

Cours accéléré Statistiques

Partie 2: Modèle linéaire

Master 2 Mathématiques et Applications

Christine Keribin

¹Laboratoire de Mathématiques d'Orsay
Université Paris-Saclay

2022-2023

université
PARIS-SACLAY

FACULTÉ
DES SCIENCES
D'ORSAY



Introduction

Références

A-Définition et
hypothèses

Exemples

Identifiabilité

B-Estimation

EMC

Propriétés

EMV

C-Lois des
estimateurs

IC

IC Préviation

RC

D-Tests

de Student

de Wald

de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Biais/variance

Sommaire

Introduction

A-Définition et hypothèses

B-Estimation

C-Lois des estimateurs

D-Tests

E-Validation

Sélection de variables et choix de modèle

Régularisation

Introduction

Références

A-Définition et hypothèses

Exemples
Identifiabilité

B-Estimation

EMC
Propriétés
EMV

C-Lois des estimateurs

IC
IC Préviation
RC

D-Tests

de Student
de Wald
de Fisher

E-Validation

Sélection de variables et choix de modèle

Biais/variance

- ▶ Expliquer ou prédire une variable aléatoire **réponse** Y en fonction d'une liste de variables **explicatives** $X = (X_1, \dots, X_p)$ **observées** sur des individus

$$Y = f(X)$$

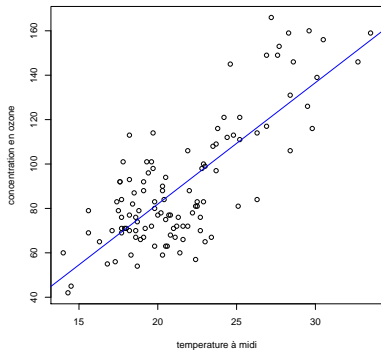
- ▶ Modéliser Y par une fonction des variables explicatives X

$$Y_{|X=x} = m(x) + \varepsilon$$

$m(x) = \mathbb{E}(Y|X = x)$ est la **fonction de régression**

- ▶ choix est guidé par la connaissance d'un phénomène physique, biologique,...
 - ▶ ou non...
 - ▶ hypothèses de modélisation guidées par l'observation
- ▶ apprentissage **supervisé**

Exemple : régression linéaire (simple) ¹



$$Y_i = \underbrace{\theta_1 + \theta_2 X_i}_{\text{déterministe}} + \underbrace{\varepsilon_i}_{\text{bruit aléatoire}}$$

1. <http://www.agrocampus-ouest.fr/math/livreR/>

Introduction

Références

A-Définition et hypothèses

Exemples
Identifiabilité

B-Estimation

EMC
Propriétés
EMV

C-Lois des estimateurs

IC
IC Prédiction
RC

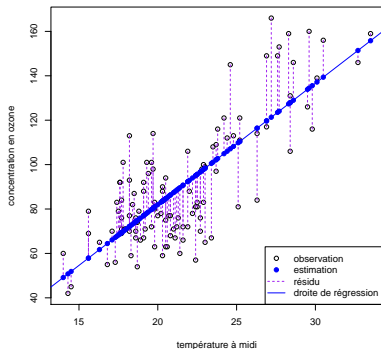
D-Tests

de Student
de Wald
de Fisher

E-Validation

Sélection de variables et choix de modèle

Exemple : régression linéaire (simple)



$\mathbb{E}(Y_i|x_i) = \theta_1 + \theta_2 x_i$ est l'équation de la **droite de régression**² : modélise le comportement moyen pour chaque condition d'expérience \hookrightarrow **estimer** $\theta = (\theta_1, \theta_2)$

2. "Régression" : Sir Galton (1822-1911)

Introduction

Références

A-Définition et hypothèses

Exemples
Identifiabilité

B-Estimation

EMC
Propriétés
EMV

C-Lois des estimateurs

IC
IC Préviation
RC

D-Tests

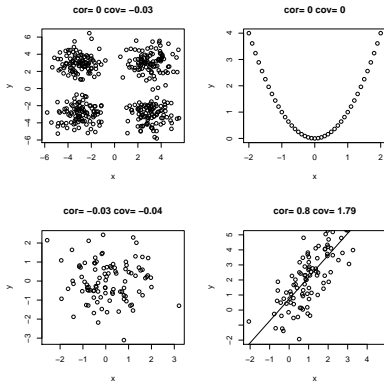
de Student
de Wald
de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Que cherche-t-on à détecter ?

La corrélation entre Y et $x...$ qui n'est que l'expression d'une **liaison linéaire**



Une corrélation nulle n'indique pas forcément une absence de liaison

Introduction

Références

A-Définition et hypothèses

Exemples
Identifiabilité

B-Estimation

EMC
Propriétés
EMV

C-Lois des estimateurs

IC
IC Préviation
RC

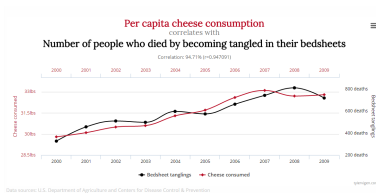
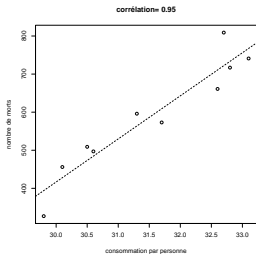
D-Tests

de Student
de Wald
de Fisher

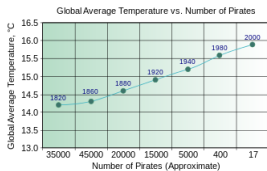
E-Validation

Sélection de variables et choix de modèle

Mais attention : corrélation \neq causalité



source : <http://tylervigen.com/spurious-correlations>



source : Wikipédia

Introduction

Références

A-Définition et hypothèses

Exemples
Identifiabilité

B-Estimation

EMC
Propriétés
EMV

C-Lois des estimateurs

IC
IC Préviation
RC

D-Tests

de Student
de Wald
de Fisher

E-Validation

Sélection de variables et choix de modèle

- ▶ **Données d'apprentissage** :

$$\mathcal{D} = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\} \quad (\text{i.i.d. } \sim \mathbf{P})$$

- ▶ **Prédicteur** : $f : \mathcal{X} \rightarrow \mathcal{Y}$ mesurable

- ▶ **Fonction de perte ou de coût** : $\ell(f(\mathbf{X}), Y)$ mesure avec laquelle qualité $f(\mathbf{X})$ "prédit" Y , d'où le **risque** :

$$\mathcal{R}(f) = \mathbf{E}\ell(Y, f(\mathbf{X}))$$

↪ souvent $\ell(f(\mathbf{X}), Y) = \|f(\mathbf{X}) - Y\|^2$ ou
 $\ell(f(\mathbf{X}), Y) = \mathbf{1}_{Y \neq f(\mathbf{X})}$

But : apprendre une règle pour construire un **classifieur** / **régresseur** $\hat{f} \in \mathcal{F}$ à partir des données d'apprentissage \mathcal{D}_n avec un **risque** $\mathcal{R}(\hat{f})$ **petit en moyenne** ou avec grande probabilité par rapport à \mathcal{D}_n .

Fonction de régression

Approcher Y par une fonction $h(X)$ de la variable X .

Risque quadratique :

$$R(h) = \mathbb{E}[(Y - h(X))^2] = \int (y - h(x))^2 dP(x, y).$$

Théorème

La fonction qui minimise le risque quadratique $R(\cdot)$ est $m(X) = \mathbb{E}(Y|X)$, l'espérance de Y conditionnellement à X

$$\forall x, m(x) = \mathbb{E}(Y|X = x).$$

*Elle est appelée **fonction de régression**.*

- ▶ PB : on ne sait pas la calculer dans le cas général
- ▶ Si X est déterministe, on travaille également à $X = x$ fixé... et les développements sont identiques.

↪ poser des hypothèses de modélisation.

Introduction

Références

A-Définition et
hypothèses

Exemples
Identifiabilité

B-Estimation

EMC
Propriétés
EMV

C-Lois des
estimateurs

IC
IC Prévision
RC

D-Tests

de Student
de Wald
de Fisher

E-Validation

Sélection de
variables et choix
de modèle

J. Pagès. *Statistique générales pour utilisateurs*. Presses Universitaires de Rennes, Rennes, 2005.

J.-M. Azais et J.-M. Bardet. *Le modèle linéaire par l'exemple*. Dunod, Paris, 2005.

P.-A. Cornillon et E. Matzner-Løber. *Régression Théorie et applications*. Springer, Paris, 2007.

P.-A. Cornillon et al. *Statistiques avec R*. Presses Universitaires de Rennes, Rennes, 2008.

Sommaire

Introduction

A-Définition et hypothèses

B-Estimation

C-Lois des estimateurs

D-Tests

E-Validation

Sélection de variables et choix de modèle

Régularisation

Introduction

Références

**A-Définition et
hypothèses**

Exemples
Identifiabilité

B-Estimation

EMC
Propriétés
EMV

**C-Lois des
estimateurs**

IC
IC Préviation
RC

D-Tests

de Student
de Wald
de Fisher

E-Validation

**Sélection de
variables et choix
de modèle**

Biais/variance

Régression linéaire : Y est une variable quantitative

à partir de n observations $(x_1, Y_1), \dots, (x_n, Y_n)$

1. Modèle à **bruit additif** posant une relation de **linéarité de l'espérance** $\mathbf{E}(Y_i|x_i)$ en un **paramètre** θ :

$$\begin{aligned} Y_i &= x_{i1}\theta_1 + \dots + x_{ip}\theta_p + \varepsilon_i \\ &= x_i\theta + \varepsilon_i, \end{aligned}$$

2. Les **résidus sont centrés** : $\mathbf{E}(\varepsilon_i|x_i) = 0$.
3. La **variance de l'erreur est constante** : $\text{var}(\varepsilon_i|x_i) = \sigma^2$.
4. Les résidus sont **décorrélés** : $\text{cov}(\varepsilon_i, \varepsilon_j|x_i, x_j) = 0$ pour $i \neq j$.
5. Eventuellement : ε_i est gaussien

$\hookrightarrow x_i$ est la valeur **fixée** des variables explicatives pour la i -ème observation. Elle n'est pas aléatoire

$\hookrightarrow m(x_i) = x_i\theta$ est appelée **fonction de régression**

- ▶ Matriciellement : $Y = X\theta + \varepsilon$

$$Y_{n \times 1} = X_{n \times p} \theta_{p \times 1} + \varepsilon_{n \times 1};$$

$$\mathbb{E}(\varepsilon_{n \times 1}) = 0; \quad \text{var}(\varepsilon_{n \times 1}) = \sigma^2 Id_n,$$

X est la **matrice du plan d'expérience**, concaténation des n vecteurs lignes x_i ou des p variables colonnes $X_j = (x_{1j}, \dots, x_{nj})'$.

Remarque : En général, $\forall i, x_{i1} = 1$, X_1 est l'**intercept**.

- ▶ Modèle linéaire gaussien

$$Y = m + \varepsilon; \quad \varepsilon \sim \mathcal{N}(0, \Sigma); \quad (\Sigma = \sigma^2 Id_n)$$

où $m \in V$, sous espace vectoriel de \mathbb{R}^n de dimension p .

