Modélisation statistique, première partie – notes de cours

Luca Ganassali*

Université Paris-Saclay

Nota bene. Ces notes de cours n'ont absolument pas vocation à remplacer le cours pris par écrit. Elles listent simplement les notions vues en cours ainsi que les résultats principaux, mais ne mentionnent pas les références, les preuves, les exemples, les remarques et développements détaillés en cours.

1. Introduction à la modélisation statistique, voyage en pays gaussien

1.1. Définitions, vocabulaire

Contexte général. On dispose de *données observées* $x := (x_1, \dots, x_n)$ (par exemple des données journalières de température) qui sont la réalisation d'un vecteur aléatoire tiré selon une loi inconnue. Le cours se concentre sur des questions :

(*i*) d'*estimation*, c'est-à-dire estimer la valeur d'un *paramètre* d'intérêt sur cette loi inconnue (par exemple la valeur moyenne de température), parfois avec plus ou moins d'erreur, donc à l'aide d'un *intervalle ou région de confiance*;

(*ii*) de *tests d'hypothèses*, c'est-à-dire être capable de décider si oui ou non telle hypothèse sur la loi inconnue est réalisée (par exemple : *d'après mes observations, la température est-elle en moyenne plus chaude que* $12\,^\circ C$ *?*). Bien sûr, dans ce genre de décision, il y a toujours un risque de se tromper, que l'on cherchera à quantifier.

Variable aléatoire, échantillon, observation. On note X qui est le vecteur aléatoire dont la réalisation est l'observation x. Lorsque X est un vecteur aléatoire de données independantes et identiquement distribuées (i.i.d.), on écrit $X = (X_1, \dots, X_n)$ et l'on parle d'échantillon pour X.

Modélisation statistique. Un *modèle statistique* est un triplet $\mathcal{M} := (\mathcal{X}, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ où $\mathcal{X}$ est l'espace des réalisations, c'est-à-dire l'espace des valeurs prises par X, $\mathcal{A}$ la tribu des évènements[1] sur $\mathcal{X}$ et $(\mathbb{P}_\theta)_{\theta\in\Theta}$ une famille de lois de probabilités sur $\mathcal{X}$. Lorsque Θ est un sous-ensemble de $\mathbb{R}^d$ pour un certain $d \geq 1$, on dit que le modèle est *paramétrique* (il dépend d'un nombre fini d de paramètres numériques). Lorsque X est un échantillon, il est formé de n variables de même loi η_θ et indépendantes et $(\mathbb{P}_\theta)_{\theta\in\Theta}$ s'écrit alors $((\eta_\theta)^{\otimes n})_{\theta\in\Theta}$.

*luca.ganassali@universite-paris-saclay.fr

[1] dans ce cours, elle sera très simple : l'ensemble des parties de $\mathcal{X}$ si $\mathcal{X}$ est au plus dénombrable, et la tribu de Lebesgue si $\mathcal{X} = \mathbb{R}^d$.

Bien savoir ce que l'on fait. Mathématiquement, modéliser, c'est supposer que ce qu'on observe (une réalisation de X) ne vient pas complètement de nulle part mais vit dans un espace donné $\mathcal{X}$ et suit une loi de probabilité, certes inconnue, mais qui appartient la famille $(\mathbb{P}_\theta)_{\theta\in\Theta}$. Sociétalement et politiquement, modéliser c'est toujours schématiser, et un choix de modèle dit donc beaucoup des hypothèses que l'on fait ! Ces hypothèses sont cruciales et doivent être explicitées, discutées, critiquées. Tout ensemble d'hypothèses a bien sûr des limites, car on schématise la réalité. Il convient de prendre garde à ne pas trop schématiser : à force de voir la réalité comme moins complexe que ce qu'elle est on peut la voir de façon totalement erronée et cela peut mener à des erreurs graves.

Exemples de modèles simples.

- On réalise un sondage sur n personnes en leur demandant si ils comptent aller voter pour le candidat Y aux prochaines élections.

  Modèle proposé : $\mathcal{M} = (\{0,1\}^n, \mathcal{P}(\{0,1\}^n), (\mathrm{Ber}(\theta)^{\otimes n})_{\theta\in[0,1]})$, où $\mathrm{Ber}(\theta)$ est la loi de Bernoulli de paramètre θ.

- On recueille le loyer mensuel de n appartements parisiens.

  Modèles proposés :

  $\mathcal{M}_1 = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), (\mathcal{N}(\mu,\sigma^2)^{\otimes n})_{(\mu,\sigma)\in\mathbb{R}\times\mathbb{R}_+})$, où $\mathcal{N}(\mu,\sigma^2)$ est la loi normale de moyenne μ et de variance σ^2.

