



Modèles de régression

OUARET Rachid

22 septembre 2022

Table of contents I

1 Objectifs du cours

2 Considérations générales

- Besoins de modélisation des données
- Démarche globale
- Objectifs de la modélisation par les données
- Quels Modèles ?

3 La régression linéaire

- Introduction
 - Vocabulaire et notations
 - Hypothèses
- Estimateurs des moindres carrés
 - Calcul des estimateurs
 - Propriétés des estimateurs
 - Résidus et variance résiduelle
 - Décomposition de la variance (ANOVA)
 - Le coefficient de détermination

Table of contents II

- Estimateurs du maximum de vraisemblance
- Prédiction
- Inférence

4 La régression linéaire multiple

5 References

Table of contents

1 Objectifs du cours

2 Considérations générales

- Besoins de modélisation des données
- Démarche globale
- Objectifs de la modélisation par les données
- Quels Modèles ?

3 La régression linéaire

- Introduction
 - Vocabulaire et notations
 - Hypothèses
- Estimateurs des moindres carrés
 - Calcul des estimateurs
 - Propriétés des estimateurs
 - Résidus et variance résiduelle
 - Décomposition de la variance (ANOVA)
 - Le coefficient de détermination
- Estimateurs du maximum de vraisemblance

Objectifs du cours

À la fin de ce cours, vous serez capable de :

- 1 Apprécier l'apport de la modélisation des données expérimentales pour les problématiques de GI ;

Objectifs du cours

À la fin de ce cours, vous serez capable de :

- 1 Apprécier l'apport de la modélisation des données expérimentales pour les problématiques de GI ;
- 2 Maîtriser les techniques d'identification paramétrique pour :

Objectifs du cours

À la fin de ce cours, vous serez capable de :

- 1 Apprécier l'apport de la modélisation des données expérimentales pour les problématiques de GI ;
- 2 Maîtriser les techniques d'identification paramétrique pour :
 - L'estimation des modèles linéaires et non-linéaire par rapport aux paramètres.

Objectifs du cours

À la fin de ce cours, vous serez capable de :

- 1 Apprécier l'apport de la modélisation des données expérimentales pour les problématiques de GI ;
- 2 Maîtriser les techniques d'identification paramétrique pour :
 - L'estimation des modèles linéaires et non-linéaire par rapport aux paramètres.
 - L'analyse et l'évaluation de ces modèles.

Objectifs du cours

À la fin de ce cours, vous serez capable de :

- 1 Apprécier l'apport de la modélisation des données expérimentales pour les problématiques de GI ;
- 2 Maîtriser les techniques d'identification paramétrique pour :
 - L'estimation des modèles linéaires et non-linéaire par rapport aux paramètres.
 - L'analyse et l'évaluation de ces modèles.
- 3 Mise en œuvre avec R/MSExcels

Table of contents

1 Objectifs du cours

2 Considérations générales

- Besoins de modélisation des données
- Démarche globale
- Objectifs de la modélisation par les données
- Quels Modèles ?

3 La régression linéaire

- Introduction
 - Vocabulaire et notations
 - Hypothèses
- Estimateurs des moindres carrés
 - Calcul des estimateurs
 - Propriétés des estimateurs
 - Résidus et variance résiduelle
 - Décomposition de la variance (ANOVA)
 - Le coefficient de détermination
- Estimateurs du maximum de vraisemblance

Besoins de modélisation des données

Pourquoi avons-nous besoin des modèles pour les données ?

- De grandes quantités de données peuvent avoir des relations et des corrélations cachées

Besoins de modélisation des données

Pourquoi avons-nous besoin des modèles pour les données ?

- De grandes quantités de données peuvent avoir des relations et des corrélations cachées
 - Seules les approches automatisées et algorithmiques peuvent être en mesure de les détecter ;

Besoins de modélisation des données

Pourquoi avons-nous besoin des modèles pour les données ?

- De grandes quantités de données peuvent avoir des relations et des corrélations cachées
 - Seules les approches automatisées et algorithmiques peuvent être en mesure de les détecter ;
- Certaines décisions à prendre sont peut être trop importantes pour les laisser uniquement aux "humains", par exemple le diagnostic médical ;

Besoins de modélisation des données

Pourquoi avons-nous besoin des modèles pour les données ?

- De grandes quantités de données peuvent avoir des relations et des corrélations cachées
 - Seules les approches automatisées et algorithmiques peuvent être en mesure de les détecter ;
- Certaines décisions à prendre sont peut être trop importantes pour les laisser uniquement aux "humains", par exemple le diagnostic médical ;
- La précision, ...

Démarche globale

Étapes d'un projet d'exploration de données dans l'industrie [3]

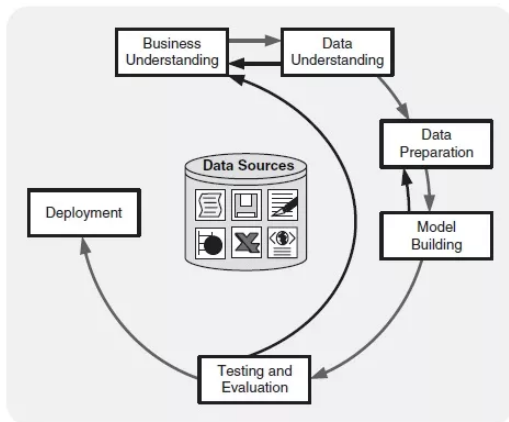


Figure: **CRISP** pour "Cross-Industry Standard Process for data mining"

Objectifs de modélisation par les données

Modéliser pour :

- Explorer ou vérifier, représenter, décrire, les variables, leurs liaisons et positionner les observations de l'échantillon ;
- Expliquer ou tester l'influence d'une variable ou facteur dans un modèle supposé connu a priori ;
- Sélectionner un meilleur ensemble de prédicteurs :
 - Par exemple dans la recherche de bio-marqueurs ;
- Prévoir le comportement d'un système et prédire ses défaillances.

Objectifs de modélisation par les données

Modéliser pour :

- Explorer ou vérifier, représenter, décrire, les variables, leurs liaisons et positionner les observations de l'échantillon ;
- Expliquer ou tester l'influence d'une variable ou facteur dans un modèle supposé connu a priori ;
- Sélectionner un meilleur ensemble de prédicteurs :
 - Par exemple dans la recherche de bio-marqueurs ;
- Prévoir le comportement d'un système et prédire ses défaillances.