Introduction

Références

A-Définition et
hypothèses

Exemples

Identifiabilité

B-Estimation

EMC

Propriétés

EMV

C-Lois des
estimateurs

IC

IC Préviation

RC

D-Tests

de Student

de Wald

de Fisher

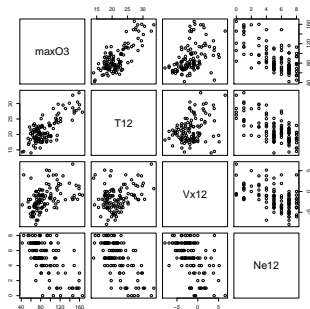
E-Validation

Sélection de
variables et choix
de modèle

Biais/variance

Régression multiple

- ▶ les covariables X sont quantitatives
- ▶ si $p = 2$ dont un intercept : régression **simple**
 $m(x_i) = \theta_1 + \theta_2 x_i$; $x_i = (1 \ x_i)$; $\theta = (\theta_1, \theta_2)'$
- ▶ si $p > 2$, régression **multiple**



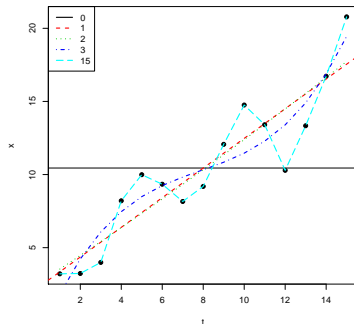
$$\text{maxO3} = \theta_1 + \text{T12 } \theta_2 + \text{Vx12 } \theta_3 + \text{Ne12 } \theta_4 + \varepsilon$$

Régression polynomiale

Un cas particulier de régression multiple : régression polynômiale

$$Y_t = \theta_1 + t\theta_2 + t^2\theta_3 + t^3\theta_4 + \varepsilon_t$$

$$X = \begin{pmatrix} 1 & t_1 & \cdots & t_1^{p-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_n & \cdots & t_n^{p-1} \end{pmatrix}$$



Un régression linéaire (en θ) peut permettre de représenter un phénomène non-linéaire (en t)

Identifiabilité

Définition : deux paramètres différents définissent deux lois différentes

Théorème (CNS d'identifiabilité)

La paramétrisation $\mathbb{E}(Y) = X\theta$ est *identifiable* si et seulement si l'une des propriétés équivalentes suivantes est vérifiée :

- ▶ les colonnes de X sont indépendantes,
- ▶ X est de rang plein,
- ▶ X est injective,
- ▶ $\text{Ker}(X) = \{0\}$.

Définition

On appelle *dimension du modèle* la dimension de $\text{Im}(X)$.

On a : $\dim(\text{Im}(X)) \leq \dim(\theta)$

\hookrightarrow On suppose dans la suite que le modèle est identifiable ou *régulier*.

Sommaire

Introduction

A-Définition et hypothèses

B-Estimation

C-Lois des estimateurs

D-Tests

E-Validation

Sélection de variables et choix de modèle

Régularisation

Introduction

Références

A-Définition et
hypothèses

Exemples
Identifiabilité

B-Estimation

EMC
Propriétés
EMV

C-Lois des
estimateurs

IC
IC Préviation
RC

D-Tests

de Student
de Wald
de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Biais/variance

Estimateur des Moindres Carrés (EMC)

Recherche un estimateur de θ sous forme d'un minimiseur de la **somme des carrés résiduels** $SCR(\theta) = \|Y - X\theta\|^2$:

$$\hat{\theta} = \arg \min_{\theta \in \Theta} \|Y - X\theta\|^2.$$

On a : $\hat{\theta} = (X'X)^{-1}X'Y$

Proposition

$\hat{Y} := \widehat{\mathbf{E}(Y)} = X\hat{\theta} = H_X Y$ est la projection orthogonale de Y sur $\text{Im}(X)$:

$$H_X = X(X'X)^{-1}X'$$

- ▶ H_X : **Hat matrix**
- ▶ \hat{Y} : **valeurs ajustées**
- ▶ $\hat{\varepsilon} = Y - \hat{Y}$: **résidus** (estimés)

Introduction

Références

A-Définition et hypothèses

Exemples

Identifiabilité

B-Estimation

EMC

Propriétés

EMV

C-Lois des estimateurs

IC

IC Prévision

RC

D-Tests

de Student

de Wald

de Fisher

E-Validation

Sélection de variables et choix de modèle

Exemple : régression linéaire (simple)

Le modèle s'écrit $Y = \theta_1 \mathbb{1} + \theta_2 X_2 + \varepsilon$, où $\mathbb{1}$ est l'intercept, et $X_2 = (x_1, \dots, x_n)'$

$$X'X = \begin{pmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix}$$

$$(X'X)^{-1} = \frac{1}{n \sum_i x_i^2 - \sum_i x_i \sum_i x_i} \begin{pmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{pmatrix}$$

soit

$$\hat{\theta} = \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix} = (X'X)^{-1} X'Y = \begin{pmatrix} \bar{Y} - \bar{x} \hat{\theta}_2 \\ \frac{(\frac{1}{n} \sum_i x_i Y_i) - \bar{x} \bar{Y}}{(\frac{1}{n} \sum_i x_i^2) - \bar{x}^2} \end{pmatrix}$$

La droite de régression **estimée** $\hat{y} = \hat{\theta}_1 + \hat{\theta}_2 x$ passe par le **point moyen** (\bar{x}, \bar{Y}) .

Introduction

Références

A-Définition et
hypothèses

Exemples

Identifiabilité

B-Estimation

EMC

Propriétés

EMV

C-Lois des
estimateurs

IC

IC Prédiction

RC

D-Tests

de Student

de Wald

de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Propriétés ne nécessitant pas l'hyp. gaussienne

Proposition

L'EMC $\hat{\theta}$ de θ est

- ▶ *sans biais* : $\mathbb{E}(\hat{\theta}) = \theta$
- ▶ *de variance* $\text{var}(\hat{\theta}) = \sigma^2(X'X)^{-1}$.

De plus, $\hat{\theta}$ vérifie la propriété de **Gauss-Markov** :

Théorème (Gauss-Markov)

Parmi les estimateurs *linéaires et sans biais* de θ , l'EMCO $\hat{\theta}$ est de variance minimum.

Estimateurs de la variance σ^2 :

- ▶ $\hat{s}^2 = \text{SCR}(\hat{\theta})/n$ est biaisé à distance finie, et asymptotiquement sans biais
- ▶ $\hat{\sigma}^2 = \text{SCR}(\hat{\theta})/(n - p)$ est sans biais

Introduction

Références

A-Définition et
hypothèses

Exemples
Identifiabilité

B-Estimation

EMC
Propriétés
EMV

C-Lois des
estimateurs

IC
IC Prévision
RC

D-Tests

de Student
de Wald
de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Estimateur du Maximum de Vraisemblance (EMV)

Si la loi de ε est connue (gaussienne), la **vraisemblance** de l'échantillon est

$$L_n(\theta, \sigma^2; Y) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\left(-\frac{1}{2\sigma^2} \underbrace{\|Y - X\theta\|^2}_{SCR(\theta)}\right).$$

L'EMV $\hat{\beta} = (\hat{\theta}, \hat{\sigma}^2)$ est

$$\hat{\beta} = \arg \max_{\beta \in \mathbb{R}^p \times \mathbb{R}^{+*}} \log L_n(\beta; Y).$$

D'où

$$\hat{\theta} = (X'X)^{-1}X'Y.$$

... identique à l'EMC. L'EMV de σ^2 est

$$\hat{\sigma}^2 = \frac{1}{n} \|Y - X\hat{\theta}\|^2 = \frac{1}{n} SCR(\hat{\theta})$$

Introduction

Références

A-Définition et
hypothèses

Exemples

Identifiabilité

B-Estimation

EMC

Propriétés

EMV

C-Lois des
estimateurs

IC

IC Prévision

RC

D-Tests

de Student

de Wald

de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Sommaire

Introduction

A-Définition et hypothèses

B-Estimation

C-Lois des estimateurs

D-Tests

E-Validation

Sélection de variables et choix de modèle

Régularisation

Introduction

Références

A-Définition et
hypothèses

Exemples
Identifiabilité

B-Estimation

EMC
Propriétés
EMV

C-Lois des
estimateurs

IC
IC Préviation
RC

D-Tests

de Student
de Wald
de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Biais/variance

Théorème

L'EMV $\hat{\beta} = (\hat{\theta}, \hat{s}^2)$ du modèle linéaire **gaussien**

$$Y = X\theta + \varepsilon, \quad \varepsilon \sim \mathcal{N}_p(0, \sigma^2 I_n), \quad \theta \in \mathbb{R}^p, \quad \sigma \in \mathbb{R}^{+*},$$

supposé régulier, suit la loi suivante :

$$\hat{\theta} \sim \mathcal{N}(\theta, \sigma^2(X'X)^{-1}); \quad \hat{s}^2 = \frac{\|Y - X\hat{\theta}\|^2}{n} \sim \frac{\sigma^2}{n} \chi^2(n-p).$$

De plus, $\hat{\theta}$ et \hat{s}^2 sont indépendants

Csq : Soit A une matrice $q \times p$

$$A\hat{\theta} \sim \mathcal{N}_q(A\theta, \sigma^2 A(X'X)^{-1}A')$$

Introduction

Références

A-Définition et
hypothèses

Exemples

Identifiabilité

B-Estimation

EMC

Propriétés

EMV

C-Lois des
estimateurs

IC

IC Prévision

RC

D-Tests

de Student

de Wald

de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Biais/variance

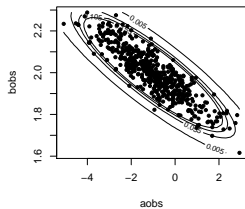
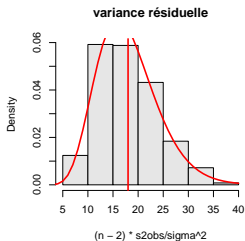
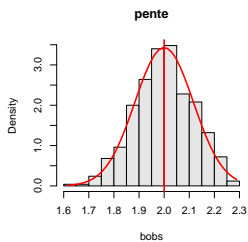
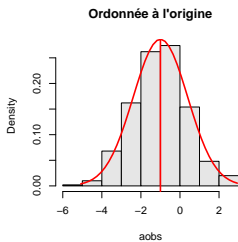
Théorème (Cochran)

Si $Y \sim \mathcal{N}(m, \sigma^2 I_n)$, et si $E_1 \oplus \dots \oplus E_r = \mathbb{R}^n$ est une décomposition de \mathbb{R}^n en p sous-espaces orthogonaux, alors les projections orthogonales $H_{E_1}(Y), \dots, H_{E_r}(Y)$ forment des vecteurs gaussiens indépendants. De plus,

$$\|H_{E_j}(Y)\|^2 \sim \sigma^2 \chi^2(\text{Dim}(E_j), \|H_{E_j}(m)\|^2).$$

Illustration simulée

500 échantillons (Y, X) de taille $n = 30$, avec $x = 1 : 30$ et $m(x) = -1 + 2x$



Théorème

Soit L de dimension $1 \times p$ définissant une forme linéaire $L\theta$ du paramètre θ d'un modèle linéaire **gaussien** identifiable. Si σ^2 est estimée par l'estimateur sans biais $\hat{\sigma}^2$, alors :

$$\frac{L\hat{\theta} - L\theta}{\sqrt{\hat{\sigma}^2 L(X'X)^{-1}L'}} \sim \mathcal{T}(n - p).$$

$\mathcal{T}(n - p)$ est la loi de **Student** à $n - p$ degrés de liberté.

Definition

Soit $N \sim \mathcal{N}(0, 1)$, et $V \sim \chi^2(t)$, indépendante de N . Alors, la variable Z suivante suit une loi de Student à t degrés de liberté :

$$Z = \frac{N}{\sqrt{V/t}} \sim \mathcal{T}(t).$$

Introduction

Références

A-Définition et
hypothèses

Exemples
Identifiabilité

B-Estimation

EMC
Propriétés
EMV

C-Lois des
estimateurs

IC
IC Prévision
RC

D-Tests

de Student
de Wald
de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Théorème

Soit A de dimension $q \times p$ de rang $r \leq q \leq p$. Alors

$$\frac{1}{r\hat{\sigma}^2}(A(\hat{\theta} - \theta))'[A(X'X)^{-1}A']^{-1}A(\hat{\theta} - \theta) \sim \mathcal{F}(r, n - p).$$

suit une loi de **Fisher** à r et $n - p$ degrés de liberté.