  $\mathcal{M}_2 = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), ((\alpha\mathcal{N}(\mu_1,\sigma_1^2) + (1-\alpha)\mathcal{N}(\mu_2,\sigma_2^2))^{\otimes n})_{(\alpha,\mu_1,\mu_2,\sigma_1,\sigma_2)\in[0,1]\times\mathbb{R}^2\times\mathbb{R}_+^2})$.

- On recueille n mesures de la consommation d'énergie en France, $y_1,\dots,y_n$ à des heures données de la journée $\mathcal{H}_1,\dots,h_n$.

  Modèle proposé : $\mathcal{M} = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \eta(\sigma,f)_{(\sigma,f)\in\mathbb{R}_+\times\mathcal{C}^0([0,24[,\mathbb{R})})$, où $\mathcal{C}^0([0,24[,\mathbb{R})$ est l'ensemble des fonctions continues de $[0,24[$ dans $\mathbb{R}$ et $\eta(\sigma,f)$ est la loi de $(y_1,\dots,y_n)$ telle que pour tout i, $y_i = f(h_i) + \mathcal{N}(0,\sigma^2)$.

1.2. Rappels sur les vecteurs aléatoires

Soit $X = (X_1,\dots,X_d)$ un vecteur aléatoire à valeurs dans $\mathbb{R}^d$, tel que les X_i admettent toutes un moment d'ordre deux fini (on dira plus simplement que X admet un moment d'ordre deux fini). définit son vecteur espérance

$$\mathbb{E}[X] := (\mathbb{E}[X_1], \dots \mathbb{E}[X_d]) \in \mathbb{R}^d$$

et sa *matrice de covariance (ou matrice de variance-covariance)* par

$$\mathrm{Var}(X) := \mathbb{E}[(X-\mu)(X-\mu)^T] \in \mathbb{R}^{d\times d}.$$

Notons que pour tous $1 \le i,j \le d$, $[\mathrm{Var}(X)]_{i,j} = \mathrm{Cov}(X_i,X_j)$: $\mathrm{Var}(X)$ est une matrice symétrique.

Proposition 1.1. Soit X un vecteur aléatoire de $\mathbb{R}^d$ de moment d'ordre deux fini. Soient $A \in \mathbb{R}^{m\times d}$ et $b \in \mathbb{R}^m$. Alors $Y = AX + b$ est un vecteur aléatoire de $\mathbb{R}^m$ qui a également un moment d'ordre deux fini, et on a :

$$\mathbb{E}[Y] = A\mathbb{E}[X] + b \quad \text{et} \quad \mathrm{Var}(Y) = A\mathrm{Var}(X)A^T.$$

Avec la propriété ci-dessus, on a que pour tout $a \in \mathbb{R}^d$, $a^T\Sigma a = \mathrm{Var}(a^TX) \ge 0$. Une matrice de covariance est donc toujours symétrique positive.

Ces rappels étant faits, nous pouvons définir les vecteurs gaussiens.

1.3. Rappels sur les vecteurs gaussiens

Cas unidimensionnel. Une variable aléatoire réelle Z est dit *gaussienne de moyenne* $\mu \in \mathbb{R}$ *et de variance* σ^2 *avec* $\sigma > 0$, et l'on note $Z \sim \mathcal{N}(\mu, \sigma^2)$, si sa loi admet sur $\mathbb{R}$ la densité suivante, par rapport à la mesure de Lebesgue :

$$f_{\mu,\sigma^2} := x \mapsto \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Si $\mu = 0$ et $\sigma = 1$, Z est dite *gaussienne centrée réduite, ou gaussienne standard.* Si $\sigma = 0$, Z est dite *gaussienne dégénérée*, c'est une v.a. p.s. égale à μ.

Proposition 1.2 (Propriétés fondamentales des gaussiennes)**.** On a les propriétés suivantes :

(*i*) *Fonction caractéristique (important)* $X \sim \mathcal{N}(\mu, \sigma^2)$ si et seulement si pour tout $t \in \mathbb{R}$, $\mathbb{E}[e^{itX}] = \exp\left(i\mu t - \frac{1}{2}\sigma^2 t^2\right)$

(*ii*) *Somme de deux vecteurs gaussiens indépendants (important)* si $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ et $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ avec X_1, X_2 indépendantes, alors $X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Cas multidimensionnel. Un vecteur aléatoire $X = (X_1, \ldots, X_d)$ à valeurs dans $\mathbb{R}^d$ est dit *gaussien* si toute combinaison linéaire de ses composantes est gaussienne (réelle unidimensionnelle). On note alors $X \sim \mathcal{N}(\mu, \Sigma)$, où $\mu = \mathbb{E}[X]$ et $\Sigma = \mathrm{Var}(X)$.

