APPRENTISSAGE SUPERVISE

Modèle linéaire gaussien
&
Introduction au modèle linéaire généralisé

HPC – BIG DATA
Mars 2023

pascal.laveau@meteo.fr

INP TOULOUSE ENSEEIHT

Introduction

1. Les statistiques dans le cursus HPC - Big Data

2. Intérêt de la modélisation statistique

3. Exemple de la prévision opérationnelle du temps

Introduction / Apprentissage statistique

1. Apprentissage supervisé :

   On cherche à expliquer une variable Y (prédictand) au moyen d'autres variables (prédicteurs), 2 situations possibles :

- Y est quantitative → modèle de régression
- Y est qualitative → modèle de discrimination

2. Apprentissage non supervisé :

   Absence de prédictand, on cherche à regrouper les données ayant des caractéristiques proches (recherche de classes) ou à analyser les données dans un sous-espace plus adapté :

   → méthodes de classification (clustering)

   → méthodes d'analyse factorielle (ACP)

Objectifs du cours

1. Connaître le cadre théorique du modèle linéaire gaussien

2. Savoir exploiter le modèle linéaire gaussien

3. Maîtriser les procédures de sélection automatique des prédicteurs et de validation du modèle

4. Avoir des notions sur le modèle linéaire généralisé et en particulier la régression logistique

Prérequis

1. Algèbre linéaire / Calcul matriciel

2. Probabilités / Vecteurs aléatoires gaussiens

3. Estimateurs / Maximisation de la vraisemblance

4. Tests statistiques / Utilisation de la p-value d'un test

Plan du cours

1. Rappels
2. Modèle linéaire gaussien
3. Inférence
4. Prédicteurs qualitatifs
5. Validation
6. Modèle linéaire généralisé (GLM)

Bibliographie

- **Statistique : la théorie et ses applications, Lejeune, Springer** (prg 1ère année)

- **Régression avec R, Cornillon & al., Springer**

- **An introduction to statistical learning, James & al., Springer** (→ moodle)

Pour aller plus loin :

-An introduction to generalized linear models, Dobson & Barnett, CRC Press

-An R companion to applied regression, Fox & Weisberg, Sage

-The elements of statistical learning, Hastie & al., Springer (→ moodle)

Plan du cours

Résultats utiles

- **Vecteur aléatoire dans IR^n**

Y_i $(i = 1,\ldots, n)$ = variables aléatoires réelles
ξ_i = $i^{ème}$ vecteur de la base canonique

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{pmatrix} = Y_1\xi_1 + Y_2\xi_2 + \cdots + Y_n\xi_n \qquad IE[Y] = \mu = \begin{pmatrix} IE[Y_1] \\ IE[Y_2] \\ IE[Y_3] \\ \vdots \\ IE[Y_n] \end{pmatrix}$$

Y vecteur aléatoire de IR^n d'espérance E[Y]

Matrice de covariance de Y : $V[Y] = \Gamma$ = matrice n×n symétrique, réelle, de terme général $\Gamma_{ij} = COV(Y_i, Y_j)$

Résultats utiles

Soit A matrice m×n non aléatoire, et soit Z = AY.

E[Z] = A E[Y] et V[Z] = A V[Y] ^{t}A

- **Loi normale dans IR^n**

Y vecteur gaussien ; toute combinaison linéaire de ses composantes suit une loi de Gauss à une dimension.
Si Γ=V[Y] est inversible, Y admet pour densité :

$$f(Y) = \frac{1}{\sqrt{(2\pi)^n \det(\Gamma)}} \exp\left(-\frac{1}{2} {}^t(Y-\mu)\Gamma^{-1}(Y-\mu) \right)$$

A matrice m×n, Z = AY, et $Y \sim N_n(\mu, \Gamma) \Rightarrow Z \sim N_m(A\mu, A\Gamma\,{}^tA)$

Résultats utiles

- **Loi normale sphérique**

La loi normale sphérique : $\Gamma = \sigma^2 I_n$
Composantes décorrélées (et donc indépendantes dans ce cadre gaussien)
Composantes de même variance σ^2

- métrique euclidienne :

$<X,Y> = \Sigma_i X_i Y_i$ produit scalaire canonique
$\|Y\|^2 = <Y,Y> = \Sigma_i Y_i^2 = {}^tYY$

- densité sphérique : $$\frac{1}{\left(\sigma\sqrt{2\pi}\right)^n}\exp\left(-\frac{1}{2\sigma^2}\|Y-\mu\|^2\right)$$

- propriété *importante* : $Y \sim N_n(\mu, \sigma^2 I_n) \Rightarrow \|Y-\mu\|^2 / \sigma^2 \sim \chi^2_n$

Résultats utiles

- **Projection orthogonale**

Q sous-espace de IR^n de dimension q.

$\Pi_Q Y$ projeté orthogonal de Y sur Q.