Objectifs de modélisation par les données

Modéliser pour :

- Explorer ou vérifier, représenter, décrire, les variables, leurs liaisons et positionner les observations de l'échantillon ;
- Expliquer ou tester l'influence d'une variable ou facteur dans un modèle supposé connu a priori ;
- Sélectionner un meilleur ensemble de prédicteurs :
 - Par exemple dans la recherche de bio-marqueurs ;
- Prévoir le comportement d'un système et prédire ses défaillances.

Objectifs de modélisation par les données

Modéliser pour :

- Explorer ou vérifier, représenter, décrire, les variables, leurs liaisons et positionner les observations de l'échantillon ;
- Expliquer ou tester l'influence d'une variable ou facteur dans un modèle supposé connu a priori ;
- Sélectionner un meilleur ensemble de prédicteurs :
 - Par exemple dans la recherche de bio-marqueurs ;
- Prévoir le comportement d'un système et prédire ses défaillances.

Objectifs de modélisation par les données

Modéliser pour :

- Explorer ou vérifier, représenter, décrire, les variables, leurs liaisons et positionner les observations de l'échantillon ;
- Expliquer ou tester l'influence d'une variable ou facteur dans un modèle supposé connu a priori ;
- Sélectionner un meilleur ensemble de prédicteurs :
 - Par exemple dans la recherche de bio-marqueurs ;
- **Prévoir le comportement d'un système et prédire ses défaillances.**

Objectifs de modélisation par les données : Exemple

299 gènes identifiés comme des facteurs de cancer

Les gènes prédictifs spécifiques au différents types de cancer.

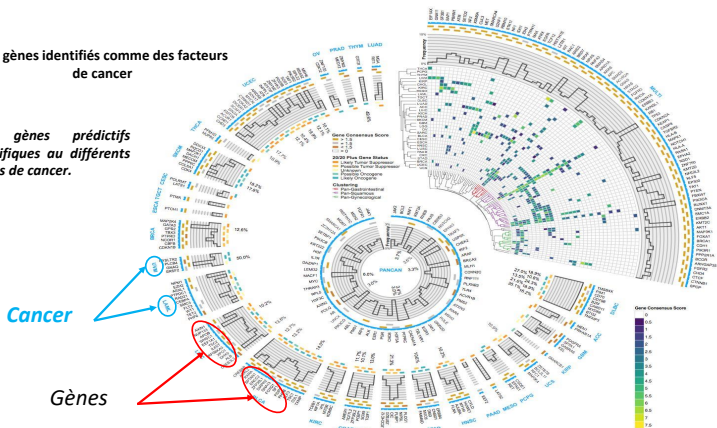


Figure: **Circos** montre les 299 gènes identifiés comme des facteurs de cancer. *Cell : Volume 173, Issue 2, 5 April 2018, Pages 371-385.e18*

Quels Modèles ?

À retenir !

On désignera sous le terme de **modèle** une équation paramétrée (ou un ensemble d'équations paramétrées) permettant de calculer la valeur de la grandeur (ou des grandeurs) à modéliser à partir des valeurs d'autres grandeurs appelées variables ou facteurs.

Quels Modèles ?

À retenir !

On désignera sous le terme de **modèle** une équation paramétrée (ou un ensemble d'équations paramétrées) permettant de calculer la valeur de la grandeur (ou des grandeurs) à modéliser à partir des valeurs d'autres grandeurs appelées variables ou facteurs.

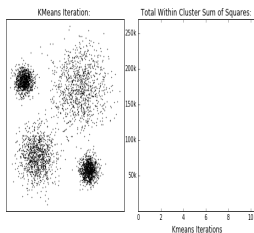
Remark (De quoi s'agit-il)

Dans ce cours, on désigne par un modèle mathématique une série d'équations ou de représentations graphiques qui décrivent des relations entre variables d'une manière précise.

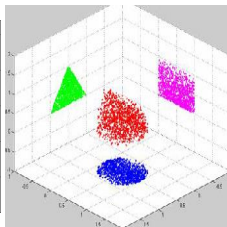
Quels Modèles ? : Tout dépend pour quel but

Modèles

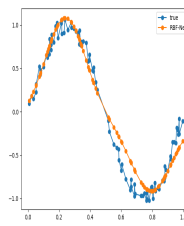
La classification



Réduction de Dimension



La régression



La prévision

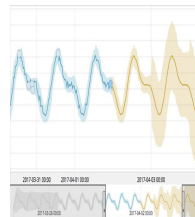


Table of contents

1 Objectifs du cours

2 Considérations générales

- Besoins de modélisation des données
- Démarche globale
- Objectifs de la modélisation par les données
- Quels Modèles ?

3 La régression linéaire

■ Introduction

- Vocabulaire et notations
- Hypothèses

■ Estimateurs des moindres carrés

- Calcul des estimateurs
- Propriétés des estimateurs
- Résidus et variance résiduelle
- Décomposition de la variance (ANOVA)
- Le coefficient de détermination

■ Estimateurs du maximum de vraisemblance

Introduction

La modélisation suppose que l'on étudie un ensemble d'objets sur lesquels on observe des caractéristiques appelées **variables**.

Introduction

La modélisation suppose que l'on étudie un ensemble d'objets sur lesquels on observe des caractéristiques appelées **variables**.

Pratique générale de la modélisation statistique

Une modélisation statistique consiste à

- 1 établir une **relation** entre variables sous forme d'équation, qu'on appelle **modèle** ;
- 2 **estimer** ce modèle sur des données observées ;
- 3 **vérifier et valider** ce modèle par des tests ;
- 4 **prédire** les nouvelles données (leurs valeurs, classes, ...) en utilisant le modèle validé.

Vocabulaire et notations

Un modèle explicatif est un modèle exprimant une variable \mathcal{Y} *appelée variable à expliquer* (ou réponse) comme une fonction d'une ou de plusieurs variables dites *variables explicatives* ou prédicteurs notées \mathcal{X} .

Vocabulaire et notations

Un modèle explicatif est un modèle exprimant une variable \mathcal{Y} appelée *variable à expliquer* (ou réponse) comme une fonction d'une ou de plusieurs variables dites *variables explicatives* ou prédicteurs notées \mathcal{X} .

