Regularization Methods for Linear Regression
Multiple Linear Regression

M1 Math et Interactions – UEVE/ENSIIE

Autumn semester 2016

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

école nationale supérieure d'informatique pour l'industrie et l'entreprise

université evry val-d'essonne

A couple of references

- 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

Outline

Outline

Multiple Regression

General Goals I

Idea/Principle

Explain the variations

- of a **quantitative** variable Y,
- by **several** quantitative variables $x = (x_1, x_2, \ldots, x_p)$

Vocabulary

- Y is the **response** variable, or **output**
- x_j are the **predictive** variables, **covariates**, **regressors** or **predictors**

Multiple Regression

General Goals II

Examples

- pesticide rate in pike = f(age, square of the age)
- diabetes progression = f(age, body mass index, blood pressure, concentration of various proteins)
- stock value at t = f(other stocks value at $t-1$)
- plant yield = f(gene expression profiles)
- [HIV] at inclusion = f(variation of genotype)

$\rightsquigarrow$ potentially **many** predictors...

Multiple Linear Regression
Model

Assume the true relationship between Y and x is linear:

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

- β_0 is the **intercept** (**constant** term)
- β_j are the **regression coefficients**
- ε is the **noise** (random)
  - $\rightsquigarrow$ uncertainties, individual variation, unexplained factor(s)

Minimal set of hypotheses

Centered with fixed and finite variance:

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

Multiple Linear Regression

Sampling and Matrix formulation

Collecting Data / Random sampling

Let $\{(Y_i, x_i)\}_{i=1}^n$ be a n-sample with $Y_i \in \mathbb{R}$ and $x_i \in \mathbb{R}^p$. We have

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

with $\{\varepsilon_i\}_{i=1}^n$ independent, identically distributed.

Notations

- Let $Y = (Y_1, \ldots, Y_n)^\intercal \in \mathbb{R}^n$ be the random vector of observations of the response variable,
- $\mathbf{y} = (y_1, \ldots, y_n)^\intercal \in \mathbb{R}^n$ the associated vector of observed values,
- $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^\intercal$ the vector of observed values associated with the jth predictor.
- $\boldsymbol{\varepsilon} = (\varepsilon_i, \ldots, \varepsilon_n)^\intercal$ the vector of noise (observed).

Multiple Linear Regression

Matrix formulation

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

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

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

To sum up,

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

Multiple Linear Regression

Matrix formulation

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

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

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

To sum up,

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

Multiple Linear Regression
Matrix formulation

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

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

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

To sum up,

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

Multiple Linear Regression

Matrix formulation

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

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

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

To sum up,

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

Multiple Linear Regression I

Linearity with respect to the parameters

Th linear model is **linear w.r.t the parameters** (not necessarily w.r.t x_j)

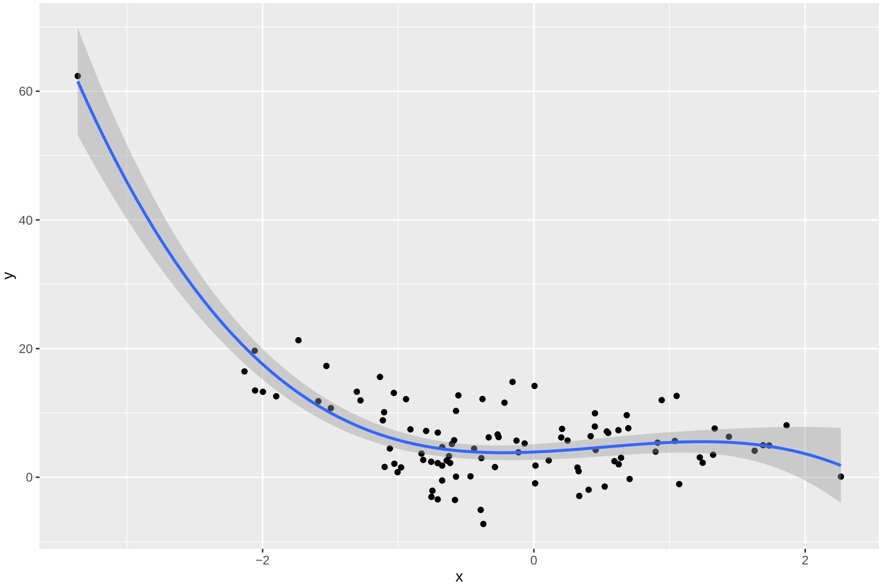
Example: polynomial regression

A multiple linear regression model that ca be plotted in 2D

```
##true paramters : third-order polynome
beta <- c(3, 1, 2, -1)
sigma <- 5
p <- length(beta)

## drawing random data / simulation
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))
```

Multiple Linear Regression II

Linearity with respect to the parameters



Multiple Linear Regression

To sum up

Statistical Goals

1. Estimate the parameters $\boldsymbol{\beta}$ and σ^2
2. Test the nullity of each coefficients $\{\beta_j\}_{j=1}^p$, i.e. the role of each preidctor x_j regarding the response
3. Predict Y_0 given a new observation x_0
4. Test the glocal relevance of the model
5. When p is large, control the model complexity

Outline

Outline

Orthogonal subspaces

Definition (Orthogonal vector subspaces)

- Subspaces V and W are orthogonal if all vectors ib=n V are orthogonal to all vectors in W.
- The set of all vectors orthogonal to V is call orthogonal of V and is denoted by $V^{\perp}$.

Theorem

Let V be a linear subspace of $\mathbb{R}^n$, then any vector in $\mathbb{R}^n$ decomposes in a unique way as a sum of two vectors from V and $V^{\perp}$.

Orthogonal Projection

Definition (Projection orthogonal)

Let V be a subspace of $\mathbb{R}^n$, the linear mapping associating to $\mathbf{u} \in \mathbb{R}^n$ the vector $\mathbf{u}^\star \in V$ such that $\mathbf{u} - \mathbf{u}^\star$ belongs to $V^\perp$ is the orthogonal projection of $\mathbf{u}$ in V.

Definition (orthogonal projector and matrix)

Let $\mathbf{X}$ be a matrix $n \times p$ with full rank, such that $n > p$.

- The orthogonal projection of $\mathbf{u} \in \mathbb{R}^n$ in the image of $\mathbf{X}$ is

$$\mathrm{proj}_{\mathbf{X}}(\mathbf{u}) = \underbrace{\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top}_{P_X}\mathbf{u}.$$

- The orthogonal projection of $\mathbf{u} \in \mathbb{R}^n$ in the kernel of $\mathbf{X}$ is

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

Orthogonal Projection

Definition (Projection orthogonal)

Let V be a subspace of $\mathbb{R}^n$, the linear mapping associating to $\mathbf{u} \in \mathbb{R}^n$ the vector $\mathbf{u}^\star \in V$ such that $\mathbf{u} - \mathbf{u}^\star$ belongs to $V^\perp$ is the orthogonal projection of $\mathbf{u}$ in V.

Definition (orthogonal projector and matrix)

Let $\mathbf{X}$ be a matrix $n \times p$ with full rank, such that $n > p$.

- The orthogonal projection of $\mathbf{u} \in \mathbb{R}^n$ in the image of $\mathbf{X}$ is

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

- The orthogonal projection of $\mathbf{u} \in \mathbb{R}^n$ in the kernel of $\mathbf{X}$ is

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

Orthogonal Projection

Definition (Projection orthogonal)

Let V be a subspace of $\mathbb{R}^n$, the linear mapping associating to $\mathbf{u} \in \mathbb{R}^n$ the vector $\mathbf{u}^\star \in V$ such that $\mathbf{u} - \mathbf{u}^\star$ belongs to $V^\perp$ is the orthogonal projection of $\mathbf{u}$ in V.

Definition (orthogonal projector and matrix)

Let $\mathbf{X}$ be a matrix $n \times p$ with full rank, such that $n > p$.

- The orthogonal projection of $\mathbf{u} \in \mathbb{R}^n$ in the image of $\mathbf{X}$ is

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

- The orthogonal projection of $\mathbf{u} \in \mathbb{R}^n$ in the kernel of $\mathbf{X}$ is

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

Outline

Gradient

Definition (gradient vector)

Let f be a mapping from $\mathbb{R}^p$ to $\mathbb{R}$. The gradient (vector) of f is the vector of partial derivatives

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

From this definition, we derive in particular the differentiation w.r.t. a vector of a linear form, a linear mapping and a quadratic form.

Differentiation with respect to a vector

Proposition (Differentiation with respect to a vector)

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

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

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{A}\mathbf{x} = \mathbf{A}$$

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

Moreover, if $\mathbf{S}$ *is symmetric, then*

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

Outline

Random vectors, expectancy, variance-covariance matrix

Let $X = (X_1, \dots, X_p)^\intercal$ be a vector of random variables the joint distribution of which is $f(\mathbf{x}) = f(x_1, \dots, x_p)$.

