UE

Modélisation

Topic 3 : MLG

Maintenant commence la partie « statistique » de l'UE. Plus précisément, le but est de vous présenter le modèle linéaire généralisé (MLG).

Uniquement pour fixer les notations et servir de rappel :

$\boxed{0}$. Rappel : Le modèle linéaire simple

① Préambule : La démarche statistique.

$\Omega / \mathbb{P}$ — Description du système étudié / Modélisation de la population étudiée

↓

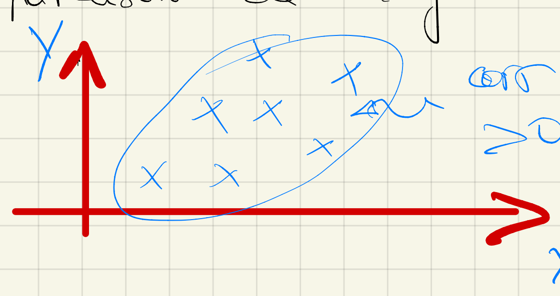
Analyse univariée — Histogrammes / Moy / Écart-type / Médiane.

↓

Analyse bivariée (X, Y) — Corrélation empirique. / Comparaison de moy

↓

Analyse multi-variée



a pour but de répondre à la question :

« Comment expliquer une variable Y en fonction de plusieurs variables explicatives $X^1, X^2, \dots, X^k$??? »

Dans ce cours, on verra

↳ le modèle linéaire

$$Y = \sum_{j=1}^{k} \underline{\theta_j} \, \underline{X^j} + \underline{\varepsilon}$$

θ_j : paramètres du modèle ; X^j : variables explicatives.

ε ↳ Bruit / Erreur

$E(\varepsilon) = 0$

$Var(\varepsilon) = \sigma^2$

Écriture matricielle $Y = X\theta + \varepsilon$

avec $X^1 = 1$ (intercept).

Remarque : On peut allègrement modifier le modèle pour inclure des non-linéarités

$$Y = \sum_{j=1}^{k} \underline{f_j}(X^j, \theta) + \varepsilon$$

$f_j(X^j; \theta) \neq \theta^j X^j$

f_j ↳ fonctions à estimer (Ex polynômes)

$\left(\text{Ex} : f_j(x, \theta) = \sum_{k=0}^{d_j} x^k \theta_{j,k} \right)$

Regression paramétrique

ou pire encore :

$$Y = \underline{F_\theta}(X^1, X^2, X^3, \dots, X^k) + \varepsilon$$

↳ Fonction à estimer

par réseau de neurone

ou plus précisément par descente de gradient

Sondage : Faire un TP de régression paramétrique ?

Oui / Non.

Sondage : Faire un TP avec réseaux de neurones.

Oui / Non.

(2) Estimation MCO (Moindres carrés ordinaires) / OLS (Ordinary Least Squares).

$$\begin{cases} Y_1 = X_1 \Theta + \varepsilon_1 \\ Y_2 = X_2 \Theta + \varepsilon_2 \\ \vdots \\ Y_n = X_n \Theta + \varepsilon_n \end{cases} \qquad \varepsilon_i \text{ iid} \qquad \Longleftrightarrow \qquad \underset{n}{\mathbb{Y}} = \underset{n \times p}{\mathbb{X}} \, \underset{p}{\Theta} + \underset{n}{\varepsilon}$$

L'estimateur MCO / OLS consiste à prendre

$$\hat{\Theta}_n^{OLS} = \underset{\Theta}{\operatorname{Argmin}} \underbrace{\| \mathbb{Y} - \mathbb{X}\Theta \|^2}_{\text{Loss / Perte quadratique}}$$

Proposition : $\hat{\Theta}_n^{OLS} = ({}^t\mathbb{X}\,\mathbb{X})^{-1}\, {}^t\mathbb{X}\,\mathbb{Y}$

⚠ $\hat{\Theta}_n^{OLS} = \hat{\Theta}_n^{MLE}$ ← Estimateur de max de vraisemblance

$$= \underset{\Theta}{\operatorname{Argmax}} \log L(\Theta)$$

lorsque $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ iid Gaussiens.

En effet $L(\theta) =$ Densité d'observer

$(y_1, \dots, y_n) = \mathbb{Y}$

$(x_1, \dots, x_n) = \mathbb{X}$.