Definition

Soient $D_1 \sim \chi^2(t_1)$ et $D_2 \sim \chi^2(t_2)$ indépendantes. Alors F

$$F = \frac{D_1/t_1}{D_2/t_2} \sim \mathcal{F}(t_1, t_2).$$

suit la loi de Fisher $\mathcal{F}(t_1, t_2)$.

Introduction

Références

A-Définition et
hypothèses

Exemples
Identifiabilité

B-Estimation

EMC
Propriétés
EMV

C-Lois des
estimateurs

IC
IC Prévision
RC

D-Tests

de Student
de Wald
de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Théorème

Soit L une matrice de dimension $1 \times p$. Un **intervalle de confiance** de niveau $1 - \alpha$ d'une forme linéaire de $L\theta$ est donné par

$$\left[L\hat{\theta} \pm q_{1-\alpha/2}^{\mathcal{T}_{n-p}} \hat{\sigma} \sqrt{L(X'X)^{-1}L'} \right],$$

où $q_{1-\alpha/2}^{\mathcal{T}_{n-p}}$ est le quantile d'ordre $1 - \alpha/2$ de la loi de Student à $n - p$ degrés de liberté.

Applications : IC d'une composante de θ , IC de $\mathbb{E}(Y|x_0)$

Introduction

Références

A-Définition et hypothèses

Exemples

Identifiabilité

B-Estimation

EMC

Propriétés

EMV

C-Lois des estimateurs

IC

IC Préviation

RC

D-Tests

de Student

de Wald

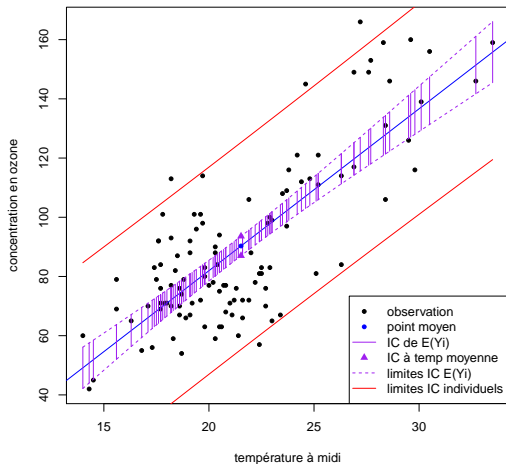
de Fisher

E-Validation

Sélection de variables et choix de modèle

Biais/variance

IC d'une valeur moyenne ou d'une prédiction individuelle



Introduction

Références

A-Définition et
hypothèses

Exemples

Identifiabilité

B-Estimation

EMC

Propriétés

EMV

C-Lois des
estimateurs

IC

IC Prédiction

RC

D-Tests

de Student

de Wald

de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Intervalle de confiance d'une prévision

A partir de n observations ($\hookrightarrow \hat{\theta}$), prédire une **nouvelle** observation **individuelle indépendante** sous la condition d'expérience $x_{n+1} = (x_{n+1,1}, \dots, x_{n+1,p})$

- ▶ modèle : $Y_{n+1} = x_{n+1}\theta + \varepsilon_{n+1}$,
- ▶ observation prédite : $\hat{Y}_{n+1}^p = x_{n+1}\hat{\theta}$
- ▶ erreur de prévision

$$\hat{\varepsilon}_{n+1}^p = Y_{n+1} - \hat{Y}_{n+1}^p,$$

$$\mathbf{E}(\hat{\varepsilon}_{n+1}^p) = 0, \quad \text{var}(\hat{\varepsilon}_{n+1}^p) = \sigma^2[x_{n+1}(X'X)^{-1}x'_{n+1} + \mathbf{1}].$$

- ▶ IC de prévision de niveau $1 - \alpha$:

$$[x_{n+1}\hat{\theta} \pm q_{1-\alpha/2}^{T_{n-p}} \hat{\sigma} \sqrt{x_{n+1}(X'X)^{-1}x'_{n+1} + \mathbf{1}}].$$

Introduction

Références

A-Définition et
hypothèses

Exemples
Identifiabilité

B-Estimation

EMC
Propriétés
EMV

C-Lois des
estimateurs

IC
IC Prévision
RC

D-Tests

de Student
de Wald
de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Théorème

Soit A une matrice de dimension $q \times p$ et de rang q . Une **région de confiance** de Wald de niveau $1 - \alpha$ de $A\theta$ est donnée par

$$RC_{\alpha}(A\theta) =$$

$$\left\{ u \in \mathbb{R}^q, \frac{1}{r\widehat{\sigma}^2} (A\widehat{\theta} - u)' [A(X'X)^{-1}A']^{-1} (A\widehat{\theta} - u) \leq f_{r, n-p, 1-\alpha} \right\}$$

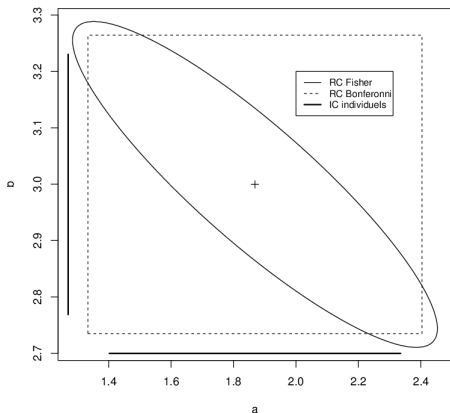
où $f_{r, n-p, 1-\alpha}$ est le quantile d'ordre $1 - \alpha$ de la loi de $\mathcal{F}(r, n - p)$.

Comparaison avec les intervalles de Bonferroni

$$\mathbb{P}((\theta_1, \theta_2) \notin RC_\alpha(\theta_1, \theta_2))$$

$$\leq \mathbb{P}(\theta_1 \notin IC_{1-\alpha/2}(\theta_1)) + \mathbb{P}(\theta_2 \notin IC_{1-\alpha/2}(\theta_2)) \leq \alpha.$$

K intervalles de confiance de Bonferroni de risque α/K forment une région de confiance de risque simultané α .



Sommaire

Introduction

A-Définition et hypothèses

B-Estimation

C-Lois des estimateurs

D-Tests

E-Validation

Sélection de variables et choix de modèle

Régularisation

Introduction

Références

A-Définition et
hypothèses

Exemples
Identifiabilité

B-Estimation

EMC
Propriétés
EMV

C-Lois des
estimateurs

IC
IC Prévision
RC

D-Tests

de Student
de Wald
de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Biais/variance

Test de Student

Test d'une **relation affine** des composantes du paramètre $L\theta = c$, avec L matrice ligne de dimension p .

- ▶ test bilatéral : $H_0 : L\theta = c$ contre $H_1 : L\theta \neq c$
- ▶ statistique de test

$$T = \frac{L\hat{\theta} - c}{\hat{\sigma} \sqrt{L(X'X)^{-1}L'}} \sim_{H_0} \mathcal{T}(n-p),$$

- ▶ région de rejet du test bilatère de risque α

$$\mathcal{R}_\alpha = \{|T| > t_{n-p, 1-\alpha/2}\}.$$

- ▶ utilisé pour tester la **significativité** (non-nullité) d'une composante de θ

Remarque : test unilatéral : $H_0 : L\theta = c$ contre $H_1 : L\theta < c$
de région de rejet à gauche :