Proposition 1.3. Si $X \sim \mathcal{N}(\mu, \Sigma)$ et si Σ est inversible, alors X admet sur $\mathbb{R}^d$ par rapport à la mesure de Lebesgue, la densité

$$f_{\mu,\Sigma} := x \mapsto \frac{1}{(2\pi)^{d/2}\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

Proposition 1.4. Soit $X \sim \mathcal{N}(\mu, \Sigma)$ avec Σ non inversible. Alors $X - \mu$ appartient p.s. à un sous espace vectoriel de dimension $\mathrm{rg}(\Sigma) < d$. En particulier, X n'a pas de densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$.

Proposition 1.5 (Propriétés fondamentales des vecteurs gaussiens)**.** On a les propriétés suivantes :

(*i*) *Fonction caractéristique (important)* $X \sim \mathcal{N}(\mu, \Sigma)$ dans $\mathbb{R}^d$ si et seulement si pour tout $s \in \mathbb{R}^d$, $\mathbb{E}[e^{is^T X}] = \exp\left(is^T\mu - \frac{1}{2}s^T\Sigma s\right)$ pour la preuve. Ce résultat est utile pour prouver $(ii), (iii)$ et (iv). Cas unidimensionnel à connaître.

(*ii*) *Somme de deux vecteurs gaussiens indépendants (important)* si $X_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$ et $Z_2 \sim \mathcal{N}(\mu_2, \Sigma_2^2)$ tous deux dans $\mathbb{R}^d$, indépendants, alors $Z_1 + Z_2 \sim \mathcal{N}(\mu_1 + \mu_2, \Sigma_1 + \Sigma_2)$.

(*iii*) *Stabilité par transformation linéaire (important)* Soit $X \sim \mathcal{N}(\mu, \Sigma)$ dans $\mathbb{R}^d$. Alors, pour toute matrice A de $\mathbb{R}^{m\times d}$ et vecteur $b \in \mathbb{R}^m$, le vecteur $AX + b$ est encore gaussien dans $\mathbb{R}^m$ et on a $AX + b \sim \mathcal{N}(A\mu + b, A\Sigma A^T)$.

(*iv*) *Lecture de l'indépendance dans les corrélations (important)* Soit $X \sim \mathcal{N}(\mu, \Sigma)$ dans $\mathbb{R}^d$. On a, pour tout $J \subseteq \{1, \ldots, d\}$, l'équivalence

$$(X_j)_{j\in J} \text{ sont mutuellement indépendants} \iff \Sigma_{J,J} \text{ est diagonale,}$$

où $\Sigma_{J,J}$ est la matrice extraite de Σ en ne gardant que les lignes et colonnes dont les indices sont les éléments de J.

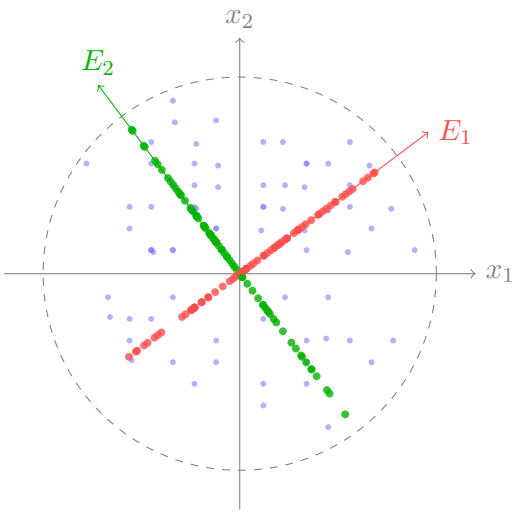


Figure 1 – *Illustration du théorème de Cochran en dimension* 2*, avec* $d_1 = d_2 = 1$*. Les nuages de points vert et rouge sont indépendants.*

1.4. Loi du khi-deux et Théorème de Cochran.

Pour parler du théorème de Cochran, il nous faut tout d'abord introduire la loi du khi-deux. C'est une loi à connaître car très classique. Elle est liée aux gaussiennes.
Loi du khi-deux. Soit $d \geq 1$ et $Z \sim \mathcal{N}(0_d, I_d)$. La *loi du khi-deux à* d *degrés de liberté, notée* $\chi^2(d)$, est la loi de la somme des carrés de ses composantes : $\|Z\|^2 = Z_1^2 + \ldots + Z_d^2 \sim \chi^2(d)$.

Theorème 1.1 (*Théorème de Cochran*)**.** Soit $Z \sim \mathcal{N}(\mu, \sigma^2 I_n)$ un vecteur gaussien de $\mathbb{R}^n$. Soient $r \geq 1$ et $E_1, \ldots, E_r$ des sous-espaces vectoriels orthogonaux deux à deux et tels que $E_1 \oplus \ldots \oplus E_r = \mathbb{R}^n$. Pour $1 \leq j \leq n$, notons Π_j le projecteur orthogonal sur E_j et d_j sa dimension. Alors,

(*i*) les vecteurs aléatoires $\Pi_1 Z, \ldots, \Pi_r Z$ sont gaussiens, mutuellement indépendants et de lois respectives $\mathcal{N}(\Pi_1\mu, \sigma^2\Pi_1), \ldots, \mathcal{N}(\Pi_r\mu, \sigma^2\Pi_r)$;

(*ii*) Les variables $\frac{\|\Pi_1(Z-\mu)\|^2}{\sigma^2}, \ldots, \frac{\|\Pi_r(Z-\mu)\|^2}{\sigma^2}$ sont mutuellement indépendantes et de lois respectives $\chi^2(d_1), \ldots, \chi^2(d_r)$.