$=$ Densité d'observer :

iid $\leadsto \varepsilon = \mathbb{Y} - \mathbb{X}\theta = \begin{pmatrix} y_1 - x_1\theta \\ \vdots \\ y_n - x_n\theta \end{pmatrix}$

$$= \prod_{j=1}^{n} \mathbb{P}_{\varepsilon_j}(y_j - x_j\theta)$$

$$= \prod_{j=1}^{n} \exp\left(-\frac{\|y_j - x_j\theta\|^2}{2\sigma^2}\right) / \sqrt{2\pi\sigma^2}$$

$$= \exp\left(-\sum_{j=1}^{n} \frac{\|y_j - x_j\theta\|^2}{2\sigma^2}\right) / (2\pi\sigma^2)^{n/2}$$

$$\leadsto \log L(\theta) = -\sum_{j=1}^{n} \frac{\|y_j - x_j\theta\|^2}{2\sigma^2} - \frac{n}{2}\log(2\pi\sigma^2)$$

Ainsi $\hat{\theta}_n^{MLE} = \underset{\theta}{\mathrm{Argmax}} \log L(\theta)$

$$= \underset{\theta}{\mathrm{Argmax}} - \sum_{j=1}^{n} \frac{\|y_j - x_j\theta\|^2}{\cancel{2\sigma^2}} \cancel{- \frac{n}{2}\log(2\pi\sigma^2)}$$

$$= \underset{\theta}{\mathrm{Argmax}} - \|\mathbb{Y} - \mathbb{X}\theta\|^2$$

$$= \underset{\theta}{\mathrm{Argmin}} \|\mathbb{Y} - \mathbb{X}\theta\|^2$$

$$= \hat{\theta}_n^{OLS} \; !$$

$\square$

Propriétés de $\hat{\Theta}_n^{OLS}$:

- $\hat{\Theta}_n^{OLS} = ({}^t\mathbb{X}\mathbb{X})^{-1}\,{}^t\mathbb{X}\,\mathbb{Y}$
  $= \Theta + ({}^t\mathbb{X}\mathbb{X})^{-1}\,{}^t\mathbb{X}\,\mathcal{E}$
- $\hat{\Theta}_n^{OLS}$ est consistant $\left(\hat{\Theta}_n^{OLS} \xrightarrow{\mathbb{P}} \Theta\right)$

et sous hypothèse $\mathcal{E}$ Gaussien

$$\hat{\Theta}_n \overset{\mathcal{L}}{=} \mathcal{N}\left(\Theta, \sigma^2 ({}^t\mathbb{X}\mathbb{X})^{-1}\right)$$

- $\hat{\Theta}_n^{OLS}$ sans biais.

③ Résidus et valeurs ajustés.

Après estimation de $\hat{\Theta}_n$ on peut former :

↳ « les valeurs ajustées » / les prédictions

$$\hat{\mathbb{Y}} = \mathbb{X}\,\hat{\Theta}_n \underset{\text{si } \mathcal{E} \text{ Gaussien}}{\overset{\mathcal{L}}{=}} \mathcal{N}\left(\mathbb{X}\Theta, \mathbb{X}({}^t\mathbb{X}\mathbb{X})^{-1}\,{}^t\mathbb{X}\right).$$

↳ « les résidus » / les erreurs de prédiction.

$$\hat{\mathcal{E}} = \mathbb{Y} - \hat{\mathbb{Y}}.$$

Propriétés : Si $\mathcal{E}$ Gaussien

$\hat{\mathbb{Y}} \perp\!\!\!\perp \hat{\mathcal{E}}$

$\hat{\mathcal{E}} \perp\!\!\!\perp \hat{\Theta}_n$

④ Estimation de σ^2.

Le bon estimateur de σ^2 est

$$\hat{\sigma}_n^2 = \frac{1}{n-k} \|\hat{\varepsilon}\|^2 = \frac{1}{n-k} \sum_{j=1}^{n} \hat{\varepsilon}_j^2$$

k nombre de facteurs

(l'intercept $X^1 = \mathbb{1}$ inclus !)

Sous hypothèse de ε Gaussien :

Propriétés :

(1) $\hat{\mathbb{Y}} \perp\!\!\!\perp \hat{\sigma}_n^2$ & $\hat{\sigma}_n^2 \perp\!\!\!\perp \hat{\theta}_n$

(2) $\dfrac{(n-k)\,\hat{\sigma}_n^2}{\sigma^2} \overset{\mathcal{L}}{=} \chi^2_{n-k}$ (Somme de $n-k$ carrés de normales)

(3) $\hat{\sigma}_n^2$ sans biais et de variance $\dfrac{2\sigma^4}{n-k} \xrightarrow[n \to +\infty]{} 0$