$$\mathcal{R}_\alpha = \{T < t_{n-p, \alpha}\}.$$

Introduction

Références

A-Définition et
hypothèses

Exemples

Identifiabilité

B-Estimation

EMC

Propriétés

EMV

C-Lois des
estimateurs

IC

IC Prédiction

RC

D-Tests

de Student

de Wald

de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Biais/variance

$$\text{Exemple : } Y = \mu + \beta_1 V1 + \beta_2 V2 + \beta_3 V3 + \varepsilon$$

```
> out=lm(Y~V1+V2+V3, data=df); summary(out)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	H0	H1
(Intercept)	-20.3129	51.8246	-0.392	0.701		
V1	-3.2241	3.6100	-0.893	0.386	~V2+V3	vs ~V1+V2+V3
V2	0.2334	0.2768	0.843	0.412	~V1+V3	vs ~V1+V2+V3
V3	20.3895	1.2662	16.103	7.1e-11 ***	~V1+V2	vs ~V1+V2+V3

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.163 on 15 degrees of freedom

F-statistic: 584.6 on 3 and 15 DF, p-value: 9.386e-16

Introduction

Références

A-Définition et
hypothèses

Exemples

Identifiabilité

B-Estimation

EMC

Propriétés

EMC

C-Lois des
estimateurs

IC

IC Prévision

RC

D-Tests

de Student

de Wald

de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Biais/variance

Test de Wald avec loi exacte

Modèle emboîté $\omega \subset \Omega$ défini par des hypothèses **affines** :
 $A\theta = c$ contre $A\theta \neq c$, A matrice $r \times p$ de rang r ,
 $\dim(\omega) = p - r = q$

$$W_{\text{exact}} = W/r = \frac{1}{r\hat{\sigma}^2} (A\hat{\theta} - c)' [A(X'X)^{-1}A']^{-1} (A\hat{\theta} - c) \sim_{(H_0)} \mathcal{F}(r, n-p)$$

- tester la significativité **globale** de la régression : tous les coeff. sont nuls sont celui de l'intercept : $(H_0) : A\theta = 0$ contre $(H_0) : A\theta \neq 0$

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

F-statistic: 584.6 on 3 and 15 DF, p-value: 9.386e-16

- tester des modèles **emboîtés**.

Petite interprétation géométrique

Soit ω un sev de $\Omega \in \mathbb{R}^n$, $\dim(\omega) = q < \dim(\Omega) = p$.

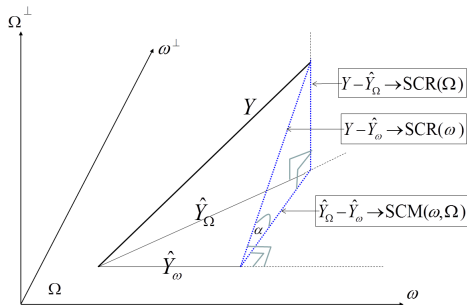
On note $\hat{Y}_\Omega = H_\Omega(Y)$. On a :

$$\|Y - \hat{Y}_\omega\|^2 = \|\hat{Y}_\Omega - \hat{Y}_\omega\|^2 + \|Y - \hat{Y}_\Omega\|^2$$

$$\text{SCR}(\omega) = \text{SCM}(\omega, \Omega) + \text{SCR}(\Omega).$$

$$F = \frac{(\text{SCR}(\omega) - \text{SCR}(\Omega))/(p - q)}{\text{SCR}(\Omega)/(n - p)} = \frac{\text{SCM}(\omega, \Omega)/(p - q)}{\text{SCR}(\Omega)/(n - p)}$$

$$\sim_\omega \mathcal{F}(p - q, n - p).$$



Théorème (Test de Fisher d'un sous-modèle linéaire)

Soit le modèle linéaire gaussien $Y = m + \varepsilon$, $m \in \Omega$,
 $\varepsilon \sim \mathcal{N}(0, I_n)$, de dimension $\dim(\Omega) < n$. Soit ω un sev de Ω
tel que $\dim(\omega) = q < p$. Le test de Fisher

$$H_0 : m \in \omega \text{ contre } H_1 : m \in \Omega \setminus \omega.$$

de région de rejet

$$\mathcal{R}_\alpha = \{F > f_{p-q, n-p, 1-\alpha}\},$$

où $f_{p-q, n-p, 1-\alpha}$ est le quantile de la loi de Fisher
 $\mathcal{F}(p - q, n - p)$, est de niveau α .

- ▶ = W_{exact} quand ω est défini par $A\theta = 0$
- ▶ = test du rapport des vraisemblances maximales avec loi exacte

Introduction

Références

A-Définition et
hypothèses

Exemples

Identifiabilité

B-Estimation

EMC

Propriétés

EMV

C-Lois des
estimateurs

IC

IC Prévision

RC

D-Tests

de Student

de Wald

de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Biais/variance

Sommaire

Introduction

A-Définition et hypothèses

B-Estimation

C-Lois des estimateurs

D-Tests

E-Validation

Sélection de variables et choix de modèle

Régularisation

Introduction

Références

A-Définition et
hypothèses

Exemples
Identifiabilité

B-Estimation

EMC
Propriétés
EMV

C-Lois des
estimateurs

IC
IC Préviation
RC

D-Tests

de Student
de Wald
de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Biais/variance

- ▶ L'étape de validation fait partie intégrante de la démarche statistique
- ▶ Elle peut remettre en cause le modèle initialement choisi, donner des idées de modifications pour repartir pour un tour...
- ▶ Outils : tests, critères, étude graphique
↔ étude des résidus

Introduction

Références

A-Définition et hypothèses

Exemples

Identifiabilité

B-Estimation

EMC

Propriétés

EMV

C-Lois des estimateurs

IC

IC Préviation

RC

D-Tests

de Student

de Wald

de Fisher

E-Validation

Sélection de variables et choix de modèle

Biais/variance

Critère du R^2

Si ω est le modèle iid :

- Coefficient de **détermination**

$$R^2 = \frac{SCM}{SCT} = \frac{\|\hat{Y} - \bar{y}\mathbf{1}\|^2}{\|Y - \bar{y}\mathbf{1}\|^2} = \cos^2 \alpha$$

Part de la variabilité expliquée par le modèle sur la variabilité totale

- Indication de la **qualité d'ajustement**, mais pas forcément de la qualité de prévision
- Pour compte de la dimension de $Im(X)$: $\hookrightarrow R^2$ ajusté :

$$R_a^2 = 1 - \frac{n-1}{n-p} \frac{\|\hat{\epsilon}\|^2}{\|Y - \bar{y}\mathbf{1}\|^2}$$

Multiple R-squared: 0.9915, Adjusted R-squared: 0.9898

- ▶ **Significativité globale** : test de Fisher du modèle iid contre le modèle d'étude.
- ▶ **Significativité d'un régresseur** : \Leftrightarrow des stratégies de sélection de variables.
- ▶ **Adéquation** : comparer le modèle d'étude Ω à un surmodèle Ω_S peu contestable qui contient Ω
- ▶ Hypothèse **gaussienne** : test de Kolmogorov-Smirnov ou test de Shapiro-Wilks.

Proposition

Dans le modèle linéaire gaussien, les résidus $\hat{\varepsilon} = Y - \hat{Y}$ possèdent les propriétés suivantes :

- ▶ *centrés* : $\mathbf{E}(\hat{\varepsilon}) = 0$.
- ▶ *hétéroscédastiques* : $\text{var}(\hat{\varepsilon}) = \sigma^2(I - H)$.
- ▶ *décorrélés avec les valeurs estimées* : $\text{cov}(\hat{Y}, \hat{\varepsilon}) = 0$.
- ▶ Si $\mathbb{I} \in \text{Im}(X)$, les résidus estimés sont *linéairement dépendants*.

Introduction

Références

A-Définition et
hypothèses

Exemples

Identifiabilité

B-Estimation

EMC

Propriétés

EMV

C-Lois des
estimateurs

IC

IC Préviation

RC

D-Tests

de Student

de Wald

de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Différents résidus

- ▶ Les **résidus normalisés** éliminent l'hétéroscédasticité

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

- ▶ **résidus standardisés**, parfois également appelés **studentisés** : de variance 1

$$t_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}.$$

- ▶ Les résidus **studentisés par validation croisée** sont définis par

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

où $\hat{\sigma}_{(i)}$ est calculé dans un échantillon privé de l'observation i . Dans le cas gaussien, $t_i^* \sim \mathcal{T}(n - 1 - p)$.

↪ points **aberrants**

Différents types de graphiques permettent une étude visuelle :

- ▶ valeurs estimées en fonction de valeurs observées,
- ▶ résidus en fonction des valeurs estimées,
- ▶ résidus en fonction d'une valeur de covariable,
- ▶ résidus de l'observation $i + 1$ en fonction du résidu de l'observation i ,
- ▶ graphe quantile/quantile des résidus.

Introduction

Références

A-Définition et
hypothèses

Exemples
Identifiabilité

B-Estimation

EMC
Propriétés
EMV

C-Lois des
estimateurs

IC
IC Prévision
RC

D-Tests

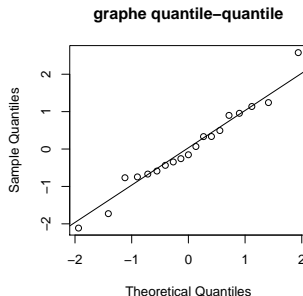
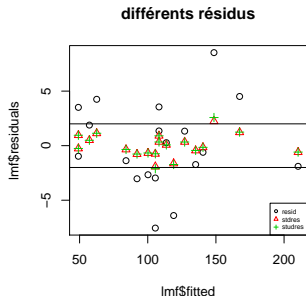
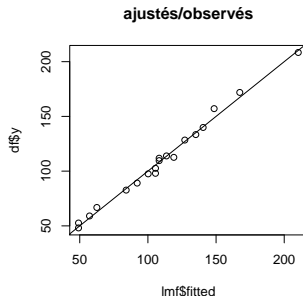
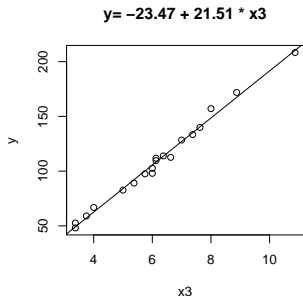
de Student
de Wald
de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Biais/variance

Exemple de diagnostics visuels



Introduction

Références

A-Définition et
hypotheses

Exemples

Identifiabilité

B-Estimation

EMC

Propriétés

EMV

C-Lois des
estimateurs

IC

IC Prévion

RC

D-Tests

de Student

de Wald

de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Influence d'une observation

- ▶ Données **influentes** ou points **levier** : h_{ii} proche de 1 :

$$\hat{y}_i = h_{ii}y_i + \sum_{j \neq i} h_{ij}y_j$$

h_{ii} représente le poids de l'observation sur son propre ajustement

- ▶ \neq Valeurs **aberrantes** : un fort résidu

$$|t_i^*| \gg t_{n-p-1}(1 - \alpha/2)$$

- ▶ **Distance de Cook** : mesure l'influence de l'observation i sur l'estimation du paramètre θ

$$C_i = \frac{(\hat{y}_i - x_i\hat{\beta}_{(i)})^2}{p\hat{\sigma}^2} = \frac{h_{ii}}{p(1 - h_{ii})^2} t_i^2$$

Influence d'une variable

La variance de $\hat{\theta}_j$ est affectée par la corrélation de X_j avec les autres covariables explicatives :

$$\text{var}(\hat{\theta}_j) = \frac{\sigma^2}{(1 - R_j^2) \sum_{i=1}^n (x_{ij} - \bar{X}_j)^2},$$

où R_j^2 est le R^2 de la régression de X_j par rapport aux autres variables explicatives

Facteur d'inflation de la variance X_j :

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

- + X_j est corrélée linéairement aux autres variables,
- + R_j est proche de 1,
- + VIF_j est grand

Introduction

Références

A-Définition et
hypothèses

Exemples

Identifiabilité

B-Estimation

EMC

Propriétés

EMV

C-Lois des
estimateurs

IC

IC Prévision

RC

D-Tests

de Student

de Wald

de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Biais/variance

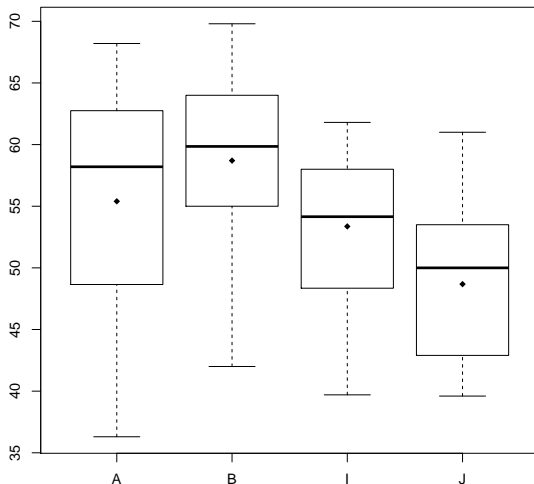
Cas de variables explicatives qualitatives

Variables explicatives qualitatives

- ▶ ANOVA(1) : un facteur explicatif qualitatif.
↔ : expliquer le poids des portées par le génotype de leur mère
- ▶ ANCOVA : un facteur explicatif qualitatif et un facteur explicatif quantitatif.
↔ : expliquer le poids des portées par le génotype et le poids de la mère
- ▶ ANOVA(2) : deux facteurs explicatifs qualitatifs.
↔ : expliquer le poids des portées par le génotype de la mère et le génotype de la portée

ANOVA1

Exemple : Poids des portées en fonction du génotype de la mère.



ANOVA1 : analyse de la variance à un facteur

1 variable explicative **qualitative** à l niveaux.

- ▶ Y_{ik} : observation du k -ème individu du groupe i .

$$Y_{ik} = \alpha_i + \varepsilon_{ik}, \quad i = 1, \dots, l; \quad k = 1, \dots, n_i; \quad n = \sum_{i=1}^l n_i$$

$$\mathbb{E}(\varepsilon|x) = 0; \quad \text{var}(\varepsilon|x) = \sigma^2 I_n$$

- ▶ vecteur colonne des observations
 $Y = (Y_{11}, \dots, Y_{1n_1}, \dots, Y_{l1}, \dots, Y_{ln_l})'$,
- ▶ paramètre $\theta = (\alpha_1, \dots, \alpha_l)'$
- ▶ matrice du plan d'expérience

$$X = \begin{pmatrix} \mathbb{1}_{n_1} & & \\ & \ddots & \\ & & \mathbb{1}_{n_l} \end{pmatrix}; \quad Y = X\theta + \varepsilon$$

où $\mathbb{1}_{n_i}$ étant un vecteur colonne de 1 de longueur n_i .

ANOVA1 : analyse de la variance à un facteur

Si on veut faire figurer explicitement un intercept

- ▶ Y_{ik} : observation du k -ème individu du groupe i .

$$Y_{ik} = \mu + \alpha_i + \varepsilon_{ik}, \quad i = 1, \dots, I; \quad k = 1, \dots, n_i; \quad n = \sum_{i=1}^I n_i$$

$$\mathbb{E}(\varepsilon|x) = 0; \quad \text{var}(\varepsilon|x) = \sigma^2 I_n$$

- ▶ vecteur colonne des observations
 $Y = (Y_{11}, \dots, Y_{1n_1}, \dots, Y_{I1}, \dots, Y_{In_I})'$,
- ▶ paramètre $\theta = (\mu, \alpha_1, \dots, \alpha_I)'$
- ▶ matrice du plan d'expérience

$$X = \begin{pmatrix} \mathbb{1}_{n_1} & \mathbb{1}_{n_1} & & \\ \vdots & & \ddots & \\ \mathbb{1}_{n_I} & & & \mathbb{1}_{n_I} \end{pmatrix}; \quad Y = X\theta + \varepsilon$$

où $\mathbb{1}_{n_i}$ étant un vecteur colonne de 1 de longueur n_i .

Introduction

Références

A-Définition et
hypothèses

Exemples
Identifiabilité

B-Estimation

EMC
Propriétés
EMV

C-Lois des
estimateurs

IC
IC Préviation
RC

D-Tests

de Student
de Wald
de Fisher

E-Validation

Sélection de
variables et choix
de modèle

- ▶ La paramétrisation ainsi définie n'est **pas identifiable** : $\theta = (\mu, \alpha_1, \dots, \alpha_I)'$ et $\tilde{\theta} = (\mu - h, \alpha_1 + h, \dots, \alpha_I + h)'$ définissent la même loi, $\dim(\text{Ker}(X)) = 1$
- ▶ Restreindre l'espace des paramètres par une **condition d'identifiabilité** pour obtenir une solution unique θ_c dans un sous-espace $\mathcal{C} = \{C\theta = 0\}$ où la restriction de X est injective
- ▶ L'unique estimateur de θ dans \mathcal{C} est

$$\hat{\theta}_c = (X'X + C'C)^{-1}X'Y$$

- ▶ La **dimension du modèle** est $\dim(\text{Im}X) < \dim(\theta)$
- ▶ Quelle que soit la CI choisie, $X\theta_c = X\theta$

Introduction

Références

A-Définition et hypothèses

Exemples

Identifiabilité

B-Estimation

EMC

Propriétés

EMV

C-Lois des estimateurs

IC

IC Prévision

RC

D-Tests

de Student

de Wald

de Fisher

E-Validation

Sélection de variables et choix de modèle

Biais/variance

ANOVA1 : Exemples de CI

- ▶ R : $\alpha_1 = 0$, soit $C = (0 \ 1 \ 0 \dots 0)$
 - ▶ μ est la valeur de la réponse pour le niveau 1
 - ▶ α_i représente la différence de réponse entre le niveau i et le niveau 1

$$\begin{aligned}\mathbb{E}(Y_{1k}) &= m_1 = \mu \\ \mathbb{E}(Y_{2k}) &= m_2 = \mu + \alpha_2 \\ \mathbb{E}(Y_{3k}) &= m_3 = \mu + \alpha_3 \\ \mathbb{E}(Y_{4k}) &= m_4 = \mu + \alpha_4\end{aligned}$$

- ▶ SAS : $\alpha_l = 0$, soit $C = (0 \ 0 \dots 0 \ 1)$
 - ▶ μ est la valeur de la réponse pour le niveau l
 - ▶ α_i représente la différence de réponse entre le niveau i et le niveau l

$$\begin{aligned}\mathbb{E}(Y_{1k}) &= m_1 = \mu + \alpha_1 \\ \mathbb{E}(Y_{2k}) &= m_2 = \mu + \alpha_2 \\ \mathbb{E}(Y_{3k}) &= m_3 = \mu + \alpha_3 \\ \mathbb{E}(Y_{4k}) &= m_4 = \mu\end{aligned}$$

- ▶ $\sum_i \alpha_i = 0$, $\sum_i n_i \alpha_i = 0$, ...

Introduction

Références

A-Définition et
hypothèses

Exemples
Identifiabilité

B-Estimation

EMC
Propriétés
EMV

C-Lois des
estimateurs

IC
IC Préviation
RC

D-Tests

de Student
de Wald
de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Exemple : comparaison de deux CI

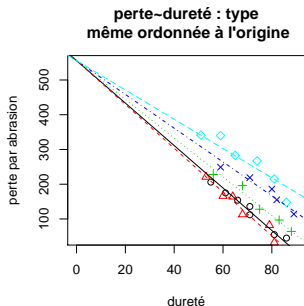
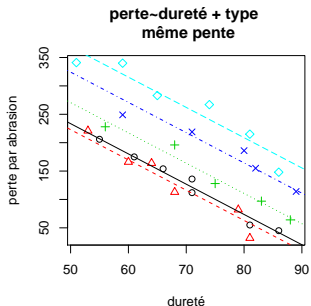
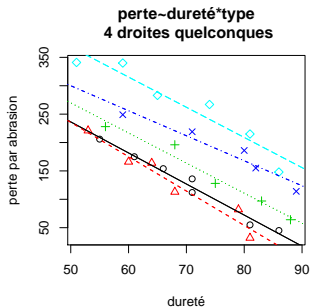
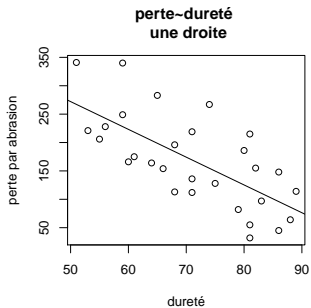
```
> anov1=lm(Wt~Mother,data=genotype); summary(anov1) # CI par défaut MotherA=0
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  55.400      1.910  28.999 <2e-16 ***
MotherB      3.300      2.797   1.180  0.2429
MotherI     -2.038      2.702  -0.754  0.4539
MotherJ     -6.720      2.746  -2.447  0.0175 *
---
Residual standard error: 7.642 on 57 degrees of freedom
Multiple R-squared:  0.1882, Adjusted R-squared:  0.1455
F-statistic: 4.405 on 3 and 57 DF,  p-value: 0.007433

