Statistiques (STA1)

Cours V – Le modèle linéaire

Luca Ganassali

Laboratoire de Mathématiques d'Orsay, Université Paris-Saclay

Jeudi 23 octobre 2025

Cadre général de l'apprentissage supervisé

Objectif : expliquer ou prédire une variable aléatoire réponse $Y \in \mathbb{R}$ en fonction de variables explicatives $X_1, \ldots, X_p \in \mathbb{R}$.

Modélisation : par exemple, modèle à bruit additif :

$$Y = f(X_1, \ldots, X_p) + \varepsilon, \quad \text{avec } \mathbb{E}[\varepsilon] = 0.$$

f est appelée fonction de régression du modèle et est inconnue.

Exemple : $Y =$ label de l'image (chat/chien), $X =$ tous les pixels de l'image. (ex: $28 \times 28 \times 3$ couleurs $\longrightarrow p = 2352$).

Remarques :

- Le choix de la fonction de régression et des variables explicatives repose sur la connaissance du phénomène (physique, biologique, empirique) et doit toujours être critiqué.
- Y est aléatoire à cause du bruit ε, mais les X sont considérées ici comme fixées.

Cadre général de l'apprentissage supervisé

Objectifs de l'apprentissage supervisé :

(i) Estimer f à partir de données observées $(x_{i,1}, \ldots, x_{i,p}, y_i)_{1 \leq i \leq n}$ (expliquer Y : train).

(ii) Prédire Y pour de nouvelles valeurs de $x_1, \ldots, x_p$ jamais vues, avec une erreur minimale (prédire Y : test).

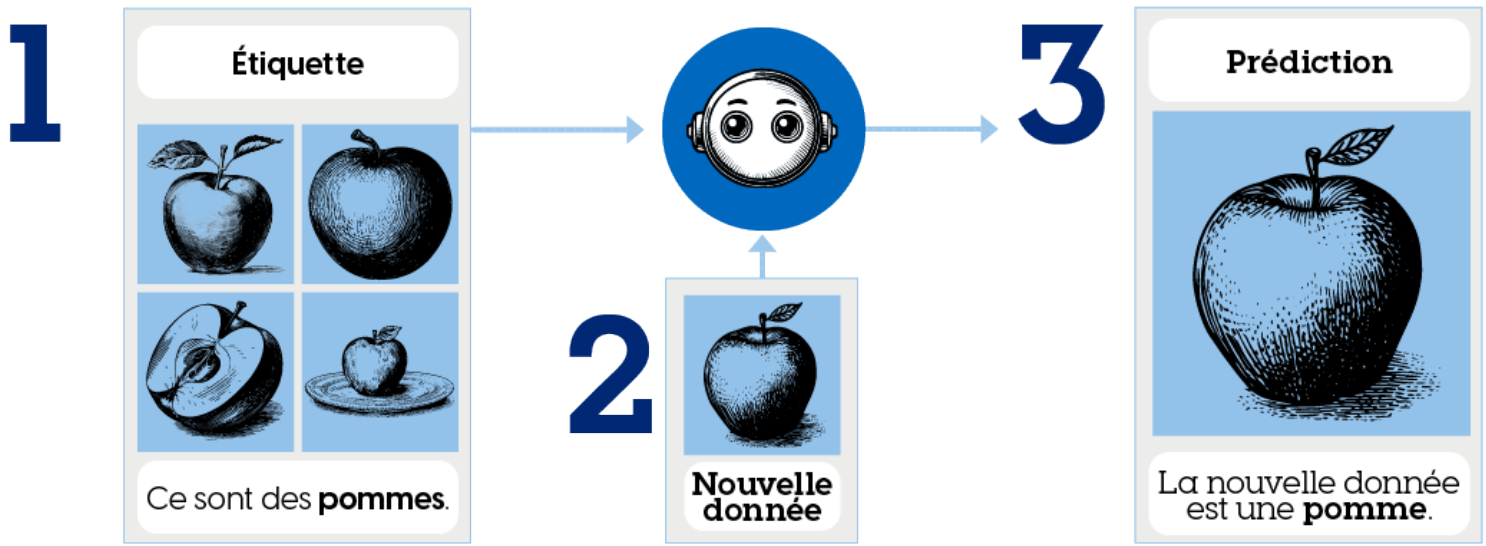


Illustration de l'apprentissage supervisé. $(i) \leftrightarrow 1$, $(ii) \leftrightarrow 2 \,\&\, 3$