Exercice : Montrer que (2) $\Longrightarrow$ (3)

et (3) $\Longrightarrow$ $\hat{\sigma}_n^2$ est un estimateur consistant.

$$\left(\hat{\sigma}_n^2 \xrightarrow[n \to +\infty]{\mathbb{P}} \sigma^2 \right).$$

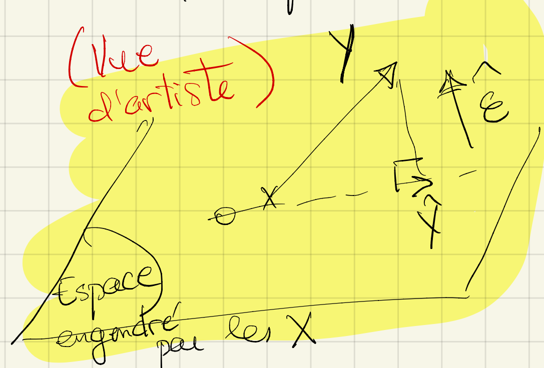
(5) Analyse et exploitation de la regression.

↳ ANOVA (Analyse de la variance).

Il est question de savoir si la regression est pertinente. Pour cela on observe la décomposition

$$\mathbb{Y} = \hat{\mathbb{Y}} + \hat{\varepsilon}$$

En fait c'est une décomposition



$\rightsquigarrow$ en variables indépendantes (Propriétés en ⑤)

donc

$$\mathrm{Var}\ Y = \mathrm{Var}\ \hat{Y} + \mathrm{Var}\ \hat{\varepsilon}$$

$\rightsquigarrow$ en vecteurs orthogonaux après centrage (indépendance géométrique)

donc

$$\| Y - \bar{Y} \|^2 = \| \hat{Y} - \bar{\hat{Y}} \|^2 + \| \hat{\varepsilon} \|^2$$

Variance expliquée

variance résiduelle

La qualité de la regression est mesurée

par le critère $R^2 = \dfrac{\| \hat{Y} - \bar{Y} \|^2}{\| Y - \bar{Y} \|^2} \in [0, 1]$

$\begin{cases} \text{si } R^2 \simeq 1, & \text{alors } X \text{ explique bien } Y. \\ \text{Si } R^2 \simeq 0, & \text{l'inverse} \end{cases}$

⑥ Tests de nullité :

$(H_0 : \theta_j = 0)$

On a : $\hat{\theta}_j \overset{\mathcal{L}}{=} \mathcal{N}(\theta_j, \sigma^2 ({}^t X X)^{-1})$

mais σ^2 inconnu !

Ainsi

$$\begin{cases} \dfrac{\hat{\theta}_j - \theta_j}{\sqrt{\sigma^2 ({}^tXX)^{-1}_{jj}}} \overset{\mathcal{L}}{=} \mathcal{N}(0, 1) \\ (n-k) \dfrac{\hat{\sigma}_n^2}{\sigma^2} \overset{\mathcal{L}}{=} \chi^2_{n-k} \end{cases}$$

$\perp\!\!\!\perp$ sous hypothèse de ε Gaussien

Par ratio, $T = \dfrac{\hat{\theta}_j - \theta_j}{\sqrt{\hat{\sigma}_n^2 \, ({}^tX X)^{-1}_{jj}}} = \dfrac{\mathcal{N}(0,1)}{\sqrt{\dfrac{\chi^2_{n-k}}{n-k}}}$ $\perp\!\!\!\perp$

$\overset{\mathcal{L}}{=}$ Student $(n-k)$

↳ définition de cette loi

Si $t_{1-\frac{\alpha}{2}}$ est le $(1-\frac{\alpha}{2})^e$ quantile de cette loi

Alors $IC_{1-\alpha}(\theta_j) = \hat{\theta}_j + \sqrt{\hat{\sigma}_n^2 ({}^tXX)^{-1}_{jj}} \times [-t_{1-\frac{\alpha}{2}} ; t_{1-\frac{\alpha}{2}}]$

est l'intervalle de confiance avec certitude α.

$\leadsto$ On rejette H_0 si $|\hat{\theta}_j| \geq t_{1-\frac{\alpha}{2}} \sqrt{\hat{\sigma}_n^2 ({}^tXX)^{-1}_{jj}}$

Le modèle linéaire généralisé

I - Introduction

Il existe plusieurs variantes du modèle linéaire (polynômes, paramétrique, non-param) :

$$\underbrace{Y}_{\text{Variable d'intérêt}} = \underbrace{X}_{\text{Facteurs}} \underbrace{\beta}_{\text{A estimer}} + \underbrace{\varepsilon}_{\text{Bruit}}$$

Toutefois, il y a des cas très simples où ceci n'a aucune chance de fonctionner.