$\Pi_Q Y$ définie par $\Pi_Q Y \in Q$ et $Y - \Pi_Q Y \perp$ à tout vecteur de Q.

Q engendré par les q vecteurs colonnes de la matrice T, avec rg(T)=q
$\Rightarrow \Pi_Q = T\,({}^tTT)^{-1}\,{}^tT$

Π_Q est une matrice symétrique idempotente : $\Pi_Q = {}^t\Pi_Q = \Pi_Q^2$

$tr(\Pi_Q) = dimQ = q$

Plan du cours

1. Rappels
2. ***Modèle linéaire gaussien***
3. Inférence
4. Prédicteurs qualitatifs
5. Validation
6. Modèle linéaire généralisé (GLM)

Modèle linéaire gaussien : formulation

$$\mathbf{Y = T\beta + e} \quad \text{avec} \quad \mathbf{e \sim N_n(0, \sigma^2 I_n)}$$

= comp. déterministe + comp. aléatoire

- **Y : vecteur aléatoire gaussien de IR^n, contenant les n observations de la variable expliquée (= prédictand)**

- **T : matrice non aléatoire de données explicatives, de dimension n×q, contenant les n observations sur les q variables explicatives (= prédicteurs)**

- **β : vecteur des q paramètres du modèle (à estimer, q=dimension du modèle)**

- **e : vecteur aléatoire gaussien de IR^n contenant les résidus du modèle. On suppose que le vecteur e suit une loi normale sphérique de vecteur espérance nul et de matrice de covariance $\sigma^2 I_n$, avec σ^2 à estimer.**

Formulation

En notant x_{ij} la $i^{ème}$coordonnée du vecteur X_j ,

pour i=1,…,n on peut écrire :

$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_q x_{iq} + e_i$$

Soit, sous forme matricielle :

$$\begin{pmatrix} Y_1 \\ . \\ . \\ . \\ Y_n \end{pmatrix} = \left(\begin{array}{c|c|c|c} . & . & & . \\ . & . & & . \\ X_1 & X_2 & \ldots & Xq \\ . & . & & . \\ . & . & & . \end{array}\right) \begin{pmatrix} \beta_1 \\ . \\ \beta_q \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ . \\ . \\ e_n \end{pmatrix}$$

Avec $T=(X_1|X_2|\ldots|X_q)$: $Y = T\beta + e$

T = matrice du plan d'expérience = *"design matrix"*

Formulation

Y est donc un vecteur aléatoire suivant une loi gaussienne sphérique, dont le vecteur espérance μ est une fonction linéaire des covariables :

$$E[Y] = \mu = T\beta \quad \text{et} \quad V[Y] = V[T\beta + e] = V[e] = \sigma^2 I_n$$

Les résidus e_i (i=1...n) sont des variables aléatoires *normales*, *centrées*, *indépendantes* et de *variance constante* (homoscédasticité).

X_1, X_2…X_q vecteurs contenant les n observations des q variables explicatives.

X_j (j=1,..,q) peut être le vecteur des observations d'une variable numérique (prédicteur quantitatif), un vecteur d'indicatrices (prédicteur qualitatif), un vecteur de coordonnées égales à 1 (estimation du terme constant du modèle).

Formulation

Remarques

1. Modèle linéaire = linéarité selon β

Les prédicteurs peuvent être eux-mêmes des fonctions non linéaires d'autres variables, de multiples prédicteurs élaborés peuvent être testés, par exemple introduire des termes d'interaction peut enrichir le modèle.
Exemple : le modèle défini par $Y = \beta_1 \ln(X1) + \beta_2 \exp(X2) + \beta_3\, X1 \,.\, X2 + e$ entre dans le cadre du modèle linéaire.

2. Vocabulaire : régression multiple et modèle linéaire

Les prédicteurs sont considérés comme non aléatoires lorsque l'on raisonne en terme de modèle linéaire.
En régression, les prédicteurs sont considérés comme étant aléatoires et le conditionnement apparaît explicitement, les 2 approches étant équivalentes.

$E[Y] \rightarrow E[Y|X_1,\ldots,X_q]$
$V[Y] \rightarrow V[Y|X_1,\ldots,X_q]$

Estimation dans le modèle linéaire gaussien

- **Estimation de β par la méthode MCO (= Moindres Carrés Ordinaires)**

$\hat{\beta}$ = **estimation de β**

ε = vecteur des résidus estimés : ε = Y-T $\hat{\beta}$

On cherche à minimiser $\Sigma e_i^2 = \|e\|^2 = \|Y-T\beta\|^2$

→ **$\hat{\beta}$ = argmin ($\|Y-T\beta\|^2$)**

$\hat{\beta}$ **tel que :** $\left(\frac{\partial \|Y - T\beta\|^2}{\partial \beta} \right)_{\beta=\hat{\beta}} = 0$

Estimation

Par dérivation vectorielle, il vient : $-2{}^tT(Y - T\beta) = 0$ pour $\beta = \hat{\beta}$

D'où $\hat{\beta}$ estimateur des moindres carrés de β, avec T de rang q :

$$\hat{\beta}_{MCO} = \left({}^tTT\right)^{-1} {}^tTY$$

Propriété **: $\hat{\beta}$ est un estimateur sans biais de β : $E[\hat{\beta}] = \beta$**

Théorème de Gauss-Markov **:**

$\hat{\beta}$ est de tous les estimateurs linéaires sans biais de β (de la forme AY) celui de variance minimale.