Définition des modèles linéaire et linéaire gaussien

Modèle linéaire : définition

Le modèle linéaire est un modèle pour chaque $Y_i \in \mathbb{R}$ étant donné des covariables $X_i = (X_{i,1}, \ldots, X_{i,p}) \in \mathbb{R}^p$. $Y_i \in \mathbb{R}$ est modélisée de la façon suivante :

$$Y_i = \theta_1 X_{i,1} + \ldots + \theta_p X_{i,p} + \varepsilon_i = X_i^T \theta + \varepsilon_i,$$

où les ε_i sont des v.a. appelées résidus ou bruits qui satisfont :

(*i*) $\mathbb{E}[\varepsilon_i] = 0$ (bruits centrés),

(*ii*) $\mathrm{Var}(\varepsilon_i) = \sigma^2$ (bruits de variance constante),

(*iii*) pour tous $i \neq j$, $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ (bruits décorrelés).

Les paramètres (inconnus) de ce modèle sont $\theta \in \mathbb{R}^p$ et la variance $\sigma^2 > 0$.

La fonction $f_\theta : x \in \mathbb{R}^p \longmapsto x^T\theta \in \mathbb{R}$ est la fonction de régression du modèle.

Ici, l'aléa dans Y_i provient uniquement de ε_i et non pas de X_i : seul ε_i est aléatoire.

Modèle linéaire : définition

Matriciellement, le modèle linéaire s'écrit :

$$\underbrace{Y}_{\in \mathbb{R}^{n\times 1}} = \underbrace{X}_{\in \mathbb{R}^{n\times p}} \cdot \underbrace{\theta}_{\in \mathbb{R}^{p\times 1}} + \underbrace{\varepsilon}_{\in \mathbb{R}^{n\times 1}} .$$

Modèle linéaire $\leftrightarrow$ Y est linéaire en X, à bruit additif près.

X = matrice du plan d'expérience (ou matrice de design).Souvent $X_{i,1} = 1$ pour tout i (X a une première colonne avec des 1). Dans ce cas, θ_1 = intercept : relation affine et non juste linéaire.

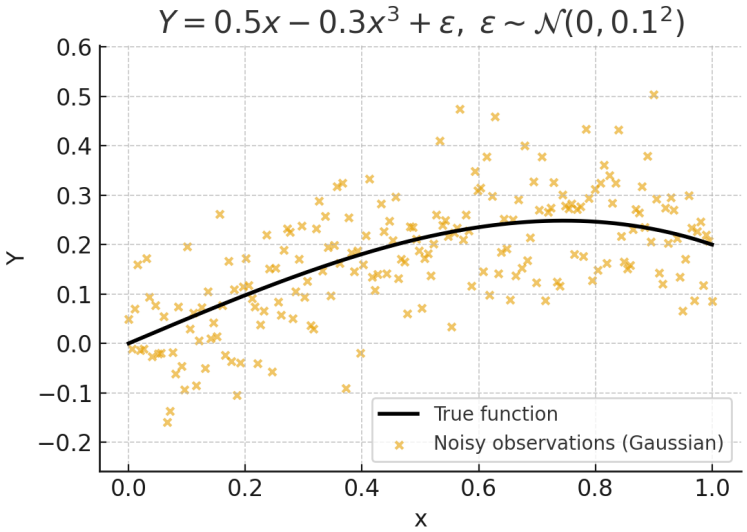
 Cela ne veut pas dire qu'on ne peut pas représenter une dépendance non linéaire de Y en une variable z : il suffit d'ajouter des termes comme z^2, z^3 dans X (régression polynomiale). La régression linéaire, ce n'est pas que des droites.