Attention !

Un tel modèle ne restituant pas nécessairement une relation directe de cause à effet. Le terme de *prédicteur* est plus approprié.

Vocabulaire et notations

Un modèle explicatif est un modèle exprimant une variable \mathcal{Y} *appelée variable à expliquer* (ou réponse) comme une fonction d'une ou de plusieurs variables dites *variables explicatives* ou prédicteurs notées \mathcal{X} .

Attention !

Un tel modèle ne restituant pas nécessairement une relation directe de cause à effet. Le terme de *prédicteur* est plus approprié.

 Mais les valeurs prises par la variable $\in \mathbb{R}$

L'environnement dans lequel les variables évoluent est *incertain*. L'entité \mathcal{Y} est considérée comme une **variable aléatoire** Y . Bien souvent, on se place dans le cadre d'un couple de v.a. (X, Y) .

Vocabulaire et notations : **taxonomie partielle des modèles**

- Modèle statique déterministe

- $Y = f(X; \theta)$

Vocabulaire et notations : **taxonomie partielle des modèles**

- Modèle statique déterministe
- Modèle statique stochastique

- $Y = f(X; \theta)$

- $Y = f(X; \theta) + \varepsilon$

Vocabulaire et notations : **taxonomie partielle des modèles**

- Modèle statique déterministe
- Modèle statique stochastique
- Modèle dynamique déterministe

- $Y = f(X; \theta)$

- $Y = f(X; \theta) + \varepsilon$

- $\frac{\partial Y}{\partial t} = f(Y(t), X(t); \theta)$

Vocabulaire et notations : **taxonomie partielle des modèles**

- Modèle statique déterministe
- Modèle statique stochastique
- Modèle dynamique déterministe
- Modèle dynamique stochastique

- $Y = f(X; \theta)$

- $Y = f(X; \theta) + \varepsilon$

- $\frac{\partial Y}{\partial t} = f(Y(t), X(t); \theta)$

- $\frac{\partial Y}{\partial t} = f(Y(t), X(t); \theta) + \varepsilon(t)$

Vocabulaire et notations

Table: Terminologie des variables pour la régression utilisée dans la littérature

Y	X
Variable dépendante	Variable indépendante
Variable à expliquer	Variable explicative
Variable endogène	Variable exogène
Variable réponse	Variable de contrôle
<i>Predicted variable (à prédire)</i>	<i>Predictor variable (prédictive)</i>
<i>Regressand</i>	<i>Regressor</i>

Vocabulaire et notations

On souhaite, par exemple étudier la variation du taux d'ozone \mathcal{Y} en fonction de la température \mathcal{X} . Pour n mesures (individus, ...) on observe un échantillon aléatoire $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ à valeurs dans \mathbb{R}^2 .

☛ Pour les **valeurs effectivement observées**

$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ l'objectif essentiel est d'étudier comment varie en "*moyenne*" les concentrations d'ozone en fonction de la température.

Vocabulaire et notations

On souhaite, par exemple étudier la variation du taux d'ozone \mathcal{Y} en fonction de la température \mathcal{X} . Pour n mesures (individus, ...) on observe un échantillon aléatoire $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ à valeurs dans \mathbb{R}^2 .

☛ Pour les **valeurs effectivement observées**

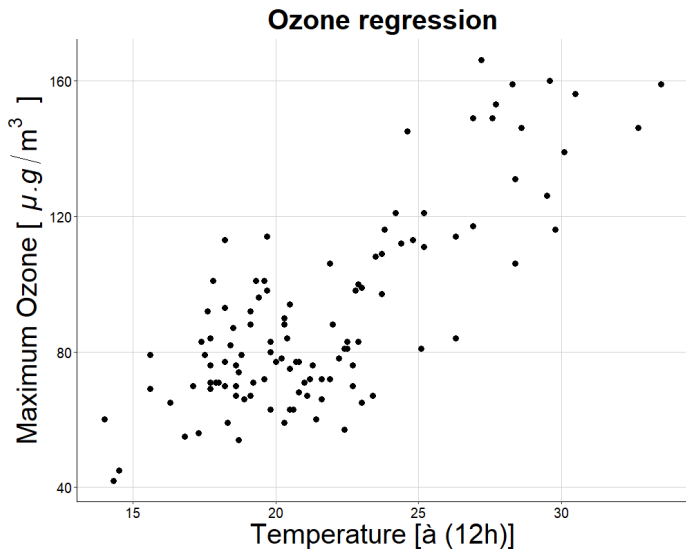
$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ l'objectif essentiel est d'étudier comment varie en "*moyenne*" les concentrations d'ozone en fonction de la température.

📖 La fonction de régression

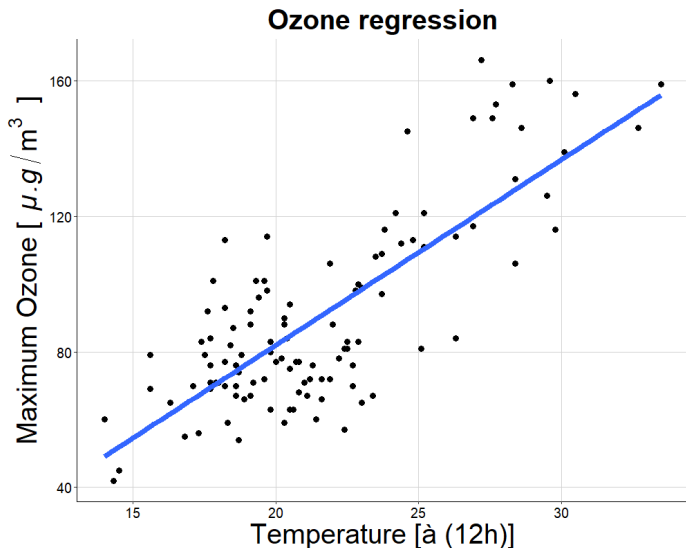
La fonction de régression $g(x)$ est alors l'**espérance mathématique de la loi conditionnelle** de Y sachant $X = x$, notée $\mathbb{E}(Y \mid X = x)$, soit :

$$g(x) = \mathbb{E}(Y \mid X = x). \quad (1)$$

Vocabulaire et notations : **visualisation**



Vocabulaire et notations : **visualisation**



Vocabulaire et notations

Supposons que la concentration d'ozone dépend linéairement de la température mais cette liaison est perturbée par un "*bruit*" :