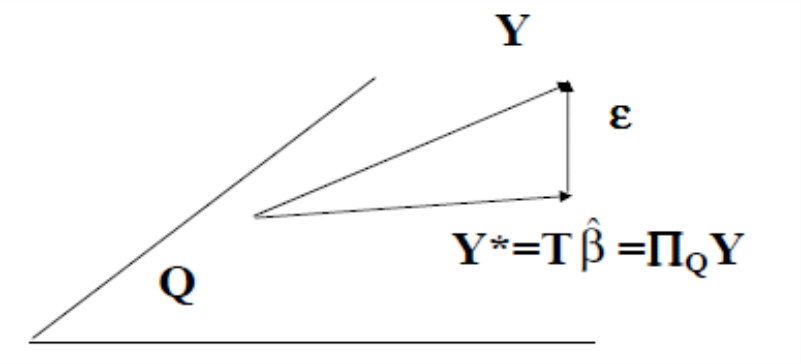
→ BLUE : Best Linear Unbiased Estimator

Estimation

- **Interprétation géométrique**

Soit Q le plan de IR^n engendré par les vecteurs colonnes de la matrice T.
Le projecteur Π_Q sur Q s'écrit : $\Pi_Q = T\,({}^tTT)^{-1}\,{}^tT$

On note Y* le vecteur des valeurs ajustées de Y par le modèle (valeurs ajustées= prévisions calculées sur l'archive d'apprentissage) :
Y* est la projection orthogonale de Y sur Q



$$Y^* = T\hat{\beta} = T\,({}^tTT)^{-1}\,{}^tTY = \Pi_Q Y$$

$Y - Y^* = \varepsilon = Y - T\hat{\beta}$ = **vecteur des résidus estimés du modèle.**

Estimation

- **Loi de $\hat{\beta}$**

Y étant un vecteur aléatoire, $\hat{\beta}$ fonction de Y est aussi un vecteur aléatoire.

Dans le cadre du modèle linéaire gaussien, $\hat{\beta}$ transformé linéaire d'un vecteur gaussien est lui-même gaussien et on a :

$$\hat{\beta} \sim N_q\left(\beta, \sigma^2\left({}^tTT\right)^{-1}\right)$$

Exercice : en utilisant $\hat{\beta} = ({}^tTT)^{-1}\,{}^tTY$ et le fait que $({}^tTT)^{-1}$ est une matrice symétrique, retrouver les expressions de $E[\hat{\beta}]$ et $V[\hat{\beta}]$.

Estimation

- **Le vecteur $\hat{\beta}$ suit une loi normale multidimensionnelle de dimension q, d'espérance β et de matrice de covariance $\sigma^2({}^tTT)^{-1}$**

- **Cette matrice de variance-covariance n'est en général pas diagonale : les estimations $\hat{\beta}_1$, $\hat{\beta}_2$,…, $\hat{\beta}_q$ ne sont alors pas indépendantes.**

Soit $\hat{\beta}_k$ (k=1,..,q) la $k^{ième}$ coordonnée de $\hat{\beta}$

$\hat{\beta}_k$ suit donc une loi normale :

- **d'espérance β_k**
- **d'écart-type $\sigma_k = \sigma \sqrt{\left({}^tTT\right)^{-1}_{kk}}$**

avec $\left({}^tTT\right)^{-1}_{kk}$: $k^{ième}$ élément diagonal de la matrice $({}^tTT)^{-1}$