Modèle linéaire : exemple

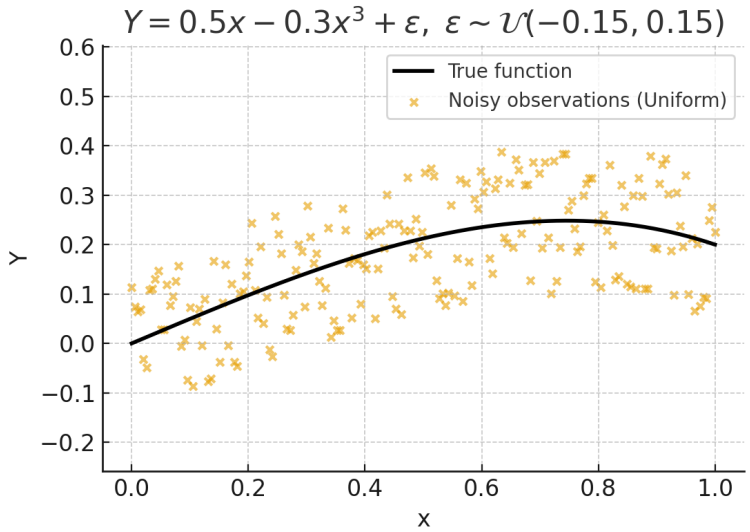
Exemple : $Y_i = \alpha + \beta x_i + \gamma x_i^3 + \varepsilon_i$, on a

$$X = \begin{pmatrix} 1 & x_1 & x_1^3 \\ 1 & x_2 & x_2^3 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^3 \end{pmatrix}, \quad \theta = \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

Ici, α = intercept.



(a) bruit gaussien: $\epsilon_i \sim \mathcal{N}(0, 0.1^2)$



(b) bruit uniforme: $\epsilon_i \sim \mathcal{U}(-0.15, 0.15)$

Modèle linéaire gaussien : définition

Le modèle linéaire gaussien est un modèle linéaire où les bruits sont de plus supposés gaussiens :

$$Y = X\theta + \varepsilon,$$

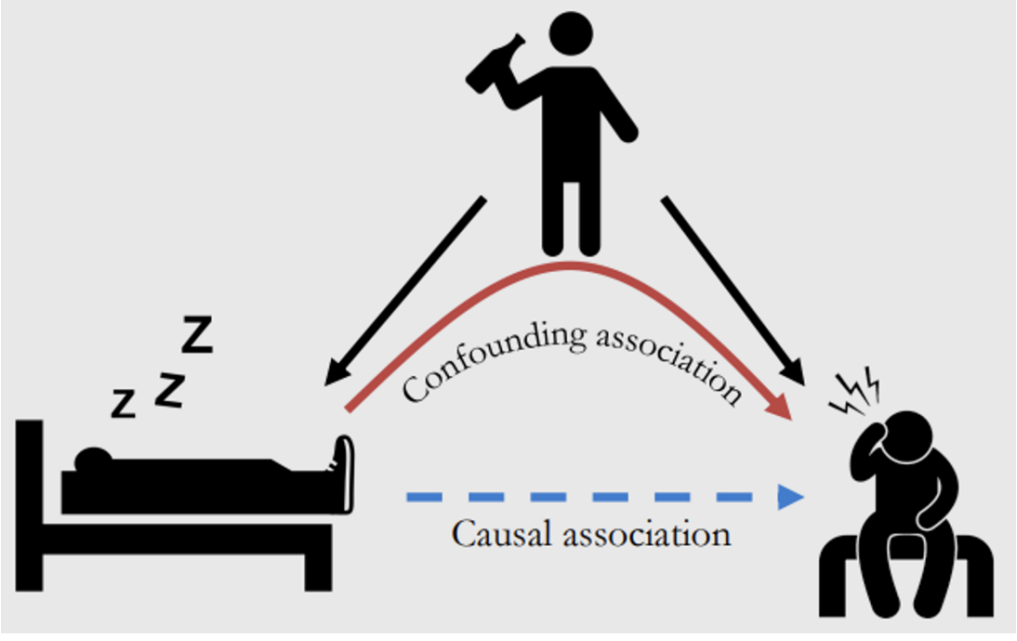
avec $\varepsilon \sim \mathcal{N}(0_n, \sigma^2 I_n)$. Dans ce cas, Y est donc un vecteur gaussien de loi

$$Y \sim \mathcal{N}(X\theta, \sigma^2 I_n)\,.$$

Modèle linéaire : ce qu'il (ne) permet (pas) de faire

Permet de détecter et quantifier via θ une corrélation entre une variable explicative Z et la réponse Y ✔