Definition (Expectancy)

The expectancy of the random vector X is the vector of expectancy of each component:

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

Definition (Variance)

The variance of X is the (variance-covariance) matrixdefined by

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

Properties

Let $\mathbf{A}$ be a $m \times p$ constant matrix, then

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

Random vectors, expectancy, variance-covariance matrix

Let $X = (X_1, \ldots, X_p)^\intercal$ be a vector of random variables the joint distribution of which is $f(\mathbf{x}) = f(x_1, \ldots, x_p)$.

Definition (Expectancy)

The expectancy of the random vector X is the vector of expectancy of each component:

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

Definition (Variance)

The variance of X is the (variance-covariance) matrixdefined by

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

Properties

Random vectors, expectancy, variance-covariance matrix

Let $X = (X_1, \dots, X_p)^\intercal$ be a vector of random variables the joint distribution of which is $f(\mathbf{x}) = f(x_1, \dots, x_p)$.

Definition (Expectancy)

The expectancy of the random vector X is the vector of expectancy of each component:

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

Definition (Variance)

The variance of X is the (variance-covariance) matrixdefined by

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

Properties

Let $\mathbf{A}$ be a $m \times p$ constant matrix, then

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

Gaussian Vector

Definition

The random vector $X \in \mathbb{R}^p$ has a multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$ if the probability density function of an observation $\mathbf{x}$ is given by

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

We denote $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ a Gaussian vector in $\mathbb{R}^p$.

Log-likelihood

Let $\mathbf{X}$ be the $n \times p$ matrix, the rows of which, denoted by $\mathbf{x}_i$, are indenpendent realization of X.

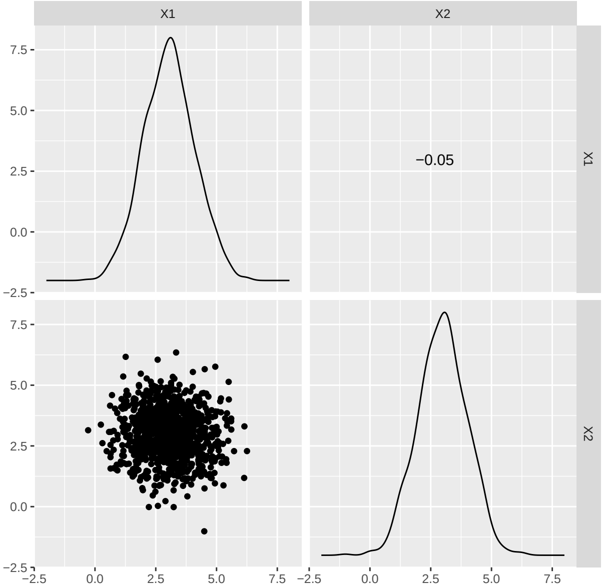
$$\log L(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X}) = -\frac{np}{2} \log(2\pi) - \frac{n}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})$$

Gaussian Vector
Bivariate Examples (I)

```
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)
```

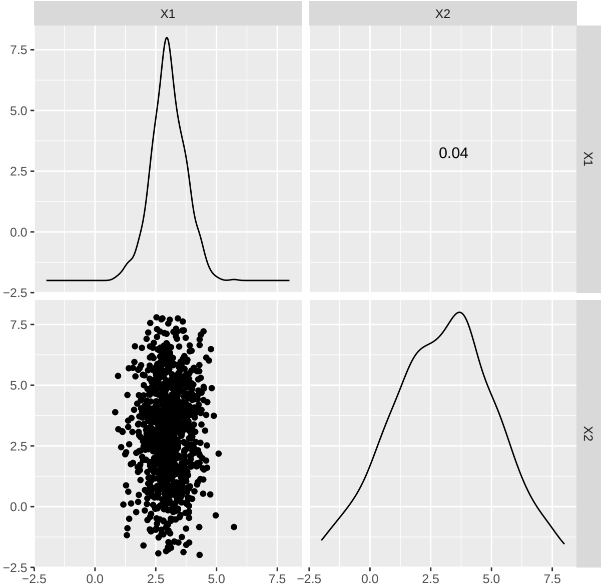
Gaussian Vector

Bivariate Examples (II)



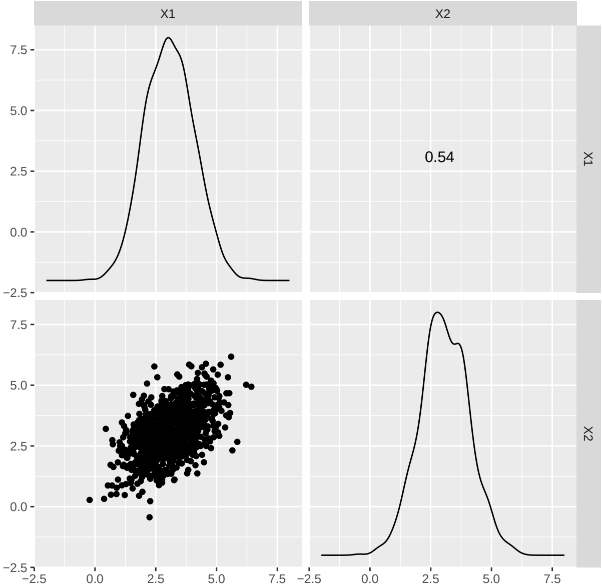
Gaussian Vector

Bivariate Examples (III)



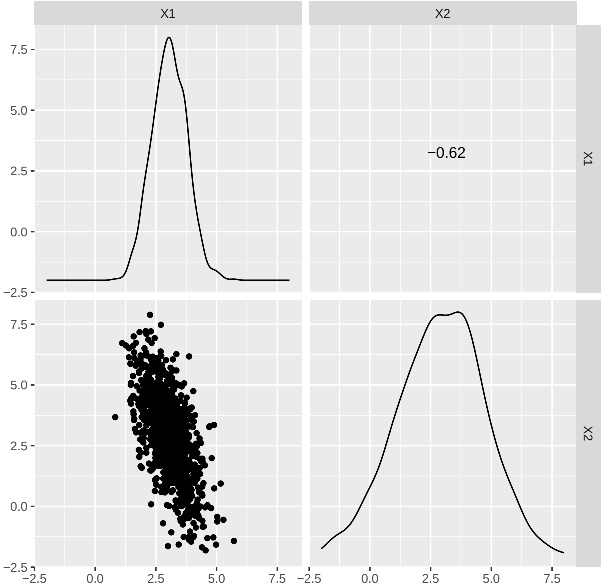
Gaussian Vector

Bivariate Examples (IV)



Gaussian Vector

Bivariate Examples (V)



Outline

Outline

Ordinary Least Squares
Intuition (I)

- The **"true"** "line/plane" of $\mathbb{R}^{p+1}$ (a hyperplane) is the closest to the points of the whole **population**.
- We look for the **closest** hyperplane to the points of the **sample**

Ordinary Least Squares
Intuition (II)

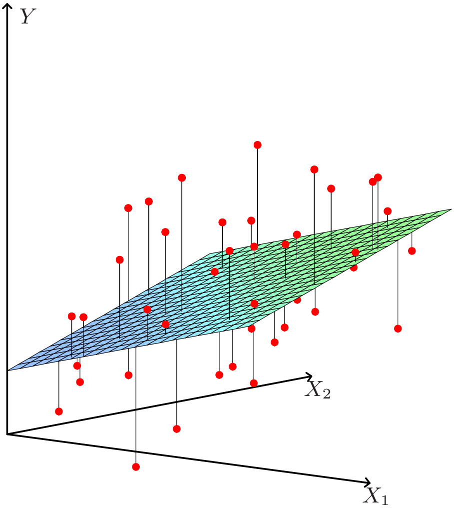


Figure: OLS: geometry in the space of the variables $\mathbb{R}^{p+1}$

Ordinary Least Squares
Criterion

Formalism

Find the hyperplan in $\mathbb{R}^{p+1}$ with the form

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

such that the distance to the sample points is as small as possible.

OLS estimator

The value estimated by the OLS (the estimate) for $\{\beta_j, j = 0, \ldots, p\}$ verify

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

Ordinary Least Squares
Criterion

Formalism

Find the hyperplan in $\mathbb{R}^{p+1}$ with the form

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

such that the distance to the sample points is as small as possible.

OLS estimator

The value estimated by the OLS (the estimate) for $\{\beta_j, j = 0, \ldots, p\}$ verify

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

Ordinary Least Squares
Interpretation in the sample Space (I)

Let $\mathbf{X}_{i\cdot} = (\mathbf{1}, \mathbf{x}_1, \dots, \mathbf{x}_p)$ be the ith row of $\mathbf{X}$. The estimated value is

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

$\rightsquigarrow$ We look for $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \in \mathrm{vec}(\mathbf{x}_1, \dots, \mathbf{x}_p)$ minimizing $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$.

