0}{\sim} \mathcal{F}_{n-\text{ddl}_\Omega, \text{ddl}_\omega - \text{ddl}_\Omega}.$$

Comparaison de modèles emboîtés

Intuition

On choisira H_1 (le modèle le plus grand Ω) si les résidus de Ω sont vraiment petits comparés au modèle ω, *i.e.*,

$$SCR_\Omega < SCR_\omega \quad \text{ou} \quad \frac{SCR_\omega - SCR_\Omega}{SCR_\Omega} \gg 1$$

Sous H_0

- $SCR_\omega - SCM_\Omega \sim \sigma \chi^2_{\text{ddl}_\omega - \text{ddl}_\Omega}$
- $SCR_\Omega \sim \sigma \chi^2_{n - \text{ddl}_\Omega}$

Statistique de test

$$F = \frac{(SCR_\omega - SCR_\Omega)}{SCR_\Omega} \times \frac{(n - \text{ddl}_\Omega)}{(\text{ddl}_\omega - \text{ddl}_\Omega)} \underset{H_0}{\sim} \mathcal{F}_{n - \text{ddl}_\Omega, \text{ddl}_\omega - \text{ddl}_\Omega}.$$

Comparaison de modèles emboîtés

Tableau de synthèse

Source	Degrés de liberté	Sommes des carrés	Carrés moyens
Modèle ω	$n - \text{ddl}_\omega$	SCR_ω	$SCR_\omega/\text{ddl}_\omega$
Modèle Ω	$n - \text{ddl}_\Omega$	SCR_Ω	$SCR_\Omega/\text{ddl}_\Omega$

$$F = \frac{(SCR_\omega - SCR_\Omega)}{SCR_\Omega} \times \frac{(n - \text{ddl}_\Omega)}{(\text{ddl}_\omega - \text{ddl}_\Omega)}$$

Plan

Objectifs du diagnostic

1. Vérification des **hypothèses du modèles**
   - linéarité/modèle adéquat
   - homoscédasticité des résidus
   - indépendance des résidus
   - normalité des résidus

2. Détection d'**observations atypiques**

Plan

Analyse des résidus

Les hypothèses du modèle sont toutes liées aux résidus

1. Résidus centrés: $\mathbb{E}(Y) = \mathbf{X}\boldsymbol{\beta}$, soit $\mathbb{E}(\varepsilon_i) = 0$
2. Résidus homoscédastiques : $\mathbb{V}(\varepsilon_i) = \sigma^2$ pour tout i,
3. Résidus indépendents, $\text{cov}(\varepsilon_i, \varepsilon_{i'}) = 0$
4. Résidus gaussiens: $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Diagnostic

À défaut de disposer de ε_i, on diagnostique $\hat{\varepsilon}_i$

1. Analyse du **graphe des résidus**, détaillé dans la suite
2. Test d'indépendance (Durbin-Watson)
3. Test de normalité (Shapiro, Kolmogorov, χ^2)

Points leviers

Définition (Levier)

La variance de la prédiction de la i^e observations vérifie

$$\mathbb{V}(\hat{Y}_i) = \sigma^2 h_i,$$

où $h_i = (\mathbf{P_X})_{ii}$ est appelé **levier** *de l'observation i.*

- Plus h_i est grand, plus l'observation y_i contribue à $\hat{Y}_i$.
- $\sum_{i=1}^n h_i = p$, donc la moyenne des leviers est p/n.

Définition (Point levier)

L'individu i est un **point levier** *si*

$$h_i > \frac{2p}{n}.$$

Points leviers

Définition (Levier)

La variance de la prédiction de la i^e observations vérifie

$$\mathbb{V}(\hat{Y}_i) = \sigma^2 h_i,$$

où $h_i = (\mathbf{P_X})_{ii}$ est appelé **levier** *de l'observation i.*

- Plus h_i est grand, plus l'observation y_i contribue à $\hat{Y}_i$.
- $\sum_{i=1}^{n} h_i = p$, donc la moyenne des leviers est p/n.

Définition (Point levier)

L'individu i est un **point levier** *si*

$$h_i > \frac{2p}{n}.$$

Résidus standardisés et studentisés

Il est utile de normaliser $\hat{\varepsilon}_i$ afin de s'affranchir des facteurs d'échelle.

Définition (Résidus standardisés)

La variance des résidus estimés s'écrit $\mathbb{V}(\hat{\varepsilon}_i) = \sigma^2(1 - h_i)$. *Ainsi, on définit la* **forme standardisée des résidus** *par*

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_i}}.$$

- $\hat{\varepsilon}_i$ n'étant pas indépendant de $\hat{\sigma}$, on ne connait pas leur distribution.
- la forme dite **studentisé** corrige ce problème.

Définition (Résidus studentisé)

On appelle **résidus studentisés** *les statistiques définies par*

$$t_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}^{(-i)}\sqrt{1 - h_i}},$$

où $\hat{\sigma}^{(-i)}$ *est la variance estimée sur les données sans la* i^e *observation.*

Résidus standardisés et studentisés

Il est utile de normaliser $\hat{\varepsilon}_i$ afin de s'affranchir des facteurs d'échelle.

Définition (Résidus standardisés)

La variance des résidus estimés s'écrit $\mathbb{V}(\hat{\varepsilon}_i) = \sigma^2(1 - h_i)$. *Ainsi, on définit la* **forme standardisée des résidus** *par*

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_i}}.$$

- $\hat{\varepsilon}_i$ n'étant pas indépendant de $\hat{\sigma}$, on ne connait pas leur distribution.
- la forme dite **studentisé** corrige ce problème.

Définition (Résidus studentisé)

On appelle **résidus studentisés** *les statistiques définies par*

$$t_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}^{(-i)}\sqrt{1 - h_i}},$$

où $\hat{\sigma}^{(-i)}$ *est la variance estimée sur les données sans la* i^e *observation.*

Résidus standardisés et studentisés

Il est utile de normaliser $\hat{\varepsilon}_i$ afin de s'affranchir des facteurs d'échelle.

Définition (Résidus standardisés)

La variance des résidus estimés s'écrit $\mathbb{V}(\hat{\varepsilon}_i) = \sigma^2(1 - h_i)$. *Ainsi, on définit la* **forme standardisée des résidus** *par*

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_i}}.$$

- $\hat{\varepsilon}_i$ n'étant pas indépendant de $\hat{\sigma}$, on ne connait pas leur distribution.
- la forme dite **studentisé** corrige ce problème.

Définition (Résidus studentisé)

On appelle **résidus studentisés** *les statistiques définies par*

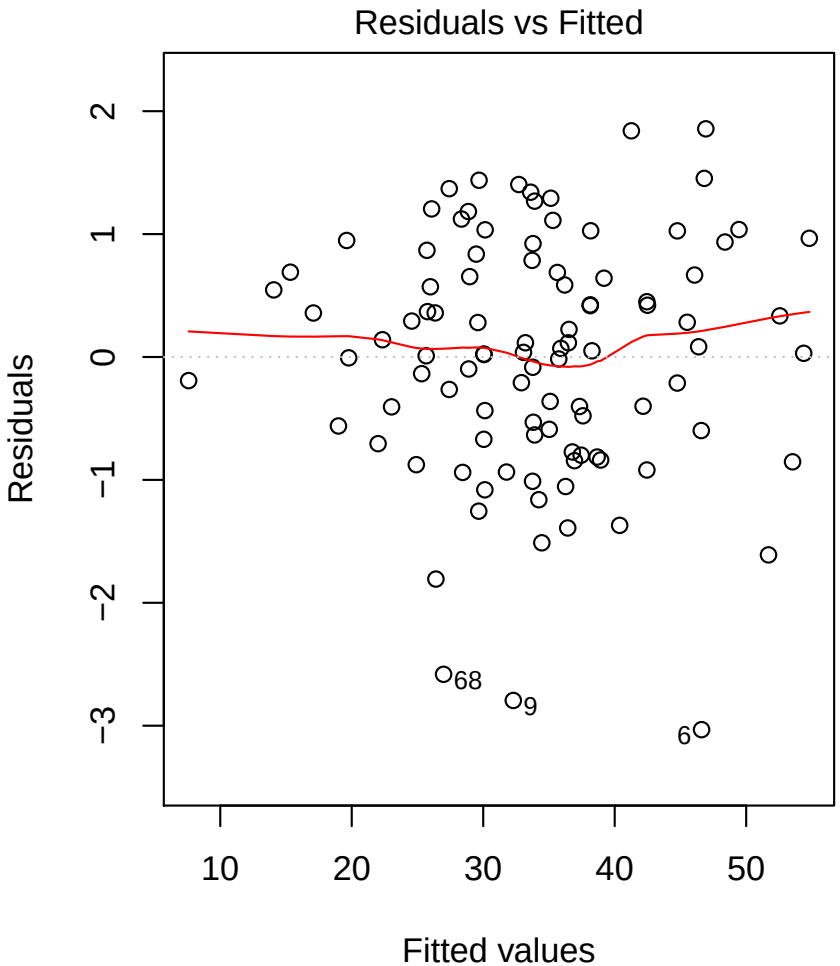
$$t_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}^{(-i)}\sqrt{1 - h_i}},$$

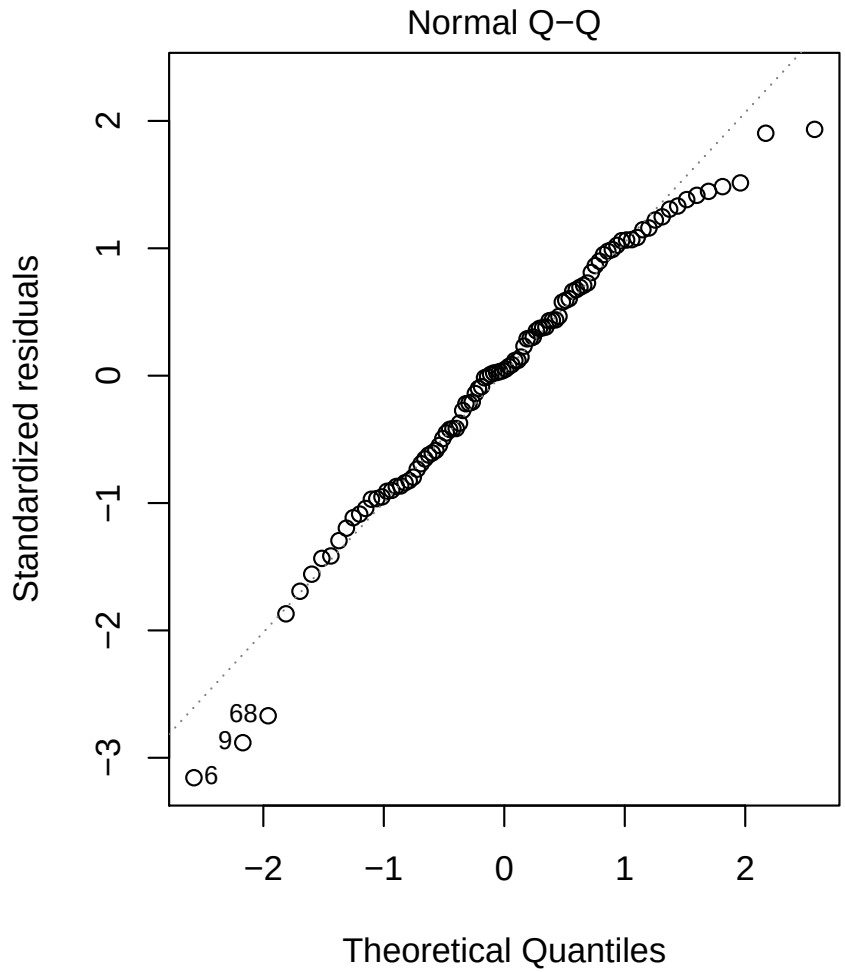
où $\hat{\sigma}^{(-i)}$ *est la variance estimée sur les données sans la* i^e *observation.*

Analyse des résidus

Cas idéal

```
n <- 100; x <- rnorm(n,10,3); y <- 5 + 3 * x + rnorm(n,0,1)
par(mfrow=c(1,2)); plot(lm(y~x), which=1:2)
```