Le modèle de régression linéaire simple

$$Y = \beta_1 + \beta_2 X + \varepsilon. \quad (2)$$

Matriciellement, on peut écrire :

$$\mathbf{y} = \beta \mathbf{x} + \varepsilon. \quad (3)$$

où

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Vocabulaire et notations

- Le terme aléatoire ε permet de résumer toute l'information qui n'est pas prise en compte dans la relation linéaire entre Y et X

Vocabulaire et notations

- Le terme aléatoire ε permet de résumer toute l'information qui n'est pas prise en compte dans la relation linéaire entre Y et X
- Après avoir estimé les paramètres β_1, β_2 de la régression, les premières vérifications portent sur l'erreur calculée sur les données

Hypothesis (sur les modèles et les variables)

1 *Sur le modèle*

Hypothesis (sur les modèles et les variables)

1 **Sur le modèle**

- *La forme est linéaire par rapport aux paramètres :*
$$Y = \beta_1 + \beta_2 X + \varepsilon.$$

Hypothesis (sur les modèles et les variables)

1 **Sur le modèle**

- *La forme est linéaire par rapport aux paramètres :*
$$Y = \beta_1 + \beta_2 X + \varepsilon.$$

2 **Sur les variables Y et X**

Hypothesis (sur les modèles et les variables)

1 **Sur le modèle**

- *La forme est linéaire par rapport aux paramètres :*

$$Y = \beta_1 + \beta_2 X + \varepsilon.$$

2 **Sur les variables Y et X**

- *X est non aléatoire*

Hypothesis (sur les modèles et les variables)

1 **Sur le modèle**

- *La forme est linéaire par rapport aux paramètres :*

$$Y = \beta_1 + \beta_2 X + \varepsilon.$$

2 **Sur les variables Y et X**

- *X est non aléatoire*
- *Y est une variable aléatoire par l'intermédiaire de ε , c.-à-d. la seule erreur que l'on a sur Y provient des insuffisances de X à expliquer ses valeurs dans le modèle.*

Hypothesis (sur les l'orgine de l'aléa dans le modèle)

Sur les perturbations aléatoires : ε_i elles sont supposées i.i.d (indépendants et identiquement distribués).

- **La spécification du modèle :** $\mathbb{E}(\varepsilon_i) = 0$, en moyenne les erreurs s'annulent.
- **Homoscédasticité et absence d'autocorrélation :**
 $\text{Var}(\varepsilon_i) = \sigma_\varepsilon^2$, la variance de ε_i est constante et finie et ε_i n'est pas corrélé avec ε_j , pour $i \neq j$. σ_ε^2 est un paramètre inconnu, à estimer. Cette hypothèse (très) restrictive, souvent levée !
- **Distribution normale :** les perturbations sont distribuées selon la loi normale, $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$.

Table of contents

1 Objectifs du cours

2 Considérations générales

- Besoins de modélisation des données
- Démarche globale
- Objectifs de la modélisation par les données
- Quels Modèles ?

3 La régression linéaire

■ Introduction

- Vocabulaire et notations
- Hypothèses

■ Estimateurs des moindres carrés

- Calcul des estimateurs
- Propriétés des estimateurs
- Résidus et variance résiduelle
- Décomposition de la variance (ANOVA)
- Le coefficient de détermination

■ Estimateurs du maximum de vraisemblance

Calcul des estimateurs

Le critère des moindres carrés consiste à minimiser la somme des carrés des écarts (des erreurs) entre les vraies valeurs de Y et les valeurs prédites avec le modèle de prédiction.

Estimateurs des MC

Les estimateurs des moindres carrés (MC) de β_1 et β_2 , notés $\widehat{\beta}_1$ et $\widehat{\beta}_2$ sont obtenus par la minimisation la quantité :

$$S(\beta_1, \beta_2) = \sum_{i=1}^n (y_i - \beta_1 - \beta_2 x_i)^2 \quad (4)$$

Les estimateurs peuvent également s'écrire sous la forme suivante :

$$(\widehat{\beta}_1, \widehat{\beta}_2) = \underset{(\beta_1, \beta_2) \in \mathbb{R} \times \mathbb{R}}{\operatorname{argmin}} \left(\sum_{i=1}^n (y_i - \beta_1 - \beta_2 x_i)^2 \right) \quad (5)$$

Calcul des estimateurs

La fonction $S(\beta_1, \beta_2)$ est strictement convexe. Si elle admet un point singulier, celui-ci correspond à l'unique minimum. Lorsque nous annulons les dérivées partielles de $S(\beta_1, \beta_2)$ par rapport aux paramètres, nous obtenons un système d'équations appelées "*équations normales*" :

$$\begin{cases} \frac{\partial S(\hat{\beta}_1, \hat{\beta}_2)}{\partial \beta_1} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0 \\ \frac{\partial S(\hat{\beta}_1, \hat{\beta}_2)}{\partial \beta_2} = -2 \sum_{i=1}^n x_i (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0 \end{cases}$$

Calcul des estimateurs

Estimateurs des MC

- *L'estimateur de l'ordonnée à l'origine :*

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x} \quad (6)$$

où $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$.

- *L'estimateur de la pente de la droite :*

$$\hat{\beta}_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{COV(X, Y)}{\hat{\sigma}_X^2} \quad (7)$$

Calcul des estimateurs : **Application directe**

Rendement de maïs et quantité d'engrais : On cherche à expliquer le rendement en maïs (en quintal) de parcelles de terrain qu'on note par la variable Y , à partir de X la quantité d'engrais (en kg). L'objectif est de modéliser la relation à travers une régression linéaire.

Table: Rendement de maïs et quantité d'engrais

Y	16	18	23	24	28	29	26	31	32	34
X	20	24	28	22	32	28	32	36	41	41

Biais et variance des estimateurs ([2], p 13)

Biais des estimateurs

$\widehat{\beta}_1$ et $\widehat{\beta}_2$ sont des estimateurs **sans biais** de β_1 et β_2 c'est-à-dire que $\mathbb{E}(\widehat{\beta}_1) = \beta_1$ et $\mathbb{E}(\widehat{\beta}_2) = \beta_2$.