Deux autres lois classiques en statistiques : lois de Student et de Fisher. Nous verrons l'utilité de ces lois très bientôt, pour faire de l'estimation dans le modèle gaussien.

Loi de Student. Soit $Z \sim \mathcal{N}(0,1)$ et $K \sim \chi^2(p)$ *indépendantes.* La *loi de Student à* p *degrés de liberté, notée* $\mathcal{T}(d)$, est la loi de la variable

$$T = \frac{Z}{\sqrt{K/p}}\,.$$

Loi de Fisher. Soient $K_1 \sim \chi^2(n_1)$ et $K_2 \sim \chi^2(n_2)$ *indépendantes.* La *loi de Fisher à* (n_1, n_2) *degrés de liberté, notée* $\mathcal{F}(n_1, n_2)$, est la loi de la variable

$$F = \frac{K_1/n_1}{K_2/n_2}\,.$$

Quelques propriétés géométriques des vecteurs gaussiens. La Figure 2 illustre la renormalisation classique d'un vecteur gaussien, bien connu en dimension 1, étendu ici en dimension

d : si $X \sim \mathcal{N}(\mu, \Sigma)$ avec Σ inversible, alors $\Sigma^{-1/2}(X - \mu) \sim \mathcal{N}(0, I_d)$. La Figure 3 montre la forme typique d'un nuage Gaussien dans $\mathbb{R}^d$. On remarquera que plus Σ a de petites valeurs propres, plus on s'approche du cas non inversible (dégénéré) et plus l'ellipsoïde a tendance à s'aplatir : c'est logique, le vecteur aléatoire finira par vivre dans un espace de dimension effective inférieure à d.

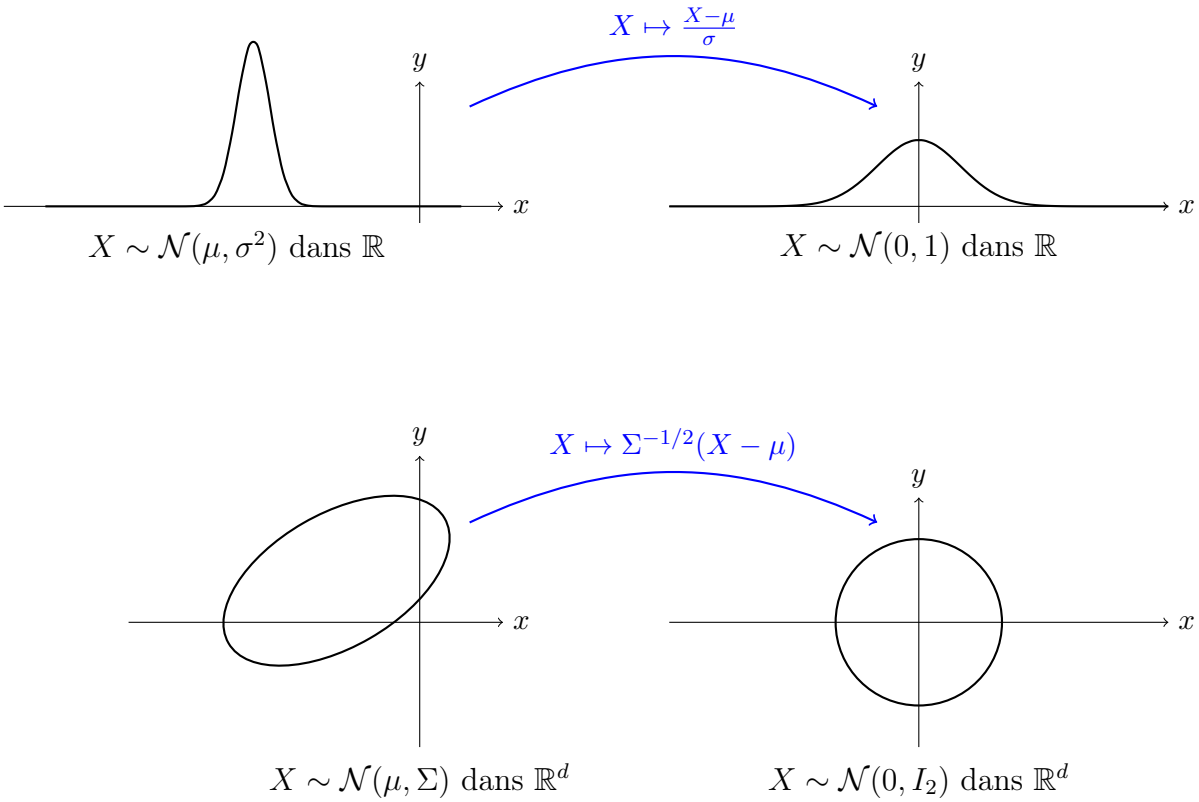


Figure 2 – *Illustration de la renormalisation gaussienne, en dimension 1 et en dimension d*

