Modèle linéaire

- 30 heures de cours
- 10 séances de 3h : 8 en septembre/octobre et 2 en décembre
- 24 heures de TD

Svetlana Gribkova

Centre de Bio-informatique (Mines ParisTech, Inserm, Institut Curie)

Mes coordonnées :

email : svetlana.gribkova@mines-paristech.fr
page personnelle : http://cbio.ensmp.fr/~sgribkova/

Plan

Introduction

Modèle linéaire simple

Motivation

- Nous nous intéressons à un phénomène modélisé par une variable Y
- Nous pensons que les variations de Y peuvent être expliquées à l'aide d'un certain nombre de variables $X_1, \ldots, X_p$ dites explicatives.

Exemple : Y – nombre d'unités vendues d'un produit de consommation

X_1 – budget publicitaire radio
X_2 – budget publicitaire télé
X_3 – budget publicitaire journaux

Questions : en observant $(Y_i, X_{i1}, X_{i2}, X_{i3})_{1 \leq i \leq n}$*, on se demande :*

1. Quel est l'impact global de publicité sur les ventes ? Peut on expliquer la quantité de produit vendu par le budget publicitaire ?
2. Quel est la publicité la plus efficace ?
3. Peut-on prédire combien de produits on va vendre à partir de notre budget publicitaire ?

Modèle de régression

Pour répondre aux questions qu'on a posées, on voudrait avoir un modèle mathématique qui explique le lien entre Y et $X_1, \ldots, X_p$:

$$Y_i = f(X_{i1}, \ldots, X_{ip}) + \varepsilon_i, \quad i \in \{1, \ldots, n\}.$$

- Y_i – valeurs de variable à expliquer (dépendante, réponse)
- $X_{i1}, \ldots, X_{ip}$ – valeurs des variables explicatives (indépendantes, prédicteurs) qui seront déterministes dans ce cours
- ε_i – erreurs aléatoires, tiennent compte d'erreurs de mesure et de facteurs qui ne sont pas pris en compte par le modèle

La fonction de régression $f(x_1, \ldots, x_p)$ modélise le lien entre la réponse et les variables explicatives.

Lorsque f est une fonction paramétrique linéaire dans ces paramètres, on parle de la régression linéaire ou du modèle linéaire.

Régression linéaire

Pour l'exemple publicitaire, le modèle linéaire s'écrit comme :

$$ventes_i = \beta_0 + \beta_1 * \text{budget(télé)}_i + \beta_2 * \text{budget(radio)}_i + \beta_3 * \text{budget(journaux)}_i + \varepsilon_i$$

Plus généralement, on écrira

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \varepsilon_i, \quad i \in \{1, \ldots, n\}.$$

- $\beta_0, \ldots, \beta_p$ – paramètres déterministes inconnus
- ε_i – erreurs aléatoires
- $p = 1$: une seule variable explicative, régression linéaire simple
- $p > 1$: régression linéaire multiple

Ajuster le modèle : trouver des estimateurs de coefficients inconnus de sorte que le modèle explique au mieux les données. Pour ça, il faut définir un critère.

Fonction de coût

Les points $(Y_i, X_{i1}, \ldots, X_{ip})_{1 \leq i \leq n}$ étant donnés, le but est donc de trouver une fonction f qui minimise l'erreur du modèle

$$Y_i = f(X_{i1}, \ldots, X_{ip}) + \varepsilon_i, \quad i \in \{1, \ldots, n\}.$$

On définit le critère à minimiser l'aide d'une fonction de coût $L(u)$:

$$\arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} L(Y_i - f(X_{i1}, \ldots, X_{ip})),$$

Exemples :

- le coût absolue $L(u) = |u|$
- le coût quadratique $L(u) = u^2$

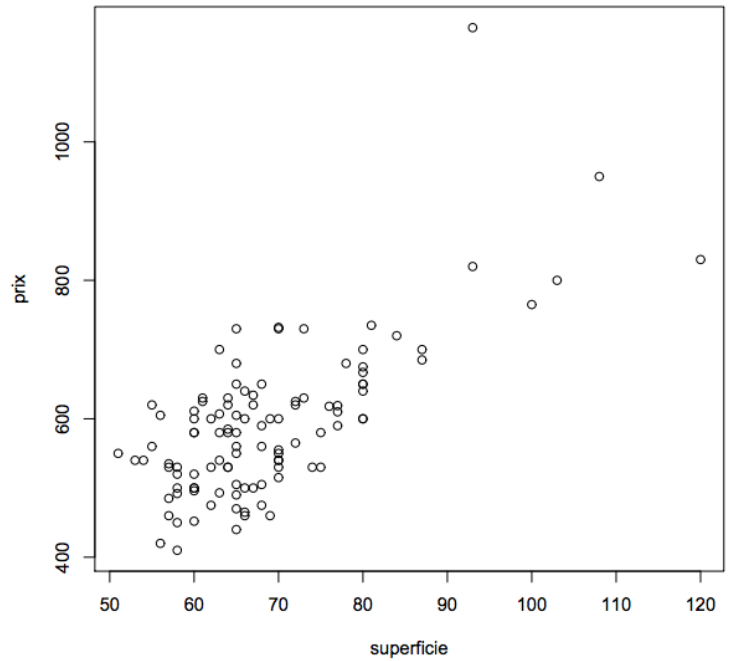
Dans le cadre du modèle linéaire avec la fonction de coût quadratique, on parle de l'estimation par moindres carrés.