☛ En moyenne sur toutes les expériences possibles de taille n , l'estimateur $\widehat{\beta}$ moyen sera égal à la valeur inconnue du paramètre β .

- Les estimations de β sont, en moyenne, autour de β ;
- Simplement, en moyenne, $\widehat{\beta}$ "tombe" sur β .

Biais et variance des estimateurs ([2], p 13)

Biais des estimateurs

$\widehat{\beta}_1$ et $\widehat{\beta}_2$ sont des estimateurs **sans biais** de β_1 et β_2 c'est-à-dire que $\mathbb{E}(\widehat{\beta}_1) = \beta_1$ et $\mathbb{E}(\widehat{\beta}_2) = \beta_2$.

☛ En moyenne sur toutes les expériences possibles de taille n , l'estimateur $\widehat{\beta}$ moyen sera égal à la valeur inconnue du paramètre β .



- Les estimations de β sont, en moyenne, autour de β ;
- Simplement, en moyenne, $\widehat{\beta}$ "tombe" sur β .

Biais et variance des estimateurs

Variances des estimateurs ([2], p 14)

Les variances et covariance des estimateurs des paramètres valent :

- $\mathbb{V}(\widehat{\beta}_1) = \frac{\sigma^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}$
- $\mathbb{V}(\widehat{\beta}_2) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$
- $Cov(\widehat{\beta}_1, \widehat{\beta}_2) = -\frac{\sigma^2 \bar{x}}{\sum (x_i - \bar{x})^2}$

-  Plus la variance est faible, plus l'estimateur sera précis
-  Avoir numérateur petit et (ou) un dénominateur grand

Biais et variance des estimateurs

📖 Variances des estimateurs ([2], p 14)

Les variances et covariance des estimateurs des paramètres valent :

- $\mathbb{V}(\widehat{\beta}_1) = \frac{\sigma^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}$
- $\mathbb{V}(\widehat{\beta}_2) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$
- $Cov(\widehat{\beta}_1, \widehat{\beta}_2) = -\frac{\sigma^2 \bar{x}}{\sum (x_i - \bar{x})^2}$

- 🐼 Plus la variance est faible, plus l'estimateur sera précis
- 🐼 Avoir numérateur petit et (ou) un dénominateur grand

Résidus et variance résiduelle

On peut en déduire l'erreur observée, appelée "résidu" de la régression par

$$\hat{\varepsilon}_i = y_i - \hat{y}_i \quad (8)$$

où \hat{y}_i est la valeur ajustée de y_i par le modèle (i.e. $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$) avec :

■ $\sum_{i=1}^n \hat{\varepsilon}_i = 0$

Variance résiduelle

l'estimateur de la variance du bruit est défini par la statistique

$$\hat{\sigma}_{\varepsilon}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{\varepsilon}_i^2 \quad (9)$$

Décomposition de la variance

La variation d'une variable y est obtenue en considérant les différences entre les valeurs observées y_i et leur moyenne \bar{y} . Or on a :

$$y_i - \bar{y} = \hat{y}_i - \bar{y} + y_i - \hat{y}_i$$

où $\hat{y}_i - \bar{y}$ est la variation expliquée (ou restituée) par le modèle, alors que $y_i - \hat{y}_i$ est la variation non expliquée par le modèle.

Décomposition de la variance

La variation d'une variable y est obtenue en considérant les différences entre les valeurs observées y_i et leur moyenne \bar{y} . Or on a :

$$y_i - \bar{y} = \hat{y}_i - \bar{y} + y_i - \hat{y}_i$$

où $\hat{y}_i - \bar{y}$ est la variation expliquée (ou restituée) par le modèle, alors que $y_i - \hat{y}_i$ est la variation non expliquée par le modèle.

Remark (Au "*sources*" de variabilité)

La somme des carrés totale (SCT) est la quantité suivante :

$$\begin{aligned} \text{SCT} &= \sum_{i=1}^n (y_i - \bar{y})^2 \\ &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 + 2 \sum_{i=1}^n (\hat{y}_i - \bar{y}) (y_i - \hat{y}_i) \end{aligned}$$

Décomposition de la variance

Dans la régression avec constante (uniquement dans ce cas), on montre que

$$2 \sum_{i=1}^n (\hat{y}_i - \bar{y}) (y_i - \hat{y}_i) = 0$$

ANOVA : ANalysis Of VAriance

$$\begin{aligned} \text{SCT} &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \text{SCE} + \text{SCR} \end{aligned}$$

Décomposition de la variance

ANOVA : ANalysis Of VAriance

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

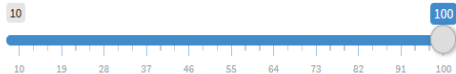
$$\text{SCT} = \text{SCE} + \text{SCR}$$

- SCT (Somme des Carrés Totale) traduit la variation totale de Y.
- SCE (Somme des Carrés Expliquée) traduit la variation expliquée par le modèle, i.e. la variation de Y expliquée par X.
- SCR (Somme des Carrés Résiduelle) traduit la variation inexpliquée par le modèle, i.e. l'écart entre les valeurs observées de Y et celles prédites par le modèle.