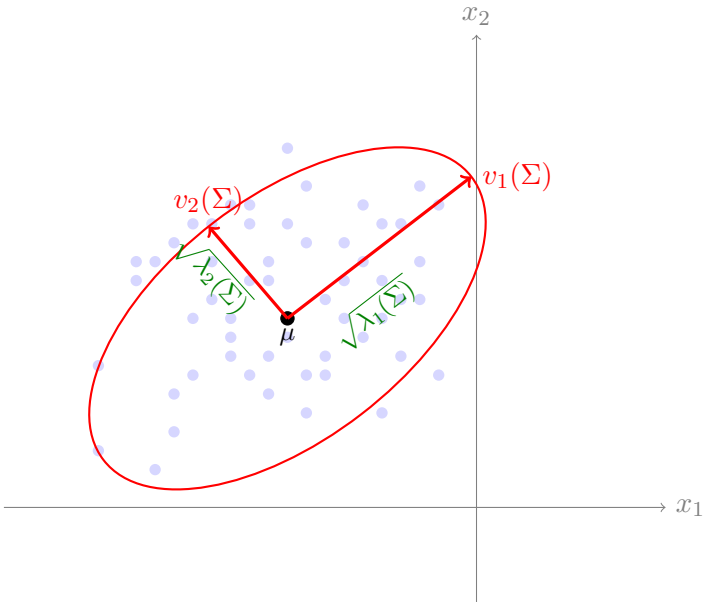


Figure 3 – *Allure typique d'un nuage Gaussien dans $\mathbb{R}^d$. L'ellipse correspond au lieu géométrique des points de densité constante.*

2. Rappels sur l'estimation, intervalles de confiance, approximation gaussienne

2.1. Estimation ponctuelle

Estimateurs. Lorsqu'on travaille avec un vecteur aléatoire X sous un modèle paramétrique avec une famille de lois $(\mathbb{P}_\theta)_{\theta\in\Theta}$, on veut pouvoir estimer θ ou une quantité $\varphi(\theta)$ qui vit dans un certain $\mathbb{R}^d$. Un *estimateur* de $\varphi(\theta)$, souvent noté $\hat{\varphi}$, est une variable aléatoire mesurable de X, c'est-à-dire de la forme $\hat{\varphi} = f(X)$ avec f mesurable.

Biais, variance, risque quadratique. Soit $\hat{\varphi}$ un estimateur de $\varphi(\theta)$ vivant dans un certain $\mathbb{R}^d$. Le *biais* de l'estimateur $\hat{\varphi}$ est défini pour tout $\theta \in \Theta$ par

$$b_\theta(\hat{\varphi}) := \mathbb{E}_\theta[\hat{\varphi} - \varphi(\theta)] = \mathbb{E}_\theta[\hat{\varphi}] - \varphi(\theta) \in \mathbb{R}^d\,.$$

La *matrice de covariance* de l'estimateur $\hat{\varphi}$ est définie pour tout $\theta \in \Theta$ par

$$\mathrm{Var}_\theta(\hat{\varphi}) := \mathbb{E}_\theta[(\hat{\varphi} - \mathbb{E}_\theta[\hat{\varphi}])^T(\hat{\varphi} - \mathbb{E}_\theta[\hat{\varphi}])] \in \mathbb{R}^{d\times d}\,.$$

Le *risque quadratique* de l'estimateur $\hat{\varphi}$ est défini pour tout $\theta \in \Theta$ par

$$R_\theta(\hat{\varphi}) := \mathbb{E}_\theta[\|\hat{\varphi} - \varphi(\theta)\|^2]\,.$$

Proposition 2.1 (Décomposition du risque quadratique : Pythagore du statisticien)**.** On a

$$R_\theta(\hat{\varphi}) = \|b_\theta(\hat{\varphi})\|^2 + \mathrm{Tr}(\mathrm{Var}_\theta(\hat{\varphi}))\,.$$