Analyse des résidus I

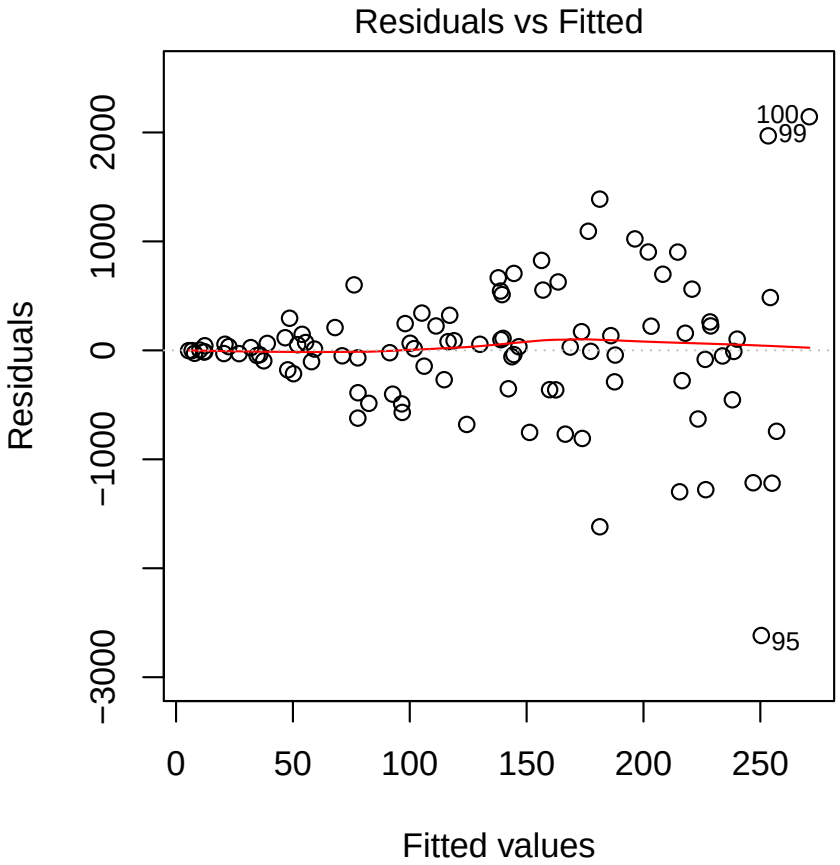
Variance proportionnelle au prédicteur

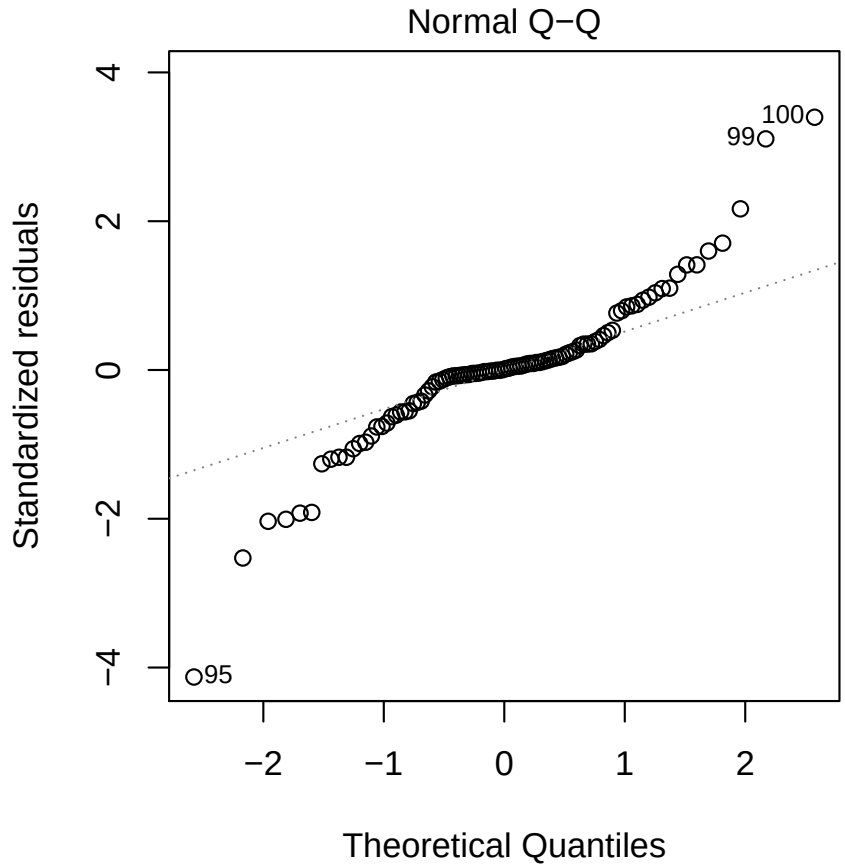
Transformer Y en log/racine peut corriger l'hétéroscédasticité

```
n <- 100; x <- (1:n + rnorm(n,0,5)); y <- 5 + 3 * x + rnorm(n,0,10)*x
par(mfrow=c(1,2));  plot(lm(y~x), which=1:2); plot(lm(sqrt(y)~x), which=1:2)
```

Analyse des résidus II

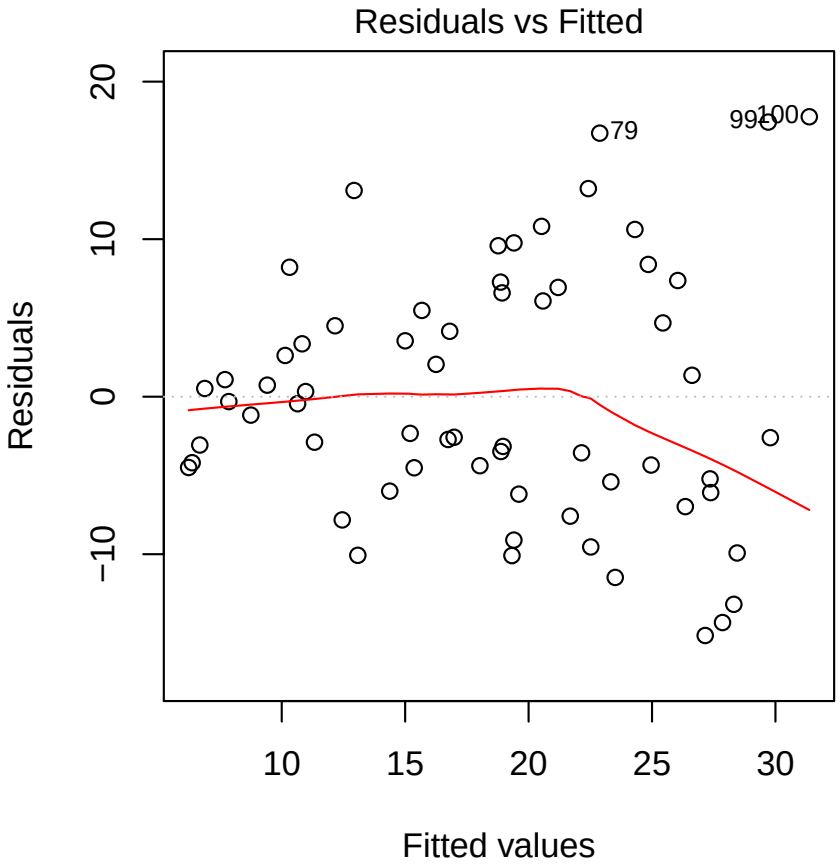
Variance proportionnelle au prédicteur

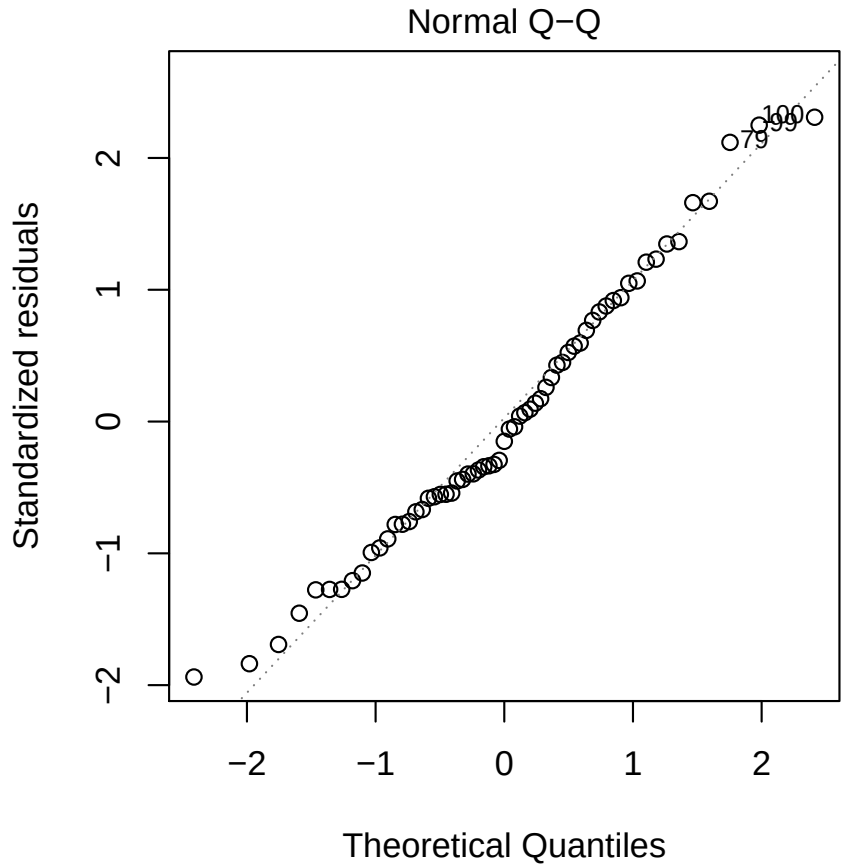




Analyse des résidus III

Variance proportionnelle au prédicteur



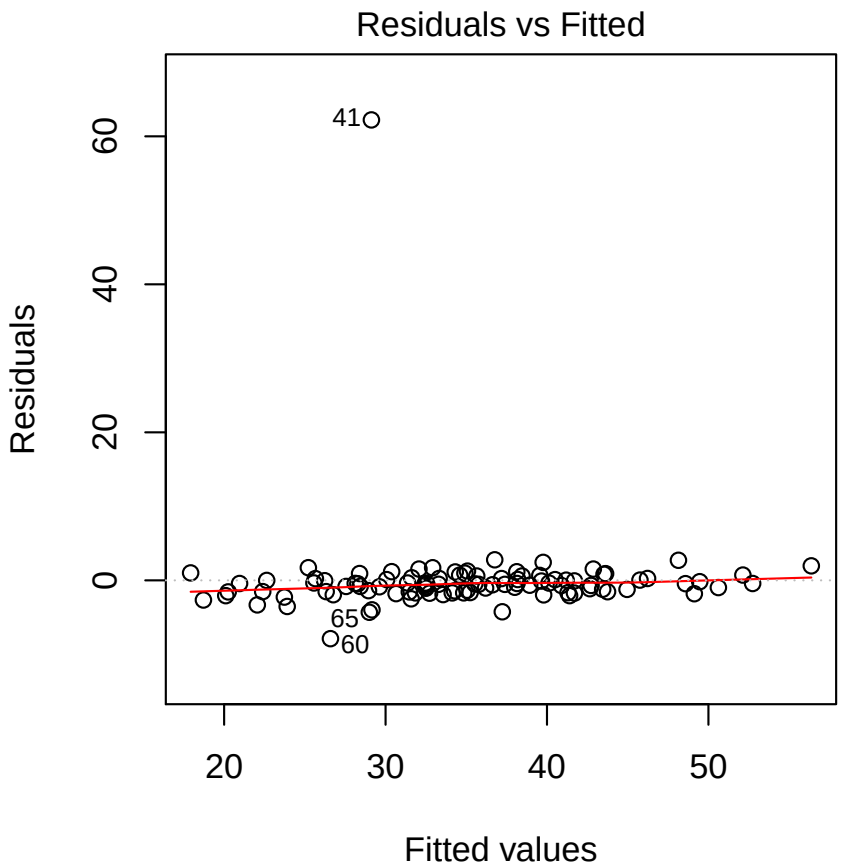


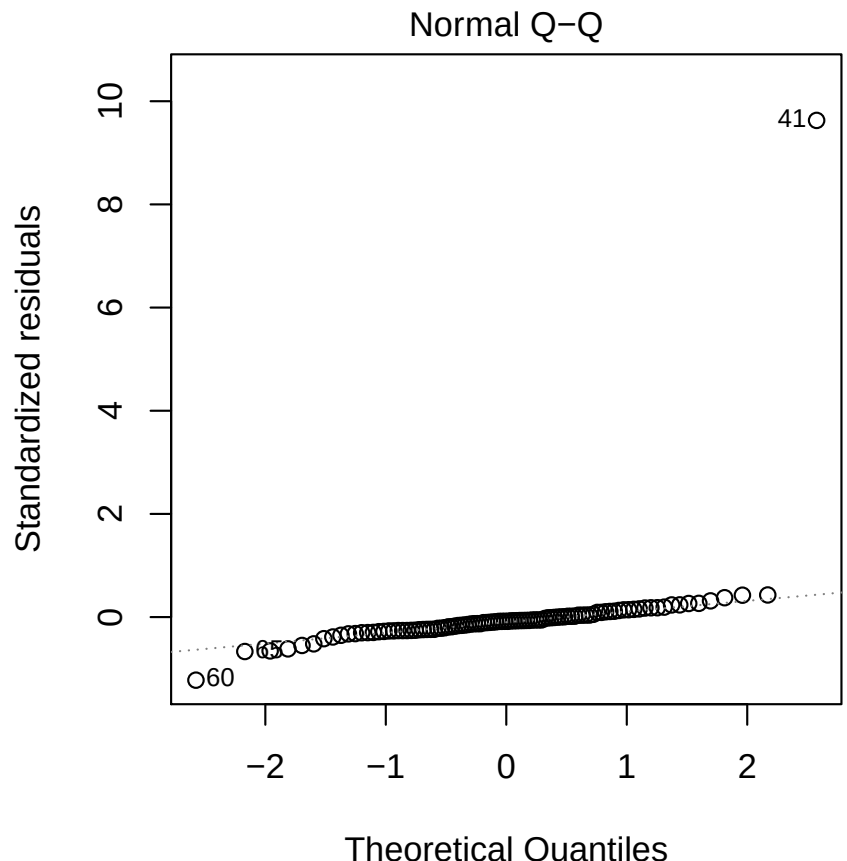
Analyse des résidus

Résidus non gaussiens

Le modèle linéaire est robuste aux résidus non gaussiens s'ils sont symétriques

```
n <- 100; x <- rnorm(n,10,3); y <- 5 + 3 * x + rt(n,2)
par(mfrow=c(1,2)); plot(lm(y~x), which=1:2)
```





Analyse des résidus I

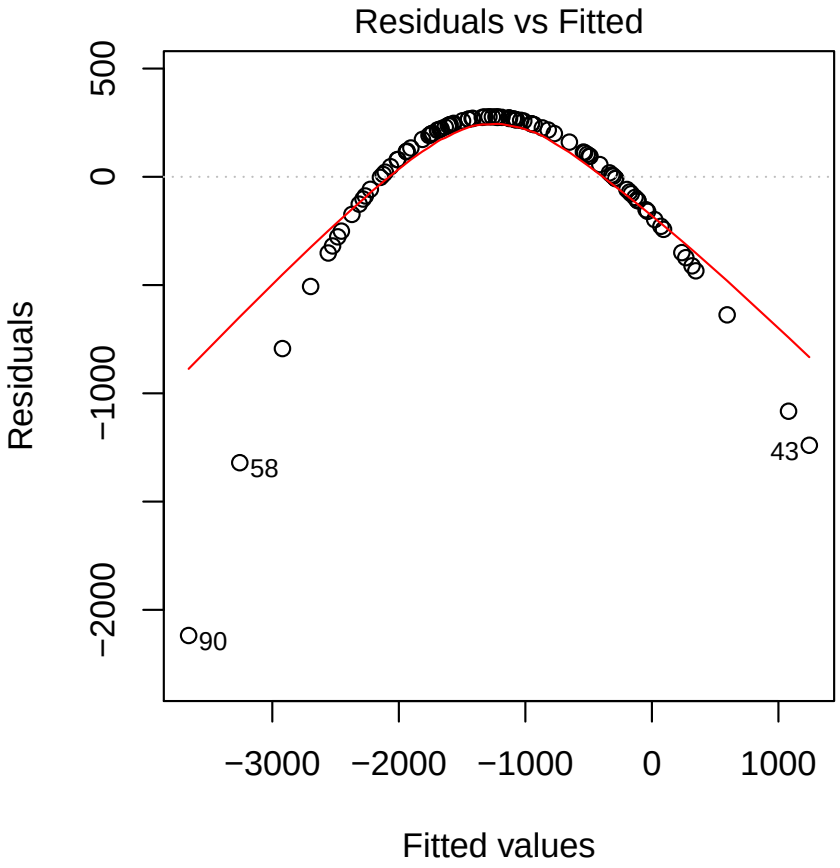
Mauvais modèle

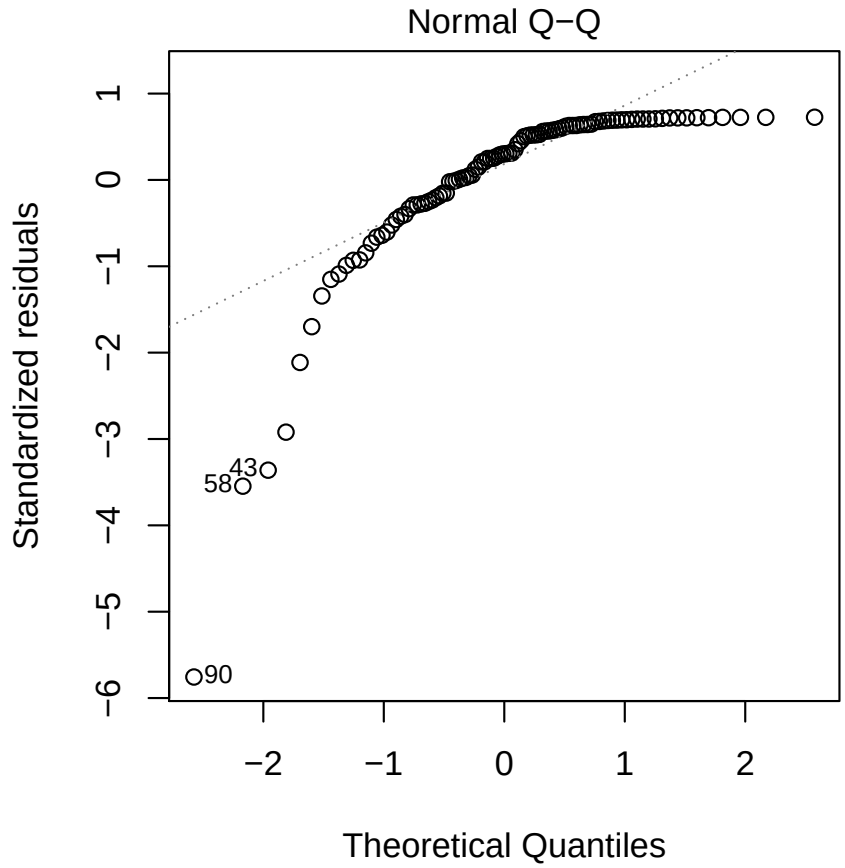
Une tendance forte dans les résidus peut indiquer un mauvais modèle.

```
n <- 100; x <- rnorm(n,10,3); y <- 5 + 3*x - x^3+rnorm(n,0,1)
par(mfrow=c(1,2)); plot(lm(y~x), which=1:2); plot(lm(y~x+I(x^3)), which=1:2)
```