Estimation

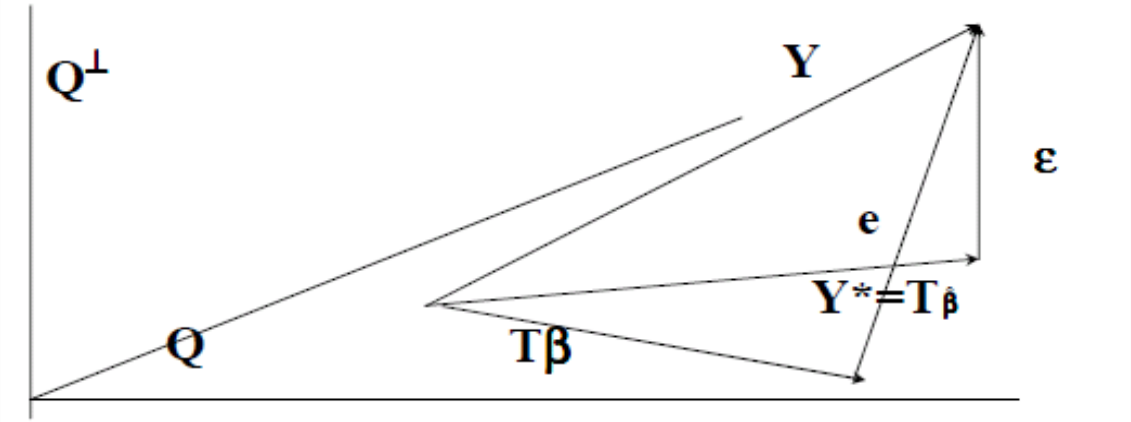
- **Estimation de σ^2**

Soient ε le vecteur des résidus estimés ($\varepsilon = Y - T\hat{\beta}$) et e le vecteur des résidus théorique ($e = Y - T\beta$).

$\varepsilon = Y - Y^* = Y - T\hat{\beta}$ est orthogonal à Q : $\varepsilon \in Q^{\perp}$

$\Pi_Q = T\,({}^tTT)^{-1}\,{}^tT$ est le projecteur sur Q

$\Pi_Q^{\perp} = I - \Pi_Q$ est le projecteur sur $Q^{\perp}$ (I désignant la matrice identité).



Formulation

On montre que : $E[\ \|Y-T\hat{\beta}\|^2\ ] = \sigma^2\ (n-q)$

On en déduit donc un estimateur de σ^2 :

$$\hat{\sigma}^2 = \frac{\left\|Y - T\hat{\beta}\right\|^2}{n-q} = \frac{\left\|Y - Y^*\right\|^2}{n-q} = \frac{\left\|\varepsilon\right\|^2}{n-q} = \frac{\sum_{i=1}^{n}\left(Y_i - Y_i^*\right)^2}{n-q}$$

On montre que $\hat{\sigma}^2$ est l'estimateur sans biais de variance minimale de σ^2.

ε = projection orthogonale de e (et de Y) sur $Q^{\perp}$

$$\frac{\left\|Y - T\hat{\beta}\right\|^2}{\sigma^2} \sim \chi^2_{n-q}$$

(Th. De Cochran)

Exercice

1. A partir de l'expression de la densité normale multidimensionnelle,

$$f(Y) = \frac{1}{\sqrt{(2\pi)^n \det(\Gamma)}} \exp\left(-\frac{1}{2} {}^t(Y-\mu)\Gamma^{-1}(Y-\mu) \right)$$

   donner la fonction de vraisemblance associée à un n vecteur aléatoire Y dans le cadre du modèle linéaire gaussien.

2. Calculer les estimateurs $\hat{\beta}_{MV}$ et $\hat{\sigma}^2_{MV}$ obtenus par la méthode de maximisation de la vraisemblance.
   Montrer en particulier que : $\hat{\beta}_{MV} = \hat{\beta}_{MCO}$

3. L'estimateur $\hat{\sigma}^2_{MV}$ est-il biaisé ?

Plan du cours

1. Rappels
2. Modèle linéaire gaussien
3. ***Inférence***
4. Prédicteurs qualitatifs
5. Validation
6. Modèle linéaire généralisé (GLM)

Tests dans le modèle linéaire gaussien

- **Test du caractère significatif de l'une des composantes de β**

$$\hat{\beta}_k \sim N(\beta_k, \sigma_k) \quad \text{avec} \quad \sigma_k = \sigma\sqrt{\left({}^tTT\right)^{-1}_{kk}}$$

$$\frac{\left(\hat{\beta}_k - \beta_k\right)}{\hat{\sigma}_k} = \frac{\left(\hat{\beta}_k - \beta_k\right)}{\sqrt{\frac{\left\|Y - T\hat{\beta}\right\|^2}{n-q} \cdot \left({}^tTT\right)^{-1}_{kk}}} \sim t_{n-q}$$

t_{n-q} désignant la variable de Student à n-q degrés de liberté.

On peut alors en déduire l'intervalle de confiance de β_k ainsi que le test du caractère significatif de β_k.