> summary( lm(Wt~Mother,data=genotype,
+           contrast=list(Mother="contr.SAS")) )
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  48.680      1.973  24.672 < 2e-16 ***
MotherA      6.720      2.746   2.447 0.017518 *
MotherB     10.020      2.840   3.528 0.000834 ***
MotherI      4.683      2.746   1.705 0.093647 .
---
Residual standard error: 7.642 on 57 degrees of freedom
Multiple R-squared:  0.1882, Adjusted R-squared:  0.1455
F-statistic: 4.405 on 3 and 57 DF,  p-value: 0.007433

> anova(anov1)
Response: Wt
      Df Sum Sq Mean Sq F value  Pr(>F)
Mother   3  771.6  257.202   4.4045 0.007433 **
Residuals 57 3328.5   58.395
```

Attention à l'interprétation des coefficients...

ANCOVA : exemple



Introduction

Références

A-Définition et
hypothèses

Exemples
Identifiabilité

B-Estimation

EMC
Propriétés
EMV

C-Lois des
estimateurs

IC
IC Préviation
RC

D-Tests

de Student
de Wald
de Fisher

E-Validation

Sélection de
variables et choix
de modèle

ANCOVA : analyse de la covariance (suite)

test de toute sous-hypothèse linéaire

- ▶ égalité des pentes ou de certaines d'entre elles
- ▶ égalité des ordonnées à l'origine ou de certaines d'entre elles
- ▶ si les pentes sont différentes, on dit qu'il y a **interaction** entre le facteur et la variable quantitative t

Introduction

Références

A-Définition et
hypothèses

Exemples
Identifiabilité

B-Estimation

EMC
Propriétés
EMV

C-Lois des
estimateurs

IC
IC Prévision
RC

D-Tests

de Student
de Wald
de Fisher

E-Validation

Sélection de
variables et choix
de modèle

ANCOVA : CI

Y variable à expliquer, F facteur explicatif à I niveaux, t variable explicative quantitative :

$$Y_{ik} = \mu + \alpha_i + (\nu + \beta_i)t_{ik} + \varepsilon_{ik}$$

$y \sim$	modèle et CI	intercept	F <i>i</i>	t	t:F <i>i</i>
F * t	I droites qcq $\alpha_1 = 0, \beta_1 = 0$	μ : ord. orig. droite 1	$\alpha_i - \alpha_1$	ν : pente droite 1	$\beta_i - \beta_1$
F+F:t	I droites qcq $\alpha_1 = 0, \nu = 0$	μ : ord. orig. droite 1	$\alpha_i - \alpha_1$	β_1 : pente droite 1	β_i : pente droite i

Sous modèles particuliers

- ▶ à droites parallèles : $y \sim F+t$
- ▶ à ordonnées à l'origine identiques : $y \sim F:t$

L'argument contrasts de lm permet de changer la CI