Ne permet pas (sans autres hypothèses) de mettre en lumière un lien de cause à effet entre une variable explicative Z et la réponse Y ✗



(source : `https://www.bradyneal.com/causal-inference-course`)

Modèle linéaire : identifiabilité

Rappel : un modèle $\mathcal{M} = (\mathcal{X}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ est identifiable si deux paramètres θ, θ' différents définissent deux lois $\mathbb{P}_\theta, \mathbb{P}_{\theta'}$ différentes, i.e. si l'application $\theta \in \Theta \longmapsto \mathbb{P}_\theta$ est injective.

Proposition (Identifiabilité du modèle linéaire). Les propositions suivantes sont équivalentes :

(*i*) Le modèle linéaire $Y = X\theta + \varepsilon$ est identifiable ;

(*ii*) Les p colonnes de X forment une famille libre de $\mathbb{R}^n$ (on dit que X est de *rang plein*) ;

(*iii*) $\mathrm{Ker}(X) = \{0_p\}$ (i.e. X est injective) ;

(*iv*) La matrice $X^T X \in \mathbb{R}^{p \times p}$ est inversible.

En particulier, $p \leq n$ est nécessaire pour l'identifiabilité du modèle linéaire.

Modèle linéaire : identifiabilité, preuve

$(ii) \iff (iii)$: c'est le théorème du rang.

$(iii) \iff (iv)$: vient de ce que $\mathrm{Ker}(X^T X) = \mathrm{Ker}(X)$.

Le seul point restant à prouver est $(i) \iff (iii)$ (noyau nul $\iff$ identifiabilité).

Supposons $\mathrm{Ker}(X) = \{0_p\}$. Soient $\theta_1, \theta_2 \in \mathbb{R}^p$ et $\varepsilon_1, \varepsilon_2$ des bruits centrés tels que, en loi, $Y = X\theta_1 + \varepsilon_1 = X\theta_2 + \varepsilon_2$. On passe a l'espérance : $X\theta_1 = X\theta_2$. On a $\theta_1 - \theta_2 \in \mathrm{Ker}(X)$ donc $\theta_1 = \theta_2$.

Prouvons la contraposée de réciproque. Supposons $\mathrm{Ker}(X) \neq \{0_p\}$ et prenons $\eta \in \mathrm{Ker}(X) \setminus \{0_p\}$. Alors $Y = X\theta + \varepsilon = X(\theta + \eta) + \varepsilon$ et pourtant $\theta \neq \theta + \eta$. Le modèle n'est pas identifiable. □

Régression linéaire : l'estimateur des moindres carrés

Estimateur des moindres carrés : définition et forme close

Dans le modèle linéaire, on observe Y et cherche à estimer θ. Pour tout $\theta \in \mathbb{R}^p$, on définit la somme des carrés résiduels :

$$\mathrm{SCR}(\theta) := \|Y - X\theta\|^2.$$

L'estimateur des moindres carrés est défini par

$$\hat{\theta}_{MC} \in \arg\min_{\theta \in \mathbb{R}^p} \mathrm{SCR}(\theta) = \arg\min_{\theta \in \mathbb{R}^p} \|Y - X\theta\|^2.$$

Intuition : on cherche la valeur de θ qui rend les résidus $Y - X\theta$ aussi petits que possible en norme euclidienne.

Proposition (Forme close de l'EMC). Dans un modèle linéaire identifiable, on a :

$$\hat{\theta}_{MC} = (X^T X)^{-1} X^T Y\,.$$

Estimateur des moindres carrés : forme close et interprétation géométrique

Interprétation géométrique : on a

$$X\hat{\theta}_{MC} = \underbrace{X(X^TX)^{-1}X^T}_{\Pi_{\mathrm{Im}(X)}} Y\,.$$

Le vecteur $\hat{Y} = X\hat{\theta}_{MC}$ est la projection orthogonale de Y sur l'espace image de X.