Analyse des résidus II

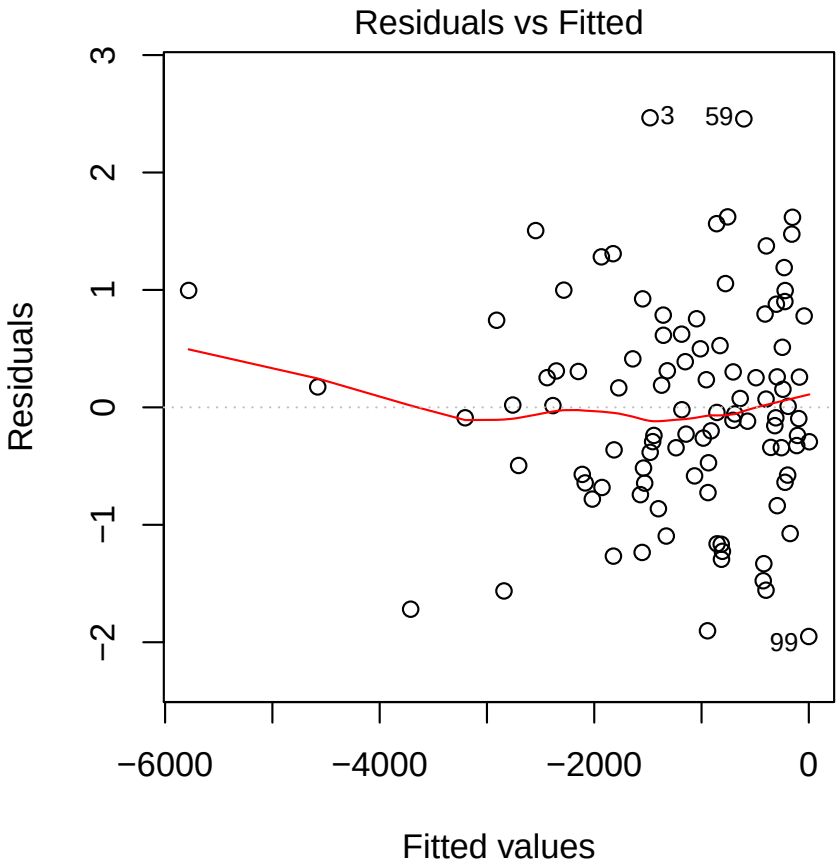
Mauvais modèle

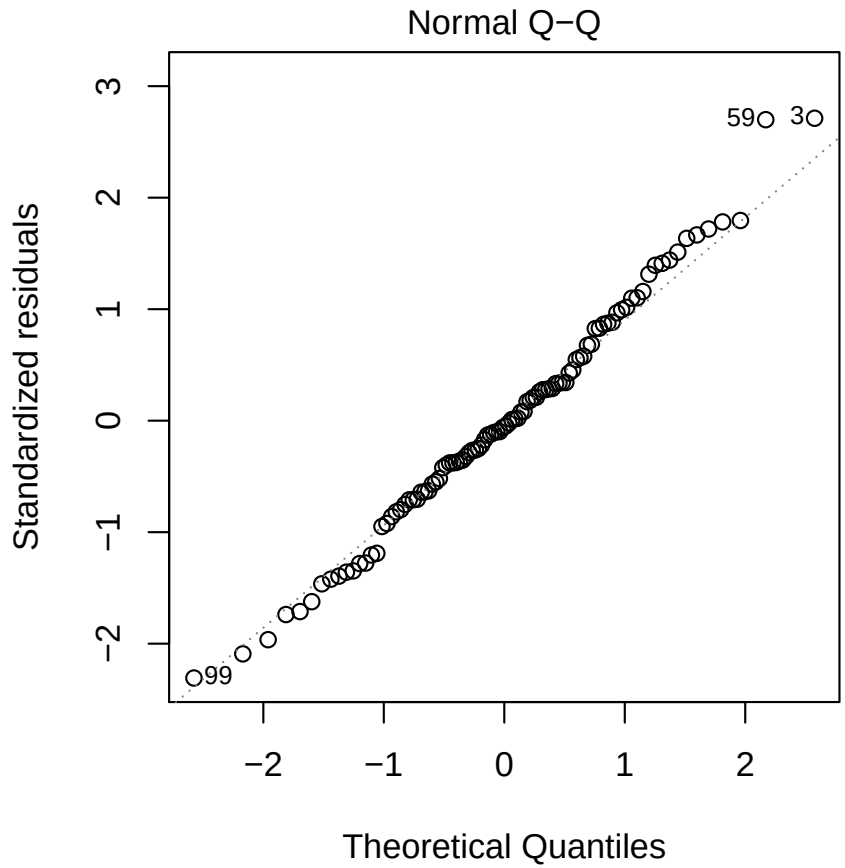




Analyse des résidus III

Mauvais modèle





Plan

Distance de Cook

Idée

Mettre en évidence l'influence "anormale" de certains points.

Définition (Distance de Cook)

La quantité D_i caractérise l'influence de l'observation i sur le résultat de la régression, une valeur élevée pouvant révéler une influence "anormale"

$$D_i = \frac{\|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}^{(-i)}\|^2}{(p+1)\hat{\sigma}^2} = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(-i)})'\mathbf{X}^\intercal\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(-i)})}{(p+1)\hat{\sigma}^2}$$

$\rightsquigarrow$ *D_i peut s'interpréter comme le carré d'une distance entre $\hat{\boldsymbol{\beta}}$ et $\hat{\boldsymbol{\beta}}^{(-i)}$.*

Proposition (Calcul pratique)

On peut calculer D_i sans réajuster de modèle car

$$D_i = \frac{\hat{\varepsilon}_i^2}{(p+1)\hat{\sigma}^2} \times \frac{h_i}{(1-h_i)^2}.$$

Distance de Cook

Idée

Mettre en évidence l'influence "anormale" de certains points.

Définition (Distance de Cook)

La quantité D_i caractérise l'influence de l'observation i sur le résultat de la régression, une valeur élevée pouvant révéler une influence "anormale"

$$D_i = \frac{\|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}^{(-i)}\|^2}{(p+1)\hat{\sigma}^2} = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(-i)})'\mathbf{X}^\intercal\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(-i)})}{(p+1)\hat{\sigma}^2}$$

$\rightsquigarrow$ *D_i peut s'interpréter comme le carré d'une distance entre $\hat{\boldsymbol{\beta}}$ et $\hat{\boldsymbol{\beta}}^{(-i)}$.*

Proposition (Calcul pratique)

On peut calculer D_i sans réajuster de modèle car

$$D_i = \frac{\hat{\varepsilon}_i^2}{(p+1)\hat{\sigma}^2} \times \frac{h_i}{(1-h_i)^2}.$$

Distance de Cook

Quelle valeur de seuil choisir ?

Règle standard

On considère qu'une valeur > 1 signifie un point aberrant.

Approche par test

On montre que D_i est une statistique de décision du test de Wald

$$H_0 : \beta = \beta_0^{-i},$$

où β_0^{-i} est la vrai valeur estimée sans la i^{e} observation.
La statistique de test est une $F_{p+1,n-p-1,1-\alpha}$

Distance de Cook

Quelle valeur de seuil choisir ?

Règle standard

On considère qu'une valeur > 1 signifie un point aberrant.

Approche par test

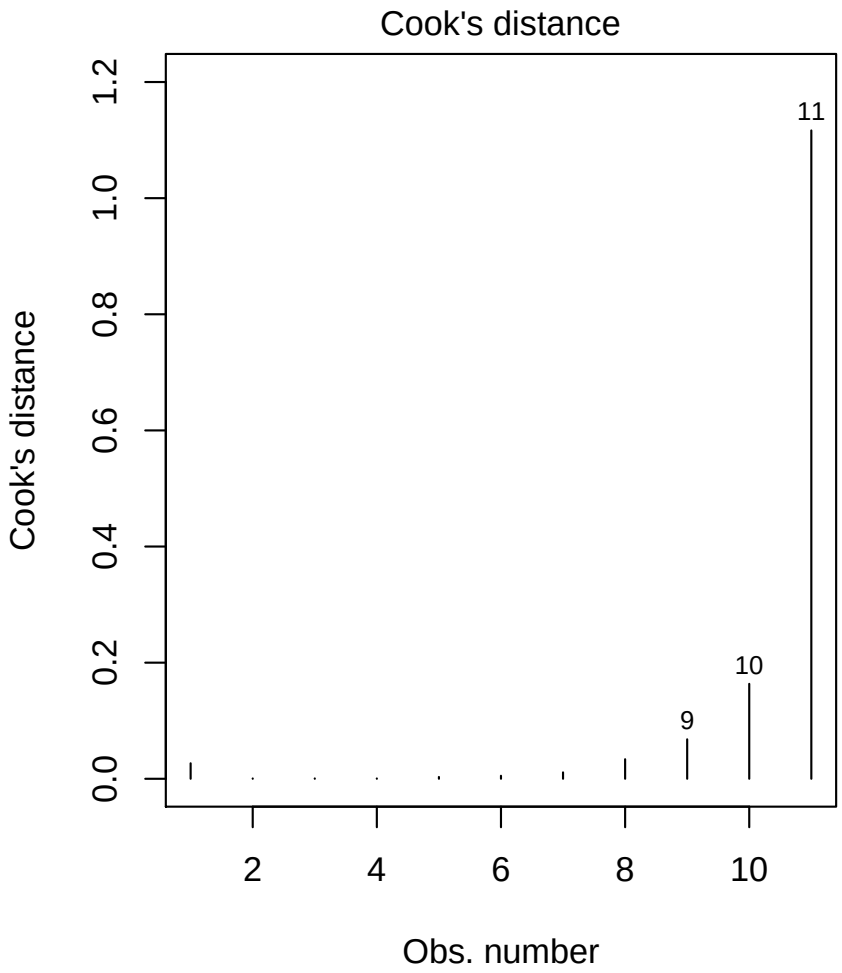
On montre que D_i est une statistique de décision du test de Wald

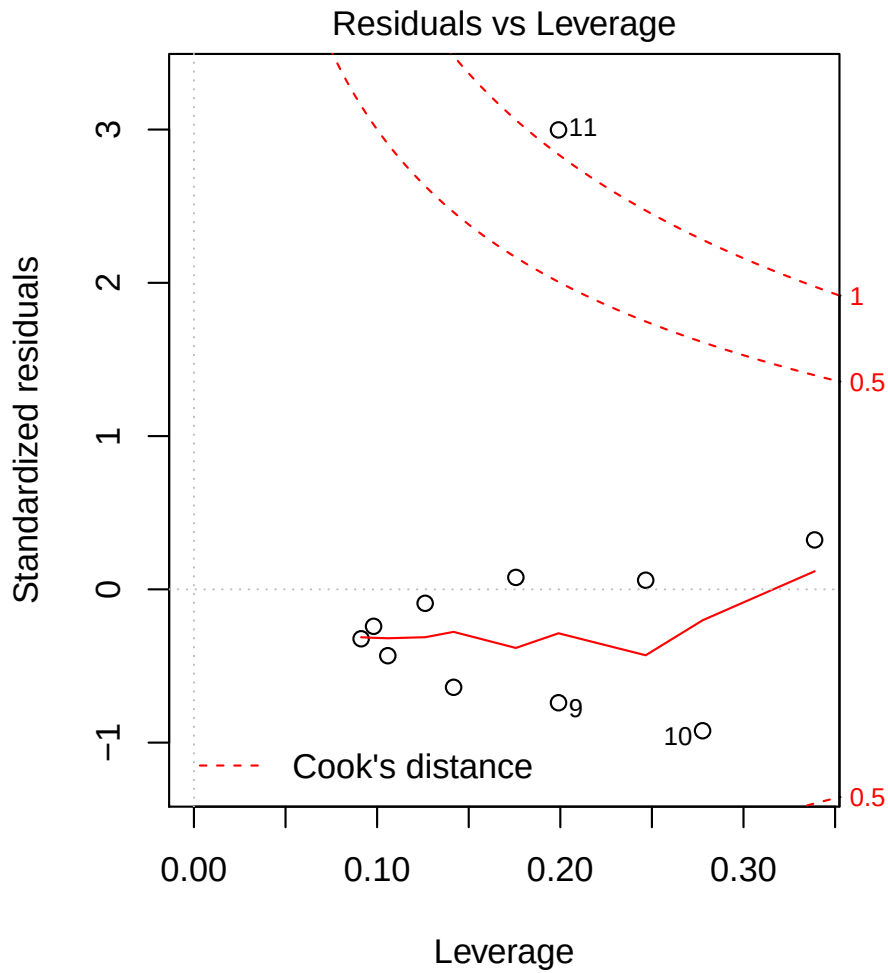
$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0^{-i},$$

où $\boldsymbol{\beta}_0^{-i}$ est la vrai valeur estimée sans la i^{e} observation.
La statistique de test est une $F_{p+1,n-p-1,1-\alpha}$