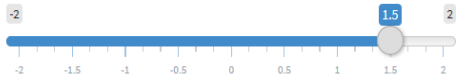
Décomposition de la variance : Exemple

Parameters

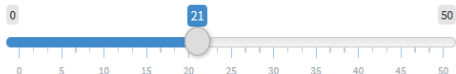
Sample size



Regression slope

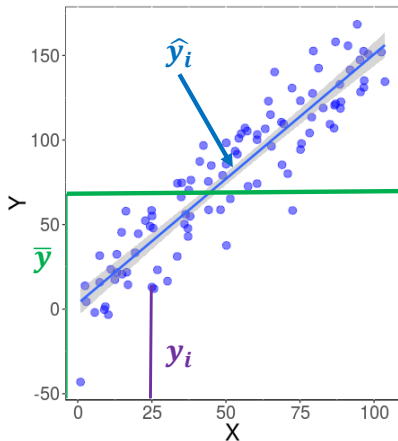


Standard deviation

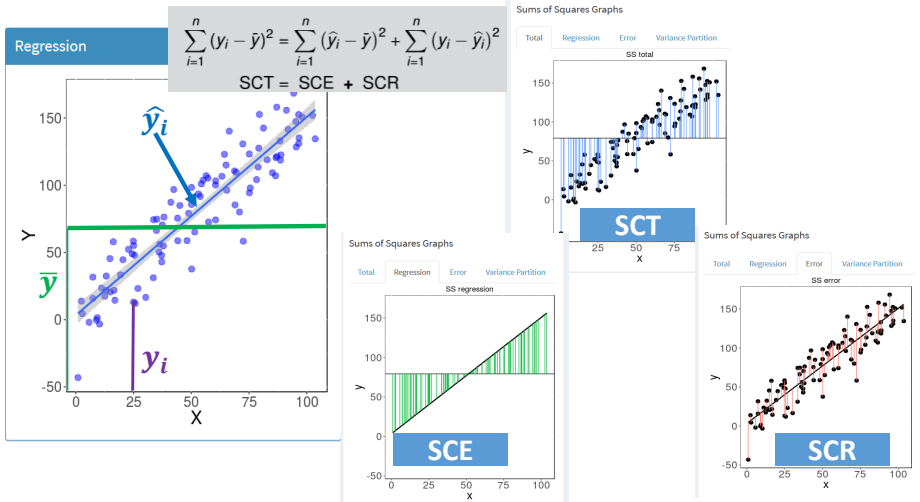


Simulate New Data

Regression



Décomposition de la variance : Exemple



Décomposition de la variance : **Exemple**

Jouyons avec la régression en suivant ce lien :

https://ouaretrachid.shinyapps.io/ANOVA_Shiny_2A_FISA_GI/

Décomposition de la variance : **deux cas extrêmes**

Pour l'interprétation, deux situations extrêmes peuvent survenir :

Décomposition de la variance : **deux cas extrêmes**

Pour l'interprétation, deux situations extrêmes peuvent survenir :

Aide à l'interprétation de la qualité du modèle

- Dans le meilleur des cas, $SCR = 0$ et donc $SCT = SCE$: les variations de Y sont complètement expliquées par celles de X . On a un modèle parfait, la droite de régression passe exactement par tous les points du nuage ($\hat{y}_i = y_i$).
- Dans le pire des cas, $SCE = 0$: X n'apporte aucune information sur Y . Ainsi, $\hat{y}_i = \bar{y}$, la meilleure prédiction de Y est sa propre moyenne.

Le coefficient de détermination

Il est possible de déduire un indicateur synthétique à partir de l'équation d'analyse de variance.

Le coefficient de détermination R^2 .

$$R^2 = \frac{SCE}{SCT} \quad (10)$$

$$= 1 - \frac{SCR}{SCT} \quad (11)$$

Il indique la proportion de variance de Y expliquée par le modèle.

Le coefficient de détermination

Ce coefficient R^2 est dans $[0,1]$, puisque : $0 < SCE < SCT$

Le coefficient de détermination

Ce coefficient R^2 est dans $[0,1]$, puisque : $0 < SCE < SCT$

Aide à l'interprétation de la qualité du modèle linéaire

- Plus R^2 sera proche de la valeur 1, meilleur sera le modèle linéaire, la connaissance des valeurs de X permet de deviner avec précision celle de Y .
- Lorsque R^2 est proche de 0, on peut dire que la variation de X n'apporte pas d'informations utiles (intéressantes) sur les variations de Y , la connaissance des valeurs de X ne nous dit rien sur celles de Y .

Le coefficient de détermination : **le risque de surinterpréter**

Il faut veiller à ne pas surinterpréter le coefficient de détermination :

 **Interpréter oui, mais avec des pincettes !**

Il faut veiller à ne pas surinterpréter le coefficient de détermination :

- Un bon ajustement linéaire se traduit par un R^2 proche de 1.
- A contrario, un R^2 proche de 1 ne traduit pas forcément un lien linéaire.
- Un R^2 proche de 0 traduit un mauvais ajustement linéaire, mais n'implique pas qu'aucune relation ne puisse être établie entre les variables.

Estimateurs du maximum de vraisemblance

Le modèle linéaire avec pour tout i , $Y_i \sim \mathcal{N}(\beta_1 + \beta_2 x_i, \sigma^2)$ contient trois paramètres inconnus : β_1 , β_2 et σ^2 .

☛ La fonction de vraisemblance des paramètres, associée aux réalisations y_1, y_2, \dots, y_n , est :

$$\mathcal{L}(\beta_1, \beta_2, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2\sigma^2} [y_i - (\beta_1 + \beta_2 x_i)]^2 \right\} \quad (12)$$

D'où la log-vraisemblance

$$\ln \mathcal{L}(\beta_1, \beta_2, \sigma^2) = -n \left(\ln \sqrt{2\pi} + \frac{1}{2} \ln \sigma^2 \right) - \frac{1}{2\sigma^2} \sum_{i=1}^n [y_i - (\beta_1 + \beta_2 x_i)]^2 \quad (13)$$

Estimateurs du maximum de vraisemblance

👉 Equations de vraisemblance

$$\begin{cases} \sum_{i=1}^n [y_i - (\beta_1 + \beta_2 x_i)] & = 0 \\ \sum_{i=1}^n x_i [y_i - (\beta_1 + \beta_2 x_i)] & = 0 \\ \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n [y_i - (\beta_1 + \beta_2 x_i)]^2 & = 0 \end{cases}$$

Estimateurs du maximum de vraisemblance

👉 Equations de vraisemblance

$$\begin{cases} \sum_{i=1}^n [y_i - (\beta_1 + \beta_2 x_i)] & = 0 \\ \sum_{i=1}^n x_i [y_i - (\beta_1 + \beta_2 x_i)] & = 0 \\ \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n [y_i - (\beta_1 + \beta_2 x_i)]^2 & = 0 \end{cases}$$

📌 Estimateurs des MV sont identiques aux EMC

- $\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}.$
- $\hat{\beta}_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{COV(X, Y)}{\hat{\sigma}_X^2}$
- $\hat{\sigma}_{MV}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$

Estimateurs du maximum de vraisemblance

Quelques notes

- $\widehat{\sigma_{MV}^2} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$ est un estimateur biaisé
- C'est $\frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2$ qui est L'Estimateur Sans Biais (ESB)
 - Au dénominateur, nous avons les degrés de liberté de la régression.
 - La valeur 2 dans $(n - 2)$ représente le nombre de paramètres estimés.
 - Dans cadre de la régression linéaire multiple avec p variables exogènes ne pose aucun problème. Le nombre de degrés de liberté sera $n - (p + 1)$.