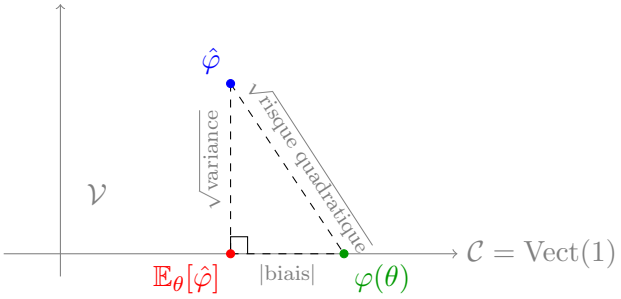


Figure 4 – *Illustration du théorème de Pythagore dans la décomposition du risque quadratique, en dimension* 1*. Ici, l'espace* $\mathcal{V}$ *est celui des v.a. réelles (définies sur* Ω*), et* $\mathcal{C} = \mathrm{Vect}(1)$ *est celui des v.a. constantes. La projection orthogonale de* $\hat{\varphi}$ *sur* $\mathcal{C}$ *est donc* $\mathbb{E}_\theta[\hat{\varphi}]$*.*

Comparaison d'estimateurs. Le critère principal pour comparer deux estimateurs $\hat{\varphi}_1$ et $\hat{\varphi}_2$ d'une même quantité $\varphi(\theta)$ est le risque quadratique. On dit que $\hat{\varphi}_1$ *est meilleur que* $\hat{\varphi}_2$ *au sens du risque quadratique* si pour tout $\theta \in \Theta$, $R_\theta(\hat{\varphi}_1) \le R_\theta(\hat{\varphi}_2)$.

Cependant, selon le contexte, on peut être amenés à comparer uniquement les biais (en norme euclidienne par exemple), ou les matrices de covariances. Pour ces dernières, on peut comparer leurs traces pour tout $\theta \in \Theta$ (comme dans le risque), mais aussi les comparer de façon plus globale avec l'ordre suivant (ordre de Löwner) : pour $A, B \in \mathbb{R}^{p\times p}$ symétriques, on dit que $A \preceq B$ si $B - A$ est symétrique positive.

Une méthode pour obtenir des estimateurs : estimation via le maximum de vraisemblance. On se place dans le cas où l'on cherche à estimer θ et où pour tout $\theta \in \Theta$, $\mathbb{P}_\theta$ est à densité f_θ par rapport à la mesure de Lebesgue, ou par rapport à la mesure de comptage. La *méthode*

du maximum de vraisemblance consiste à estimer θ en maximisant la (log-)vraisemblance de l'échantillon, comme suit :

$$\hat{\theta}_{MV} := \arg\max_{\theta\in\Theta} \log f_\theta(X_1,\dots,X_n)\,.$$

Dans le cas où X est un échantillon de taille n avec X_1 de densité g_θ,

$$\hat{\theta}_{MV} := \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \log g_\theta(X_i)\,.$$

2.2. Régions et intervalles de confiance (rappels)

On reste dans le contexte où l'on veut estimer une certaine fonction de θ, $\varphi(\theta) \in \mathbb{R}^d$. Dans le cas $d \geq 2$, une *région de confiance de probabilité de couverture* $1-\alpha$ *pour* $\varphi(\theta)$ est un ensemble aléatoire $C_\alpha(X)$ dépendant de X tel que pour tout $\theta \in \Theta$,

$$\mathbb{P}_\theta(\varphi(\theta) \in C_\alpha(X)) \geq 1-\alpha\,.$$

Si $C_\alpha(X) = C_{\alpha,n}(X)$ dépend du nombre d'observations n et vérifie seulement

$$\liminf_n \mathbb{P}_\theta(\varphi(\theta) \in C_{\alpha,n}(X)) \geq 1-\alpha,$$

on parlera de *région de confiance asymptotique de probabilité de couverture* $1-\alpha$ *pour* $\varphi(\theta)$.

Dans le cas où $d=1$, on parle plutôt d'*intervalle de confiance de probabilité de couverture* $1-\alpha$ *pour* $\varphi(\theta)$. C'est un intervalle $[\hat{\varphi}_{\min}, \hat{\varphi}_{\max}]$ dont les bornes sont des variables aléatoires mesurables de X, et telles que pour tout $\theta \in \Theta$,

$$\mathbb{P}_\theta(\hat{\varphi}_{\min} \leq \varphi(\theta) \leq \hat{\varphi}_{\max}) \geq 1-\alpha\,.$$

Une méthode pour en obtenir : méthode pivotale. La méthode pivotale est utilisée pour trouver une région de confiance d'un niveau donné. Cette méthode est assez générale, illustrons la dans le cas unidimensionnel. A partir d'un estimateur de $\varphi(\theta)$, on calcule sa loi $\mathbb{P}_\theta$, puis on en déduit une quantité réelle $T_n(\varphi(\theta),\theta)$ dont la loi ne dépend plus de θ : cette quantité est dite pivotale. On exprimera les bornes de l'IC en fonction de T_n et ses quantiles.