Distance de Cook

```
x <- seq(1,10,len=10); y <- 5+.4*x+rnorm(10,0,1); x <- c(x,9); y <- c(y,100)
par(mfrow=c(1,2)); plot(lm(y~x), which=4:5)
```





Plan

Plan

Données de processionnaires du pin I

Données

On dispose de 33 échantillons de parcelles forestière de 10 hectares. Chaque parcelle est coupée en placette de 5 ares sur lesquelles sont calculées les moyennes des mesures suivantes

```
chenilles <- read.table(file='Chenilles.txt',header=TRUE)
colnames(chenilles)

##  [1] "Altitude" "Pente"    "NbPins"   "Hauteur"  "Diametre" "Densite"
##  [7] "Orient"   "HautMax"  "NbStrat"  "Melange"  "NbNids"
```

Objectif

Prédire le **nombre de nids** par les autres variables.

source:https://www.agroparistech.fr/IMG/pdf/ExemplesModeleLineaire-AgroParisTech.pdf

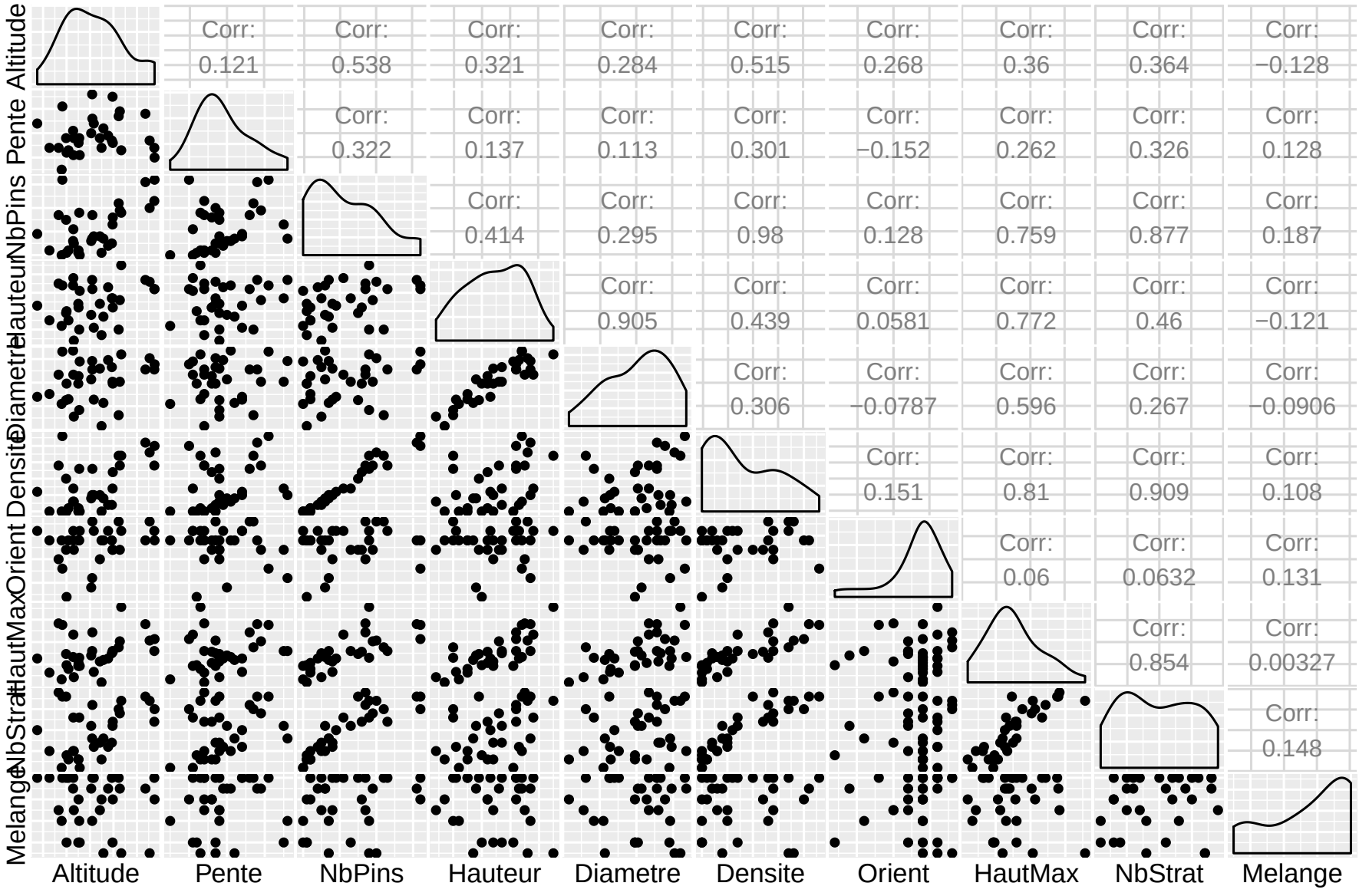
Données de processionnaires du pin II

L'entête du tableau de données donne

```
head(chenilles)
```

```
##   Altitude Pente NbPins Hauteur Diametre Densite Orient HautMax NbStrat
## 1     1200    22      1     4.0     14.8     1.0    1.1     5.9     1.4
## 2     1342    28      8     4.4     18.0     1.5    1.5     6.4     1.7
## 3     1231    28      5     2.4      7.8     1.3    1.6     4.3     1.5
## 4     1254    28     18     3.0      9.2     2.3    1.7     6.9     2.3
## 5     1357    32      7     3.7     10.7     1.4    1.7     6.6     1.8
## 6     1250    27      1     4.4     14.8     1.0    1.7     5.8     1.3
##   Melange NbNids
## 1     1.4   2.37
## 2     1.7   1.47
## 3     1.7   1.13
## 4     1.6   0.85
## 5     1.3   0.24
## 6     1.4   1.49
```

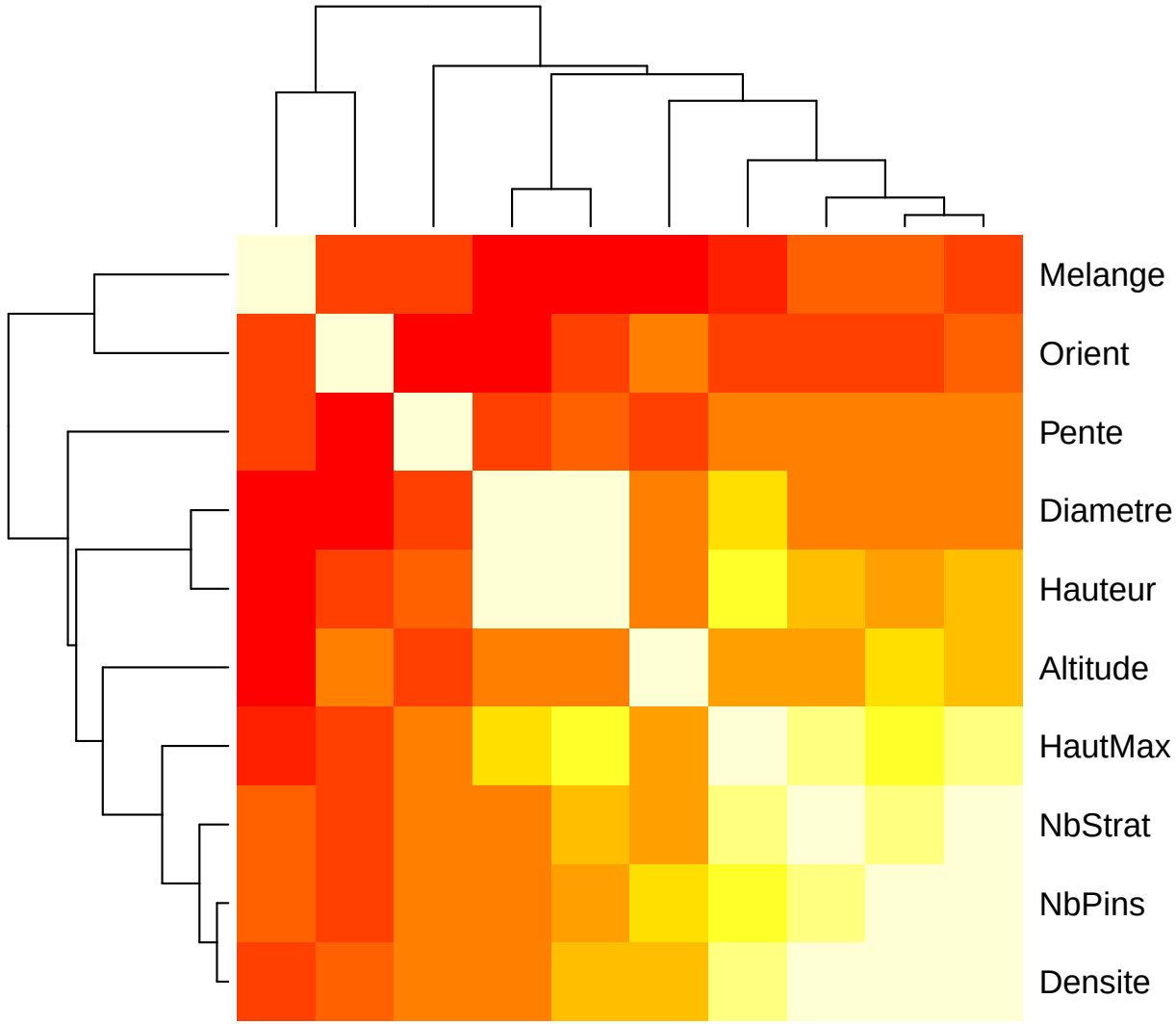
Données de processionnaires du pin III



Corrélation entre prédicteurs

De fortes corrélations induisent une estimation difficile des paramètres corrélés

```
heatmap(cor(chenilles[, -ncol(chenilles)]), symm=TRUE)
```



Plan

Moindres carrés ordinaires

Simple vérification

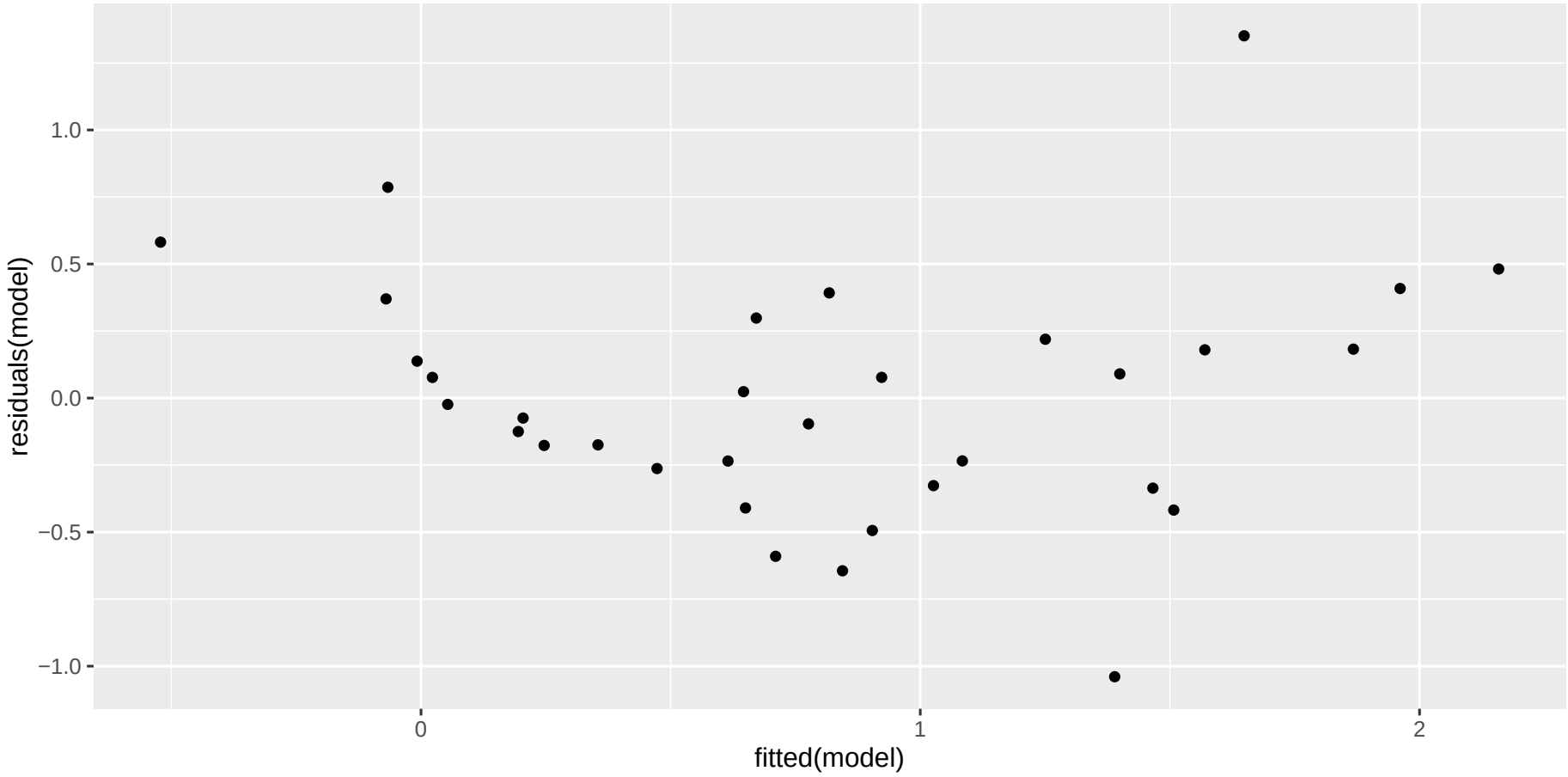
```
X <- cbind(1, as.matrix(chenilles[, -ncol(chenilles)]))
y <- chenilles[,  ncol(chenilles)]
beta.ols <- solve(crossprod(X), crossprod(X,y))
print(t(beta.ols))

##                   Altitude       Pente     NbPins    Hauteur  Diametre
## [1,] 8.561849 -0.002956282 -0.03482086 0.03538525 -0.5015637 0.1087387
##          Densite     Orient    HautMax    NbStrat    Melange
## [1,] -0.03271541 -0.2039587 0.02818019 -0.8624094 -0.4481242

coefficients(lm(NbNids~., data=chenilles)) ## sanity check

##  (Intercept)     Altitude        Pente       NbPins      Hauteur
##  8.561848740 -0.002956282 -0.034820858  0.035385252 -0.501563729
##     Diametre      Densite       Orient      HautMax      NbStrat
##  0.108738715 -0.032715407 -0.203958683  0.028180190 -0.862409366
##      Melange
## -0.448124198
```

Modèle linéaire multiples "brute"

Graphe des résidus

Le graphe des résidus suggèrent une transformation logarithmique de la réponse.

