Statistiques (STA1)

Cours VI – Le modèle linéaire, suite et fin

Luca Ganassali

Laboratoire de Mathématiques d'Orsay, Université Paris-Saclay

Jeudi 6 novembre 2025

Rappels : le modèle linéaire, jusqu'a présent

Le modèle linéaire s'écrit :

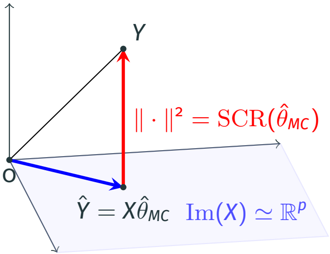
$$\underbrace{Y}_{\in\mathbb{R}^{n\times 1}} = \underbrace{X}_{\in\mathbb{R}^{n\times p}} \cdot \underbrace{\theta}_{\in\mathbb{R}^{p\times 1}} + \underbrace{\varepsilon}_{\in\mathbb{R}^{n\times 1}} .$$

avec ε vecteur de buits centrés, décorélés et de même variance σ^2.

Dans le modèle linéaire gaussien on suppose de plus $\varepsilon \sim \mathcal{N}(0_n, \sigma^2 I_n)$.

Identifiabilité du modèle linéaire (en θ) ssi $X^T X \in \mathbb{R}^{p\times p}$ est inversible.

Estimateur des moindres carrés : $\hat{\theta}_{MC} \in \arg\min_{\theta\in\mathbb{R}^p} \|Y - X\theta\|^2$. Dans le cas identifiable, $\hat{\theta}_{MC} = (X^T X)^{-1} X^T Y$.



Rappels : le modèle linéaire, jusqu'a présent

Dans un modèle linéaire identifiable, l'estimateur des moindres carrés $\hat{\theta}_{MC}$ est sans biais, de variance $\sigma^2(X^TX)^{-1}$ et $\hat{\sigma}^2 = \frac{\|Y - X\hat{\theta}_{MC}\|^2}{n-p}$ est sans biais pour l'estimation de σ^2.

De plus, dans le modèle linéaire gaussien $\hat{\theta}_{MC}$ est l'estimateur du max de vraisemblance, celui de σ^2 étant $\frac{n-p}{n}\hat{\sigma}^2$. De plus, $\hat{\theta}_{MC}$ et $\hat{\sigma}^2$ sont indépendants.

Tests et intervalles de confiance dans le modèle Gaussien

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 $a \in \mathbb{R}^p$:

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

et

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

Proposition (Conséquence 1). Un IC de niveau $1-\alpha$ pour $a^T\theta$ est :

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

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

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 : $a = e_i - e_j$.

Proposition (Conséquence 2). Avec ce qui précède, on peut tester H_0 : $a^T\theta = c$ contre H_1 : $a^T\theta \neq c$ en considérant la statistique

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

dont la loi sous $\mathcal{H}_0$ est $\mathcal{T}(n-p)$. La zone de rejet associée au test pour un niveau $1-\alpha$ est

$$\left\{|T| > t_{1-\alpha/2}^{(n-p)}\right\},$$

avec $t_\beta^{(n-p)}$ quantile de $\mathcal{T}(n-p)$ d'ordre β. C'est le test de Student.

Modèles emboîtés et test de Fisher : un exemple

On considère le modèle linéaire gaussien identifiable $Y = X\theta + \varepsilon$ et on veut tester la nullité des $q > 0$ derniers paramètres du modèle. On note $p_0 = p - q$. On teste donc :

$$\mathcal{H}_0 : \theta_{p_0+1} = \ldots = \theta_p = 0 \quad \text{contre} \quad \mathcal{H}_1 : \exists j \in \{p_0 + 1, \ldots, p\}, ; \theta_j \neq 0 \, .$$

En terme de modèle, si $\theta_{p_0+1} = \ldots = \theta_p = 0$, le modèle devient

$$Y = X_0\theta_0 + \varepsilon$$

avec $X_0 \in \mathbb{R}^{n \times p_0}$ matrice extraite de X (p_0 premières colonnes), de rang p_0.

On est parti d'un modèle avec $\mathbb{E}[Y] \in \Omega = \mathrm{Im}(X)$ de dimension p, et sous $\mathcal{H}_0$, $\mathbb{E}[Y] \in \omega$, avec $\omega = \mathrm{Im}(X_0)$, sous espace de Ω de dimension $p_0 < p$.

Modèles emboîtés et test de Fisher : idée générale

Notons :

- $\hat{\theta} = (X^T X)^{-1} X^T Y$ l'EMC de θ pour le grand modèle et $\hat{Y} = X\hat{\theta}$;
- $\hat{\theta}_0 = (X_0^T X_0)^{-1} X_0^T Y$ l'EMC de θ dans le petit modèle et $\hat{Y}_0 = X_0 \hat{\theta}_0$.