Prévision d'une nouvelle valeur du prédictand

Soit t_k une nouvelle observation des q prédicteurs : ${}^t t_k = (x_{k1},...,x_{kq})$

$$Y_k^* = \hat{\beta}_1 x_{k1} + \hat{\beta}_2 x_{k2} + \hat{\beta}_3 x_{k3} + \dots + \hat{\beta}_q x_{kq} = {}^t t_k \hat{\beta}$$

Y_k^* transformé linéaire de $\hat{\beta} \Rightarrow Y_k^* \sim N({}^t t_k \beta , \sigma^2 {}^t t_k ({}^t TT)^{-1} t_k)$

Par studentisation, puisque σ^2 doit être estimé, et Y_k aléatoire, on trouve :

$$\frac{Y_k - Y_k}{\hat{\sigma}\sqrt{1 + {}^t t_k \left({}^t TT\right)^{-1} t_k}} \sim t_{n-q}$$

$\Rightarrow$ intervalle de prévision de niveau $(1 - \alpha)$ sur Y_k :

$$Y_k^* - t_{n-q,1-\alpha/2}\ \hat{\sigma}\sqrt{1 + {}^t t_k \left({}^t TT\right)^{-1} t_k} < Y_k < Y_k^* + t_{n-q,1-\alpha/2}\ \hat{\sigma}\sqrt{1 + {}^t t_k \left({}^t TT\right)^{-1} t_k}$$

avec $t_{n-q,1-\alpha/2}$ = quantile d'ordre $1-\alpha/2$ de la loi de Student à n-q degrés de liberté

Coefficient de corrélation multiple R

R est le coefficient de corrélation entre valeurs observées et ajustées par le modèle statistique, c'est à dire entre les vecteurs Y et Y*.

Son carré R^2, appelé coefficient de détermination, mesure la qualité du modèle et s'interpréte en termes de variance expliquée :

$$R^2 = \frac{\text{variance expliquée}}{\text{variance totale de Y}} = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$$

Formule de décomposition de la variance :

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - Y_i^*\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\left(Y_i^* - \overline{Y}\right)^2$$

Variance totale (SCT) = Variance résiduelle (SCR) + Variance expliquée (SCE)

Test de significativité du R^2 (test de Fisher global)

Soit $H_0 : R^2 = 0$ **(hypothèse de non-régression)**
$H' : R^2 > 0$

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - q}{q - 1} = \frac{\sum_{i=1}^{n} \left(Y_i^* - \overline{Y} \right)^2}{\sum_{i=1}^{n} \left(Y_i - Y_i^* \right)^2} \cdot \frac{n - q}{q - 1}$$

Sous H_0, F suit une loi de Fischer-Snédécor à q-1 et n-q degrés de liberté.

Au niveau α , on rejette H_0 si $F > f_{q-1,n-q,\alpha}$ (p-value < α)

Ce test compare le modèle complet au modèle ne comportant que le terme constant et vérifiant $E[Y] = \beta_0$

Sélection de variables : test de Fisher

Principe de parcimonie :

A qualité d'information égale, on préfère toujours un modèle simple à un modèle plus compliqué.

- Test de Fisher :

R^2 augmente « mécaniquement » avec la dimension du modèle, même si l'information apportée n'est pas pertinente. En effet R^2 ne tient pas compte de la dimension du sous espace de projection Q.

→ On teste donc l'accroissement de R^2 :

Modèle complet à k+p prédicteurs, et modèle restreint à k prédicteurs

$H_0 : R^2_k = R^2_{k+p}$
$H' : R^2_k < R^2_{k+p}$

$$F = \frac{n-k-p}{p} \cdot \frac{R^2_{k+p} - R^2_k}{1 - R^2_{k+p}}$$

Sous H_0, on montre que F suit une loi de Fisher à p et n-k-p degrés de liberté.
On rejettera donc H_0 au niveau α si $F > f_{p,n-k-p,\alpha}$ (p-value < α)

Remarque **: autres critères de sélection courants :**
AIC, BIC, Cp de Mallows, R^2 ajusté (cf.TP)

Sélection de variables

Algorithmes de sélection automatique des prédicteurs :

Procédure optimale :
Comparaison de toutes les combinaisons possibles de prédicteurs.
Mais il faut alors estimer 2^q modèles,
…soit près de 10^{30} modèles avec une centaine de prédicteurs potentiels.

- **Sélection descendante : on retire du modèle complet les variables inutiles une à une, avec une règle d'arrêt (baisse significative du R^2, AIC minimal…)**

- **Sélection ascendante : on introduit dans le modèle à chaque étape la variable la plus utile, avec une règle d'arrêt (accroissement de R^2 non significatif par exemple) : à privilégier si nb important de prédicteurs.**

- **Sélection avec remise en cause : on introduit dans le modèle à chaque étape une variable supplémentaire, et on teste si toutes les variables sont utiles, avec une règle d'arrêt.**

Plan du cours

1. Rappels
2. Modèle linéaire gaussien
3. Inférence
4. ***Prédicteurs qualitatifs***
5. Validation
6. Modèle linéaire généralisé (GLM)

Prédicteurs qualitatifs : ANOVA

L'analyse de variance (ANOVA en anglais) est une technique qui permet de quantifier l'effet de variables *qualitatives* sur une variable quantitative.

- **Analyse de variance à un facteur**

Soit Y la variable quantitative à expliquer. On considère ici l'effet moyen d'une variable qualitative (appelée facteur) à p modalités.