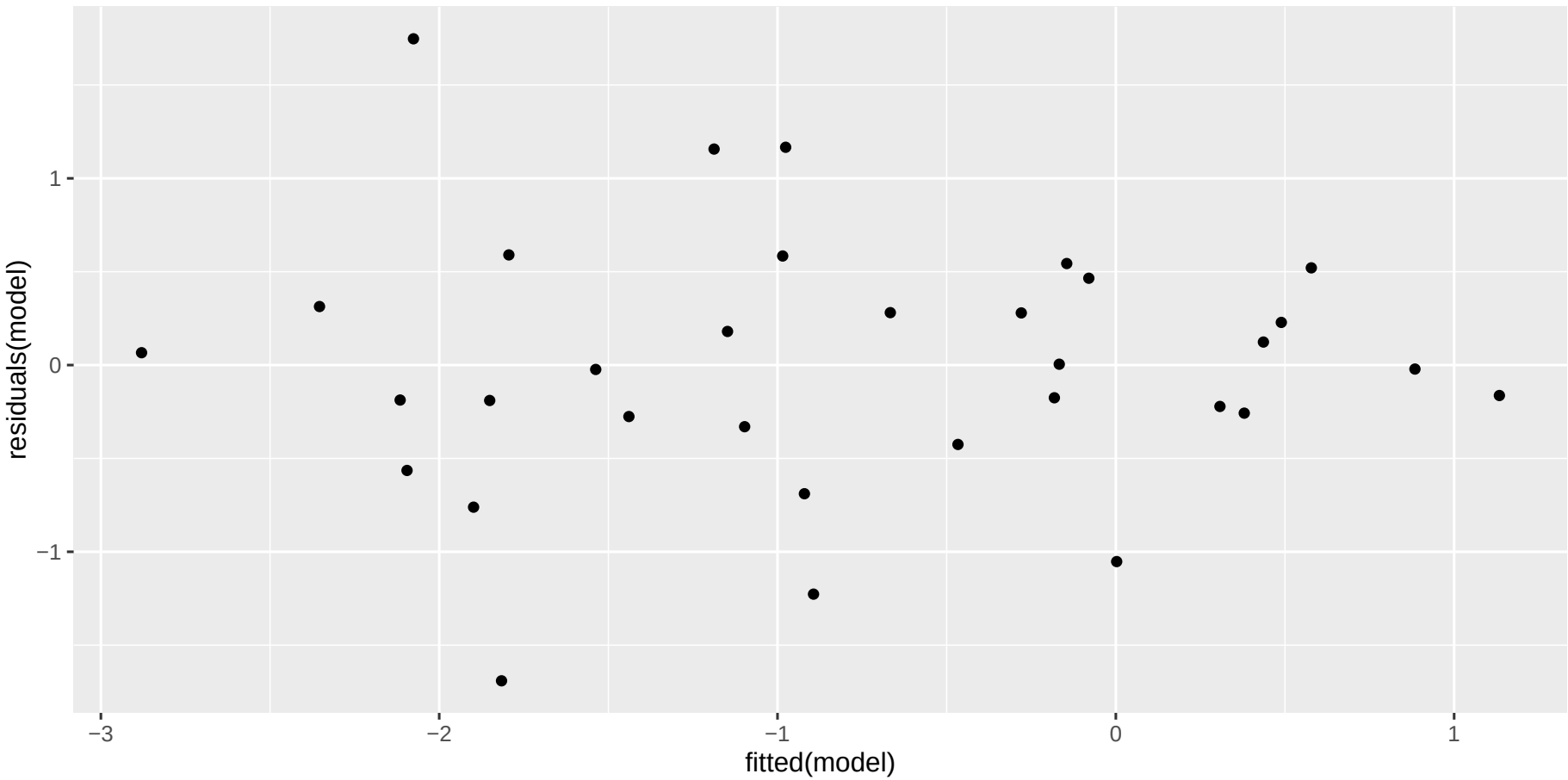
```
model <- lm(NbNids~.,data=chenilles)
qplot(fitted(model),residuals(model), geom='point')
```



Modèle log-transformé

Graphe des résidus

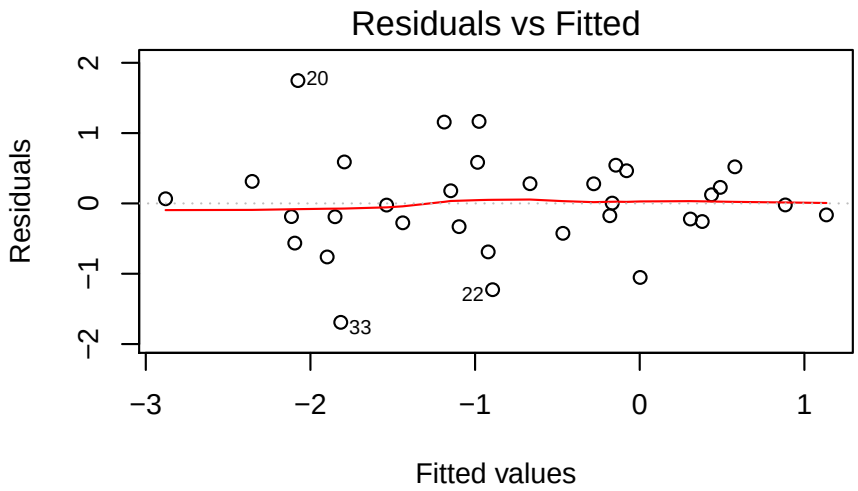
```
model <- lm(log(NbNids)~.,data=chenilles)
qplot(fitted(model),residuals(model), geom='point')
```

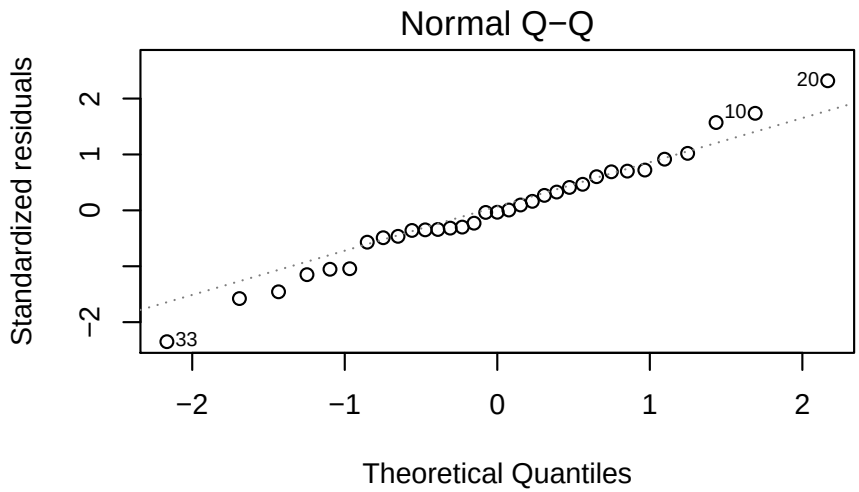


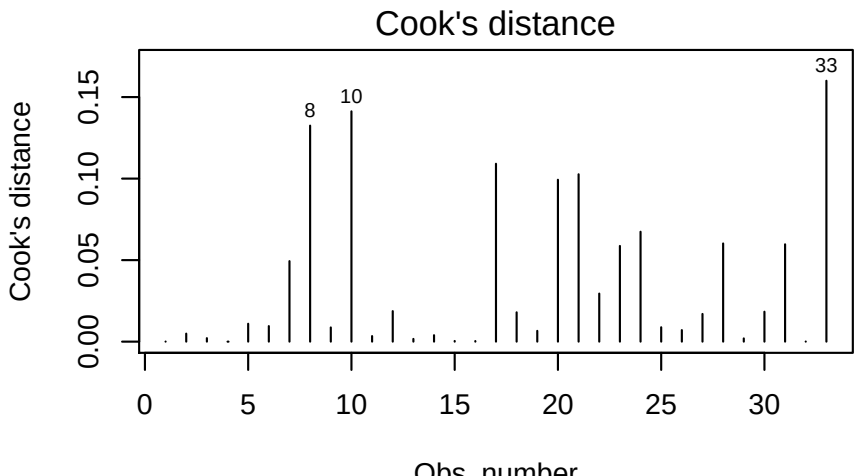
Modèle log-transformé

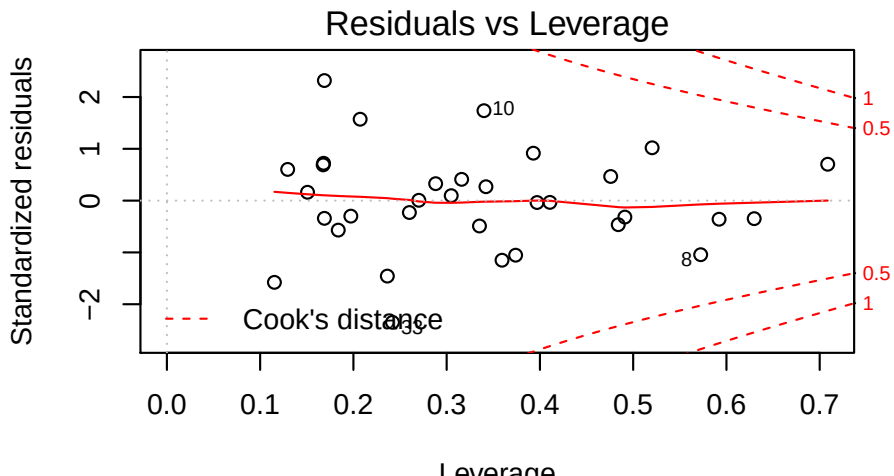
Le diagnostic complet

```
par(mfrow=c(2,2)); plot(model, which=c(1,2,4,5))
```









Modèle log-transformé

Lnormalité des résidus

```
shapiro.test(residuals(model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model)
## W = 0.97572, p-value = 0.6517
```

Modèle log-transformé
indépendance des résidus

```
library(car)
durbinWatsonTest(model)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      -0.1208374      2.051547   0.906
##  Alternative hypothesis: rho != 0
```

Modèle log-transformé

Test des paramètres

```
summary(model)$coefficients
```

```
##                 Estimate  Std. Error    t value    Pr(>|t|)
## (Intercept) 11.300912256 3.156550408  3.5801463 0.001669442
## Altitude    -0.004505222 0.001563014 -2.8823938 0.008647574
## Pente       -0.053605957 0.021842576 -2.4541957 0.022502117
## NbPins       0.074581111 0.100232834  0.7440786 0.464702763
## Hauteur     -1.328276893 0.570060846 -2.3300616 0.029375766
## Diametre     0.236101193 0.104611127  2.2569415 0.034280797
## Densite     -0.451118399 1.572915841 -0.2868039 0.776946247
## Orient      -0.187809689 1.007950218 -0.1863283 0.853894734
## HautMax      0.185636485 0.236343928  0.7854506 0.440566985
## NbStrat     -1.266028388 0.861235074 -1.4700149 0.155715201
## Melange     -0.537203283 0.773372382 -0.6946243 0.494561933
```

Modèle log-transformé

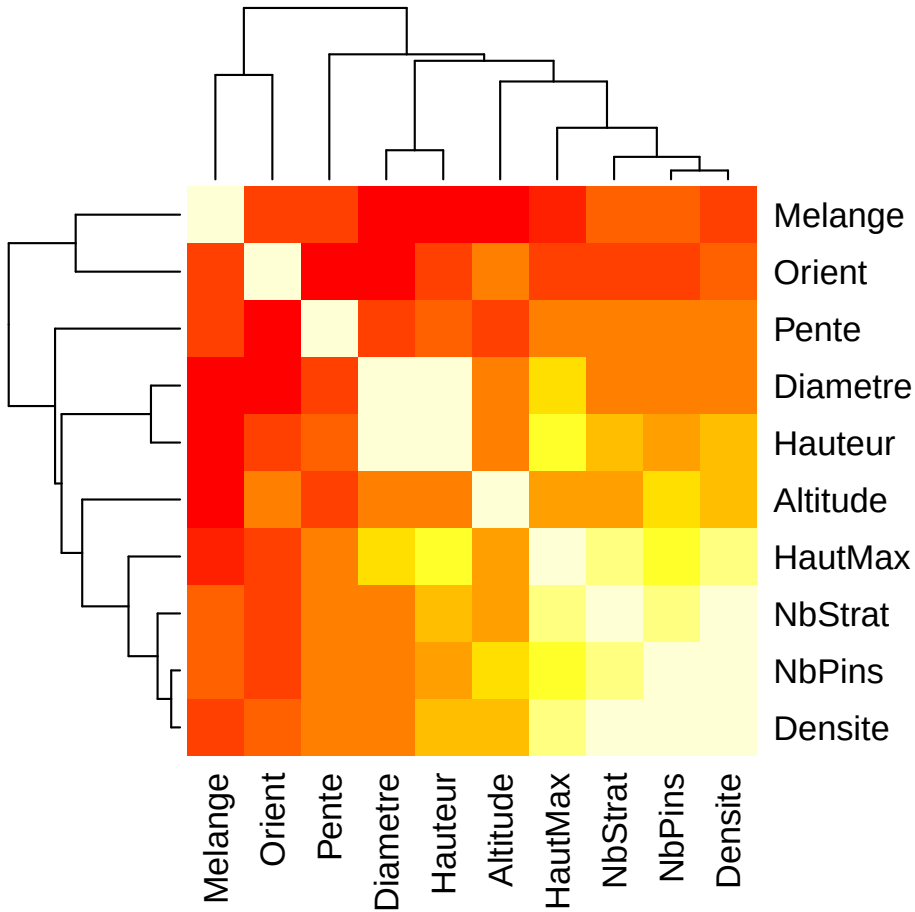
Test du modèle

```
anova(lm(log(NbNids)~1,chenilles), model)

## Analysis of Variance Table
##
## Model 1: log(NbNids) ~ 1
## Model 2: log(NbNids) ~ Altitude + Pente + NbPins + Hauteur + Diametre +
##     Densite + Orient + HautMax + NbStrat + Melange
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     32 49.596
## 2     22 15.039 10    34.557 5.0553 0.0007441 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Modèle log-transformé et prédicteurs normalisés

Prédicteurs corrélés

```
chenilles.scaled <- data.frame(scale(chenilles[,-ncol(chenilles)]),NbNids=chenilles
model.scaled <- lm(log(NbNids)~., chenilles.scaled)
```



Modèle log-transformé et prédicteurs normalisés I
Test du modèle

Constat

- les paramètres mal estimés (grande variance) sont ceux dont les corrélations sont élevées (densité, nb pins, nb strates, hauteur)
  $\rightsquigarrow$ S'il y a un effet, il est caché par la redondance

- les variables faiblement corrélées (pente, orientation, mélange) sont mieux estimées
  $\rightsquigarrow$ On peut conclure sur leur effets sur le nombre de nids.

$\rightsquigarrow$ attention, ce constat n'est possible que sur données normalisées car on compare les variances.

Modèle log-transformé et prédicteurs normalisés II

Test du modèle