Prédiction

Un des buts de la régression est de proposer des prévisions pour la variable à expliquer Y .

👉 Equations de prédiction

Notons par x_{n+1} la nouvelle valeur (réalisation) de la variable X et voulons prédire la valeur y_{n+1} correspondante. Nous pouvons prédire la valeur correspondante grâce au modèle estimé :

$$\widehat{y_{n+1}^p} = \widehat{\beta_1} + \widehat{\beta_2}x_{n+1}. \quad (14)$$

- 1 ❶ La valeur pour laquelle nous effectuons la prédiction, ici la $(n+1)^{ème}$, n'a pas servi dans le calcul des estimateurs.

Prédiction

Un des buts de la régression est de proposer des prévisions pour la variable à expliquer Y .

👉 Equations de prédiction

Notons par x_{n+1} la nouvelle valeur (réalisation) de la variable X et voulons prédire la valeur y_{n+1} correspondante. Nous pouvons prédire la valeur correspondante grâce au modèle estimé :

$$\widehat{y_{n+1}^p} = \widehat{\beta_1} + \widehat{\beta_2}x_{n+1}. \quad (14)$$

- 1 ❶ La valeur pour laquelle nous effectuons la prédiction, ici la $(n+1)^{\text{ème}}$, n'a pas servi dans le calcul des estimateurs.
- 2 ❷ Deux types d'erreurs vont entacher notre prévision, l'une due à la nonconnaissance de ε_{n+1} et l'autre due à l'estimation des paramètres.

Prédiction

Variance de la prévision $\widehat{y_{n+1}^p}$ exercice

La variance de la valeur prévue de $\widehat{y_{n+1}^p}$ vaut

$$\mathbb{V} \left(\widehat{y_{n+1}^p} \right) = \sigma^2 \left(\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \quad (15)$$

- 1 ❶ Le calcul de la variance nous donne une idée sur la stabilité de l'estimation.

Prédiction

Variance de la prévision $\widehat{y_{n+1}^p}$ exercice

La variance de la valeur prévue de $\widehat{y_{n+1}^p}$ vaut

$$\mathbb{V} \left(\widehat{y_{n+1}^p} \right) = \sigma^2 \left(\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \quad (15)$$

- 1 **❶** Le calcul de la variance nous donne une idée sur la stabilité de l'estimation.
- 2 **❷** On s'intéresse dans la prévision ponctuelle aux erreurs que l'on commet entre la vraie valeur à prévoir y_{n+1} et celle que l'on prévoit $\widehat{y_{n+1}^p}$

Prédiction

Erreur de prévision **complément**

L'erreur de prévision, définie par $\widehat{\varepsilon_{n+1}^p} = y_{n+1} - \widehat{y_{n+1}^p}$ satisfait les propriétés suivantes :

$$\mathbb{E} \left(\widehat{\varepsilon_{n+1}^p} \right) = 0$$

$$\mathbb{V} \left(\widehat{\varepsilon_{n+1}^p} \right) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right).$$

Table of contents I

1 Objectifs du cours

2 Considérations générales

- Besoins de modélisation des données
- Démarche globale
- Objectifs de la modélisation par les données
- Quels Modèles ?

3 La régression linéaire

■ Introduction

- Vocabulaire et notations
- Hypothèses

■ Estimateurs des moindres carrés

- Calcul des estimateurs
- Propriétés des estimateurs
- Résidus et variance résiduelle
- Décomposition de la variance (ANOVA)
- Le coefficient de détermination

Table of contents II

- Estimateurs du maximum de vraisemblance
- Prédiction
- Inférence

4 La régression linéaire multiple

5 References

Inférence

Rappelons brièvement les étapes suivies jusqu'à présent :

- Nous avons pu, en choisissant **une fonction de coût quadratique**, ajuster un **modèle de régression** : calculer $\hat{\beta}_1$ et $\hat{\beta}_2$.
- grâce aux coefficients estimés, nous pouvons donc prédire, pour chaque nouvelle valeur x_{n+1} une valeur de la variable à expliquer $\widehat{y_{n+1}^p}$ qui est tout simplement le point sur la droite ajustée correspondant à l'abscisse x_{n+1}
- Avec l'hypothèse d'homoscédasticité des erreurs (centrées, de même variance), nous avons pu calculer l'espérance et la variance des estimateurs. Nous avons pu, également calculer l'espérance et la variance de la valeur prédite $\widehat{y_{n+1}^p}$.

Inférence

Nous souhaitons en général connaître la loi des estimateurs afin de calculer des intervalles ou des régions de confiance ou effectuer des tests. C'est là qu'intervient l'hypothèse de la normalité des perturbations :

$$\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$$

Pourquoi l'intervalle de confiance ?

Un paramètre estimé n'a presque "*aucune*" valeur si la **précision** de l'estimation réalisée n'est pas connue. Ceci peut être réalisé soit par :

- le calcul de l'erreur standard ;
- la détermination autour de la valeur estimée, un intervalle dont on a de bonnes raisons de croire qu'il contient la "vraie" valeur du paramètre recherché : *un intervalle de confiance*.

Inférence

Tests de la significativité de β_1 et β_2

Pour $j = \{1, 2\}$, on teste :

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

On utilise comme statistique de test :

$$T_j = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}}$$

Plus cette quantité est grande, plus on est enclin à rejeter l'hypothèse de nullité du paramètre.

Inférence

Notons que :

$$\hat{\sigma}_{\hat{\beta}_1} = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2},$$

$$\hat{\sigma}_{\hat{\beta}_2} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

la loi de T_j

On peut montrer que, sous H_0 , T_j admet comme loi la loi de Student à $n - 2$ degrés de liberté ($\mathcal{T}(n - 2)$).

$$T_j \sim \mathcal{T}(n - 2)$$

Inférence

On décide du rejet de H_0 au niveau de test α si

$$|T_j| > t_{n-2, 1-\frac{\alpha}{2}} \quad (16)$$

où $t_{n-2, 1-\frac{\alpha}{2}}$ désigne le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi $\mathcal{T}(n-2)$.