Ordinary Least Squares

Interpretation in the sample Space (II)

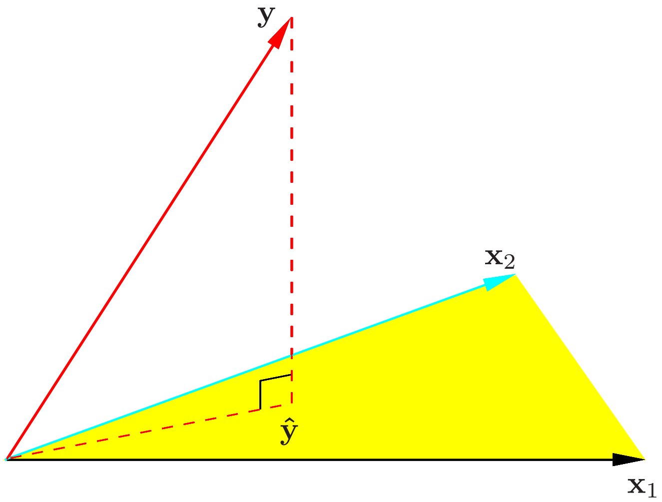


Figure: OLS: geometry in the space of the observations (sample) $\mathbb{R}^n$

Theorem

The OLS estimators verify the **normal equations** :

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

If $\mathbf{X}^\intercal \mathbf{X}$ is not singular, then

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

Proof

- Show that $\hat{\boldsymbol{\beta}}^{\text{ols}}$ is such that $\mathbf{X}\hat{\boldsymbol{\beta}}^{\text{ols}} = \text{proj}_{\mathbf{X}}(Y)$
- Use the orthogonality between $Y - \mathbf{X}\hat{\boldsymbol{\beta}}^{\text{ols}}$ and $\mathbf{x}_j$, for all $j = 1, \ldots, p$.

Theorem

The OLS estimators verify the **normal equations** :

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

If $\mathbf{X}^\intercal \mathbf{X}$ is not singular, then

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

Proof

- Show that $\hat{\boldsymbol{\beta}}^{\text{ols}}$ is such that $\mathbf{X}\hat{\boldsymbol{\beta}}^{\text{ols}} = \mathsf{proj}_{\mathbf{X}}(Y)$
- Use the orthogonality between $Y - \mathbf{X}\hat{\boldsymbol{\beta}}^{\text{ols}}$ and $\mathbf{x}_j$, for all $j = 1, \ldots, p$.

Orthogonal Projection and the hat matrix

Orthogonal Projection in the image of $\mathbf{X}$

If $\mathbf{X}^\intercal\mathbf{X}$ is not singular, the predicted value is

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

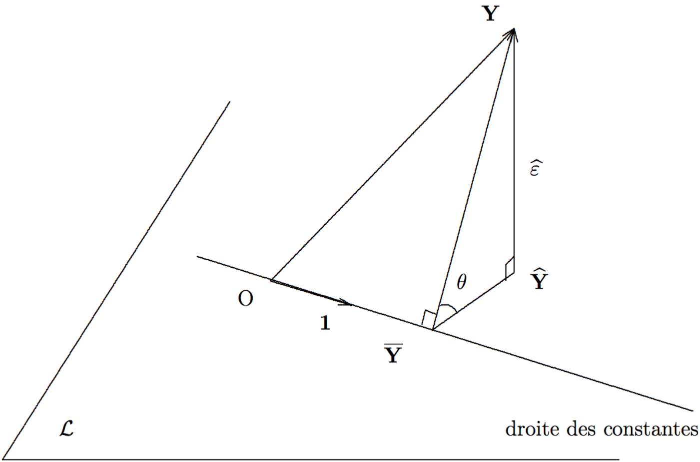
$\mathbf{P}_{\mathbf{X}}$ is sometimes denoted by $\mathbf{H}$ and called the "hat matrix" (*since it puts a hat on* $\mathbf{y}$).

Orthogonal Projection in the kernel of $\mathbf{X}$

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

$\rightsquigarrow$ The projectors $\mathbf{P}_X$ and $\mathbf{P}_X^{\perp}$ are idempotent. They ease the calculus and the interpretation!

Geometrical view of OLS



Proposition

The vector of residuals is orthogonal to the line of constant $\mathbf{1}_n$. Then

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

Moreover, $\hat{Y} \perp \hat{\varepsilon}$.

Corollary

- $\frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}$.
- The orthogonal projection of Y on $\mathbf{1}_n$ has for coordinate $\bar{Y}$:

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

Ordinary least squares
Remarks

Purely Geometrical

- Do not rely on the Gaussian assumption
- Do not say a thing on **the residual variance σ^2...**

Requirement for non singularity of $\mathbf{X}^\intercal\mathbf{X}$

A necessary and sufficient condition is that $\mathbf{X}$ has full rank.

$\rightsquigarrow$ No column is a linear combination of the other columns.

$\rightsquigarrow$ Each variable must bring "some original information".

$\rightsquigarrow$ Strong correlations induce numerical instabilities.

Outline

Maximum likelihood
criterion

Formalism

With the assumption that $\varepsilon_i \sim \mathcal{N}(0, \sigma)$,

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

MLE

The values estimated by ML for β and σ verify

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

Maximum likelihood
criterion

Formalism

With the assumption that $\varepsilon_i \sim \mathcal{N}(0, \sigma)$,

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

MLE

The values estimated by ML for β and σ verify

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

Maximum likelihood

criterion

Formalism

With the assumption that $\varepsilon_i \sim \mathcal{N}(0, \sigma)$,

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

MLE

The values estimated by ML for $\boldsymbol{\beta}$ and σ verify

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

Theorem

When $n > p$, the MLE have the following expression:

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

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

Proof:

By zeroing the derivatives of the objective function, which is concave.

Maximum likelihood

Practical estimation of the residual variance

Theorem

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

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

Corollary

An unbiased estimator of the residual variance is

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

Vocabulary

The quantity $n - p - 1$ is the **number of residual degrees of freedom**, equal to the rank of $\mathbf{P}_{\mathbf{X}}^{\perp}$.

Maximum likelihood

Practical estimation of the residual variance

Theorem

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

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

Corollary

An unbiased estimator of the residual variance is

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

Vocabulary

The quantity $n - p - 1$ is the **number of residual degrees of freedom**, equal to the rank of $\mathbf{P}_{\mathbf{X}}^{\perp}$.

Maximum likelihood

Practical estimation of the residual variance

Theorem

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

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

Corollary

An unbiased estimator of the residual variance is

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

Vocabulary

The quantity $n - p - 1$ is the **number of residual degrees of freedom**, equal to the rank of $\mathbf{P}_{\mathbf{X}}^{\perp}$.

Outline

Parameters Estimation

Properties of the estimators (I)

General case

$\hat{\boldsymbol{\beta}}$ are unbiased estiamtors of $\boldsymbol{\beta}$ with variance

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

Gaussian case

If the noise is Gaussian, i.e. $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, then

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

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

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

Parameters Estimation

Properties of the estimators (I)

General case

$\hat{\boldsymbol{\beta}}$ are unbiased estiamtors of $\boldsymbol{\beta}$ with variance

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

Gaussian case

If the noise is Gaussian, i.e. $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, then

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

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

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

Parameters Estimation

Properties of the estimators (II)

Gauss-Markov theorem

- Gaussian case: $\hat{\boldsymbol{\beta}}^{\mathrm{ols}}$ is the best unbiased estimators (i.e. with minimal variance).
- General case: $\hat{\boldsymbol{\beta}}^{\mathrm{ols}}$ is the best **linear** unbiased estimators.

$\rightsquigarrow$ We say that $\hat{\boldsymbol{\beta}}^{\mathrm{ols}}$ is the **BLUE** (best linear unbiased estimator)

Outline

Test and confidence interval for the parametersβ_j

Under the Gaussian assumption

Testing the nullity of β_j

Does the jth variable bring additional significant information for predicting the response?

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

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

Test Statistic and Decision rule

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

Confidence interval on $\hat{\beta}_j$

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

Test and confidence interval for the parameters β_j

Under the Gaussian assumption

Testing the nullity of β_j

Does the jth variable bring additional significant information for predicting the response?

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

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

Test Statistic and Decision rule

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

Confidence interval on $\hat{\beta}_j$

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

Test and confidence interval for the parameters β_j

Under the Gaussian assumption

Testing the nullity of β_j

Does the jth variable bring additional significant information for predicting the response?

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

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

Test Statistic and Decision rule

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

Confidence interval on $\hat{\beta}_j$

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

Outline

Residuals and Prediction

Let $\mathbf{x}_0 \in \mathbb{R}^p$ be a new observation and $\hat{Y}_0 = \mathbf{x}_0\hat{\boldsymbol{\beta}}$ the associated predictor.

Proposition

Let $\hat{\varepsilon}_0 = Y_0 - \hat{Y}_0$ the prediction noise at the new point. We have:

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

Confidence interval

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

Residuals and Prediction

Let $\mathbf{x}_0 \in \mathbb{R}^p$ be a new observation and $\hat{Y}_0 = \mathbf{x}_0\hat{\boldsymbol{\beta}}$ the associated predictor.