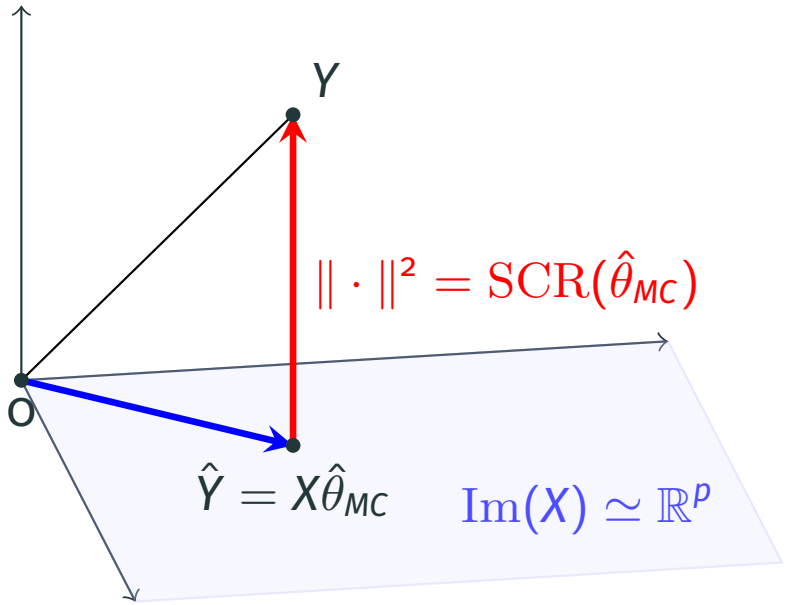


Illustration de la régression linéaire : les vecteurs $X\hat{\theta}_{MC}$ (bleu) et $Y - X\hat{\theta}_{MC}$ (rouge) sont orthogonaux.

Estimateur des moindres carrés : preuve de la forme close

On prend le gradient de $\theta \mapsto \mathrm{SCR}(\theta) = \|Y - X\theta\|^2 = (Y - X\theta)^T(Y - X\theta)$:

$$\begin{aligned}\nabla_\theta \mathrm{SCR}(\theta) &= \nabla_\theta (Y - X\theta)^T (Y - X\theta) \\ &= \nabla_\theta \left(Y^T Y - 2(X^T Y)^T \theta + \theta^T X^T X \theta \right) \\ &= -2X^T Y + 2X^T X\theta.\end{aligned}$$

Puis $\nabla_\theta \mathrm{SCR}(\theta) = 0 \iff X^T X \hat{\theta}_{MC} = X^T Y \iff \hat{\theta}_{MC} = (X^T X)^{-1} X^T Y$, car $X^T X$ est inversible.

La hessienne de $\theta \mapsto \mathrm{SCR}(\theta)$ vaut $2X^T X$ et est symétrique définie positive. La fonction est strictement convexe : le point critique est donc un minimum global. □

Estimateur des moindres carrés : propriétés générales

Proposition (Propriétés générales de $\hat{\theta}_{MC}$). Dans un modèle linéaire identifiable, l'estimateur des moindres carrés de θ:

(*i*) est toujours sans biais : $\mathbb{E}_\theta[\hat{\theta}_{MC}] = \theta$.

(*ii*) a pour matrice de covariance $\mathrm{Var}_\theta(\hat{\theta}_{MC}) = \sigma^2(X^TX)^{-1}$.

(*iii*) De plus, la somme des carrés résiduels en $\hat{\theta}_{MC}$, $\mathrm{SCR}(\hat{\theta}_{MC})$, peut être utilisée pour estimer σ^2. On a $\mathrm{SCR}(\hat{\theta}_{MC})/n$ qui est biaisé, et $\mathrm{SCR}(\hat{\theta}_{MC})/(n-p)$ qui est sans biais.

Interprétation de (*ii*) : on paye le bruit en σ^2 (logique) + la transformation par X dans Y (qu'il faut en quelque sorte inverser).

Remarque : Supposons que les p colonnes de X sont orthogonales, et tous les coefficients sont d'ordre 1. Alors X^TX est proche de nI_p, et

$$\mathrm{Tr}(\sigma^2(X^TX)^{-1}) \simeq \sigma^2 p/n\,.$$

Si la dimension p est constante, pas de souci. Mais si p scale avec n, on a des problèmes pour estimer θ...