```
contrasts = list(F = "contr.SAS") # CI alpha_I=0
```


Sommaire

Introduction

A-Définition et hypothèses

B-Estimation

C-Lois des estimateurs

D-Tests

E-Validation

Sélection de variables et choix de modèle

Régularisation

Introduction

Références

A-Définition et
hypothèses

Exemples
Identifiabilité

B-Estimation

EMC
Propriétés
EMV

C-Lois des
estimateurs

IC
IC Préviation
RC

D-Tests

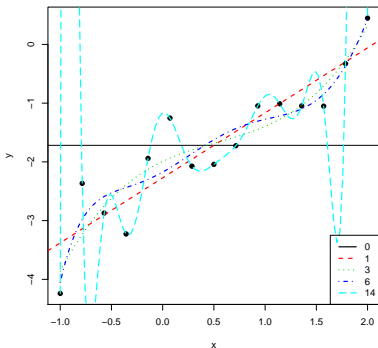
de Student
de Wald
de Fisher

E-Validation

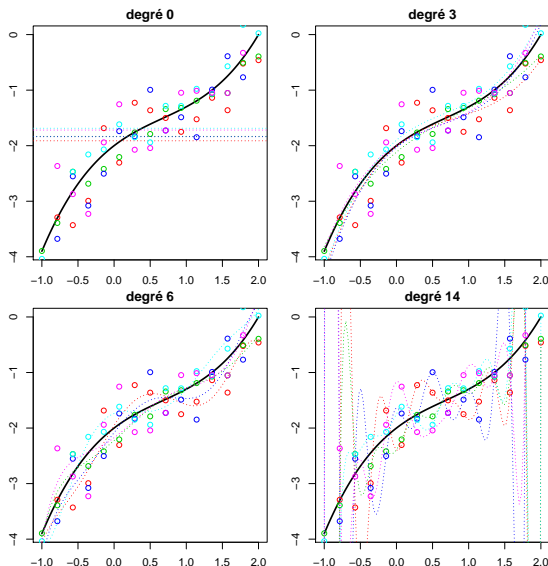
Sélection de
variables et choix
de modèle

Apprentissage supervisé

- ▶ as so far : recherche d'un modèle minimisant l'erreur empirique sur un ensemble donné d'observations
- ▶ Choix entre
 - ▶ \neq modélisations (choix de \mathcal{L} , de la fonction de régression m , ...)
 - ▶ plusieurs régresseurs (sélection de variables)
 - ▶ \neq règles de prédiction/classification



Choix de variable



Introduction

Références

A-Définition et
hypothèses

Exemples

Identifiabilité

B-Estimation

EMC

Propriétés

EMV

C-Lois des
estimateurs

IC

IC Prévision

RC

D-Tests

de Student

de Wald

de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Comprendre ou prévoir ?

► Objectifs

- **Description** : qualité de l'ajustement $SCR = \|Y - \widehat{Y}\|^2$
- **Estimation** : $EQM(\widehat{\theta})$, $EQM(\widehat{m(\theta, X)})$
- **Prédiction** : $EQMP = \mathbb{E}\|Y - \widehat{Y}^p\|^2$

► Comprendre : démarche statistique historique

↔ **estimateur**

► Prévoir : se concentrer sur l'élaboration d'algorithmes prédictifs

↔ **règle d'apprentissage**

L'objectif n'est pas d'estimer un paramètre de la loi, mais plutôt de faire aussi bien que la meilleure des règles d'apprentissage en terme de risque

► Définir et établir la **performance** du modèle

- ▶ On dispose d'un échantillon de résultats d'une expérience $(x_1, Y_1), \dots, (x_n, Y_n)$ concernant une variable d'intérêt Y et p variables explicatives (disons quantitatives pour simplifier)

- ▶ On appelle **modèle** un ensemble de fonctions

Ex : régression linéaire simple, régression polynomiale d'ordre 3, modèle logistique avec lien probit et 3 covariables, fonctions étagées

- ▶ On appelle **famille de modèles**, une collection \mathcal{F} de modèles

Ex : cas de sélection de variables on peut identifier les variables par un numéro :

- ▶ famille exhaustive : $\mathcal{F} = \mathcal{P}(1 \cup \{2, \dots, p\})$
- ▶ famille emboîtée : $\mathcal{F} = \{\{1, \dots, j\} \mid 1 \leq j \leq p\}$

Introduction

Références

A-Définition et hypothèses

Exemples

Identifiabilité

B-Estimation

EMC

Propriétés

EMV

C-Lois des estimateurs

IC

IC Préviation

RC

D-Tests

de Student

de Wald

de Fisher

E-Validation

Sélection de variables et choix de modèle

- Supposons que les données ont été générées à partir d'une fonction d'un modèle \mathcal{M} ayant généré les observations, par ex

$$Y = f_*(X_*, \theta_*) + \varepsilon \text{ où } \varepsilon \sim \mathcal{N}(0, \sigma_*^2)$$

- On modélise

$$\mathbb{E}(Y|X = x) = f(x_\omega, \theta_\omega)$$

- On définit un critère de performance (par exemple, risque quadratique, perte 0/1, ...)

Introduction

Références

A-Définition et hypothèses

Exemples

Identifiabilité

B-Estimation

EMC

Propriétés

EMV

C-Lois des estimateurs

IC

IC Prévision

RC

D-Tests

de Student

de Wald

de Fisher

E-Validation

Sélection de variables et choix de modèle

Conséquence d'une sur-paramétrisation ³

Pas d'oubli de variables, mais certaines sont **superflues**

(ω) : $\mathbb{E}(Y) = X_1\theta_1$ (vrai modèle, inconnu),

$$\dim(\theta_1) = q, \dim(X_1) = n \times q$$

(Ω) : $\mathbb{E}(Y) = X\theta$ (modèle de travail), avec $\theta = (\theta'_1, \theta'_2)'$ de dimension p et $X = [X_1 X_2]$ de dimension $n \times p$.

Le vrai paramètre dans Ω est $\theta_* = (\theta'_1, 0)'$, est estimé par $\hat{\theta}_\Omega = (X'X)^{-1}X'Y$:

▶ $\hat{\theta}_\Omega, \hat{Y}_\Omega = X\hat{\theta}_\Omega$ et $\hat{\sigma}_\Omega^2 = SCR/(n-p)$ sont **sans biais**

▶ avec une **variance plus forte**

$$\underbrace{[(X'_1 X_1 - X'_1 X_2 (X'_2 X_2)^{-1} X'_2 X_1)]^{-1}}_{[\text{var}(\hat{\theta}_\Omega)]_{1:q \times 1:q}} \geq \underbrace{(X'_1 X_1)^{-1}}_{\text{var}(\hat{\theta}_\omega)}$$

Introduction

Références

A-Définition et
hypothèses

Exemples
Identifiabilité

B-Estimation

EMC
Propriétés
EMV

C-Lois des
estimateurs

IC
IC Préviation
RC

D-Tests

de Student
de Wald
de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Conséquence d'une sous-paramétrisation ⁴

Oubli de covariables.

(ω) : $\mathbb{E}(Y) = X_1\theta_1$ (modèle de travail)

(Ω) : $\mathbb{E}(Y) = X\theta$ (vrai modèle, inconnu), avec $\theta = (\theta'_1, \theta'_2)'$
de dimension p et $X = [X_1 X_2]$

$\hat{\theta}_\omega$, \hat{Y}_ω , et $\hat{\sigma}_\omega^2$ en général **biaisés** et de plus faible variance.

Erreur quadratique moyenne (EQM)

Risque quadratique d'utiliser $\hat{\theta}$ à la place de θ_*

$$EQM(\hat{\theta}) = \mathbb{E}(\|\theta_* - \hat{\theta}\|^2)$$

Risque quadratique d'utiliser $\hat{F} = F(\hat{\theta}_\omega, X_\omega)$ à la place de $F_*(\theta_*, X)$

$$\begin{aligned}EQM(\hat{F}) &= \|\hat{F} - F_*\|^2 \\ &= \|\text{Biais}(\hat{F})\|^2 + \text{tr}(\text{var}(\hat{F})) \\ &= \|\text{Biais}(X_\omega \hat{\theta}_\omega)\|^2 + |\omega| \sigma^2\end{aligned}$$

dans le cas du Mod. lin. de dim $|\omega|$

↪ EQM réalise un **compromis biais/variance**

Introduction

Références

A-Définition et
hypothèses

Exemples

Identifiabilité

B-Estimation

EMC

Propriétés

EMV

C-Lois des
estimateurs

IC

IC Préviation

RC

D-Tests

de Student

de Wald

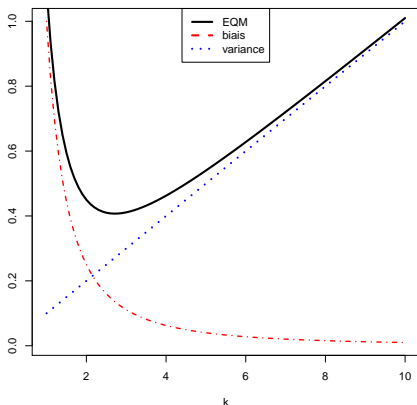
de Fisher

E-Validation

Sélection de
variables et choix
de modèle

EQM réalise un compromis Biais Variance

← fort biais, faible variance faible biais, forte variance →



Mais **PB** : pas calculable car dépend de θ_*

Introduction

Références

A-Définition et
hypothèses

Exemples

Identifiabilité

B-Estimation

EMC

Propriétés

EMV

C-Lois des
estimateurs

IC

IC Prévision

RC

D-Tests

de Student

de Wald

de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Erreur quadratique moyenne de prévision (EQMP)

Risque quadratique des **estimations** \widehat{F}_ω^v dans le modèle (ω)
de **nouvelles** observations (X^v, Y^v)

$$\begin{aligned} EQMP(\widehat{F}_\omega^v) &= \mathbb{E}(\|\widehat{F}_\omega^v - Y^v\|^2) \\ &= \mathbb{E}(\|\widehat{F}_\omega^v - F_*^v + F_*^v - Y^v\|^2) \\ &= EQM(\widehat{F}_\omega^v) + \mathbb{E}(\|F_*^v - Y^v\|^2) \\ &\quad - 2\mathbb{E}[(X_\omega^v \widehat{\theta}_\omega - X^v \theta)^v (F_*^v - Y^v)] \end{aligned}$$

Si l'échantillon (X^v, Y^v) est indépendant de (X, Y) :

- ▶ L'espérance du **double produit** est **nulle**
- ▶ $EQMP(\widehat{F}_\omega^v) = EQM(\widehat{F}_\omega^v) + n^v \sigma^2$
- ▶ $EQMP$ réalise le **compromis biais/variance** et est facile à estimer à partir de données...
- ▶ mais **attention** ! pas directement avec celles ayant pas servi à calculer $\widehat{\theta}_\omega$...

Introduction

Références

A-Définition et
hypothèses

Exemples
Identifiabilité

B-Estimation

EMC
Propriétés
EMV

C-Lois des
estimateurs

IC
IC Prévision
RC

D-Tests

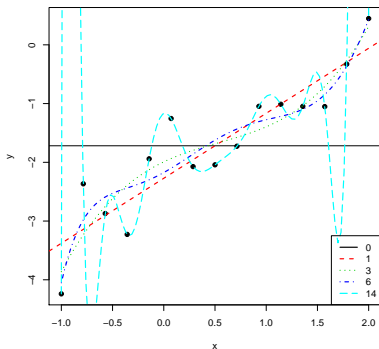
de Student
de Wald
de Fisher

E-Validation

Sélection de
variables et choix
de modèle

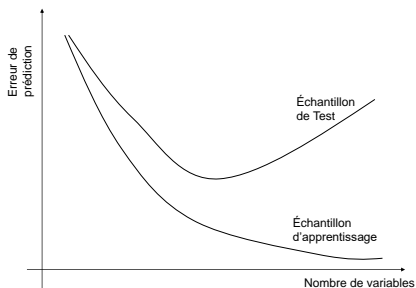
Estimer sur l'échantillon d'apprentissage ?

- Estimer une erreur moyenne de prédiction par SCR sur les données d'apprentissage :



- Ne peut réaliser le compromis biais variance, problème de **sur-apprentissage**

- Calculer une erreur de prédiction sur de nouvelles observations pour établir la **performance de généralisation** du modèle



- Définir des **critères** qui corrigent le biais d'apprentissage

Introduction

Références

A-Définition et
hypothèses

Exemples

Identifiabilité

B-Estimation

EMC

Propriétés

EMV

C-Lois des
estimateurs

IC

IC Prévision

RC

D-Tests

de Student

de Wald

de Fisher

E-Validation

Sélection de
variables et choix
de modèle

- ▶ si les données sont suffisamment nombreuses :
↔ les scinder en échantillon d'**apprentissage**,