Proposition

Let $\hat{\varepsilon}_0 = Y_0 - \hat{Y}_0$ the prediction noise at the new point. We have:

$$\mathbb{E}(\hat{\varepsilon}_0) = 0$$

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

Prediction interval

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

Outline

Decomposing the variance

Theorem of total variance

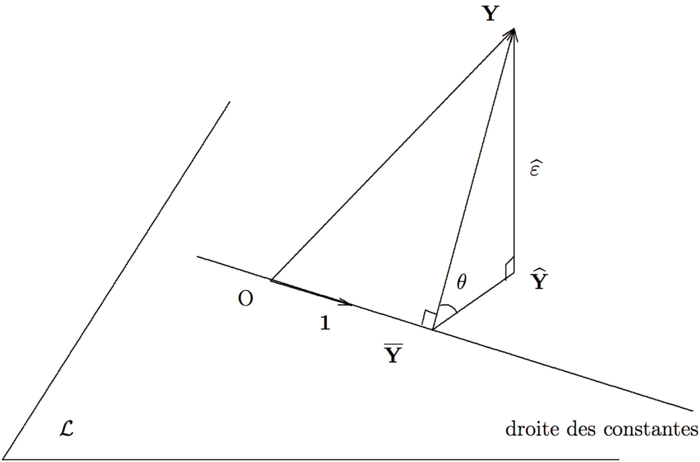
Since $\hat{\varepsilon} = \mathbf{Y} - \hat{\mathbf{Y}}$ is orthogonal to $\hat{\mathbf{Y}} - \bar{\mathbf{Y}}$, we have

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

with

- TSS = Total Sum of Squares
  $\rightsquigarrow$ Total variability to explain
- ESS = Explained Sum of Squares
  $\rightsquigarrow$ variability explained by the model
- RSS = Residual Sum of Squares
  $\rightsquigarrow$ Residual variability, not explained by the model

Reminder: geometrical interpretation



Coefficient of determination
Definition

R^2

The coefficient of determination is defined by

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

adjusted R^2

The adjusted coefficient of determination is defined by

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

Remark

The coefficient of determination can be interpreted as the percentage of variance explained by the model.

Testing the relevance of the model (I)

Tested Hypothesis

$$\begin{cases} \mathcal{M}_0 : & \text{more simple model} \\ \mathcal{M}_1 : & \text{more complex model} \end{cases} \Leftrightarrow \begin{cases} \mathcal{M}_0 : & Y_i = \beta_0 + \varepsilon_i \\ \mathcal{M}_1 : & Y_i = \mathbf{X}\boldsymbol{\beta} + \varepsilon_i \end{cases}$$

Distributions of Sums of Squares under H_0

- $SCR = \hat{\varepsilon}^\intercal \hat{\varepsilon}$, hence $SCR = (n-p-1)\hat{\sigma}^2 \sim \sigma^2 \chi^2_{n-p-1}$.
- $SCM = \|\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\|^2 = \|\mathrm{proj}_{\mathbf{X}}(\mathbf{Y}) - \mathrm{proj}_{\mathbf{1}}(\mathbf{Y})\|^2$, then $SCM \overset{H_0}{\sim} \sigma^2 \chi^2_p$.
- Since $SCT = \|\mathbf{Y} - \bar{\mathbf{Y}}\|^2$, we have $SCT \overset{H_0}{\sim} \sigma^2 \chi^2_{n-1}$

Testing the relevance of the model (I)

Tested Hypothesis

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

Distributions of Sums of Squares under H_0

- $SCR = \hat{\varepsilon}^\intercal\hat{\varepsilon}$, hence $SCR = (n-p-1)\hat{\sigma}^2 \sim \sigma^2\chi^2_{n-p-1}$.
- $SCM = \|\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\|^2 = \|\mathsf{proj}_{\mathbf{X}}(\mathbf{Y}) - \mathsf{proj}_{\mathbf{1}}(\mathbf{Y})\|^2$, then $SCM \overset{H_0}{\sim} \sigma^2\chi^2_p$.
- Since $SCT = \|\mathbf{Y} - \bar{\mathbf{Y}}\|^2$, we have $SCT \overset{H_0}{\sim} \sigma^2\chi^2_{n-1}$

Testing the relevance of the model (II)

Test Statistic: Fisher

We reject when F, measuring the part of variability explained by the model, is “large”:

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

Decision rule

We reject H_0 if $F \geq f_{p,n-p-1;1-\alpha}$

$p-$value

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

Analysis of variance

Summary table

Source	Degrees of freedom	Sum of squares	mean of squares	F
Model	p	ESS	ESS/p	$F = \frac{(n-p-1)ESS}{RSS/p}$
Residual	$n-p-1$	RSS	$\frac{RSS}{(n-p-1)}$	
Total	$n-1$	TSS		

Model comparison

A natural question considering a series of model related to the same set of p predictors is

What model is the more relevant?

A natural answer consists in considering the following test

$$\begin{cases} \mathcal{M}_\omega : & \text{a model} \\ \mathcal{M}_\Omega : & \text{a more complex model} \end{cases},$$

where $\mathcal{M}_\omega \subset \mathcal{M}_\Omega$: the model are "nested".

Model comparison
Geometrical view

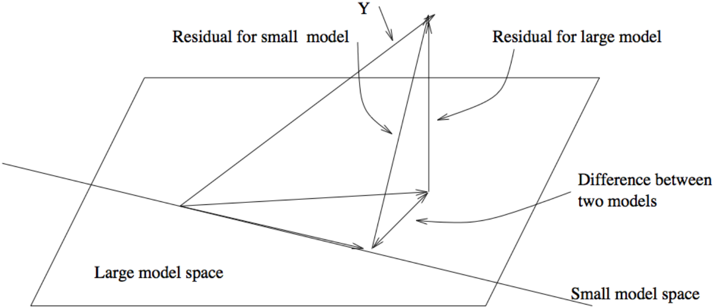


Figure: Source: *Pratical regression and anova using R*, J. Faraway

Comparing nested model

Intuition

We choose H_1 (i.e. the more complex model Ω) if the residuals of Ω are significantly smaller compared to the ones of the simple model ω, *i.e.*,

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

Under H_0

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

Test Statistic

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

Comparing nested model

Intuition

We choose H_1 (i.e. the more complex model Ω) if the residuals of Ω are significantly smaller compared to the ones of the simple model ω, *i.e.*,

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

Under H_0

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

Test Statistic

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

Comparing nested model

Intuition

We choose H_1 (i.e. the more complex model Ω) if the residuals of Ω are significantly smaller compared to the ones of the simple model ω, *i.e.*,

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

Under H_0

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

Test Statistic

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

Comparing nested models

Summary table

Source	Degrees of freedom	Sums of squares	Means of squares
Model ω	$n - \mathsf{ddl}_\omega$	RSS_ω	$RSS_\omega/\mathsf{ddl}_\omega$
Model Ω	$n - \mathsf{ddl}_\Omega$	RSS_Ω	$RSS_\Omega/\mathsf{ddl}_\Omega$

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

Outline

Goals of the diagnostic

1. Check the **hypotheses of the model**
    - linearity/model appropriate
    - Homoscedasticity of the noise
    - Independence of the noise
    - Gaussianity of the noise

2. Detecting **outliers**

Outline

Residual analysis

Hypotheses of the model are mostly related to the noise

1. Centered: $\mathbb{E}(Y) = \mathbf{X}\boldsymbol{\beta}$, soit $\mathbb{E}(\varepsilon_i) = 0$
2. Homoscedastic: $\mathbb{V}(\varepsilon_i) = \sigma^2$ for all i,
3. Independent, $\operatorname{cov}(\varepsilon_i, \varepsilon_{i'}) = 0$ for all $i \neq i'$,
4. Gaussian: $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Diagnostic

We do not observe ε_i, so we the residual $\hat{\varepsilon}_i$ for the diagnostic

1. Analysis of the **Residual graph**
2. Testing the independency (Durbin-Watson)
3. Testing the normality (Shapiro, Kolmogorov, χ^2)

Leverage points

Definition (Leverage)

The variance of the prediction of the ith observation verifies

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

where $h_i = (\mathbf{P_X})_{ii}$ is called **leverage** of observation i.

- The larger h_i, the larger the contribution of y_i to $\hat{Y}_i$.
- $\sum_{i=1}^n h_i = p$, hence the mean of the leverage is p/n.

Definition (Leverage point)

Individual i is a **leverage point** if

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

Leverage points

Definition (Leverage)

The variance of the prediction of the ith observation verifies

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

where $h_i = (\mathbf{P}_{\mathbf{X}})_{ii}$ is called **leverage** of observation i.

- The larger h_i, the larger the contribution of y_i to $\hat{Y}_i$.
- $\sum_{i=1}^{n} h_i = p$, hence the mean of the leverage is p/n.

Definition (Leverage point)

Individual i is a **leverage point** if

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

Standardized residuals and studentized residuals

To remove any scale factor, it is useful to normalize $\hat{\varepsilon}_i$

Definition (Standardized residuals)

The variance of the residuals can be written as $\mathbb{V}(\hat{\varepsilon}_i) = \sigma^2(1 - h_i)$. Hence, we define the **standardized residuals** by

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

- $\hat{\varepsilon}_i$ is not independent of $\hat{\sigma}$, and we do not know their distribution
- The **studentized form** fixes this issue.