2.3. Approximation gaussienne: le Théorème Central Limite

Theorème 2.1 (Rappel : loi des grands nombres)**.** Soit $(X_n)_{n\geq 1}$ un échantillon de variables aléatoires réelles i.i.d. de moyenne finie μ. Alors, on a la convergence presque sûre (et en probabililité, et en loi) :

$$\bar{X}_n \underset{n\to\infty}{\longrightarrow} \mu\,.$$

Theorème 2.2 (Théorème Central Limite, cas unidimensionnel)**.** Soit $(X_n)_{n\geq 1}$ un échantillon de variables aléatoires réelles i.i.d. de moyenne finie μ et variance finie σ^2. Alors, on a la convergence en loi :

$$\sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma^2} \underset{n\to\infty}{\longrightarrow} \mathcal{N}(0,1)\,.$$

Theorème 2.3 (Théorème Central Limite, cas multidimensionnel)**.** Soit $(X_n)_{n\geq 1}$ un échantillon de variables aléatoires i.i.d. dans $\mathbb{R}^d$ de moyenne finie μ et covariance finie Σ. Alors, on a la convergence en loi :

$$\sqrt{n}\,(\bar{X}_n - \mu) \underset{n\to\infty}{\longrightarrow} \mathcal{N}(0,\Sigma)\,.$$

Proposition 2.2 (Lemme de Slutsky)**.** Si on a les convergences en loi $X_n \underset{n\to\infty}{\longrightarrow} X$ et $Y_n \underset{n\to\infty}{\longrightarrow} c$ où c est constante, alors $(X_n, Y_n) \underset{n\to\infty}{\longrightarrow} (X,c)$ en tant que couple, et en particulier, $X_n +$

$Y_n \underset{n\to\infty}{\longrightarrow} X + c$, $X_n Y_n \underset{n\to\infty}{\longrightarrow} cX$ et si c est réel et $c \neq 0$ alors $X_n/Y_n \underset{n\to\infty}{\longrightarrow} X/c$.

2.4. Parenthèse : enjeu de la dimension

Un des problèmes le plus classique auquel le statisticien est confronté est celui de la grande dimension. Pour illustrer cette question fondamentale, prenons un $n-$échantillon de vecteurs aléatoires $(X_1, \ldots, X_n)$ dont chacun est de moyenne finie $\mu \in \mathbb{R}^p$ et de matrice de covariance Σ finie, avec chaque coordonnées de X_1 de variance 1. L'estimateur de la moyenne empirique $\bar{X}$ étant sans biais, son risque quadratique est donné par le terme de variance uniquement :

$$\mathrm{Tr}(\mathrm{Var}(\bar{X})) = \frac{\mathrm{Tr}\,\Sigma}{n} = \frac{p}{n}\,.$$

Lorsque $p \ll n$, ce risque tend bien vers 0, mais si p devient comparable à n, on ne sait plus estimer proprement μ. Ce phénomène est appelé *fléau de la grande dimension (curse of dimensionality)*.

3. Le modèle linéaire

Cadre général de l'apprentissage supervisé. On veut *expliquer ou prédire* une variable (ou un vecteur) aléatoire réponse Y en fonction d'observables (dites parfois variables explicatives) $X_1, \ldots, X_p \in \mathbb{R}$. On *modélise* Y comme une fonction de ces variables, avec éventuellement du bruit, par exemple, via un modèle à bruit additif :

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

L'*apprentissage supervisé* consiste à

(*i*) estimer f à partir d'un échantillon $(X_{i,1}, \ldots, X_{i,p}, Y_i)_{1\leq i\leq n}$; (training).

(*ii*) prédire Y pour une autre valeur de X, jamais vue auparavant, et ce avec le moins d'erreur possible (testing).

3.1. Définition des modèles linéaire et linéaire gaussien

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

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

où ε_i est une v.a. appélée *résidu* ou *bruit* qui satisfait :

(*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).

La fonction $f_\theta : x \in \mathbb{R}^p \to x^T\theta \in \mathbb{R}$ est la *fonction de régression du modèle.*
On notera que l'aléa de Y_i provient uniquement de ε_i et non pas X_i : seul ε_i est aléatoire.
Matriciellement, le modèle s'écrit

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

X est la *matrice du plan d'expérience ou matrice de design*, et souvent $X_{i,1} = 1$ pour tout i, θ_1 est alors appelé intercept (ordonnée à l'origine qui garantit une relation affine et non juste linéaire).

Modèle linéaire gaussien C'est le même modèle linéaire que ci dessus $Y = X\theta + \varepsilon$ avec l'hypothèse supplémentaire : $\varepsilon \sim \mathcal{N}(0_n, \sigma^2 I_n)$.

Qu'est-ce que ce modèle permet de faire ? Détecter une corrélation entre une variable explicative et la réponse. La quantifier. .