Le modèle s'écrit : $Y_{ik} = \mu + \alpha_i + e_{ik}$

Avec : $i=1,\ldots,p \quad k=1,\ldots,n_i \qquad \sum_{i=1}^{p} n_i = n \qquad e \sim N(0, \sigma^2 I_n)$

μ est la valeur moyenne de Y, i la $i^{ème}$ modalité du facteur α, α_i l'effet moyen de la $i^{ème}$ modalité de α, et Y_{ik} la $k^{ième}$ observation de Y dans la catégorie i.

Prédicteurs qualitatifs : ANOVA

Modèle linéaire où les variables prédicteurs sont remplacées par le tableau d'indicatrices des modalités du facteur α.

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{31} \\ Y_{32} \end{pmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{21} \\ e_{22} \\ e_{23} \\ e_{31} \\ e_{32} \end{pmatrix}$$

PROBLEME : T n'est pas de plein rang.

Prédicteurs qualitatifs : ANOVA

→ **On pose alors une contrainte d'identification :**

$$\sum_{i=1}^{p} \alpha_i = 0 \quad \Leftrightarrow \quad \alpha_p = -\sum_{i=1}^{p-1} \alpha_i$$

(Remarque : d'autres contraintes peuvent être utilisées, le choix de la contrainte n'impacte pas le modèle lui-même mais l'interprétation des paramètres finalement estimés → cf. TD & TP)

Sous cette dernière contrainte, le modèle alors identifiable s'écrit :

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{31} \\ Y_{32} \end{pmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{bmatrix} \cdot \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{21} \\ e_{22} \\ e_{23} \\ e_{31} \\ e_{32} \end{pmatrix}$$

Prédicteurs qualitatifs : ANCOVA

• **Analyse de covariance (ANCOVA)**

Analyse de covariance = mélange de prédicteurs qualitatifs et quantitatifs

- Pas de difficultés particulières

- Une contrainte sera alors imposée pour chaque prédicteur qualitatif

- Par exemple, si la variable Y dépend des variables X_1, X_2 et des facteurs α et ν, on peut écrire :

$$Y_{ijk} = \mu + \beta_1 X_{1ijk} + \beta_2 X_{2ijk} + \alpha_i + \nu_j + e_{ijk}$$

Plan du cours

1. Rappels
2. Modèle linéaire gaussien
3. Inférence
4. Prédicteurs qualitatifs
5. ***Validation***
6. Modèle linéaire généralisé (GLM)

Diagnostics et validation

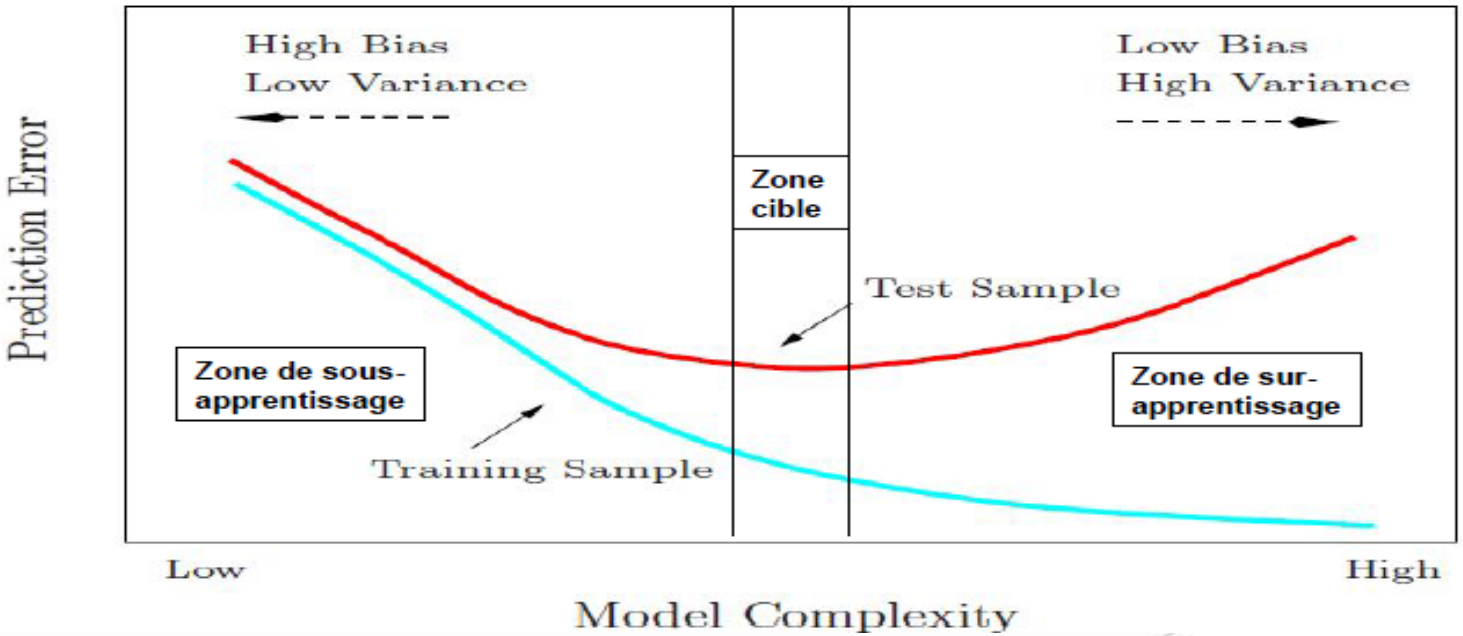
Généralités :