Definition (Studentized residuals)

We call **Studentized residuals** the statistics defined by

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

where $\hat{\sigma}^{(-i)}$ is the variance estimated on the data deprived of i.

Standardized residuals and studentized residuals

To remove any scale factor, it is useful to normalize $\hat{\varepsilon}_i$

Definition (Standardized residuals)

The variance of the residuals can be written as $\mathbb{V}(\hat{\varepsilon}_i) = \sigma^2(1 - h_i)$.
Hence, we define the **standardized residuals** by

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

- $\hat{\varepsilon}_i$ is not independent of $\hat{\sigma}$, and we do not know their distribution
- The **studentized form** fixes this issue.

Definition (Studentized residuals)

We call **Studentized residuals** the statistics defined by

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

where $\hat{\sigma}^{(-i)}$ is the variance estimated on the data deprived of i.

Standardized residuals and studentized residuals

To remove any scale factor, it is useful to normalize $\hat{\varepsilon}_i$

Definition (Standardized residuals)

The variance of the residuals can be written as $\mathbb{V}(\hat{\varepsilon}_i) = \sigma^2(1 - h_i)$.
Hence, we define the **standardized residuals** by

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

- $\hat{\varepsilon}_i$ is not independent of $\hat{\sigma}$, and we do not know their distribution
- The **studentized form** fixes this issue.

Definition (Studentized residuals)

We call **Studentized residuals** the statistics defined by

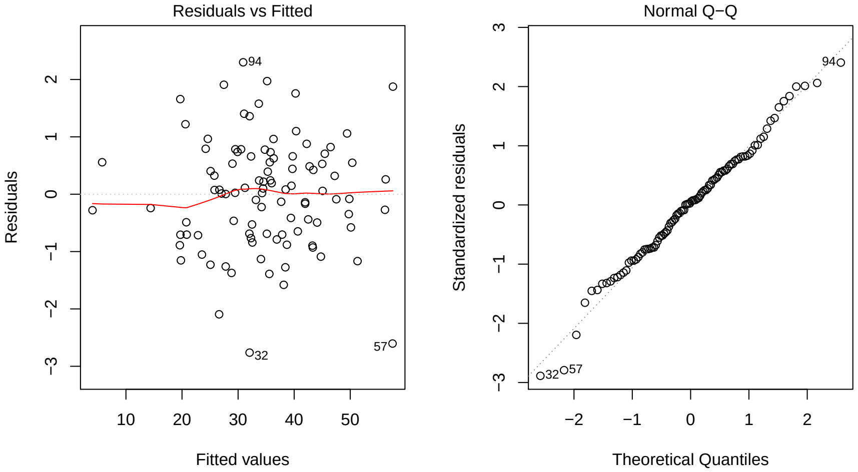
$$t_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}^{(-i)}\sqrt{1 - h_i}},$$

where $\hat{\sigma}^{(-i)}$ is the variance estimated on the data deprived of i.

Residual analysis

Ideal case

```
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)
```



Residual analysis I

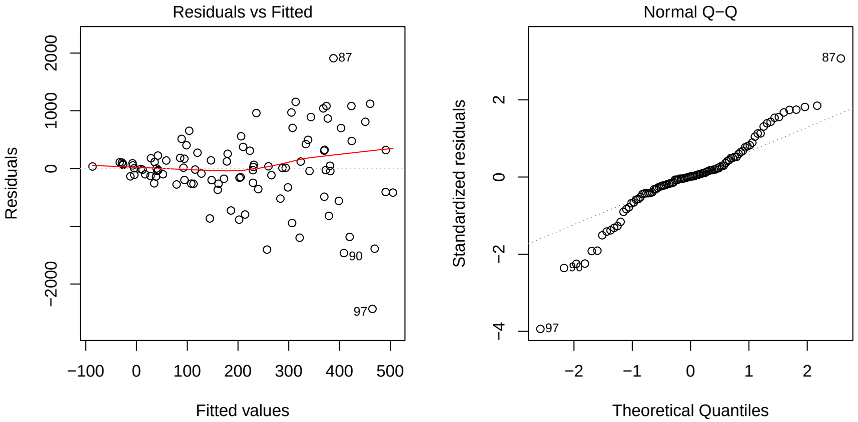
Variance proportional to a predictor

Transforming Y with log/sqrt may fix heteroscedasticity

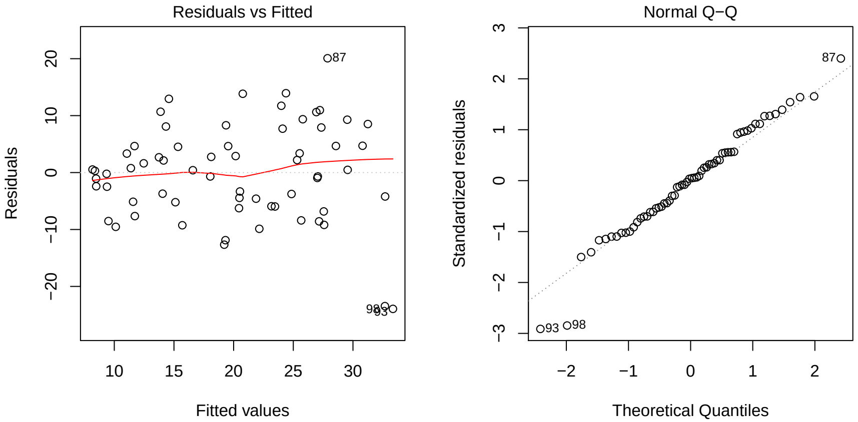
```
n <- 100; x <- (1:n + rnorm(n,0,5)); y <- 5 + 3 * x + rnorm(n,0,10)*x
par(mfrow=c(1,2));  plot(lm(y~x), which=1:2); plot(lm(sqrt(y)~x), which=1:2)
```

Residual analysis II

Variance proportional to a predictor



Residual analysis III
Variance proportional to a predictor

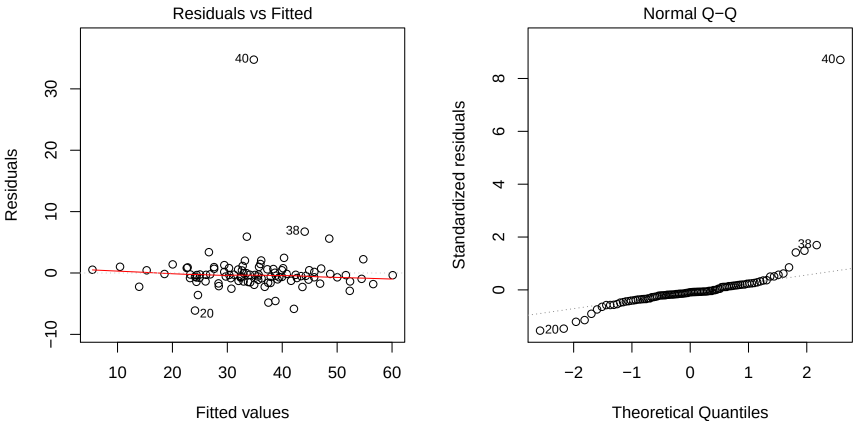


Residual analysi

Non Gaussian residuals

The linear model is relatively robust to non Gaussian residuals if their distirbution remains symmetric.

```
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)
```



Residual analysis I

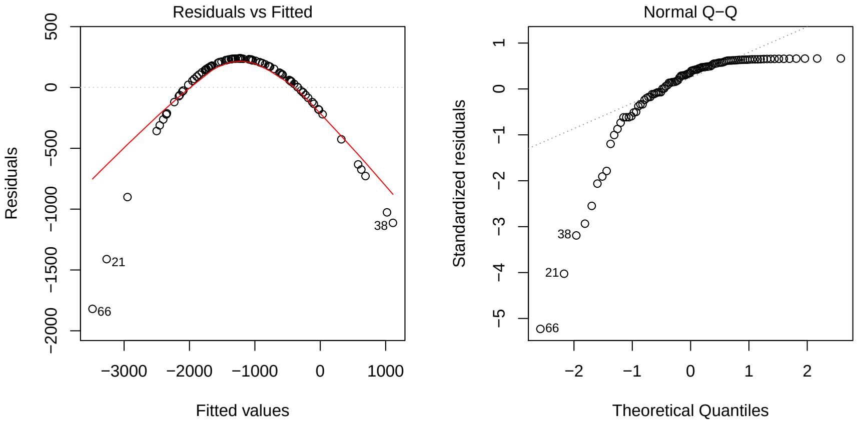
Wrong model

A strong tendency in the residuals suggests a model mispecification.

```
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)
```

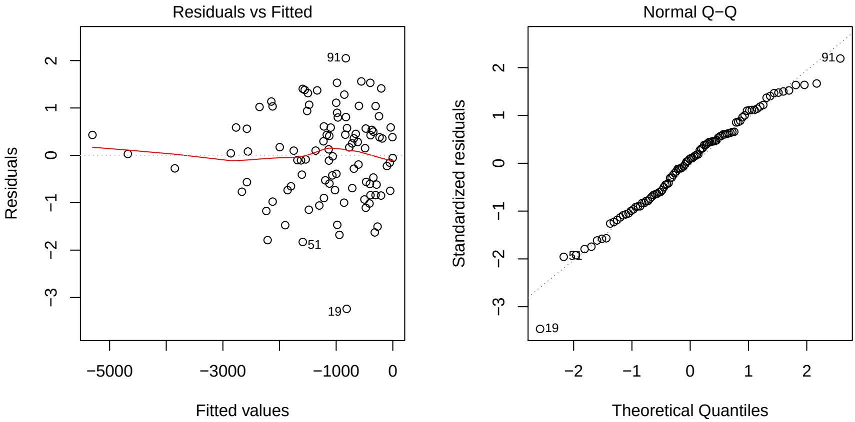
Residual analysis II

Wrong model



Residual analysis III

Wrong model



Outline

Cook distance

Idea

Unraveling the influence or “abnormality” of certain points.