```
summary(model.scaled)$coefficients
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) -0.81328069  0.1439262 -5.6506788 1.107569e-05
## Altitude    -0.58134027  0.2016866 -2.8823938 8.647574e-03
## Pente       -0.39151731  0.1595298 -2.4541957 2.250212e-02
## NbPins       0.71123631  0.9558617  0.7440786 4.647028e-01
## Hauteur     -1.38242983  0.5933018 -2.3300616 2.937577e-02
## Diametre     1.01583758  0.4500948  2.2569415 3.428080e-02
## Densite     -0.32361332  1.1283435 -0.2868039 7.769462e-01
## Orient      -0.03514548  0.1886212 -0.1863283 8.538947e-01
## HautMax      0.43658971  0.5558462  0.7854506 4.405670e-01
## NbStrat     -0.71719038  0.4878797 -1.4700149 1.557152e-01
## Melange     -0.13358672  0.1923151 -0.6946243 4.945619e-01
```

Modèle log-transformé et prédicteurs normalisés III

Test du modèle

```
anova(model.scaled)

## Analysis of Variance Table
##
## Response: log(NbNids)
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Altitude   1 14.1222 14.1222 20.6589 0.0001593 ***
## Pente      1  6.7095  6.7095  9.8152 0.0048376 **
## NbPins     1  1.4175  1.4175  2.0736 0.1639516
## Hauteur    1  1.8035  1.8035  2.6383 0.1185567
## Diametre   1  8.0480  8.0480 11.7732 0.0023866 **
## Densite    1  0.1353  0.1353  0.1979 0.6608026
## Orient     1  0.0385  0.0385  0.0563 0.8146664
## HautMax    1  0.0001  0.0001  0.0001 0.9910625
## NbStrat    1  1.9528  1.9528  2.8567 0.1051153
## Melange    1  0.3298  0.3298  0.4825 0.4945619
## Residuals 22 15.0389  0.6836
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Modèle log-transformé I

Comparaison de sous-modèles

```
M0 <- lm(log(NbNids)~1, chenilles)
M11 <- lm(log(NbNids)~Pente, chenilles)
anova(M0, M11)

## Analysis of Variance Table
##
## Model 1: log(NbNids) ~ 1
## Model 2: log(NbNids) ~ Pente
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     32 49.596
## 2     31 40.450  1    9.1464 7.0097 0.01263 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Modèle log-transformé II

Comparaison de sous-modèles

```
M12 <- lm(log(NbNids)~Altitude, chenilles)
anova(M0, M12)

## Analysis of Variance Table
##
## Model 1: log(NbNids) ~ 1
## Model 2: log(NbNids) ~ Altitude
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     32 49.596
## 2     31 35.474  1    14.122 12.341 0.001384 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Modèle log-transformé III

Comparaison de sous-modèles

```
M13 <- lm(log(NbNids)~Diametre, chenilles)
anova(M0, M13)

## Analysis of Variance Table
##
## Model 1: log(NbNids) ~ 1
## Model 2: log(NbNids) ~ Diametre
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     32 49.596
## 2     31 47.594  1    2.0025 1.3043 0.2622
```

Modèle log-transformé IV

Comparaison de sous-modèles

```
M21 <- lm(log(NbNids)~Altitude+Pente, chenilles)
anova(M0, M12, M21)

## Analysis of Variance Table
##
## Model 1: log(NbNids) ~ 1
## Model 2: log(NbNids) ~ Altitude
## Model 3: log(NbNids) ~ Altitude + Pente
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     32 49.596
## 2     31 35.474  1   14.1222 14.7288 0.0005951 ***
## 3     30 28.764  1    6.7095  6.9978 0.0128642 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Modèle log-transformé V

Comparaison de sous-modèles

```
M22 <- lm(log(NbNids)~Altitude+Diametre, chenilles)
anova(M0, M12, M22)

## Analysis of Variance Table
##
## Model 1: log(NbNids) ~ 1
## Model 2: log(NbNids) ~ Altitude
## Model 3: log(NbNids) ~ Altitude + Diametre
##   Res.Df    RSS Df Sum of Sq       F   Pr(>F)
## 1     32 49.596
## 2     31 35.474  1   14.1222 11.9877 0.001632 **
## 3     30 35.342  1    0.1322  0.1122 0.739932
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Modèle log-transformé VI

Comparaison de sous-modèles

```
M3 <- lm(log(NbNids)~Altitude+Diametre+Pente, chenilles)
anova(M22, M3)

## Analysis of Variance Table
##
## Model 1: log(NbNids) ~ Altitude + Diametre
## Model 2: log(NbNids) ~ Altitude + Diametre + Pente
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     30 35.342
## 2     29 28.742  1    6.5994 6.6586 0.0152 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Modèle log-transformé VII

Comparaison de sous-modèles

```
anova(M21, M3)

## Analysis of Variance Table
##
## Model 1: log(NbNids) ~ Altitude + Pente
## Model 2: log(NbNids) ~ Altitude + Diametre + Pente
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     30 28.764
## 2     29 28.742  1  0.022081 0.0223 0.8824
```

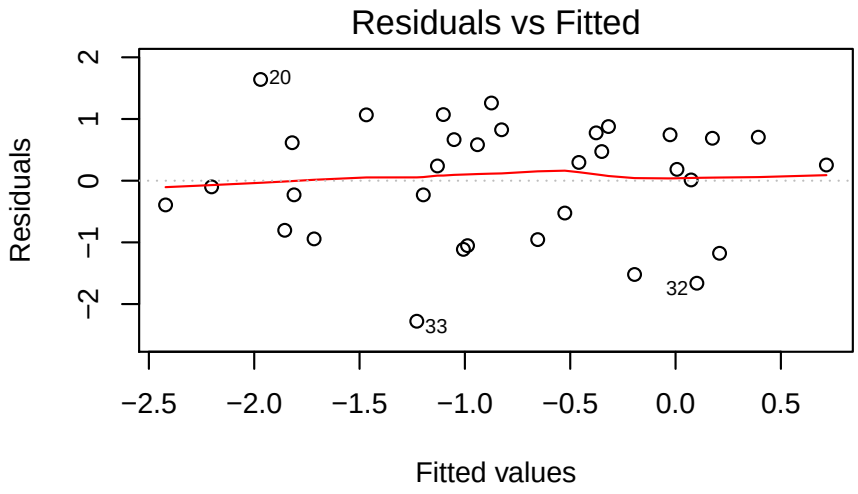
Modèle final I

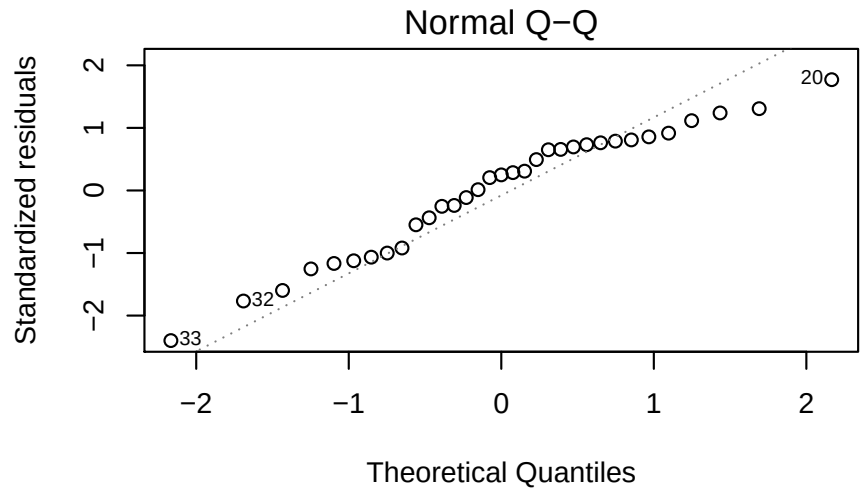
```
summary(M21)