Estimateur des moindres carrés : preuve des propriétés générales

Pour (i), on utilise la linéarité de l'espérance:

$$\mathbb{E}_\theta[\hat{\theta}_{MC}] = \mathbb{E}_\theta[(X^TX)^{-1}X^TY] = (X^TX)^{-1}X^T\mathbb{E}_\theta[Y] = (X^TX)^{-1}X^TX\theta = \theta.$$

Pour (ii), par bilinéarité de la covariance :

$$\begin{aligned}\mathrm{Var}_\theta(\hat{\theta}_{MC}) &= \mathrm{Var}_\theta((X^TX)^{-1}X^TY) = (X^TX)^{-1}X^T\mathrm{Var}_\theta(Y)\left((X^TX)^{-1}X^T\right)^T \\ &= (X^TX)^{-1}X^T\sigma^2 I_n X(X^TX)^{-1} = \sigma^2(X^TX)^{-1}.\end{aligned}$$

Pour (iii), on calcule $\mathbb{E}_\theta[\|Y - X\hat{\theta}_{MC}\|^2] = \mathbb{E}_\theta[\|(I_n - \Pi_{\mathrm{Im}(X)})Y\|^2]$. En notant Π le projecteur orthogonal sur $\mathrm{Im}^\perp(X)$:

$$\begin{aligned}\mathbb{E}_\theta[\|Y - X\hat{\theta}_{MC}\|^2] &= \mathbb{E}_\theta[\|\Pi Y\|^2] = \mathbb{E}_\theta[\|\Pi(X\theta + \varepsilon)\|^2] \\ &= \mathbb{E}_\theta[\|\Pi\varepsilon\|^2] = \mathbb{E}_\theta[\varepsilon^T\Pi^T\Pi\varepsilon] \\ &= \sigma^2\,\mathrm{Tr}(\Pi^T\Pi) = \sigma^2\,\mathrm{Tr}(\Pi) = (n-p)\sigma^2. \quad \square\end{aligned}$$

Résultats spécifiques au modèle linéaire gaussien

Proposition (Moindres carrés dans le modèle gaussien). Dans le modèle linéaire gaussien, supposé identifiable, l'estimateur du MV de (θ, σ^2) noté $(\hat{\theta}_{MV}, \hat{\sigma}^2_{MV})$ est donné par

$$\hat{\theta}_{MV} = \hat{\theta}_{MC}, \quad \text{et} \quad \hat{\sigma}^2_{MV} = \frac{\|Y - X\hat{\theta}_{MC}\|^2}{n} = \frac{\mathrm{SCR}(\hat{\theta}_{MC})}{n}.$$

De plus, $\hat{\theta}_{MV}$ et $\hat{\sigma}^2_{MV}$ sont indépendants, et

$$\hat{\theta}_{MV} \sim \mathcal{N}(\theta, \sigma^2(X^T X)^{-1}) \quad \text{et} \quad \hat{\sigma}^2_{MV} \sim \frac{\sigma^2}{n}\chi^2(n-p).$$

En particulier, l'estimateur du max de vraisemblance de θ à l'élégance de coincider avec l'EMC.

Résultats spécifiques au modèle linéaire gaussien, preuve

Modèle linéaire gaussien, $Y \sim \mathcal{N}(X\theta, \sigma^2 I_n)$, de densité

$$f_{\theta,\sigma^2}(y) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\sigma^2 I_n)}} \exp\left(-\frac{1}{2}(y - x\theta)^T \frac{1}{\sigma^2} I_n (y - x\theta)\right)$$
$$= \text{cste} \times (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\|y - x\theta\|^2\right) .$$

Ainsi l'estimateur du MV est

$$\hat{\theta}_{MV} := \underset{(\theta,\sigma^2)\in\mathbb{R}^p\times\mathbb{R}_+}{\arg\max} \log f_{\theta,\sigma^2}(Y) = \underset{(\theta,\sigma^2)\in\mathbb{R}^p\times\mathbb{R}_+}{\arg\min} \left(\frac{n}{2}\log(\sigma^2) + \frac{1}{2\sigma^2}\|Y - X\theta\|^2\right) .$$

Pour tout σ^2, on minimise d'abord en θ :

$$\hat{\theta}_{MV} = \arg\min_{\theta\in\mathbb{R}^p} \|Y - X\theta\|^2 = \hat{\theta}_{MC} .$$

Puis, on dérive la fonction

$$g : \sigma^2 \mapsto \frac{n}{2}\log(\sigma^2) + \frac{1}{2\sigma^2}\mathrm{SCR}(\hat{\theta}_{MC}) .$$

On a $g'(\sigma^2) = \frac{n}{2\sigma^2} - \frac{1}{2\sigma^4}\mathrm{SCR}(\hat{\theta}_{MC})$, $g'(\sigma^2) > 0 \iff \sigma^2 > \frac{\mathrm{SCR}(\hat{\theta}_{MC})}{n} = \sigma^2_{MV}$.