Definition (Distance de Cook)

D_i characterizes the influence of observation i on the regression fit: a high value may unravel an unusual influence

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

$\rightsquigarrow$ D_i can be interpreted as the square distance between $\hat{\boldsymbol{\beta}}$ et $\hat{\boldsymbol{\beta}}^{(-i)}$.

Proposition (Practical Computation)

We can compute D_i without fitting a new model because

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

Cook distance

Idea

Unraveling the influence or “abnormality” of certain points.

Definition (Distance de Cook)

D_i characterizes the influence of observation i on the regression fit: a high value may unravel an unusual influence

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

$\rightsquigarrow$ D_i can be interpreted as the square distance between $\hat{\boldsymbol{\beta}}$ et $\hat{\boldsymbol{\beta}}^{(-i)}$.

Proposition (Practical Computation)

We can compute D_i without fitting a new model because

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

Cook distance

What threshold?

Rule of thumb

We consider that a value greater than 1 corresponds to an outlier.

Hypothesis testing

One can show that D_i is a test statistic from the Wald test for

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

where $\boldsymbol{\beta}_0^{-i}$ is the true value estimated without observation i.
The test statistic follows a $F_{p+1,n-p-1,1-\alpha}$ under H_0.

Cook distance
What threshold?

Rule of thumb

We consider that a value greater than 1 corresponds to an outlier.

Hypothesis testing

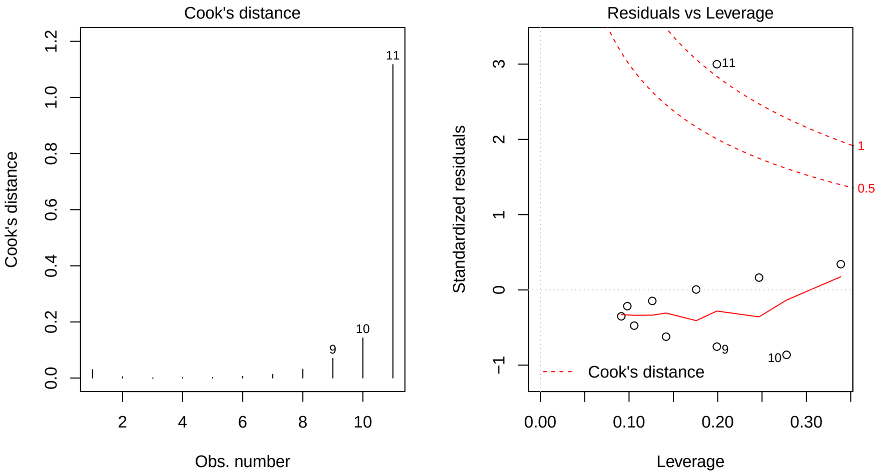
One can show that D_i is a test statistic from the Wald test for

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

where $\boldsymbol{\beta}_0^{-i}$ is the true value estimated without observation i.
The test statistic follows a $F_{p+1,n-p-1,1-\alpha}$ under H_0.

Cook distance

```
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)
```



Outline

Outline

Pine processionary (caterpillar) data set I

Data set

Consider 33 samples of 10 hectares forest plots. Each plot is cut into small squares of 5 acres on which the average of the following measures are calculated

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

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

Goal

Predict the **number of nests** from the other variables.

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

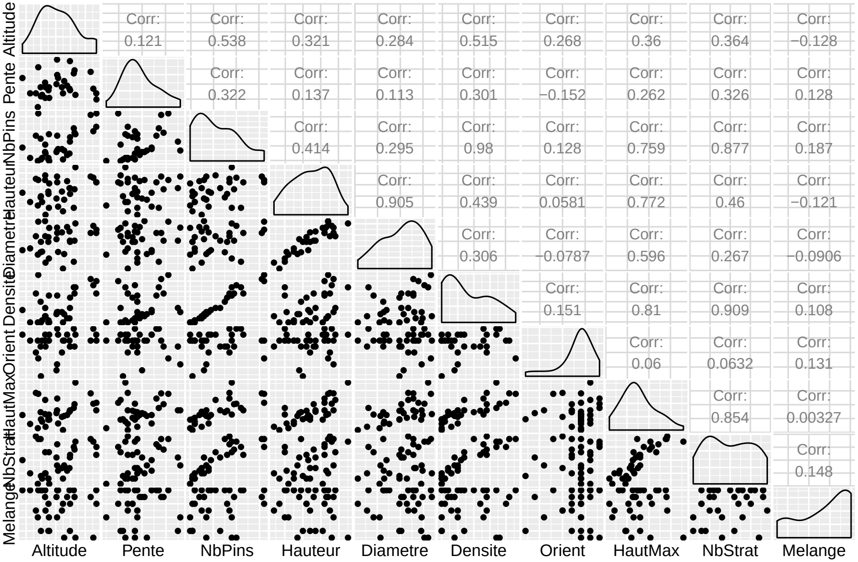
Pine processionary (caterpillar) data set II

The data frame header looks like

```r
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
```

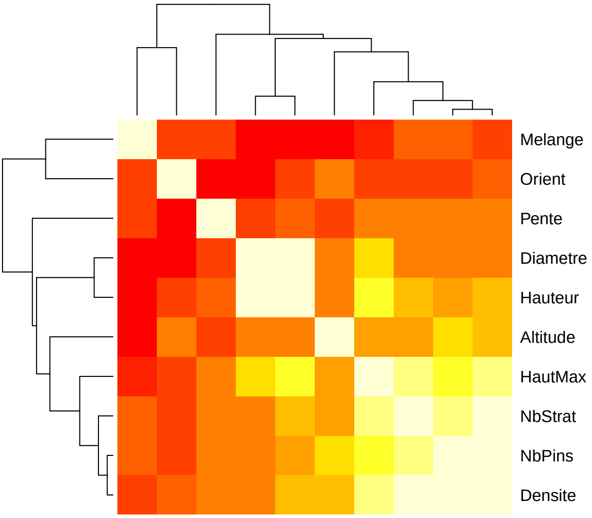
Pine processionary (caterpillar) data set III



Correlations between predictors

Strong correlations between variables induced bad estimates of the corresponding parameters

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



Outline

OLS

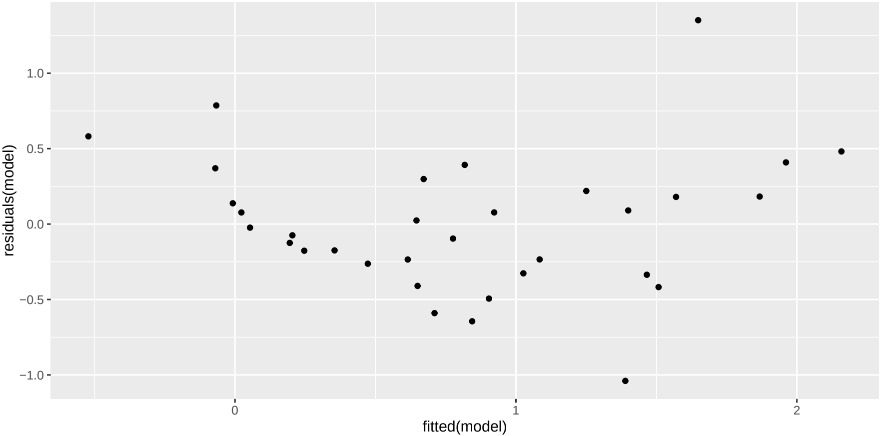
Simple sanity check

```
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
```