Plan

Introduction

Modèle linéaire simple

Exemple : prix d'appartement et la superficie



On commence par représenter les données à l'aide d'un nuage de points (scatterplot en anglais).

D'après la visualisation, on peut envisager la modélisation par une droite :

$$prix_i = \beta_1 + \beta_2 * superficie_i + \varepsilon_i$$

Ceci est un exemple de modèle linéaire simple.

Modèle de régression linéaire simple : définition

La régression linéaire simple est définie par une équation

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad \forall i \in \{1, \dots, n\}$$

Les erreurs ε_i sont aléatoires et satisfont deux hypothèses :

$$\begin{cases} \mathcal{H}_1 : E[\varepsilon_i] = 0 & \text{pour} \quad i \in \{1, \dots, n\} \\ \mathcal{H}_2 : cov(\varepsilon_i, \varepsilon_j) = \sigma^2 \delta_{ij} & \text{pour} \quad i, j \in \{1, \dots, n\} \end{cases}$$

- $X = [X_1, \dots, X_n]^T$ est un vecteur déterministe
- β_0 et β_1 sont les paramètres inconnus (pas aléatoires) du modèle
- $Y = [Y_1, \dots, Y_n]^T$ est donc un vecteur aléatoire avec

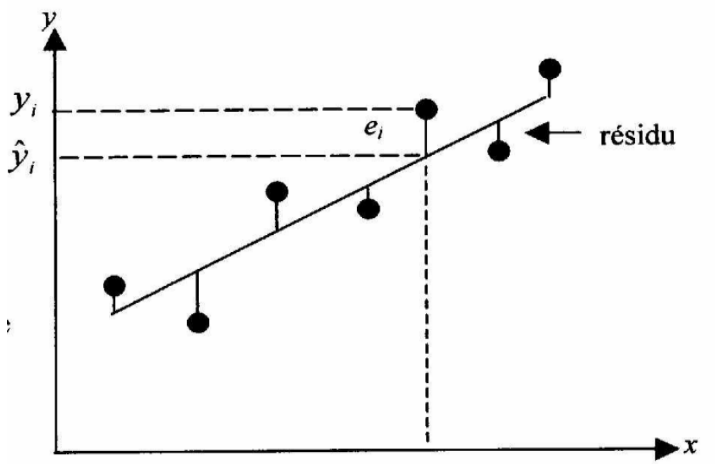
$$E[Y_i | X_i] = \beta_0 + \beta_1 X_i, \qquad i \in \{1, \dots, n\}$$

Estimateurs des Moindres Carrés Ordinaires

On définit les estimateurs des Moindres Carrés Ordinaires $\hat{\beta}_0, \hat{\beta}_1$ comme les valeurs qui minimisent

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 = \sum_{i=1}^{n} \varepsilon_i^2$$

$S(\beta_0, \beta_1)$ représente la somme des carrés des écarts entre les valeurs observées Y_i et les valeurs $\beta_0 + \beta_1 X_i$ ajustées, sur la droite $y = \beta_0 + \beta_1 x$.



Calcul des estimateurs des MCO

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg \min_{\beta_0, \beta_1 \in \mathbb{R}} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Théorème (expressions des estimateurs des moindres carrés)

Les estimateurs des moindres carrés ordinaires sont donnés par

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\ \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \end{aligned}$$

Proposition (une autre expression pour $\hat{\beta}_1$)

$\hat{\beta}_1$ *peut être écrit sous une autre forme, qui nous sera utile par la suite :*

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(X_i - \bar{X})\varepsilon_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Propriétés des estimateurs des MCO

Théorème (les estimateurs des MCO sont sans biais)

Sous les hypothèses $\mathcal{H}_1$ et $\mathcal{H}_2$, on a

$$E[\hat{\beta}_0] = \beta_0 \qquad et \qquad E[\hat{\beta}_1] = \beta_1$$

Théorème (variance et covariance des estimateurs des MCO)

Sous les hypothèses $\mathcal{H}_1$ et $\mathcal{H}_2$, les variances des estimateurs sont

$$\begin{aligned} var(\hat{\beta}_0) &= \frac{\sigma^2 \sum X_i^2}{n \sum (X_i - \bar{X})^2} = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2} \right) \\ var(\hat{\beta}_1) &= \frac{\sigma^2}{\sum (X_i - \bar{X})^2} \end{aligned}$$

et leur covariance est donnée par

$$cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2 \bar{X}}{\sum (X_i - \bar{X})^2}$$