échantillon de **validation**

$$\widehat{EQMP}(\omega) = \frac{1}{n^v} \|Y^v - F(\hat{\theta}_\omega, X^v)\|^2$$

Estimation du risque en pratique

- ▶ si les données sont peu nombreuses :
↪ **Validation croisée** d'ordre B (*B-fold cross validation*)

$$\widehat{EQMP}(\omega) = \frac{1}{n} \sum_{b=1}^B \|Y^b - F(\hat{\theta}_{\omega}^{(-b)}, X^b)\|^2$$

↪ **cas particulier** $B = n$, *leave one out*

$$\text{LOO} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{F}_{(-i)})^2 \quad \underbrace{=}_{\text{en régr. lin.}} \quad \frac{1}{n} \sum_{i=1}^n \left(\frac{Y_i - \hat{F}_i}{1 - h_i} \right)^2 \quad \text{:= } \underline{\text{PRESS}}_{\text{estimateurs}}$$

Une alternative : les critères de choix

Pénaliser une estimation directe de performance sur l'échantillon d'apprentissage pour contrecarrer le biais d'optimisme

- ▶ à partir du R^2
↔ critère R^2 ajusté
- ▶ à partir de SCR
↔ critère de Mallows
- ▶ Considérer la distance de Kullback comme étant le critère de performance à estimer sans biais sur l'échantillon
↔ critère AIC
- ▶ Considérer la vraisemblance intégrée
↔ critère BIC

Permettent de comparer des modèles, et de faire de la sélection

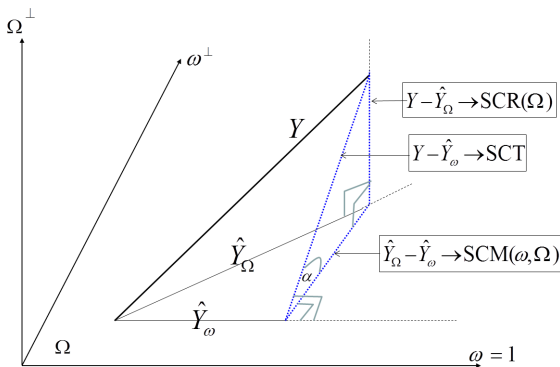
Critère : coefficient de détermination R^2

- pour comparer des modèles Ω de même dimension

$$R^2(\Omega) = \frac{SCM(\mathbb{I}, \Omega)}{SCR(\mathbb{I})} = 1 - \frac{SCR(\Omega)}{SCT}$$

- pour prendre en compte différentes dimensions

$$R_a^2(\Omega) = 1 - \frac{n-1}{n-|\Omega|} (1 - R^2(\Omega)) = 1 - \frac{n-1}{n-|\Omega|} \frac{\|Y - \hat{Y}^2\|}{\|Y - \bar{y}\mathbb{I}\|^2}$$



Estimation de EQM : le C_p de Mallows

- ▶ d'une part : $EQM(\hat{m}_\omega) = \|\mathbf{E}(\hat{m}_\omega) - \mathbf{E}(Y)\|^2 + |\omega|\sigma^2$
- ▶ d'autre part : $\mathbf{E}(SCR(\omega))$

$$\begin{aligned} &= \mathbf{E}(\|Y - \hat{m}_\omega\|^2) \\ &= \mathbf{E}(\|Y - \hat{m}_\omega - \mathbf{E}(Y - \hat{m}_\omega) + \mathbf{E}(Y - \hat{m}_\omega)\|^2) \\ &= \mathbf{E}(\|(I - H_\omega)(Y - \mathbf{E}(Y))\|^2) + \|\mathbf{E}(Y) - \mathbf{E}(\hat{m}_\omega)\|^2 \\ &= (n - |\omega|)\sigma^2 + \|\mathbf{E}(Y) - \mathbf{E}(\hat{m}_\omega)\|^2 \end{aligned}$$

- ▶ d'où $EQM(\widehat{\hat{m}_\omega}) = SCR(\omega) - (n - 2|\omega|)\sigma^2$.
- ▶ Soit $\frac{EQM(\widehat{\hat{m}_\omega})}{\hat{\sigma}^2} = \frac{SCR(\omega)}{\hat{\sigma}^2} + 2|\omega| - n$

Minimiser

$$C_p(\omega) = \frac{SCR(\omega)}{\hat{\sigma}^2} + 2|\omega| - n$$

avec $\hat{\sigma}^2$ est calculé en général dans le modèle complet.

Attention ! : à calculer sur un échantillon **indépendant** de la sélection...

Introduction

Références

A-Définition et
hypothèses

Exemples
Identifiabilité

B-Estimation

EMC
Propriétés
EMV

C-Lois des
estimateurs

IC
IC Préviation
RC

D-Tests

de Student
de Wald
de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Critères : la vraisemblance pénalisée

La qualité de l'ajustement d'un modèle peut être évaluée par la **log-vraisemblance maximale** $\log L(\omega)$ calculée dans le modèle (ω) : **pénalisation** pour éviter le surajustement
Choix du modèle (ω) minimisant

$$\text{critère } (\omega) = -2 \log L(\omega) + |\omega| \text{pen}(n)$$

- **AIC** (Akaike Information Criterion) : $\text{pen}(n) = 2$

$$\begin{aligned} AIC(\omega) &= -2 \log L(\omega) + 2|\omega| \\ &= n \log \frac{SCR(\omega)}{n} + 2|\omega| + cte \quad (\text{Cas gaussien}) \end{aligned}$$

- **BIC** (Bayesian Information Criterion) : $\text{pen}(n) = \log(n)$

$$\begin{aligned} BIC(\omega) &= -2 \log L(\omega) + |\omega| \log(n) \\ &= n \log \frac{SCR(\omega)}{n} + |\omega| \log(n) + cte \quad (\text{Cas gaussien}) \end{aligned}$$

Recherche de sous-ensembles de variables

Limite de la recherche **exhaustive** : $2^{|\omega|-1}$ modèles à comparer suivant un **critère**

↪ méthodes (sous-optimales) **explorant** l'espace des modèles

- ▶ Méthode descendante (**backward**) :
à partir du modèle complet, élimination une à une des variables dont la suppression améliore le plus le critère
- ▶ Méthode ascendante (**forward**) :
A partir du modèle le plus simple, ajout une à une de variables dont l'ajout améliore le plus le critère
- ▶ Méthode mixte (**stepwise**)
enchaînement des étapes ascendantes et descendantes

Introduction

Références

A-Définition et hypothèses

Exemples

Identifiabilité

B-Estimation

EMC

Propriétés

EMV

C-Lois des estimateurs

IC

IC Prévision

RC

D-Tests

de Student

de Wald

de Fisher

E-Validation

Sélection de variables et choix de modèle

Biais/variance

Pour aller plus loin...

- ▶ Les critères pénalisés sont des critères **asymptotiques**
- ▶ Ils reposent sur le fait que le nombre de modèles n'est pas trop grand (et le nombre de paramètres dans un modèle non plus)
- ▶ une autre approche non asymptotique et prenant en compte la **complexité** de la famille de modèle a été développée (cf. Vapnik, Birgé et Massart)

- ▶ Pour estimer **estimer** la performance **sans biais**
 - ↔ apprentissage/validation
 - ↔ validation croisée
- ▶ **choix de modèle** : réaliser un compromis biais (erreur d'approximation) / variance (erreur d'estimation)
 - ↔ par calcul de l'erreur de prédiction (A/V ou VC),
 - ↔ par calcul de critères réalisant un compromis biais/variance
- ▶ Comment concilier ces deux objectifs ?
 - ↔ apprentissage/**validation**/**test** ou **double VC**

Introduction

Références

A-Définition et
hypothèses

Exemples

Identifiabilité

B-Estimation

EMC

Propriétés

EMV

C-Lois des
estimateurs

IC

IC Préviation

RC

D-Tests

de Student

de Wald

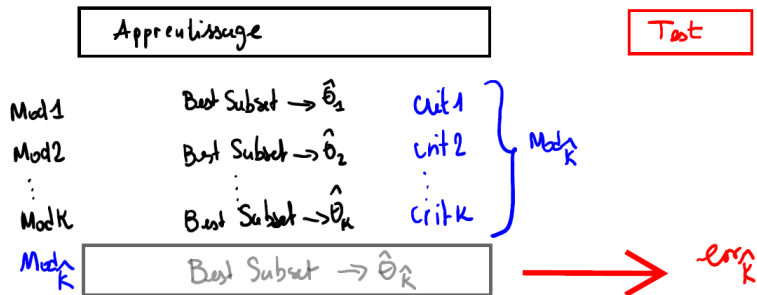
de Fisher

E-Validation

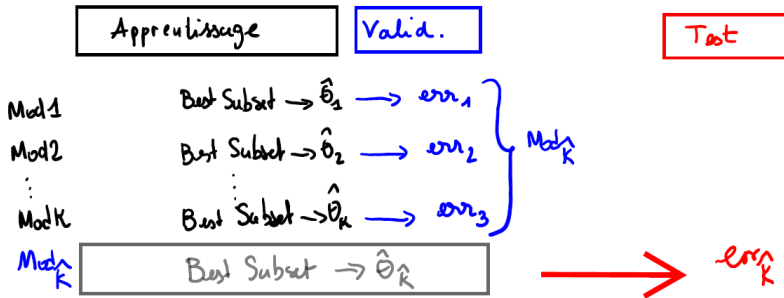
Sélection de
variables et choix
de modèle

Biais/variance

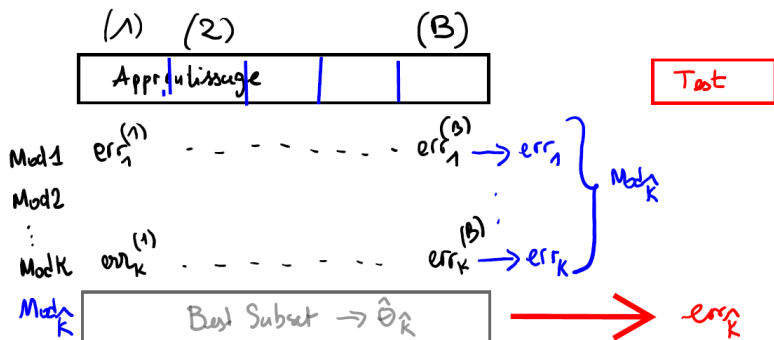
Sélection par critère et erreur de généralisation



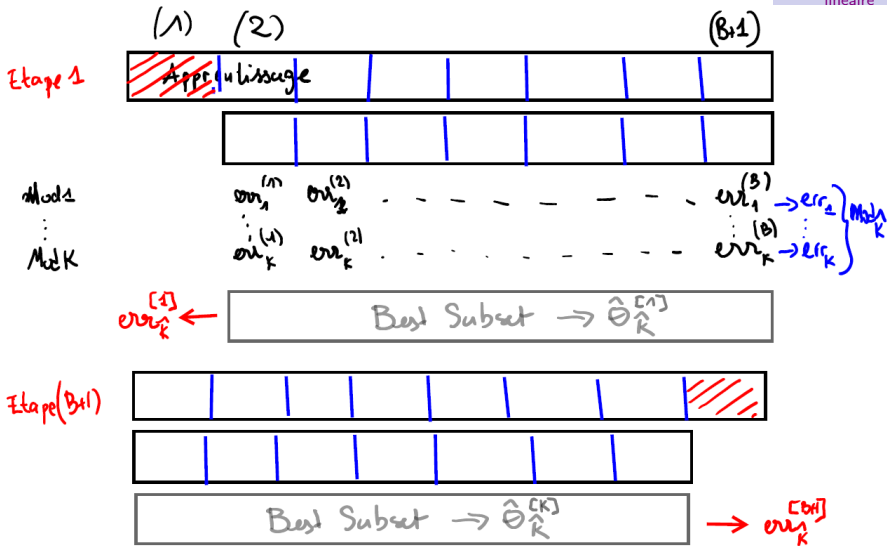
Sélection par A/V et erreur de généralisation



Sélection par VC et erreur de généralisation



Double validation croisée



Quelle taille pour les segments (*folds*) ?

Pour le calcul de l'erreur :

- ▶ LOO :
peu de biais, mais de la variance,
attention au temps d'exécution !
- ▶ Apprentissage/validation :
potentiellement + biaisé, mais moins de variance
temps d'exécution maîtrisé
attention à la stratification
- ▶ K segments est une solution intermédiaire

Introduction

Références

A-Définition et
hypothèses

Exemples
Identifiabilité

B-Estimation

EMC
Propriétés
EMV

C-Lois des
estimateurs

IC
IC Prévision
RC

D-Tests

de Student
de Wald
de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Sommaire

Introduction

A-Définition et hypothèses

B-Estimation

C-Lois des estimateurs

D-Tests

E-Validation

Sélection de variables et choix de modèle

Régularisation

Introduction

Références

A-Définition et
hypothèses

Exemples

Identifiabilité

B-Estimation

EMC

Propriétés

EMV

C-Lois des
estimateurs

IC

IC Préviation

RC

D-Tests

de Student

de Wald

de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Biais/variance

Cas où X n'est pas injective (ou pas loin de ne pas l'être) :

- ▶ variables qualitatives : \hookrightarrow contraintes d'identifiabilité
- ▶ $n \geq p$, mais les colonnes (X_1, \dots, X_p) forment une famille liée de R^n : variables explicatives très "corrélées"

\hookrightarrow la variance de $\hat{\theta}$ peut fortement augmenter si $X'X$ est proche d'une matrice non inversible, donc pb de **précision**

- ▶ $n < p$, le nombre de variables est supérieur au nombre d'observations : "grande dimension"

\hookrightarrow méthodes pour pallier le pb de dégénérescence du rang de X : réduction de dimension, régularisation

Introduction

Références

A-Définition et hypothèses

Exemples
Identifiabilité

B-Estimation

EMC
Propriétés
EMV

C-Lois des estimateurs

IC
IC Prévision
RC

D-Tests

de Student
de Wald
de Fisher

E-Validation

Sélection de variables et choix de modèle

- ▶ Contraindre les estimations des paramètres
- ▶ Réduit significativement la variance
- ▶ régression **ridge** et régression **Lasso**

Introduction

Références

A-Définition et
hypothèses

Exemples

Identifiabilité

B-Estimation

EMC

Propriétés

EMV

C-Lois des
estimateurs

IC