Raw multiple linear regression
Residual analysis

The Residual graph suggest a lograithmic transformation of the response

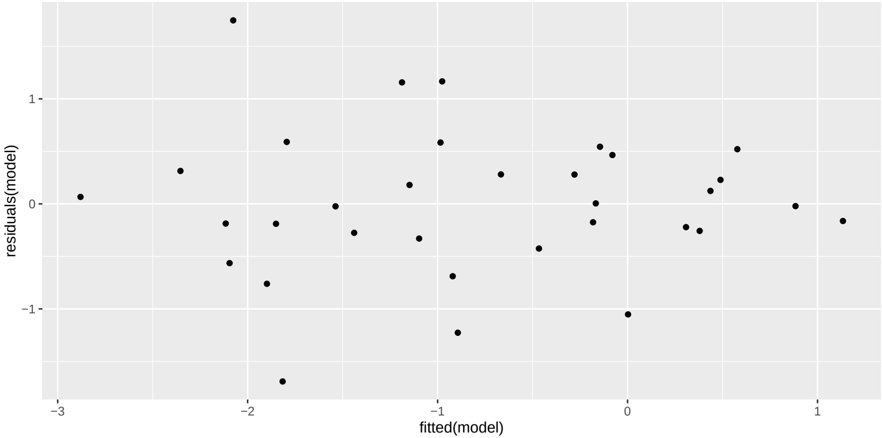
```
model <- lm(NbNids~.,data=chenilles)
qplot(fitted(model),residuals(model), geom='point')
```



Log-transformed model

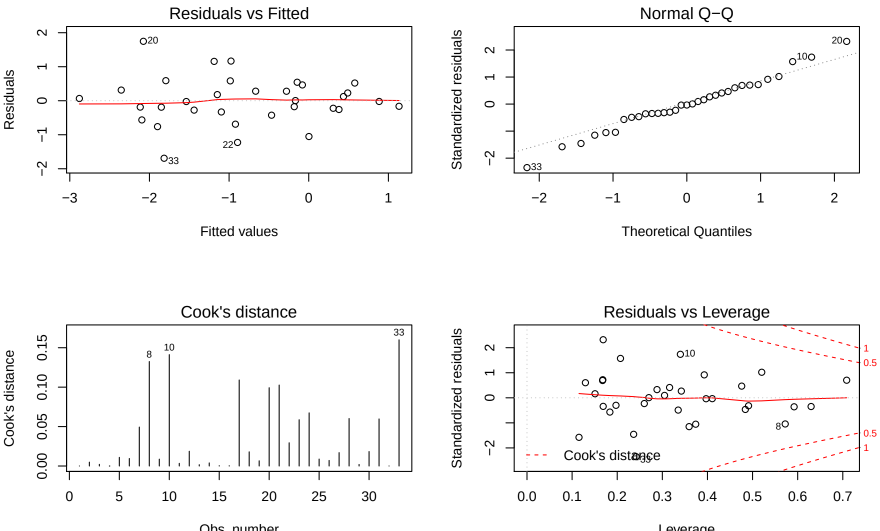
Residual analysis

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



Log-transformed model
Complete diagnostic

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



Log-transformed model
Residual normality

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

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

Log-transformed model

Residual independency

```
library(car)
durbinWatsonTest(model)

##  lag Autocorrelation D-W Statistic p-value
##    1      -0.1208374      2.051547   0.948
##  Alternative hypothesis: rho != 0
```

Log-transformed model

Testing the parameters

```r
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
```

Log-transformed model

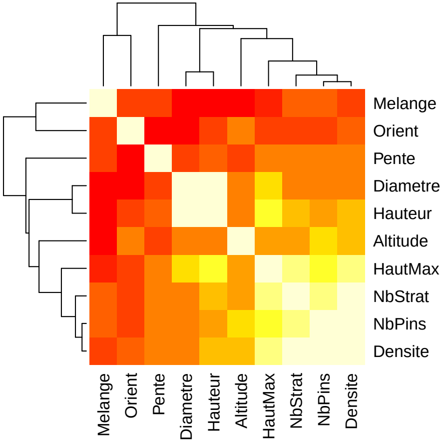
Testing the model

```
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
```

Log-transformed model and normalized predictors

Correlated predictors

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



Constat

- The parameters which are badly estiamted (i.e. with large variance) are the one with high correlation (densité, nb pins, nb strates, hauteur)
  $\rightsquigarrow$ IF there is an effect, it is hidden due to the redundancy between the variables

- Weakly correlated variables (pente, orientation, mélange) are better estimated
  $\rightsquigarrow$ We can conclude about their effect on the number of nests.

$\rightsquigarrow$ This statement can only be made on normalized data, to put the variances on the same scale

Log-transformed model and normalized predictors II

Testing the model

```
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
```

Log-transformed model and normalized predictors III

Testing the model

```
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
```

Log-transformed model I

Nested models comparison

```
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
```

Log-transformed model II

Nested models comparison

```
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
```

Log-transformed model III

Nested models comparison

```
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
```

Log-transformed model IV

Nested models comparison

```
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
```

Log-transformed model V

Nested models comparison

```
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
```

Log-transformed model VI

Nested models comparison

```
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
```

Log-transformed model VII

Nested models comparison

```
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
```

Final model I

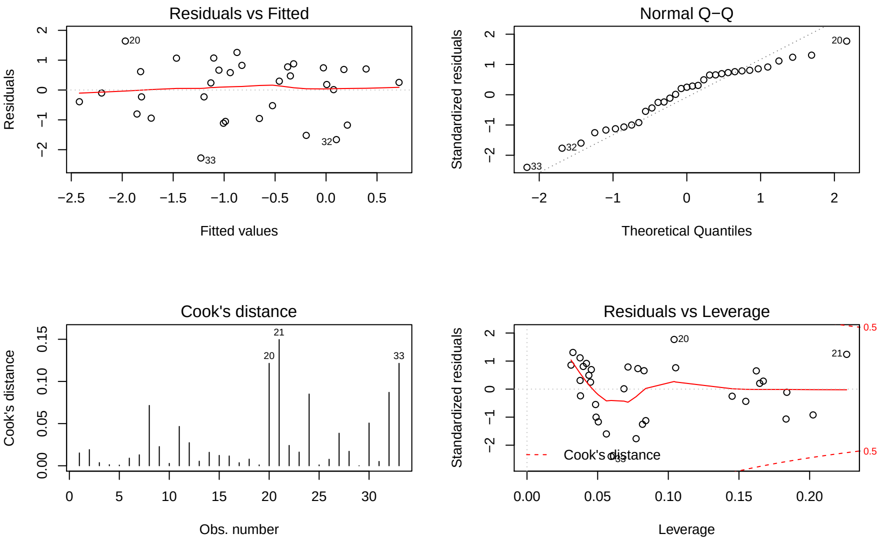
```
summary(M21)
```

```
##
## 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
```

Final model II

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

Final model III

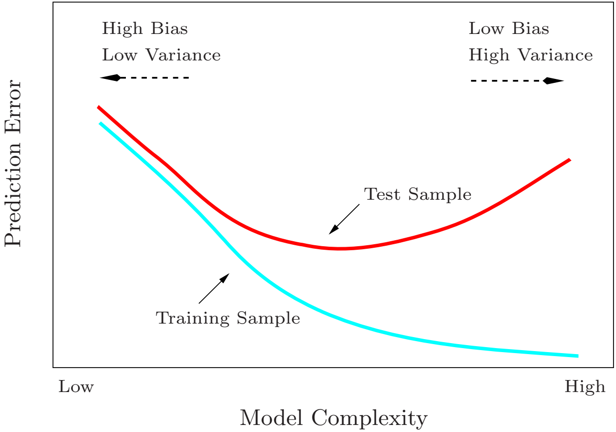


Outline

Motivation : Bias/Variance tradeoff

At a new point $X = x$,

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



Linear regression

Prediction Error

For fixed $\mathbf{X}$, we have

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

Reminder: Gauss-Markov

$\hat{Y} = X^\intercal \hat{\boldsymbol{\beta}}^{\text{ols}}$ is the BLUE

$\rightsquigarrow$ Are there situations where we should trade some bias for less variance ?

Variable Selection

Problematic

With many regressor,

- we integrate more and more information in the model ;
- we have more and more parameters to estimate and $\mathbb{V}(\hat{Y}_i) \nearrow$.

Idea

Look for a (small) set $\mathcal{S}$ with k variables among p such that

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

Ingredients

To find this tradeoff, we need

1. a criterion to evaluate the performance ;
2. an algorithm to determine the subset of k variables optimising the criterion.

Variable Selection

Problematic

With many regressor,

- we integrate more and more information in the model ;
- we have more and more parameters to estimate and $\mathbb{V}(\hat{Y}_i) \nearrow$.

Idea

Look for a (small) set $\mathcal{S}$ with k variables among p such that

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

Ingredients

To find this tradeoff, we need

1. a criterion to evaluate the performance ;
2. an algorithm to determine the subset of k variables optimising the criterion.

Penalized Criterion
Principle

Idea

Rather than estimating the prediciton error with the test error, we estimate how much the training error under estimate the true prediction error.

General form

Based on the available model fit, compute

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

Remarks

- “penalize”to much complex models

Penalized Criterion

Principle

Idea

Rather than estimating the prediciton error with the test error, we estimate how much the training error under estimate the true prediction error.

General form

Based on the available model fit, compute

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

Remarks

- “penalize”to much complex models

Penalized Criteria

The most Popular in linear regression

Let k be the size of the current model (i.e. the current number of predictors).