Par exemple, si la variable d'intérêt est

- binaire ($\in \{0,1\}$) et que l'on s'intéresse à la probabilité

$$p = \mathbb{P}(Y = 1)$$

- le comptage ($\in \mathbb{N}$).

Ainsi nous verrons dans ce chapitre une généralisation dans deux directions

- Utilisation de lois plus générales

dans le cadre des lois de la famille exponentielle.
En effet, le modèle linéaire et son étude sont très « gaussiens ».

↳ Incorporation d'une transformation pour passer de $\mathbb{R}$ à $[0,1]$; $\mathbb{N}$; $\{0,1\}$

Le modèle linéaire généralisé se décompose en 3 composantes.

① La composante aléatoire Y (≡ Variable d'intérêt / de réponse) à laquelle est associée une loi de proba.

② La composante déterministe composée d'une combinaison linéaire de $X^1, X^2, \ldots, X^p$ (variables explicatives) ⇝ $X\beta$ où $X = (X^1, \ldots, X^p)$

③ Un lien c'est-à-dire une transformation entre $\begin{cases} \text{la composante déterministe} \\ \text{un paramètre de la composante aléatoire.} \end{cases}$

Ex: $Y \overset{\mathcal{L}}{=} \mathcal{N}(\underset{\underset{X\beta}{\parallel}}{\mu}, 1)$ ⟵ Lien

II - Les 3 composantes

II - 1 - La composante déterministe

variables fixées par l'exp.

C'est juste la combinaison lin. $X\beta = \sum \beta_i X^i$

Autre nom : Prédicteur linéaire.

I-2- La composante aléatoire.

On dispose de n observations $(y_1, y_2, \dots, y_n)$ de la v.a Y. On choisit une loi dans la famille exponentielle en fonction de la nature de Y.

↳ $Y \in \mathbb{R}$, valeur quantitative.

Ex naturel ⇝ $Y \overset{\mathcal{L}}{=} \mathcal{N}(\mu, \sigma^2)$, et $\mu = \mathbb{E}(Y) \in \mathbb{R}$

↳ $Y \in \{0,1\}$ (Succès / Echec).

Ex naturel ⇝ $Y \overset{\mathcal{L}}{=} \text{Bernoulli}(\pi)$ où $\pi = \mathbb{E}(Y) \in [0,1]$

↳ $Y \in \mathbb{N}$ (Comptage)

Exemple naturel ⇝ $Y \overset{\mathcal{L}}{=} \text{Poisson}(\lambda)$

où $\lambda = \mathbb{E}(Y) \in \mathbb{R}_+$

II-3- La fonction de lien

Définition : On définit $g : I$ intervalle $\to \mathbb{R}$ donnant $g(\mathbb{E}(Y)) = X\beta$ où $\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$

$X = (X^1, X^2, \dots, X^k)$, $X^1 = 1$

Dit simplement

Partie déterministe $X\beta$ $\xleftrightarrow[\text{fonction de lien}]{g}$ $\mathbb{E}(Y)$ $\xleftarrow{\mathbb{E}}$ Partie aléatoire Y

Exemples : $\mapsto$ Si $Y \overset{\mathcal{L}}{=} \mathcal{N}(\mu, \sigma^2) = \mu + \sigma \mathcal{N}(0,1)$

Alors $g(\mu) = \mu$.

$$\Longrightarrow \quad Y = \overset{\mu}{\overset{\shortparallel}{X\beta}} + \underbrace{\sigma \mathcal{N}(0,1)}_{\varepsilon}$$

C'est le modèle linéaire.

$\mapsto$ Si $Y \overset{\mathcal{L}}{=} \mathcal{P}(\lambda)$

Alors $X\beta = g(\mathbb{E}(Y)) = g(\lambda) \overset{\text{choix}}{=} \log \lambda$

donne $Y \overset{\mathcal{L}}{=} \mathcal{P}(\lambda = e^{X\beta})$

(Modèle log-linéaire).

$\mapsto$ Si $Y \overset{\mathcal{L}}{=} \text{Bernoulli}(\pi)$

Alors $X\beta = g(\mathbb{E}(Y)) = g(\pi) \overset{\text{choix}}{=} \log\left(\dfrac{\pi}{1-\pi}\right)$

$= \text{logit}(\pi)$

donne

$$Y \overset{\mathcal{L}}{=} \text{Bernoulli}\left(\text{logit}^{\langle -1 \rangle}(X\beta)\right)$$

(Modèle de régression logistique).