Résultats spécifiques au modèle linéaire gaussien, preuve (suite)

Ensuite, comme $Y \sim \mathcal{N}(X\theta, \sigma^2 I_n)$, $\hat{\theta}_{MC} = (X^T X)^{-1} X^T Y$ est encore gaussien, de moyenne et de variance déjà établie auparavant.

De plus, par le Théorème de Cochran, $X\hat{\theta}_{MC}$ (proj. de Y sur $\mathrm{Im}(X)$) et $Y - X\hat{\theta}_{MC}$ (proj. de Y sur $\mathrm{Im}(X)$) (proj. de Y sur $\mathrm{Im}(X)^{\perp}$) sont indépendantes, et $\|Y - X\hat{\theta}_{MC}\|^2 \sim \sigma^2 \chi^2(\dim(\mathrm{Im}(X)^{\perp})) = \sigma^2 \chi^2(n-p)$.

Il reste à voir que le σ^2_{MV} est une fonction mesurable de $Y - X\hat{\theta}_{MC}$ (c'est clair) et que $\hat{\theta}_{MC}$ est une fonction mesurable de $X\hat{\theta}_{MC}$: c'est le cas car $\hat{\theta}_{MC} = ((X^T X)^{-1} X^T) X\hat{\theta}_{MC}$. □

Tests et intervalles de confiance dans le modèle Gaussien (1/2)

Intervalles de confiance de formes linéaires en θ

On se place dans le modèle linéaire gaussien identifiable, et on note $\hat{\theta} = \hat{\theta}_{MC}$, et $\hat{\sigma^2} = \frac{\mathrm{SCR}(\hat{\theta})}{n-p}$ l'estimateur de σ^2 débiaisé.

Objectif : tester ou estimer la valeur de composantes ou combinaisons linéaires de θ.

Résultats clés (à savoir retrouver) : Pour tout $a \in \mathbb{R}^p$:

$$(a^T(X^TX)^{-1}a)^{-1/2}\frac{a^T\hat{\theta} - a^T\theta}{\sigma^2} \sim \mathcal{N}(0,1)$$

et

$$(a^T(X^TX)^{-1}a)^{-1/2}\frac{a^T\hat{\theta} - a^T\theta}{\hat{\sigma}} \sim \mathcal{T}(n-p)\,.$$

Conséquence : un IC de niveau $1-\alpha$ pour $a^T\theta$ est :

$$\left[a^T\hat{\theta} \pm t^{(n-p)}_{1-\alpha/2}\,\hat{\sigma}\sqrt{a^T(X^TX)^{-1}a}\right],$$

avec $t^{(n-p)}_{\beta}$ quantile de $\mathcal{T}(n-p)$ d'ordre β.

Intervalles de confiance de formes linéaires en θ

Remarque : on a aussi une solution pour tester $H_0 : a^T\theta = c$ contre $H_1 : a^T\theta \neq c$ en considérant la statistique

$$T := (a^T(X^TX)^{-1}a)^{-1/2}\frac{a^T\hat{\theta} - c}{\hat{\sigma}}$$

et en appliquant les résultats ci-dessus. C'est le test de Student.

Exemples :

- IC/test pour un coefficient particulier θ_j : $a = e_j$. Dans ce cas, $(a^T(X^TX)^{-1}a)^{-1/2} = \frac{1}{\sqrt{[(X^TX)^{-1}]_{j,j}}}$
- IC/test pour la différence entre deux coefficients : $a = e_i - e_j$.

Merci !

Rdv en TD pour les questions et la pratique de ces notions.

(contenu du cours disponible sur ma page web : `lganassali.github.io`)