Idée : si $\mathcal{H}_0$ est vraie, $\mathbb{E}[Y]$ appartient à un sous-espace $\omega \subset \Omega$, donc $\hat{Y}$ doit être "proche" de $\hat{Y}_0$. Réciproquement, si $\hat{Y}_0$ est proche de $\hat{Y}$, le modèle plus simple (avec moins de coefficients) explique presque aussi bien les données.

Question : que veut dire "proche" ici ? Proche par rapport à quoi ? $\longrightarrow$ On compare en fait $\|\hat{Y} - \hat{Y}_0\|^2$ à la somme des carrés des résidus du grand modèle $\|Y - \hat{Y}\|^2$.

Cette comparaison n'est pas vraiment juste. En effet, le vecteur aléatoire $Y - \hat{Y}$ vit dans $\mathrm{Im}(X)^{\perp}$, de dimension $n - p$, et $\hat{Y} - \hat{Y}_0$ vit dans $\mathrm{Im}(X_0)^{\perp} \cap \mathrm{Im}(\mathrm{X})$, de dimension $p - p_0$...

Modèles emboîtés et test de Fisher : idée générale

On suit l'idée précédente, mais on normalise par les dimensions qui sont les degrés de liberté.

Proposition. Pour tester l'appartenance de $\mathbb{E}[Y]$ au sous-modèle ω (par exemple la nullité des $q = p - p_0$ coefficients) dans le modèle gaussien identifiable, on se base sur la statistique

$$F := \frac{\|\hat{Y} - \hat{Y}_0\|^2/(p - p_0)}{\|Y - \hat{Y}\|^2/(n - p)} \,.$$

Sous $\mathcal{H}_0$, $F \sim \mathcal{F}(p - p_0, n - p) = \mathcal{F}(q, n - p)$. La zone de rejet pour un niveau $1 - \alpha$ est donc

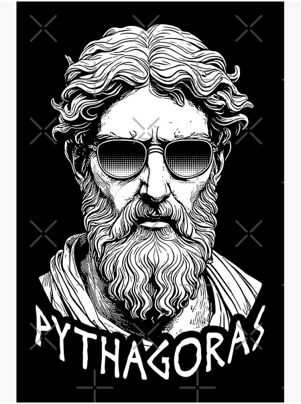
$$\left\{F > f_{1-\alpha}^{(p-p_0,n-p)}\right\},$$

où $f_\beta^{(d_1,d_2)}$ est le quantile d'ordre β de la loi $\mathcal{F}(d_1, d_2)$. C'est le test de Fisher.

Modèles emboîtés et test de Fisher : yet another...

On a $Y - \hat{Y} \in \mathrm{Im}(X)^{\perp}$ et $\hat{Y} - \hat{Y}_0 \in \mathrm{Im}(X_0)^{\perp} \cap \mathrm{Im}(\mathrm{X})$, donc $Y - \hat{Y} \perp \hat{Y} - \hat{Y}_0$, donc :

$$\underbrace{\|Y - \hat{Y}_0\|^2}_{\text{erreur du petit modèle } \omega} = \underbrace{\|\hat{Y} - \hat{Y}_0\|^2}_{\text{erreur du grand modèle } \Omega} + \underbrace{\|Y - \hat{Y}\|^2}_{\text{erreur entre } \Omega \text{ et } \omega}.$$



Modèles emboîtés et test de Fisher : yet another Pythagore

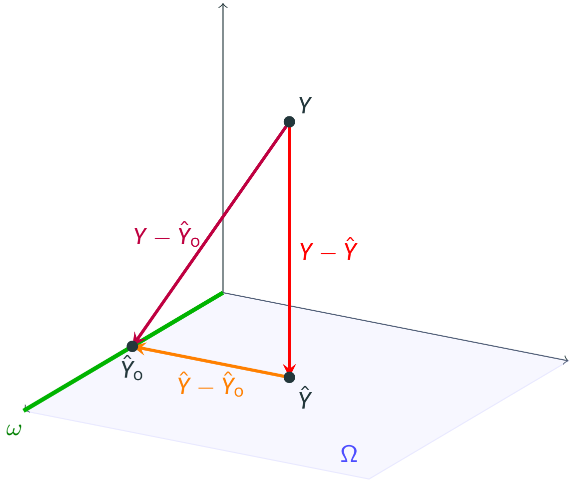


Illustration du théorème de Pythagore dans les modèles emboités. On va rejeter l'hypothèse nulle $\mathbb{E}[Y] \in \omega$ si l'erreur supplémentaire entre les modèles, $\|\hat{Y} - \hat{Y}_0\|^2$, n'est pas négligeable par rapport à l'erreur du grand modèle, $\|Y - \hat{Y}\|^2$.