- **L'apprentissage du modèle se fera sur un sous-échantillon dédié, une partie des données ayant été préalablement isolée de l'archive afin d'évaluer les performances du modèle en conditions opérationnelles (phase de test).**

- **Un modèle trop complexe manquera de robustesse (incapacité à généraliser à de nouvelles données), on parle alors de *sur-apprentissage* (cf. TP).**

Illustration du phénomène de sur-apprentissage

Graphe extrait de l'ouvrage The elements of statistical learning, Hastie & al.



Diagnostics et validation

- **Multicolinéarité et calcul explicite de $({}^tTT)^{-1}$**

Si ${}^tTT$ est mal conditionnée, son déterminant est proche de zéro.

$\Downarrow$

variance élevée des estimateurs

→ estimateurs instables, conclusions peu fiables

Diagnostic : facteur d'inflation de la variance (VIF) : $V_j = \dfrac{1}{1 - R_j^2}$

R_j^2 coefficient de détermination multiple de la régression de la $j^{ème}$ variable avec toutes les autres. De fortes valeurs du VIF permettent d'identifier et de retirer les variables posant problème (VIF > 10), on peut également combiner ces variables pour créer un prédicteur élaboré.

Régression en composantes principales :

Permet de résoudre les problèmes de multicolinéarité lorsque l'on tient à conserver toutes les variables.

Diagnostics et validation

- **Etude des résidus**

L'étude des résidus après estimation est *FONDAMENTALE*.
Elle permet de *vérifier les hypothèses de base* du modèle linéaire, de repérer les données pathologiques, et de mettre en évidence les inadéquations au modèle.

<u>Avant estimation</u> : analyses uni et bivariées, afin d'identifier des problèmes sur les distributions de chacune des variables (dissymétrie, valeurs atypiques) ou sur les liaisons des variables prises deux par deux (non linéarité de la liaison → création d'un prédicteur élaboré plus adapté aux données).

<u>Après estimation</u> : détection des violations d'hypothèses (linéarité, indépendance, homoscédasticité) ou de points influents dans le contexte multidimensionnel (cf. TP).

Diagnostics et validation

1. Normalité non respectée

- **estimations de β et σ^2 restent sans biais**
- **$\hat{\beta}$ n'est plus exactement normal, ni optimal**
- **les tests de Student et Fisher sont biaisés**

Les théorèmes centraux-limites généralisés relativisent l'influence de la non-normalité des données :

Echantillon suffisamment grand + loi des données pas trop dissymétrique
→ propriétés de normalité de$\hat{\beta}$ et exactitude des tests à peu près conservées.

Vérification :
diagrammes quantiles/quantiles, tests de Kolmogorov-Smirnov, Shapiro…

Diagnostics et validation

2. Non-indépendance

- Les estimateurs restent non biaisés mais ne sont plus optimaux

- Concerne les données issues de séries chronologiques

- Décorrélation totale : souvent difficile à obtenir

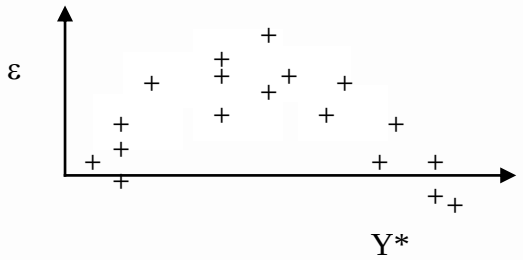
Tant que la corrélation temporelle reste faible, ce n'est pas très gênant.

En cas de problèmes importants :
modèle autorégressif ou estimation par Moindres Carrés Généralisés.

Vérification : corrélogramme des résidus, test de Durbin-Watson

Diagnostics et validation

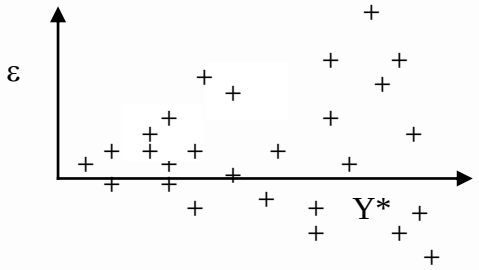
3. Inadéquation



Le graphe des résidus en fonction de Y* permet de vérifier les propriétés d'adéquation et d'homoscédasticité du modèle.