⚠ $\text{logit}^{\langle -1 \rangle}(y) = \dfrac{1}{1+e^{-y}} = \dfrac{e^y}{1+e^y}$ (Fonction sigmoïde)

$$y = \text{logit}(x) \iff e^y = \frac{x}{1-x} \iff e^{-y} = \frac{1-x}{x}$$

$$\iff 1 + e^{-y} = \frac{1}{x}$$

$$\iff x = \frac{1}{1+e^{-y}}$$

III – Familles exponentielles.

Large famille qui contient

↳ Bernoulli, Binômiale, Poisson (1 param).

↳ Normale, Gamma, Binômiale négative (2 param)

↳ Multinômiale (à + de 2 param).

La densité utilise deux paramètres $\psi = \begin{pmatrix} \theta \\ \phi \end{pmatrix}$

où $\begin{cases} \theta : \text{Paramètre principal lié à } \mathbb{E}(Y) \\ \phi : \text{Paramètre de nuisance lié à } Var(Y) \end{cases}$

(Prendre 0 si loi à 1 param)

Y appartient à la famille exponentielle avec param (θ, ϕ) lorsque

$$\mathbb{P}(Y \in dy) = dy \, f(y; \theta, \phi) \quad \text{(Continue)}$$

ou

$$\mathbb{P}(Y = y) = f(y; \theta, \phi) \quad \text{(discret)}$$

avec

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

Les expressions exactes de $c(y, \phi)$ & $a(\phi)$ ne sont pas importantes et sont identifiées au cas par cas. On a tout de même les propriétés suivantes, que l'on peut établir en toute généralité :

Thm:
$$\begin{cases} b'(\theta) = \mathbb{E}(Y) \iff \theta = (b')^{<-1>}(\mathbb{E}(Y)) \\ Var(Y) = a(\phi)\, b''(\theta) \end{cases}$$

Si un seul paramètre au modèle, on prend $a(\phi) = 1$

Lien avec MLG : Dans toute famille exponentielle utilisée pour MLG

$Y \overset{\mathcal{L}}{=}$ Famille exp (θ, ϕ)

$$X\beta = g(\mathbb{E}(Y)) \quad \text{par def de MLG}$$
$$= g(b'(\theta))$$

Poser $\theta = X\beta$ comme paramètre de la loi

donne $\theta = g \circ b'(\theta) \iff g = (b')^{<-1>}$

Preuve : Montrons que $\mathbb{E}(Y) = b'(\theta)$ en toute généralité.

Preuve amusante de statisticien :

Je vais calculer l'estimateur de max de vraisemblance $\hat{\theta}_n$ pour un tirage iid. Par consistance $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$

Calcul: $$L(y; \theta, \phi) = \prod_{i=1}^{n} \exp\left(\frac{\theta y_i - b(\theta)}{a(\phi)} + c(y_i, \phi) \right)$$

où $y = (y_1, \ldots, y_n)$ échantillon observé.

Donc
$$\hat{\theta}_n = \underset{\theta}{\text{Argmax}} \ \log L(y_i, \theta, \phi)$$
$$= \underset{\theta}{\text{Argmax}} \ \sum_{i=1}^{n} \left(\frac{\theta y_i - b(\theta)}{\cancel{a(\phi)}} \cancel{+ c(y_i, \phi)} \right)$$
$$= \underset{\theta}{\text{Argmax}} \ \theta \left(\sum_{i=1}^{n} y_i \right) - n\, b(\theta)$$
$$= \underset{\theta}{\text{Argmax}} \ \theta \left(\frac{\sum_{i=1}^{n} y_i}{n} \right) - b(\theta)$$

d'où
$$\frac{\partial}{\partial \theta} = 0 \iff \frac{\sum_{i=1}^{n} y_i}{n} = b'(\theta)$$
$$\iff \theta = (b')^{\langle -1 \rangle} \left(\frac{\sum_{i=1}^{n} y_i}{n} \right)$$

Donc
$$\hat{\theta}_n = (b')^{\langle -1 \rangle} \left(\frac{\sum_{i=1}^{n} y_i}{n} \right) \xrightarrow[n \to +\infty]{} \theta = (b')^{\langle -1 \rangle}(\mathbb{E}(Y))$$

$\boxed{//}$

⚠ Ce résultat est difficile à montrer en calculant $\mathbb{E}(Y)$ à partir de la densité !