##
## Call:
## lm(formula = log(NbNids) ~ Altitude + Pente, data = chenilles)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2783 -0.8041  0.2387  0.7057  1.6412
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.225158   1.836220   3.935 0.000457 ***
## Altitude    -0.004717   0.001351  -3.491 0.001512 **
## Pente       -0.063155   0.023874  -2.645 0.012864 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9792 on 30 degrees of freedom
## Multiple R-squared:   0.42,Adjusted R-squared:  0.3814
## F-statistic: 10.86 on 2 and 30 DF,  p-value: 0.0002826
```

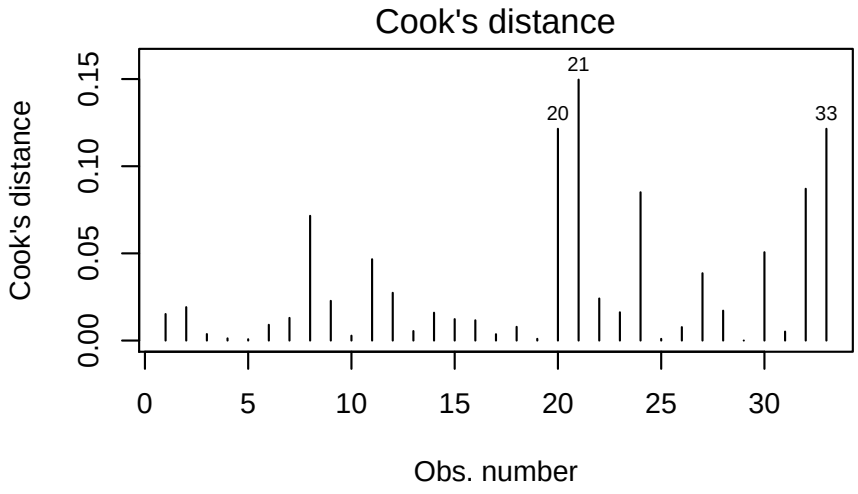
Modèle final II

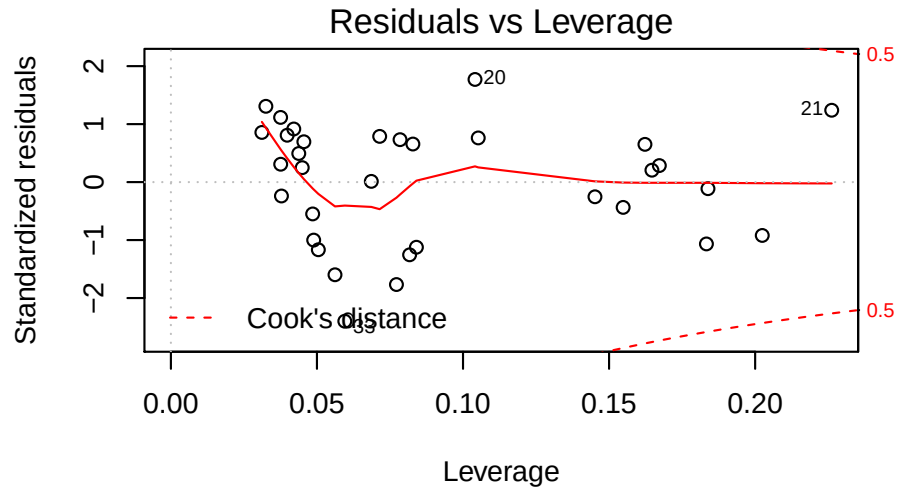
```
par(mfrow=c(2,2)); plot(M21, which=c(1,2,4,5))
```

Modèle final III







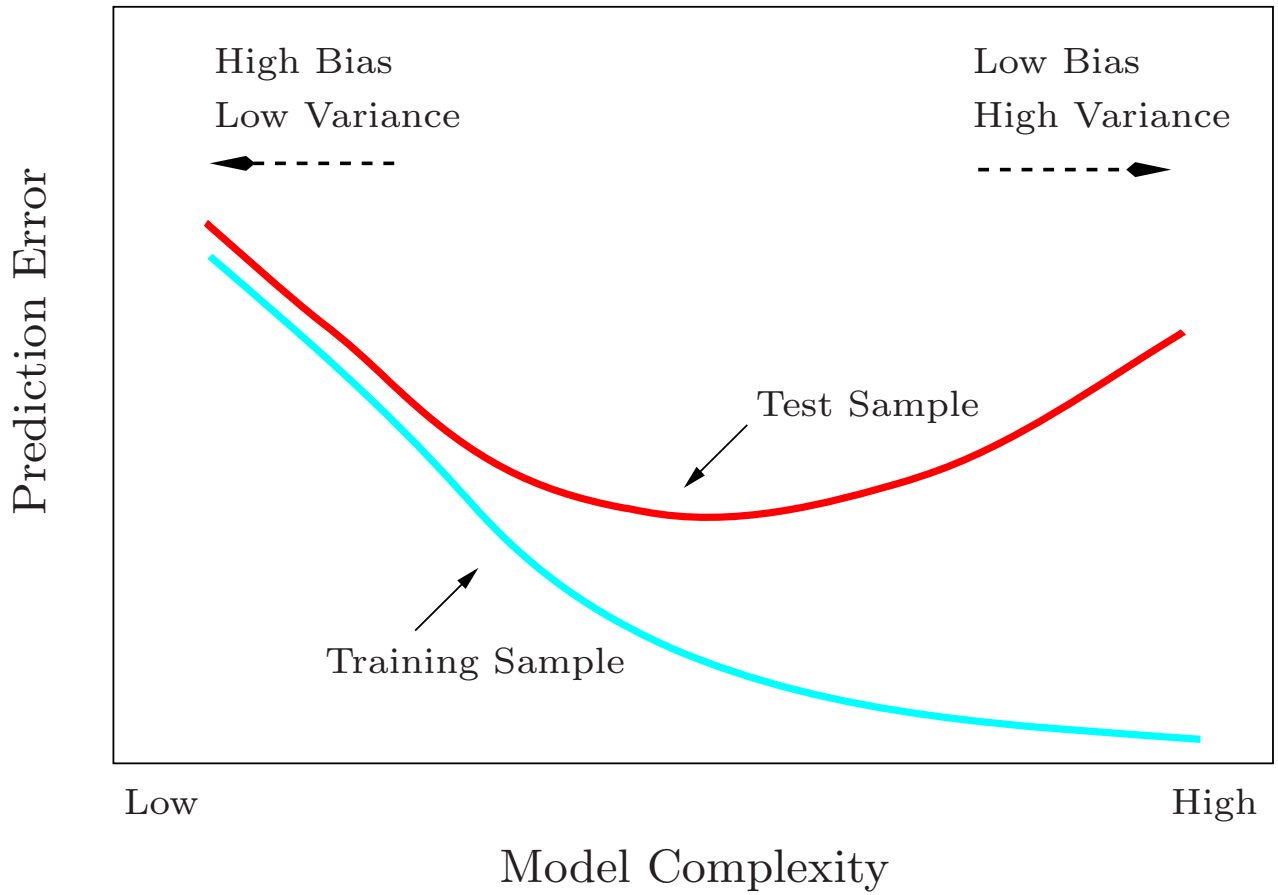


Plan

Motivation : compromis Biais/Variance

À un nouveau point $X = x$,

$$\mathrm{err}(\hat{f}(x)) = \underbrace{\sigma^2}_{\substack{\text{incompressible} \\ \text{error}}} + \underbrace{\mathsf{bias}^2(\hat{f}(x)) + \mathbb{V}(\hat{f}(x))}_{\mathrm{MSE}(\hat{f}(x))}.$$



Cas de la régression linéaire

Erreur de prédiction

On peut montrer pour $\mathbf{X}$ fixé que

$$\hat{\mathrm{err}}(\mathbf{X}\hat{\boldsymbol{\beta}}^{\mathrm{ols}}) = \sigma^2 \frac{(p+1)}{n} + \sigma^2.$$

Thérorème de Gauss-Markov

$\hat{Y} = X^\intercal \hat{\boldsymbol{\beta}}^{\mathrm{ols}}$ est le meilleur modèle (i.e. de plus faible variance) pour les estimateurs sans biais de $\boldsymbol{\beta}$.

$\rightsquigarrow$ Y a-t-il des situations où l'on a intérêt à utiliser un **estimateur biaisé de plus faible variance** ?

Sélection de variable

Problématique

En augmentant le nombre de variables

- on intègre de plus en plus d'information dans le modèle ;
- on augmente le nombre de paramètres à estimer et $\mathbb{V}(\hat{Y}_i) \nearrow$.

Idée

On recherche un (petit) ensemble $\mathcal{S}$ de k variables parmi p telles que

$$Y \approx X_{\mathcal{S}}^T \hat{\boldsymbol{\beta}}_{\mathcal{S}}.$$

Ingrédients

Pour trouver un compromis, on a besoin

1. d'un critère pour évaluer la qualité du modèle;
2. d'un algorithme pour déterminer les k variables optimisant le critère.

Sélection de variable

Problématique

En augmentant le nombre de variables

- on intègre de plus en plus d'information dans le modèle ;
- on augmente le nombre de paramètres à estimer et $\mathbb{V}(\hat{Y}_i) \nearrow$.

Idée

On recherche un (petit) ensemble $\mathcal{S}$ de k variables parmi p telles que

$$Y \approx X_{\mathcal{S}}^T \hat{\boldsymbol{\beta}}_{\mathcal{S}}.$$

Ingrédients

Pour trouver un compromis, on a besoin

1. d'un critère pour évaluer la qualité du modèle;
2. d'un algorithme pour déterminer les k variables optimisant le critère.

Critères pénalisés

Principe général

Idée

Plutôt que d'estimer l'erreur de prédiction par l'erreur de test, on estime de combien l'erreur d'entraînement sous-estime la vraie erreur.

Forme générique des critères

Sans ajuster d'autres modèles, on calcule

$$\hat{\text{err}} = \text{err}_{\mathcal{D}} + \text{"optimisme"}.$$

Remarques

- beaucoup moins coûteux que la validation croisée
- revient à "pénaliser"les modèles trop complexes.

Critères pénalisés

Principe général

Idée

Plutôt que d'estimer l'erreur de prédiction par l'erreur de test, on estime de combien l'erreur d'entraînement sous-estime la vraie erreur.

Forme générique des critères

Sans ajuster d'autres modèles, on calcule

$$\hat{\text{err}} = \text{err}_{\mathcal{D}} + \text{"optimisme"}.$$

Remarques

- beaucoup moins coûteux que la validation croisée
- revient à "pénaliser"les modèles trop complexes.

Critères pénalisés

Les plus populaires en régression

Soit k la dimension du modèle (le nombre de prédicteurs utilisés).

Critères pour le modèle de régression linéaire σ connue

On choisit le modèle de taille k minimisant un des critères suivants.

- C_p **de Mallows**

$$C_p = \frac{\text{err}_{\mathcal{D}}}{\sigma^2} - n + 2\frac{k}{n}$$

- **Akaïke Information Criteria** équivalent au C_p quand σ est connue

$$\text{AIC} = -2\text{loglik} + 2k = \frac{n}{\sigma^2}\text{err}_{\mathcal{D}} + 2k.$$

- **Bayesian Information Criterion**

$$\text{BIC} = -2\text{loglik} + k\log(n) = \frac{n}{\sigma^2}\text{err}_{\mathcal{D}} + k\log(n).$$

Critères pénalisés

Les plus populaires en régression

Soit k la dimension du modèle (le nombre de prédicteurs utilisés).

Critères pour le modèle de régression linéaire σ inconnue

On choisit le modèle de taille k minimisant un des critères suivants.

- C_p **de Mallows** σ estimée par l'estimateur sans biais $\hat{\sigma}$

$$C_p = \frac{\mathrm{err}_{\mathcal{D}}}{\hat{\sigma}^2} - n + 2\frac{k}{n}$$

- **Akaïke Information Criteria** σ^2 estimée par $\mathrm{err}_{\mathcal{D}}/n$

$$\mathrm{AIC} = -2\mathrm{loglik} + 2k = n\log(\mathrm{err}_{\mathcal{D}}) + 2k.$$

- **Bayesian Information Criterion** σ^2 estimée par $\mathrm{err}_{\mathcal{D}}/n$

$$\mathrm{BIC} = -2\mathrm{loglik} + k\log(n) = n\log(\mathrm{err}_{\mathcal{D}}) + k\log(n).$$

C_p/AIC: preuve

L'idéal serait de minimiser l'espérance de la distance entre le vrai modèle $\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\mu}$ et celui de l'OLS. La distance se décompose comme suit:

$$\begin{aligned}\|\boldsymbol{\mu} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\text{ols}}\|^2 =& \|\mathbf{y} - \varepsilon - \mathbf{P}_\mathbf{X}\mathbf{y}\|^2 \\ =& \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \|\varepsilon\|^2 - 2\varepsilon^\intercal(\mathbf{y} - \mathbf{P}_\mathbf{X}\mathbf{y}) \\ =& n\text{err}_\mathcal{D} + \|\varepsilon\|^2 - 2\varepsilon^\intercal(\mathbf{I} - \mathbf{P}_\mathbf{X})(\boldsymbol{\mu} + \varepsilon) \\ =& n\text{err}_\mathcal{D} - \|\varepsilon\|^2 + 2\varepsilon^\intercal\mathbf{P}_\mathbf{X}\varepsilon - 2\varepsilon^\intercal(\mathbf{I} - \mathbf{P}_\mathbf{X})\boldsymbol{\mu}\end{aligned}$$

En espérance, on a

- $\mathbb{E}[\|\varepsilon\|^2] = n\sigma^2$
- $\mathbb{E}[\varepsilon^\intercal(\mathbf{I} - \mathbf{P}_\mathbf{X})\boldsymbol{\mu}] = 0$
- $\mathbb{E}[2\varepsilon^\intercal\mathbf{P}_\mathbf{X}\varepsilon] = 2\mathbb{E}[\text{trace}(\varepsilon^\intercal\mathbf{P}_\mathbf{X}\varepsilon)] = 2\text{trace}(\mathbf{P}_\mathbf{X})\sigma^2$

Si k est la dimension de l'espace où l'on projette, on trouve

$$\mathbb{E}\|\boldsymbol{\mu} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\text{ols}}\|^2 = n\text{err}_\mathcal{D} - n\sigma^2 + 2k\sigma^2$$

Il suffit alors de diviser par $n\sigma^2$.

Plan

Recherche exhaustive (best-subset)

Algorithme

Pour $k = 0, \ldots, p$, trouver le sous-ensemble de k variables qui donne le plus petit SCR parmi les 2^k modèles.

Propriétés

- Peut être généralisé à d'autres critères (R^2, AIC, BIC...)
- Existence d'un algorithme efficace ("Leaps and Bound")
- impossible dès que $p > 30$.

Sélection avant (Forward regression)

Algorithme

1. Commencer avec $\mathcal{S} = \emptyset$
2. À l'étape k trouver la variable qui ajoutée à $\mathcal{S}$ donne le meilleur modèle

2'. À l'étape k trouver le meilleur modèle lorsqu'une variable est ajoutée ou enlevée.

3 etc. jusqu'au modèle à p variables

Propriétés

- le meilleur modèle est compris en terme de SCR ou R^2, AIC, BIC...
- approprié lorsque p est grand
- biais important, mais variance/complexité contrôlée.
- algorithme dit "glouton" (greedy)

Sélection avant Pas à pas (Forward-stepwise)

Algorithme

1. Commencer avec $\mathcal{S} = \emptyset$
2. À l'étape k trouver la variable qui ajoutée à $\mathcal{S}$ donne le meilleur modèle

2'. À l'étape k trouver le meilleur modèle lorsqu'une variable est ajoutée ou enlevée.

3 etc. jusqu'au modèle à p variables

Propriétés

- le meilleur modèle est compris en terme de SCR ou R^2, AIC, BIC...
- approprié lorsque p est grand
- biais important, mais variance/complexité contrôlée.
- algorithme dit "glouton" (greedy)

Sélection arrière

Algorithm

1. Commencer avec le modèle plein $\mathcal{S} = \{1, \ldots, p\}$
2. À l'étape k, enlever la variable ayant le moins d'influence sur l'ajustement.
3. etc. jusqu'au modèle nul.

Propriétés

- le meilleur modèle est compris en terme de SCR ou R^2, AIC, BIC...
- ne fonctionne pas si $n < p$
- algorithme dit "glouton" (greedy)

Plan

Recherche exhaustive I

```
library(leaps)
```

On calcule tous les modèles possibles

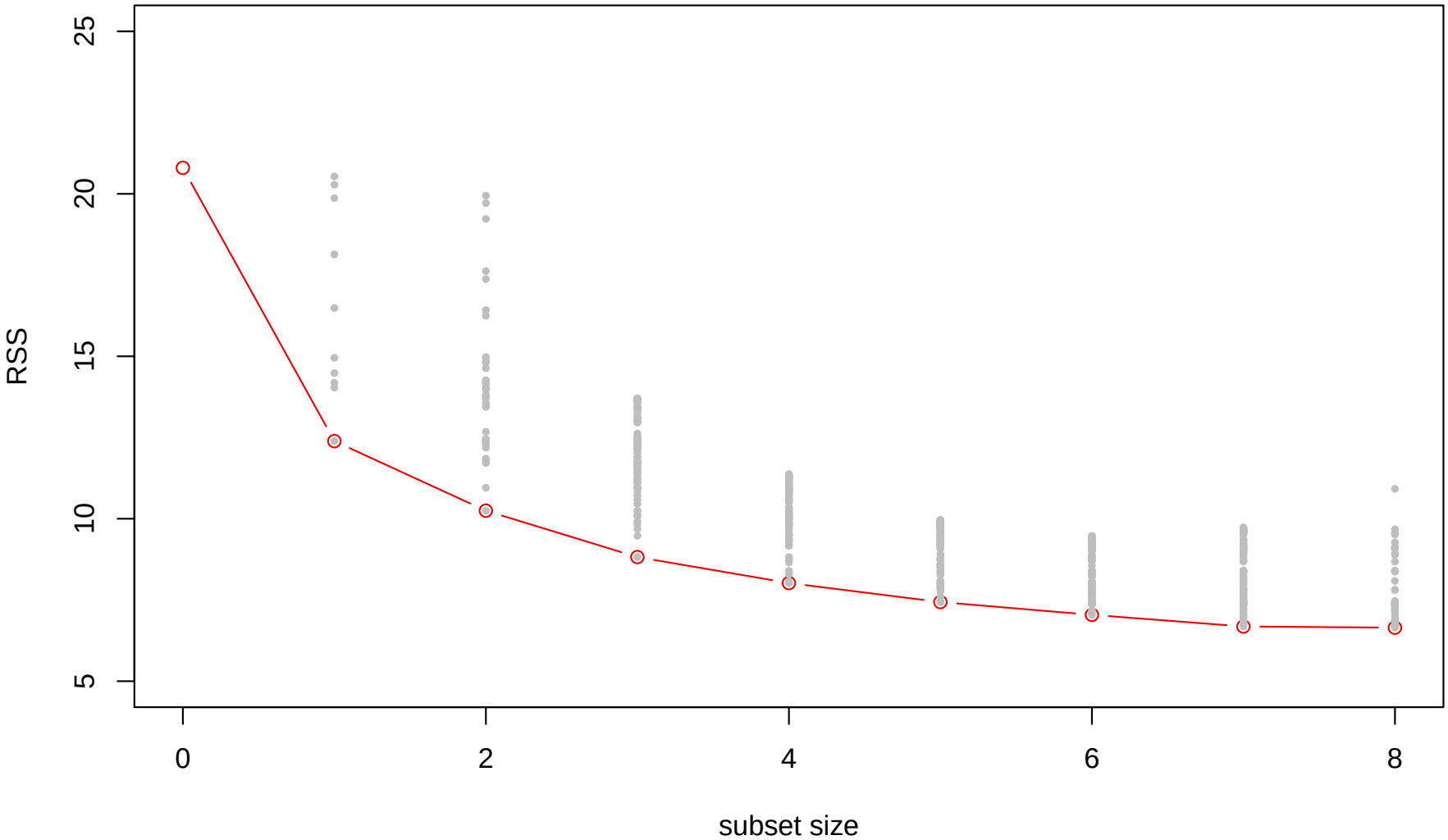
```
out <- regsubsets(NbNids ~ . , data=chenilles,
                  nbest=100, really.big=TRUE)
bss <- summary(out)
```

Extraction de la taille et des SCR. Ajout du modèle nul (juste l'intercept)

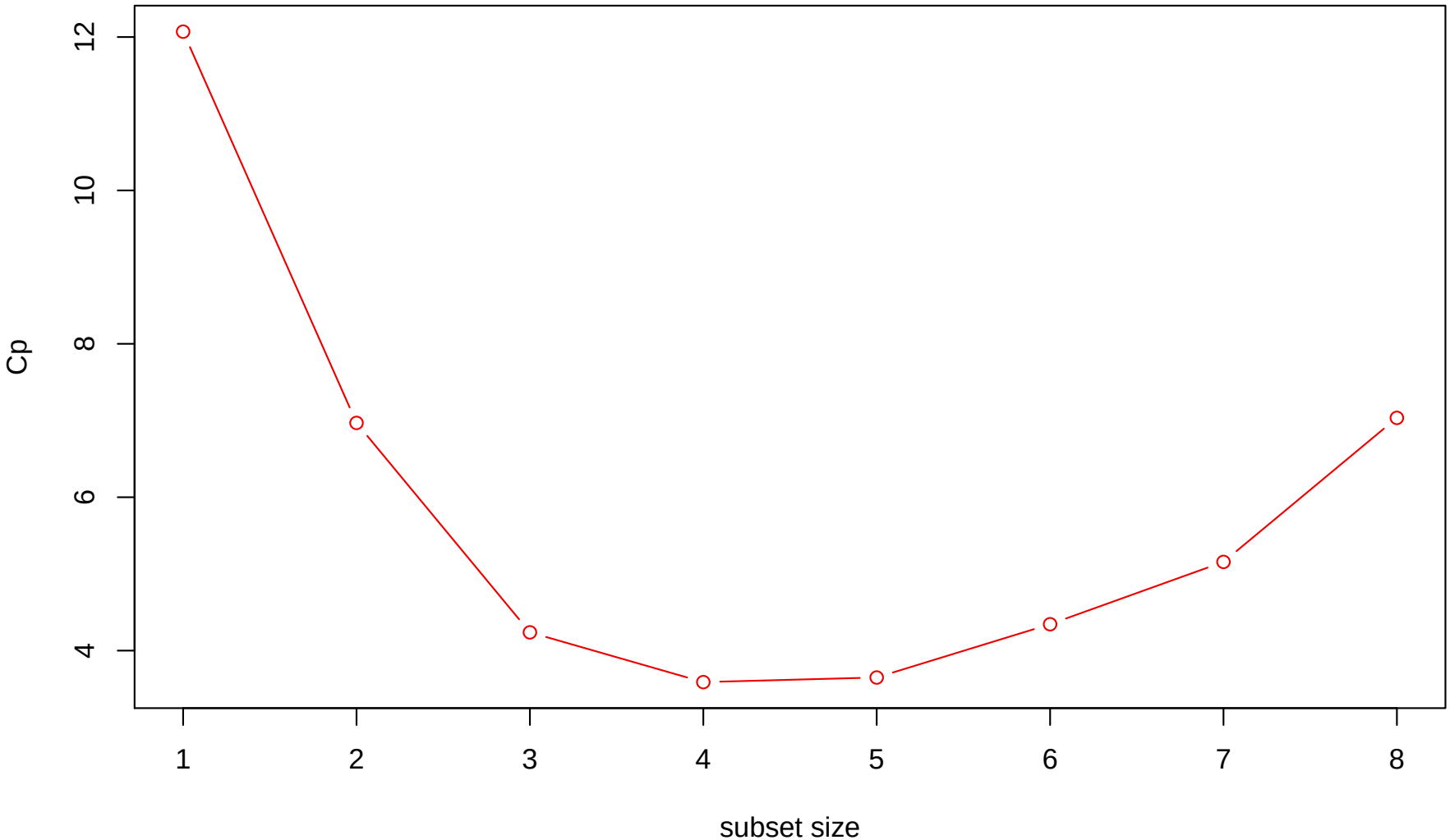
```
bss.size  <- as.numeric(rownames(bss$which))
intercept <- lm(NbNids ~ 1, data=chenilles)
bss.best.rss  <- c(sum(resid(intercept)^2), tapply(bss$rss  , bss.size, min))
```