La relation entre Y et Y* n'est pas linéaire. L'analyse du problème doit être améliorée, en proposant d'autres prédicteurs ou en les transformant, ce qui peut être fait sans précautions particulières.

4. Hétéroscédasticité



La structure de la variance est la pathologie la plus problématique.
Seules les propriétés de non biais et de convergence de $\hat{\beta}$ sont conservées.

Nécessite un changement de variable , ou l'emploi de techniques de modèle linéaire généralisé.
(Vérification : graphique ou tests de Bartlett, Levene)

Diagnostics et validation

5. Observations influentes

Certaines observations peuvent avoir une plus grande influence que d'autres, par effet levier.

On utilise généralement la distance de Cook pour détecter de telles observations (lorsque $D_i > 1$, la ième observation est considérée comme pathologique).
Une distance entre $\hat{\beta}$ et $\hat{\beta}_i$ est calculée, $\hat{\beta}_i$ étant le vecteur estimé en retirant l'observation i.

Ces observations (points leviers ou valeurs aberrantes) influencent grandement les estimations et doivent donc être retirées.
Cette distance est disponible dans la plupart des logiciels statistiques.

Plan du cours

1. Rappels
2. Modèle linéaire gaussien
3. Inférence
4. Prédicteurs qualitatifs
5. Validation
6. ***Modèle linéaire généralisé (GLM)***

Modèle linéaire généralisé et fonction de lien

- **Modèle linéaire gaussien** :

  L'espérance du prédictand Y est modélisée linéairement et conditionnellement aux p prédicteurs : $E[Y / X_1,...,X_p] = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$

- **Modèle linéaire généralisé** : (généralisation à des prédictands non gaussiens)

  Une fonction de l'espérance du prédictand Y est modélisée linéairement et conditionnellement aux p prédicteurs : $g(E[Y / X_1,...,X_p]) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$

  La fonction g, nommée fonction de lien canonique, dépend de la loi suivie par le prédictand. La théorie du modèle linéaire généralisé montre qu'il existe une fonction de lien pour chaque loi de la famille exponentielle.

  Remarque : le modèle linéaire gaussien est donc un cas particulier de modèle linéaire généralisé pour lequel la fonction de lien est la fonction identité.

Famille exponentielle

- La famille exponentielle regroupe de nombreuses loi courantes : Normale, Exponentielle, Gamma, Bêta, Poisson, Bernoulli, Binomiale...

- La loi de Y appartient à la famille exponentielle si sa densité peut s'exprimer sous la forme :

$$f(Y) = \exp\left[\frac{Y\theta - b(\theta)}{a(\phi)} + c(Y,\phi)\right]$$

avec a, b, c trois fonctions arbitraires,

θ paramètre naturel (ou canonique) de la loi, fonction de E[Y] :

θ = g(E[Y]) avec g = fonction de lien canonique,

ϕ paramètre d'échelle (absent si la loi ne comporte qu'un paramètre).

Modèle généralisé particulier : la régression logistique

Dans le cas de la régression logistique, le prédictand est une variable binaire Y qui suit une loi de Bernoulli de paramètre : P(Y=1)=E[Y]

La fonction de lien canonique associée à la loi de Bernoulli est la fonction logit : g(x) = ln(x/(1-x)) (→ résultat à montrer en exercice)

Un modèle de régression logistique aura donc pour expression :

$$g(E[Y/X_1,...,X_q]) = \ln\left(\frac{P(Y=1/X_1,...,X_q)}{1-P(Y=1/X_1,...,X_q)}\right) = \beta_0 + \sum_{j=1}^{q}\beta_j X_j$$

L'estimation des paramètres se fera par maximisation de la vraisemblance, on obtiendra ensuite une estimation de l'espérance conditionnelle ainsi (fonction logistique = fct réciproque de la fct logit) :

$$\hat{P}(Y=1/X_1,...,X_p) = g^{-1}(\hat{\beta}_0 + \sum_{j=1}^{p}\hat{\beta}_j X_j)$$

$$= \frac{1}{1+\exp\left[-(\hat{\beta}_0 + \sum_{j=1}^{p}\hat{\beta}_j X_j)\right]}$$

Intérêt de la régression logistique

Un modèle de régression logistique va permettre de modéliser l'occurrence d'un phénomène (orage, brouillard…) mais également le dépassement d'un seuil d'un prédictand quantitatif.

Un tel modèle permet donc la discrimination des deux modalités d'un prédictand binaire : occurrence (codée 1) et non occurrence (codée 0).

Le modèle calcule des estimations de probabilités d'occurrence du phénomène étudié, il va donc falloir post-traiter ces probabilités prévues pour définir la prévision binaire (occurrence ou non occurrence prévue).

→ Un seuil de probabilité va donc devoir être exploité pour passer d'une prévision probabiliste à une prévision déterministe (cf.TP et courbe ROC)