Cas du sous-modèle constant : critère du R^2

Cas du sous-modèle constant : critère du R^2

On veut tester un sous-modèle très particulier : le sous-modèle constant $\omega = \mathrm{Vect}(\mathbf{1})$. Dans ce modèle, tous les coefficients sont nuls sauf l'intercept.

On a dans ce cas (projection sur les vecteurs de coordonnées constantes) :

$$\hat{Y}_0 = \bar{y}\mathbf{1}.$$

Pythagore se réécrit :

$$\underbrace{\|Y - \bar{y}\mathbf{1}\|^2}_{\text{variance totale}} = \underbrace{\|\hat{Y} - \bar{y}\mathbf{1}\|^2}_{\text{variance expliquée par le modèle}} + \underbrace{\|Y - \hat{Y}\|^2}_{\text{variance résiduelle (SCR)}}$$

Coefficient de détermination du R^2

Le coefficient de détermination du R^2 est :

$$R^2 = \frac{\text{variance expliquée}}{\text{variance totale}} = 1 - \frac{\|Y - \hat{Y}\|^2}{\|Y - \bar{y}\mathbf{1}\|^2} = 1 - \frac{\mathrm{SCR}}{\text{variance totale}}.$$

- $R^2 \in [0, 1]$
- R^2 élevé $\Rightarrow$ le modèle explique bien les données

Le R^2 une mesure empirique de mesure de qualité du modèle : si $R^2 = 0.75$, les covariables X expliquent 75% de la variance de Y.

On a aussi le R^2 ajusté (prend en compte les dimensions) :

$$R_a^2 := 1 - \frac{\|Y - \hat{Y}\|^2/(n-p)}{\|Y - \bar{y}\mathbf{1}\|^2/(n-1)} = 1 - \frac{n-1}{n-p}(1 - R^2) \leq R^2.$$

Modèle linéaire avec une variable qualitative

Modèle linéaire avec une variable qualitative

Supposons que nous ayons une variable qualitative G à k modalités. Par exemple :

$$G \in \{\text{bleu, rouge, jaune, vert}\}.$$

On veut écrire un modèle linéaire prenant en compte la covariable G.
Problème : elle n'a pas de valeur numérique propre.

Idée : remplacer G par k variables indicatrices:

$$x_{\ell,i} = \begin{cases} 1 & \text{si l'individu } i \text{ appartient au groupe } \ell, \\ 0 & \text{sinon.} \end{cases}$$

L'idée serait d'écrire :

$$Y_i = \theta_0 + \theta_1 x_{1,i} + \cdots + \theta_k x_{k,i} + \varepsilon_i.$$

Modèle linéaire avec une variable qualitative

Exemple pour 3 groupes :

$$X = \begin{pmatrix} 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 \end{pmatrix}$$

Problème : $x_1 + \cdots + x_k = \mathbf{1}$.

X n'est pas de rang plein : le modèle n'est pas identifiable.

Solution : on choisit un groupe de référence, par exemple le groupe $\ell_0 = 1$, et on impose $\theta_{\ell_0} = 0$. Le modèle devient :

$$Y_i = \theta_0 + \sum_{j=2}^{k} \theta_j x_i^{(j)} + \varepsilon_i,$$

qui est identifiable.

Modèle linéaire avec une variable qualitative

Interprétation : θ_ℓ mesure l'effet du groupe ℓ relativement au groupe ℓ_0.

On dit que les groupes ℓ et ℓ' sont statistiquement équivalents si on ne rejette pas $\mathcal{H}_0$ dans le test de Student pour $\theta_\ell = \theta_{\ell'}$

Attention : l'équivalence statistique n'est pas transitive ! On peut avoir $\ell \simeq \ell_0$ et $\ell_0 \simeq \ell'$ mais $\ell \not\simeq \ell'$.

Lire dans `R` les résultats d’une régression linéaire

Lorsqu'on ajuste un modèle linéaire gaussien dans R via la commande :

```
> summary(lm(Y ~ X1 + X2 + X3))
```

on obtient notamment un tableau de coefficients tel que :

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.3451    0.4567    5.134 0.00012 ***
X1           0.7823    0.1875    4.173 0.00157 **
X2          -0.3128    0.1452   -2.154 0.04010 *
X3           0.0914    0.1021    0.895 0.37760
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
```

Modalités de l'examen

- **Date :** Vendredi 14 novembre, de 10h à 12h (tiers-temps : 12h30).
- **Durée :** 2 heures, sur feuille
- **Séance de questions/réponses :** juste avant, de 9h à 9h45.
- **Sont autorisés :**
  - une feuille A4 manuscrite, recto uniquement
  - la calculatrice
- **Un spoiler :** il y aura peut-être des questions de cours.

Merci !

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

(cours, TD et quizz disponibles sur ma page `lganassali.github.io`)