Inférence

On décide du rejet de H_0 au niveau de test α si

$$|T_j| > t_{n-2, 1-\frac{\alpha}{2}} \quad (16)$$

où $t_{n-2, 1-\frac{\alpha}{2}}$ désigne le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi $\mathcal{T}(n-2)$.

Remark

Dans la pratique, nous analysons souvent la p-valeur de la sortie des logiciels (notamment R) : elle représente la probabilité de faire une erreur de type 1, ou de **rejeter l'hypothèse nulle si elle est vraie**. Plus la valeur de p est petite, plus la probabilité de faire une erreur en rejetant l'hypothèse nulle est faible. Une valeur limite de 0,05 est souvent utilisée.

☞ Autrement dit, vous pouvez rejeter l'hypothèse nulle si la valeur de p est inférieure à 0,05.

Inférence

intervalles de confiance pour de β_1 et β_2

Pour $j \in \{1, 2\}$, le paramètre β_j admet comme intervalle de confiance de niveau $1 - \alpha$:

$$\left[\hat{\beta}_j - t_{n-2, 1-\frac{\alpha}{2}} \hat{\sigma}_{\hat{\beta}_j} ; \hat{\beta}_j + t_{n-2, 1-\frac{\alpha}{2}} \hat{\sigma}_{\hat{\beta}_j} \right]$$

Exemple

Imaginons que, pour une moyenne d'échantillon de 10 g, nous calculions que son intervalle de confiance à 95% aille de 5 à 15 g. Cela signifie qu'il y a 95% de chance que la vraie valeur de la moyenne de la population soit comprise entre 5 et 15 g, et que sa valeur la plus probable (sur base des données expérimentales observées) est 10 g.

La régression linéaire multiple

On souhaite cette fois expliquer, de manière linéaire, une variable \mathbf{Y} (variable à expliquer), aléatoire en fonction de p variables $(\mathbf{x}_1, \dots, \mathbf{x}_p)$.

Definition

Modèle de régression multiple : *Un modèle de régression linéaire général est défini par une équation de la forme :*

$$\mathbf{Y}_{n \times 1} = \mathbf{X}_{n \times p} \boldsymbol{\beta}_{p \times 1} + \boldsymbol{\varepsilon}_{n \times 1}$$

où

- \mathbf{Y} est un vecteur aléatoire de dimension n et \mathbf{X} est une matrice de taille $n \times p$
- $\boldsymbol{\beta}$ est le vecteur de dimension p des paramètres inconnus
- $\boldsymbol{\varepsilon}_{n \times 1}$ est le vecteur centré, de dimension n , des erreurs.

Estimation du modèle la régression linéaire multiple

Theorem

Soit \mathbf{X} une matrice de plein rang, l'estimateur des MC $\hat{\beta}$ de β vaut

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \quad (17)$$

l'estimateur $\hat{\beta}$ des MC est un estimateur sans biais de β est sa variance vaut $\mathbb{V}(\hat{\beta}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$.

Les **valeurs ajustées** (ou valeurs estimées) sont obtenues à partir de la formule suivante :

$$\hat{\mathbf{Y}} = \mathbf{X} \hat{\beta} \quad (18)$$

Régression non-linéaire

Supposons que nous ayons n observations (\mathbf{x}_i, y_i) , $i = 1, 2, \dots, n$, provenant d'un modèle de régression non linéaire avec **une relation fonctionnelle connue**.

$$y_i = f(\mathbf{x}_i; \theta^*) + \varepsilon_i, \quad (i = 1, 2, \dots, n), \quad (19)$$

où $\mathbb{E}(\varepsilon_i) = 0$, \mathbf{x}_i est un $k \times 1$ vecteur et la vraie valeur θ^* de θ avec $\theta^* \in \Theta \subset \mathbb{R}^p$. L'estimation par les moindres carrés de θ^* notée par $\hat{\theta}$ minimise la somme des carrés des erreurs :

$$S(\theta) = \sum_{i=1}^n [y_i - f(\mathbf{x}_i; \theta)]^2 \quad (20)$$

Régression non-linéaire

Lorsque chaque $f(\mathbf{x}_i; \theta)$ est différentiable par rapport à θ , et que $\hat{\theta}$ est à l'intérieur de \mathcal{D} , alors $\hat{\theta}$ satisfera

$$\left. \frac{\partial \mathcal{S}(\theta)}{\partial \theta_k} \right|_{\hat{\theta}} = 0 \quad (k = 1, 2, \dots, p) \quad (21)$$

L'estimation paramétrique dans ce cas est souvent complexe : pas de solution analytique.

- Sensibilité à l'initialisation
- Algorithme numérique approprié (cf. vos cours d'Analyse Numérique / Optimisation)
- Importance du conditionnement
- Estimation des variances (intervalles de confiance) plus compliqué

Régression non-linéaire

Exemple :

Considérons le modèle suivant :

$$y_i = \alpha x_i^\beta + \varepsilon_i, \quad (i = 1, 2, \dots, n), \quad (22)$$

Les équation normales son :

$$\begin{cases} \sum_{i=1}^n (y_i - \alpha x_i^\beta) \alpha x_i^\beta &= 0 \\ \sum_{i=1}^n (y_i - \alpha x_i^\beta) \alpha x_i^\beta \log x_i &= 0 \end{cases}$$

Ces équations n'admettent pas de solutions analytiques pour α et β .

Régression non-linéaire par linéarisation




Les modèles algébriques explicitement non linéaires peuvent parfois être transformés en des modèles explicitement linéaires. Par exemple on peut linéariser les fonctions

- $y = e^{\alpha + \beta x}$ en : $\log(y) = \alpha + \beta x$.

- $y = \alpha x^\beta$ en $\log(y) = \log(\alpha) + \beta \log(x)$.

Attention : Les transformations utilisées pour linéariser le problème modifient, par propagation, les variances des mesures utilisées dans le problème d'estimation.

References I

-  J Rawlings et al.
Applied Regression Analysis.
Springer-Verlag, 1998.
-  Xin Yan and Xin Yan.
Linear Regression Analysis.
World Scientific, 2009.
-  CRISP, IBM SPSS Modeler CRISP-DM Guide.
Cross-Industry Standard Process for Data Mining, 2011.
https://inseaddataanalytics.github.io/INSEADAnalytics/CRISP_DM.pdf



OUARET Rachid

rachid.ouaret@toulouse-inp.fr



Modèles de régression

Certificat Science des données
Big-Data