IC Préviation

RC

D-Tests

de Student

de Wald

de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Biais/variance

Régression ridge (Hoerl et Kennard (1970))

- ▶ Les valeurs propres λ_j de $X'X$ sont ≥ 0
- ▶ S'il y a dégénérescence du rang, les λ_j ordonnées par ordre décroissant sont quasi-nulles à partir d'un certain rang r .
- ▶ $X'X$ et $X'X + \kappa Id_p$ ont les mêmes vecteurs propres, mais des valeurs propres différentes : λ_j et $\lambda_j + \kappa$ respectivement, $j = 1, \dots, p$
- ▶ d'où l'estimateur ridge

$$\hat{\theta}_{ridge}(\kappa) = (X'X + \kappa Id_p)^{-1} X'Y$$

Introduction

Références

A-Définition et
hypothèses

Exemples

Identifiabilité

B-Estimation

EMC

Propriétés

EMV

C-Lois des
estimateurs

IC

IC Prévision

RC

D-Tests

de Student

de Wald

de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Régression ridge

En régression linéaire, on optimise

$$SCR(\theta) = \sum_{i=1}^n \left(Y_i - \sum_{j=1}^p \theta_j x_{ij} \right)^2$$

En régression ridge, on minimise un critère pénalisé

$$SCR(\theta) + \kappa \sum_{j=1}^p \theta_j^2 = \sum_{i=1}^n \left(Y_i - \sum_{j=1}^p \theta_j x_{ij} \right)^2 + \kappa \sum_{j=1}^p \theta_j^2$$

où $\kappa \geq 0$ est un paramètre de réglage (**tuning**) à déterminer séparément

- ▶ $SCR(\theta)$ diminue avec p croissant
- ▶ $\kappa \sum_{j=1}^p \theta_j^2$ augmente avec p

Régression ridge = contrainte ℓ_2

- ▶ Minimisation du critère pénalisé

$$\hat{\theta}_{ridge}(\kappa) = \arg \min_{\theta \in \mathbb{R}^p} \|Y - X\theta\|^2 + \kappa \|\theta\|^2$$

- ▶ Revient à un pb de **minimisation sous contraintes**

$$\tilde{\theta}(\delta) = \arg \min_{\theta \in \mathbb{R}^p; \|\theta\|^2 \leq \delta} \|Y - X\theta\|^2$$

- ▶ N'a d'intérêt que si δ est **petit**
- ▶ Choix de κ (ou δ)?

Introduction

Références

A-Définition et
hypothèses

Exemples

Identifiabilité

B-Estimation

EMC

Propriétés

EMV

C-Lois des
estimateurs

IC

IC Préviation

RC

D-Tests

de Student

de Wald

de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Régression ridge : propriétés

► Biais : $\mathbb{E}(\hat{\theta}_{ridge}) - \theta = -\kappa(X'X + \kappa Id_p)^{-1}\theta$

► Variance :

$$\text{var}(\hat{\theta}_{ridge}) = \sigma^2(X'X + \kappa Id_p)^{-1}X'X(X'X + \kappa Id_p)^{-1}$$

► $EQM(\hat{\theta}_{ridge}) =$

$$\begin{aligned} \text{tr} [(X'X + \kappa Id_p)^{-1}[\kappa^2\theta\theta' + \sigma^2X'X](X'X + \kappa Id_p)^{-1}] \\ = \sum_{j=1}^p \frac{\sigma^2\lambda_j + \kappa^2[P'\theta]_j^2}{(\lambda_j + \kappa)^2} \end{aligned}$$

► $EQM(\hat{\theta}_{MC}) = \sigma^2 \text{tr}[(X'X)^{-1}] = \sigma^2 \sum_{j=1}^p \lambda_j^{-1}$.

Proposition

La régression ridge est plus précise pour l'estimation des paramètres que celle des MC si $\kappa \leq 2\sigma^2/\theta'\theta$

Introduction

Références

A-Définition et hypothèses

Exemples

Identifiabilité

B-Estimation

EMC

Propriétés

EMV

C-Lois des estimateurs

IC

IC Prévision

RC

D-Tests

de Student

de Wald

de Fisher

E-Validation

Sélection de variables et choix de modèle

Biais/variance

Centrer / réduire les régresseurs ?

- ▶ Avec les MC, l'estimation de $X\theta$ n'est pas affectée par la standardisation des variables. Si une variable est divisée par c , son coefficient est multiplié par c et $X\hat{\theta}$ est le même.
- ▶ Mais l'estimation peut considérablement changer dans le cas de la régression ridge, à cause de la pénalisation
- ▶ Il est donc recommandé de **centrer et réduire les variables** en régression ridge
- ▶ La pénalité ne s'applique pas en général l'intercept, qui capture la réponse quand les covariables sont nulles
 - ▶ En général on **centre** donc aussi Y qui est remplacé par $Y - \bar{Y}\mathbf{1}_n$ et on n'introduit plus d'intercept
 - ▶ On peut aussi **réduire** Y

Régression ridge : en pratique

- ▶ Centrer et réduire les variables X_j (et Y)

$$\tilde{X}_j = (X_j - \bar{X}_j \mathbf{1}_n) / \hat{\sigma}_j$$

- ▶ A κ fixé :

$$\hat{\theta}_{ridge}(\kappa) = (\tilde{X}'\tilde{X} + \kappa Id_p)^{-1} \tilde{X}'\tilde{Y}$$

- ▶ valeurs ajustées

$$\hat{Y}_{ridge}(\kappa) = \hat{\sigma}_Y^2 [\tilde{X} \hat{\theta}_{ridge}(\kappa)] + \bar{Y} \mathbf{1}_n$$

Introduction

Références

A-Définition et
hypothèses

Exemples

Identifiabilité

B-Estimation

EMC

Propriétés

EMV

C-Lois des
estimateurs

IC

IC Prévision

RC

D-Tests

de Student

de Wald

de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Régr. ridge : Choix de κ à partir des données

- ▶ **graphiquement** : tracer $\hat{\theta}_{ridge}(\kappa)$ en fonction de κ
 $\hookrightarrow \tilde{\kappa}$ est la plus petite valeur avant laquelle les coefficients "plongent" vers 0

- ▶ **analytiquement** (Hoerl, 1975) : $\tilde{\kappa} = \frac{p\hat{\sigma}_{MC}^2}{\hat{\theta}'_{MC}\hat{\theta}_{MC}}$

- ▶ **apprentissage/validation**

- ▶ Séparer en (X, Y) en (X_a, Y_a) et (X_v, Y_v)
- ▶ Centrer (X_a, Y_a)
- ▶ Sur (X_a, Y_a) , régression ridge pour une grille de $\kappa \in [0; \kappa_{max}]$, $\hookrightarrow \hat{\theta}_{ridge}$ et \hat{Y}_{ridge}

$$\hat{Y}_{v,ridge}(\kappa) = \hat{\sigma}_{a,Y} \sum_{j=1}^p \frac{X_{v,j} - \bar{X}_{a,j} \mathbb{1}_{n_v}}{\sigma_{a,j}} \hat{\theta}_{ridge,j}(\kappa) + \bar{Y}_a \mathbb{1}_{n_v}$$

- ▶ $\tilde{\kappa}$ minimise de l'erreur quadratique de prévision observée

$$PRESS(\kappa) = \|\hat{Y}_{v,ridge}(\kappa) - Y_v\|^2$$

- ▶ **VC** : $\tilde{\kappa} = \arg \min_{\kappa} \sum_i \left(\frac{y_i - \hat{y}_{ridge,i}(\kappa)}{1 - H_{jj}(\kappa)} \right)^2$

Introduction

Références

A-Définition et
hypothèses

Exemples
Identifiabilité

B-Estimation

EMC
Propriétés
EMV

C-Lois des
estimateurs

IC
IC Prévision
RC

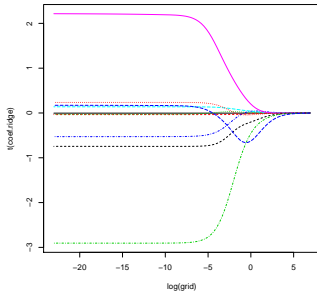
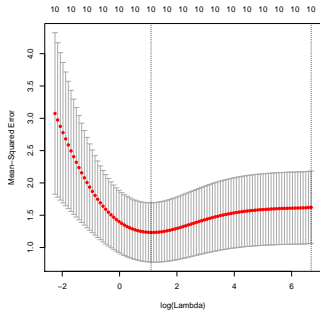
D-Tests

de Student
de Wald
de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Exemple



Introduction

Références

A-Définition et
hypothèses

Exemples
Identifiabilité

B-Estimation

EMC
Propriétés
EMV

C-Lois des
estimateurs

IC
IC Prévision
RC

D-Tests

de Student
de Wald
de Fisher

E-Validation

Sélection de
variables et choix
de modèle

On utilise la norme ℓ_1 : $\|\theta\|_1 = \sum_j |\theta_j|$

$$\tilde{\theta} = \arg \min_{\theta \in \mathbb{R}^p; \|\theta\|_1 \leq \delta} \|Y - X\theta\|^2$$

soit, résoudre le problème de pénalisation

$$\hat{\theta}(\tau) = \arg \min_{\theta \in \mathbb{R}^p} \|Y - X\theta\|^2 + \tau \|\theta\|_1$$

- ▶ Si $\tau \geq \|X'Y\|_\infty$, $\tilde{\theta}(\tau) = 0$
- ▶ au fur et à mesure, des variables sont ajoutées, et les coefficients des variables déjà utilisées sont modifiés
- ▶ ... mais certaines variables peuvent être désélectionnées
- ▶ Méthode de **sélection** de variable

Introduction

Références

A-Définition et
hypothèses

Exemples

Identifiabilité

B-Estimation

EMC

Propriétés

EMV

C-Lois des
estimateurs

IC

IC Prévision

RC

D-Tests

de Student

de Wald

de Fisher

E-Validation

Sélection de
variables et choix
de modèle

- ▶ Centrer et réduire les variables (X, Y) ($\hookrightarrow \tilde{X}, \tilde{Y}$) et ajuster le modèle sans coefficient constant
- ▶ Le modèle de prévision est

$$\hat{Y}(\tilde{\tau}) = \hat{\sigma}_Y^2 \tilde{X} \hat{\theta}(\tilde{\tau}) + \bar{Y} \mathbb{1}_n$$

- ▶ $\tilde{\tau}$ (ou $\tilde{\delta}$) sont choisis grâce aux données, de façon graphique, par méthode analytique ou par VC

Introduction

Références

A-Définition et
hypothèses

Exemples

Identifiabilité

B-Estimation

EMC

Propriétés

EMV

C-Lois des
estimateurs

IC

IC Prévision

RC

D-Tests

de Student

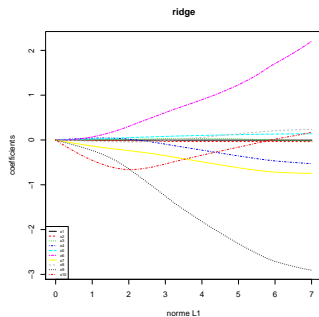
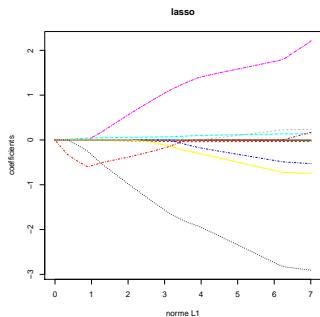
de Wald

de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Comparaison Lasso/Ridge : exemple



Introduction

Références

A-Définition et
hypothèses

Exemples

Identifiabilité

B-Estimation

EMC

Propriétés

EMV

C-Lois des
estimateurs

IC

IC Prévision

RC

D-Tests

de Student

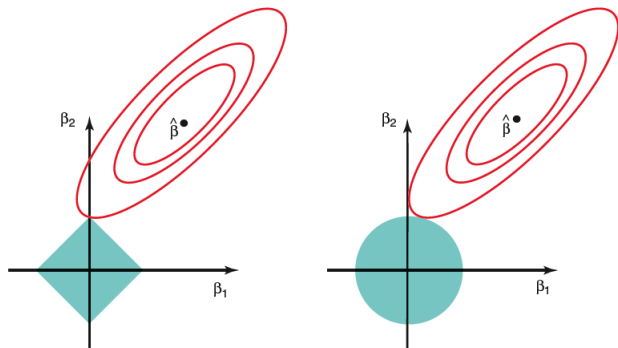
de Wald

de Fisher

E-Validation

Sélection de
variables et choix
de modèle

Comparaison Lasso/Ridge



An Introduction to Statistical Learning with Applications in
R. James, Witten, Hastie, Tibshirani (Springer 2013)