```
plot(0:8, bss.best.rss, ylim=c(5, 25), type="b",
     xlab="subset size", ylab="RSS", col="red2" )
points(bss.size, bss$rss, pch=20, col="gray", cex=0.7)
```

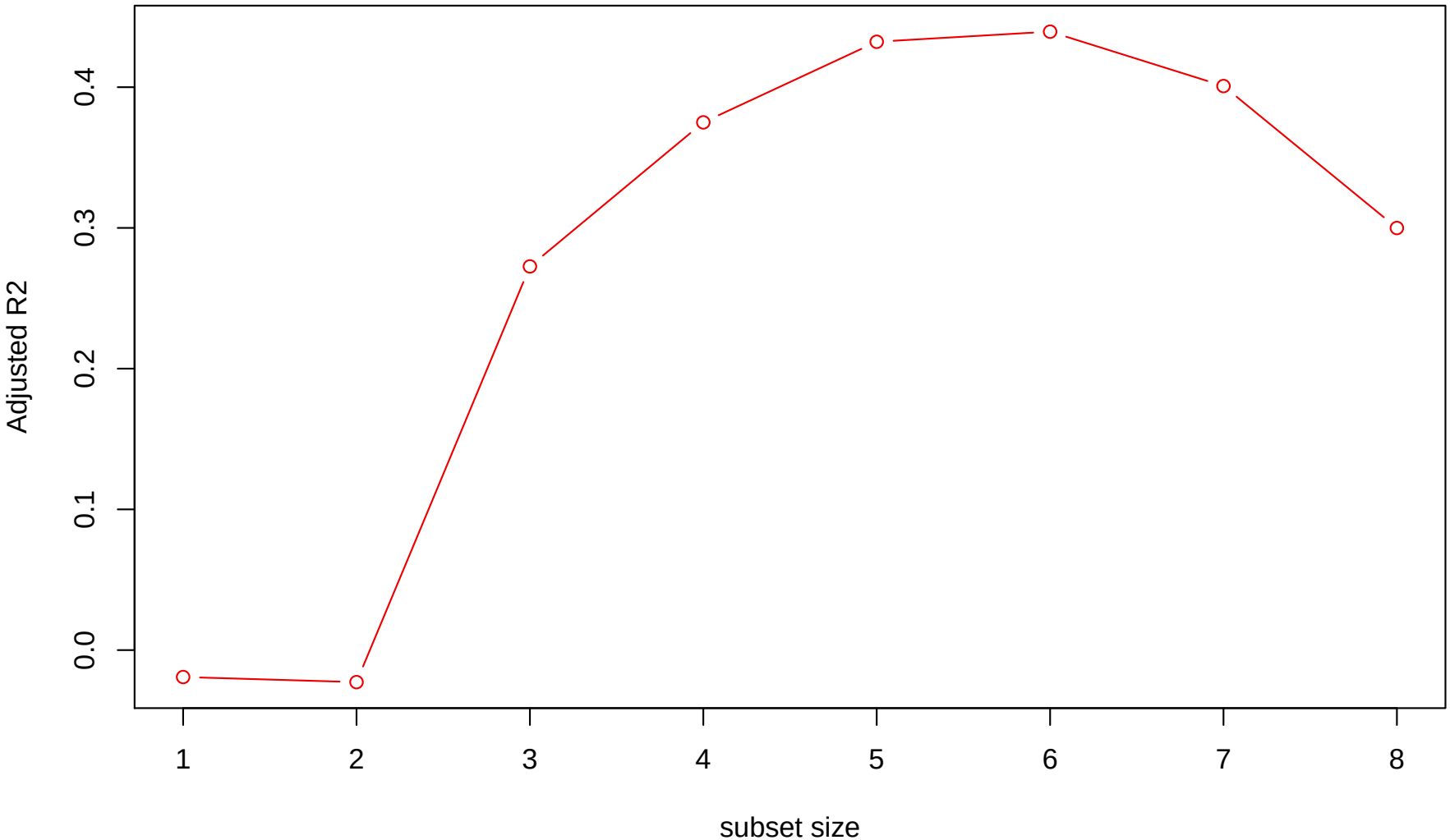
Recherche exhaustive II



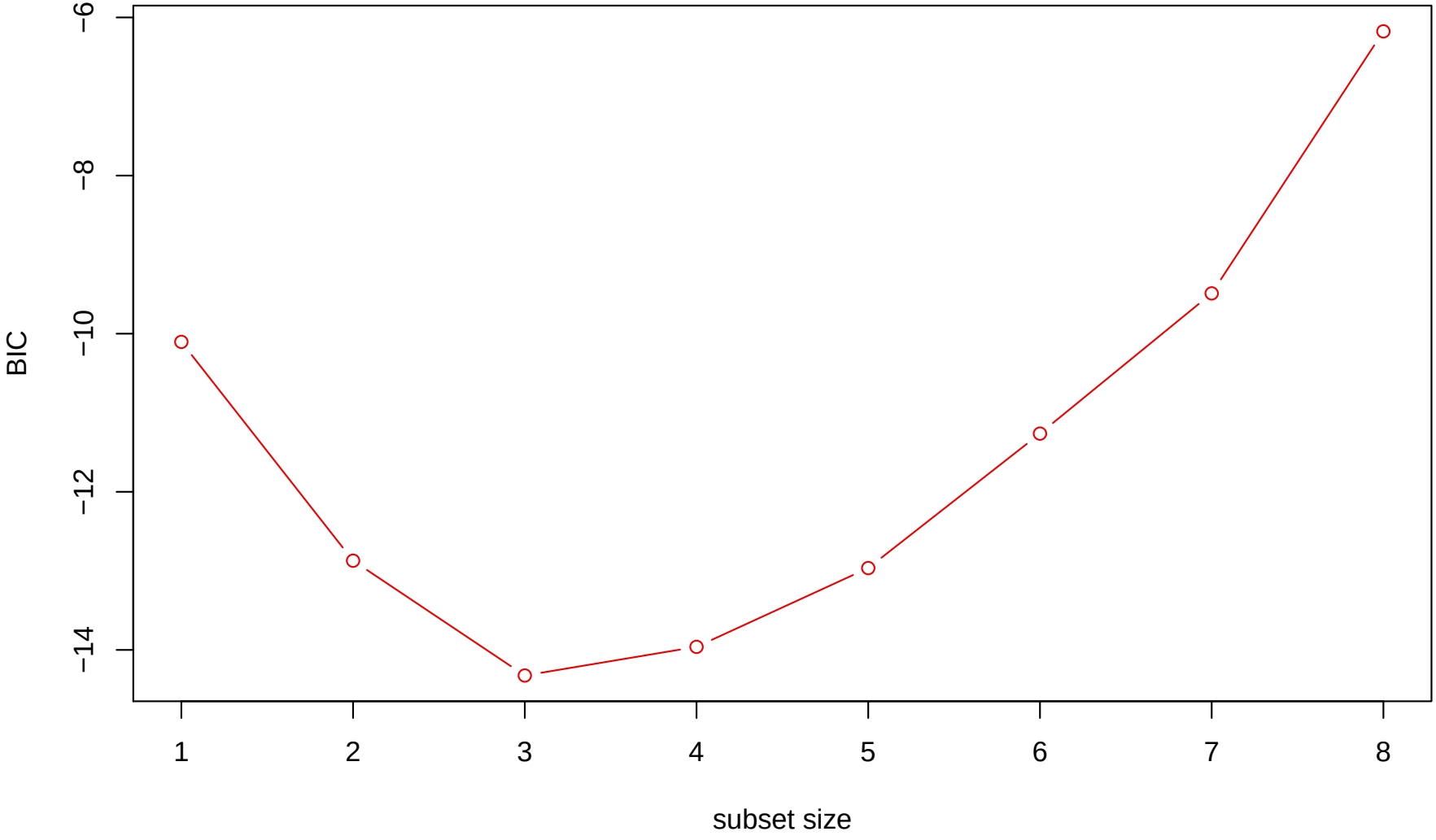
Recherche exhaustive III



Recherche exhaustive VI



Recherche exhaustive V



Forward-Stepwise dans R (I)

Création du modèle nul et du modèle plein

```
null  <- lm(NbNids ~ 1, data=chenilles)
full  <- lm(NbNids ~ ., data=chenilles)
```

Création de l'ensemble des modèles à parcourir ("scope")

```
lower <- ~1
upper <- ~Altitude+Pente+NbPins+Hauteur+Diametre+Densite+Orient+HautMax+NbStrat+Mel
scope <- list(lower=lower,upper=upper)
```

Stepwise avec AIC: forward, backward, both

```
fwd  <- step(null, scope, direction="forward" , trace=FALSE)
bwd  <- step(full, scope, direction="backward", trace=FALSE)
both <- step(null, scope, direction="both"    , trace=FALSE)
```

Forward regression

```
fwd
```

```
##
## Call:
## lm(formula = NbNids ~ NbStrat + Altitude + Pente + Densite +
##     Orient, data = chenilles)
##
## Coefficients:
## (Intercept)      NbStrat     Altitude        Pente      Densite
##    7.898605    -1.286964    -0.002612    -0.034727     0.660826
##      Orient
##   -0.770365
```

```
fwd$anova
```

```
##         Step Df  Deviance Resid. Df Resid. Dev       AIC
## 1             NA        NA        32  20.800152 -13.23106
## 2  + NbStrat -1 8.4101815        31  12.389970 -28.32747
## 3 + Altitude -1 2.1421673        30  10.247803 -32.59166
## 4    + Pente -1 1.4271671        29   8.820636 -35.54065
## 5  + Densite -1 0.7991552        28   8.021480 -36.67469
## 6   + Orient -1 0.5851813        27   7.436299 -37.17443
```

Backward regression

```
bwd
```

```
##
## Call:
## lm(formula = NbNids ~ Altitude + Pente + Hauteur + Diametre +
##     NbStrat, data = chenilles)
##
## Coefficients:
## (Intercept)     Altitude        Pente      Hauteur     Diametre
##    5.998179    -0.002292    -0.033809    -0.521596     0.124145
##     NbStrat
##   -0.384935
```

```
bwd$anova
```

```
##         Step Df     Deviance Resid. Df Resid. Dev       AIC
## 1            NA           NA        22   6.636926 -30.92734
## 2 - Densite  1 0.0002957245        23   6.637222 -32.92587
## 3 - HautMax  1 0.0101799535        24   6.647402 -34.87529
## 4  - Orient  1 0.0367720062        25   6.684174 -36.69324
## 5 - Melange  1 0.4016781476        26   7.085852 -36.76745
## 6  - NbPins  1 0.3522123842        27   7.438064 -37.16660
```

Stepwise regression

```
both
```

```
##
## Call:
## lm(formula = NbNids ~ NbStrat + Altitude + Pente + Densite +
##     Orient, data = chenilles)
##
## Coefficients:
## (Intercept)      NbStrat     Altitude        Pente      Densite
##    7.898605    -1.286964    -0.002612    -0.034727     0.660826
##      Orient
##   -0.770365
```

```
both$anova
```

```
##          Step Df  Deviance Resid. Df Resid. Dev       AIC
## 1             NA        NA        32  20.800152 -13.23106
## 2  + NbStrat -1 8.4101815        31  12.389970 -28.32747
## 3 + Altitude -1 2.1421673        30  10.247803 -32.59166
## 4    + Pente -1 1.4271671        29   8.820636 -35.54065
## 5  + Densite -1 0.7991552        28   8.021480 -36.67469
## 6   + Orient -1 0.5851813        27   7.436299 -37.17443
```

Stepwise en R: modification pour le BIC

Modèle plus parcimonieux

```
BIC <- step(null, scope, k=log(n <- nrow(chenilles)), trace=FALSE)
BIC
```

```
##
## Call:
## lm(formula = NbNids ~ NbStrat + Altitude + Pente, data = chenilles)
##
## Coefficients:
## (Intercept)      NbStrat     Altitude        Pente
##    5.711169    -0.598567    -0.002148    -0.030582
```