Criterion for the Linear regression model σ known

We choose the model with size k minimizing one of the following

- **Mallows C_p**

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

- **Akaïke Information Criteria** equivalent to C_p when σ is known

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

- **Bayesian Information Criterion**

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

Penalized Criteria

The most Popular in linear regression

Let k be the size of the current model (i.e. the current number of predictors).

Criterion for the Linear regression model σ unknown

We choose the model with size k minimizing one of the following

- **Mallows C_p** σ estimated by the unbiased estimator $\hat{\sigma}$

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

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

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

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

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

C_p/AIC: proof

Ideally, we would like to minimize the error of the mean distance between the true model $\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\mu}$ and the OLS. This diustance splits as follows

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

On average we get

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

If k is the dimension of the space of the projection, we find

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

We then just have to divide by $n\sigma^2$.

Outline

Model

Background

Estimation

Residuals and Prediction

Analysis of Variance

Diagnostic

A full example: pine processionary

Variable Selection

Exhaustive search (best-subset)

Algorithm

For $k = 0, \dots, p$, find the subset with k variables with the smallest SCR among 2^k models.

Properties

- Generalize to any criterion (R^2, AIC, BIC...)
- Efficient algorithm with pruning ("Leaps and Bound")
- impossible as soon as $p > 30$.

(Forward regression)

Algorithm

1. Begin with $\mathcal{S} = \emptyset$
2. at step k find the variable which, added to $\mathcal{S}$, gives the best model

2'. At step k find the best model by either adding or removing one variable.

3 etc. until p variables enter the model

Properties

- Best model is understood as SCR or R^2, AIC, BIC...
- useful when p is large
- large bias, but variance/complexity controlled.
- "greedy" algorithm

Forward-stepwise

Algorithm

1. Begin with $\mathcal{S} = \emptyset$

2. at step k find the variable which, added to $\mathcal{S}$, gives the best model

2'. At step k find the best model by either adding or removing one variable.

3 etc. until p variables enter the model

Properties

- Best model is understood as SCR or R^2, AIC, BIC...
- useful when p is large
- large bias, but variance/complexity controlled.
- "greedy" algorithm

Backward regression

Algorithm

1. Start with the full model $\mathcal{S} = \{1, \dots, p\}$
2. At step k, remove the less influent variable.
3. etc. until $\mathcal{S}$ is empty.

Properties

- Best model is understood as SCR or R^2, AIC, BIC...
- does not work when $n < p$
- large bias, but variance/complexity controlled.
- "greedy" algorithm

Outline

Model

Background

Estimation

Residuals and Prediction

Analysis of Variance

Diagnostic

A full example: pine processionary

Variable Selection

Exhaustive search I

```
library(leaps)
```

All possible models

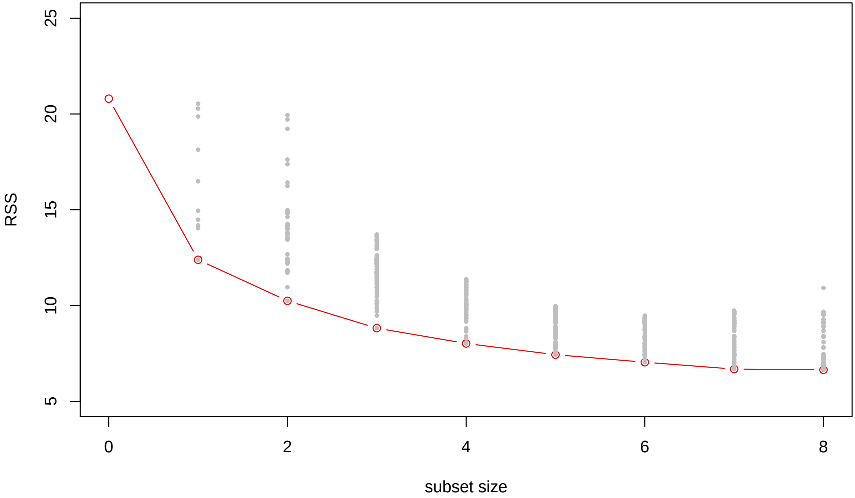
```
out <- regsubsets(NbNids ~ . , data=chenilles,
                  nbest=100, really.big=TRUE)
bss <- summary(out)
```

Extract model sizes and SCR. Add the null model (just the intercept)

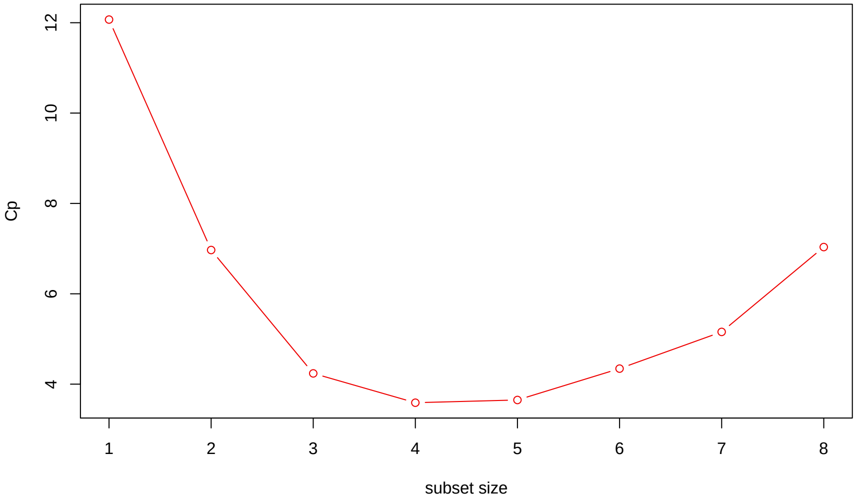
```
bss.size  <- as.numeric(rownames(bss$which))
intercept <- lm(NbNids ~ 1, data=chenilles)
bss.best.rss  <- c(sum(resid(intercept)^2), tapply(bss$rss  , bss.size, min))
```

```
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)
```

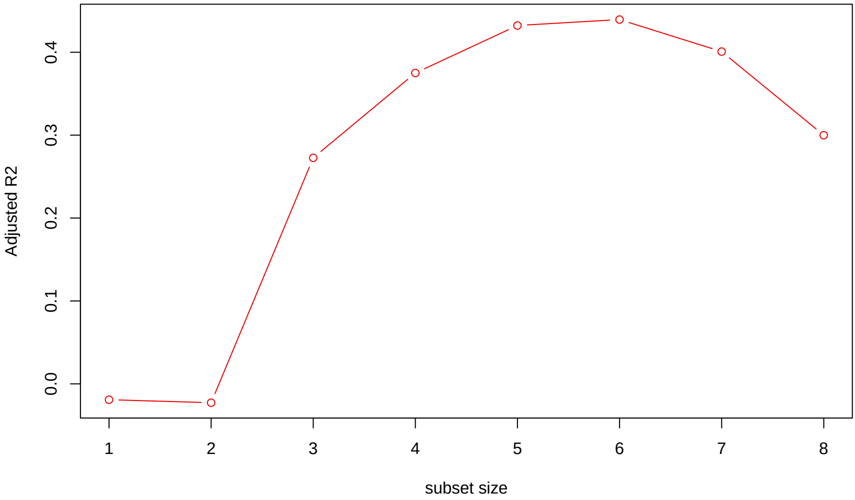
Exhaustive search II



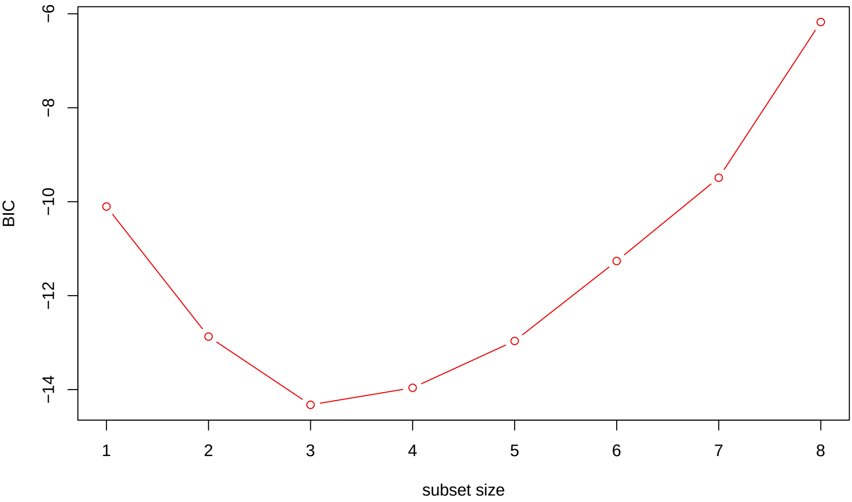
Exhaustive search III



Exhaustive search VI



Exhaustive search V



Forward-Stepwise with R (I)

Crete the null model and the full one

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

Create the scope of model to consider

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

Stepwise 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 with R: BIC

Keep the sparsest model

```
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
```