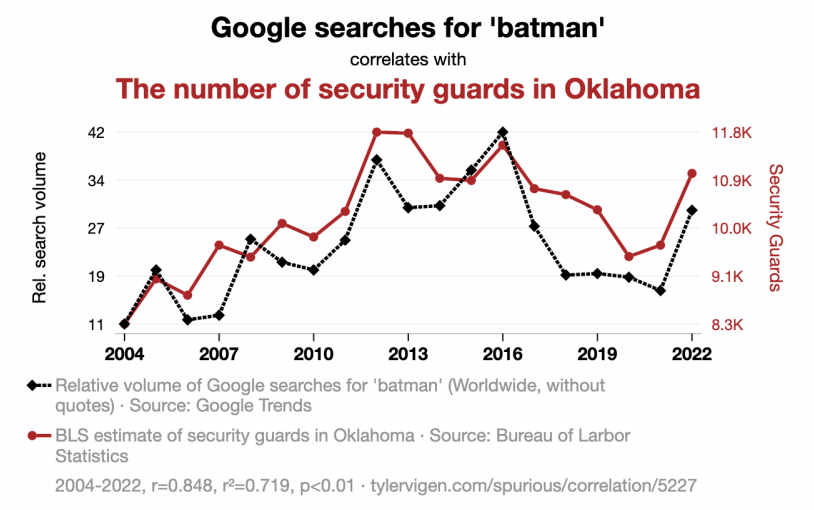
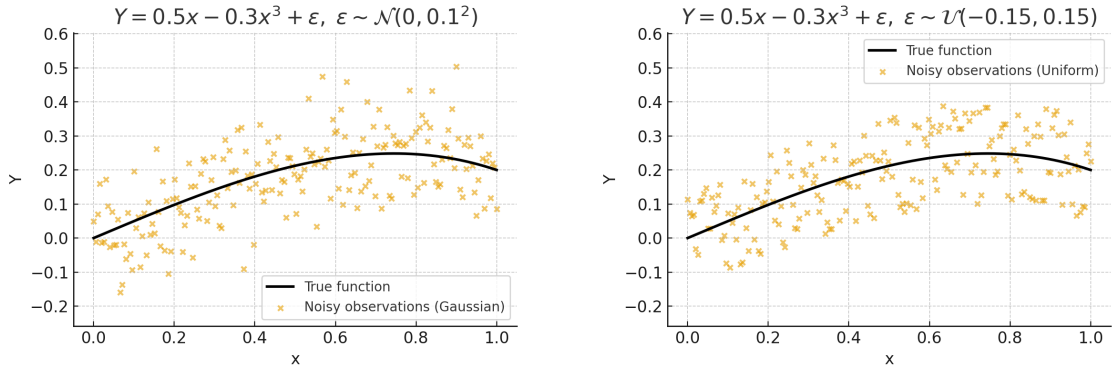


Figure 5 – *Causalité ou corrélation ?*



(a) – *bruit gaussien:* $\epsilon \sim \mathcal{N}(0, 0.1^2)$ (b) – *bruit uniforme:* $\epsilon \sim \mathcal{U}(-0.15, 0.15)$

Figure 6 – *Le modèle linéaire* $Y = 0.5x - 0.3x^3 + \epsilon$ *avec différentes distributions de bruits.*

Identifiabilité Un modèle $\mathcal{M} = (\mathcal{X}, \mathcal{A}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ est dit *identifiable* si deux paramètres θ, θ' différents définissent deux lois $\mathbb{P}_\theta, \mathbb{P}_{\theta'}$ différentes, i.e. si $\theta \in \Theta \mapsto \mathbb{P}_\theta$ est injective.

Dans le cas du modèle linéaire, on a le résultat d'identifiabilité suivant.

Proposition 3.1. 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*) $X^T X \in \mathbb{R}^{p \times p}$ est inversible.

3.2. Régression linéaire, estimateur des moindres carrés

Dans le modèle linéaire, 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 et un minimiseur de la fonction ci-dessus :

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

Forme close de l'EMC et propriétés générales

Proposition 3.2 (Forme close de $\hat{\theta}_{MC}$ et interprétation)**.** Supposons que le modèle est identifiable. On a alors

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

et $X\hat{\theta}_{MC}$ est la projection orthogonale de Y sur $Im(X)$ dont la dimension est p. . En effet

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

On appelle *vecteur des résidus* le vecteur $Y - X\hat{\theta}_{MC}$: sa norme est la distance de Y à l'espace $Im(X) \subset \mathbb{R}^n$ qui est de dimension p.

Proposition 3.3. 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 $\mathrm{Var}_\theta(\hat{\theta}_{MC}) = \sigma^2 (X^T X)^{-1}$.

(*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.

3.3. Résultats spécifiques au modèle linéaire gaussien

Lien entre l'EMC et l'EMV dans le cas gaussien On se place dans le cas gaussien, avec nos paramètres $\theta \in \mathbb{R}^p$ et $\sigma^2 > 0$.

Proposition 3.4. 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

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

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

Dans la suite, on travaillera souvent avec un estimateur débiaisé de la variance $\hat{\sigma}^2$ défini par

$$\hat{\sigma}^2 = \frac{n-p}{n}\hat{\sigma}^2_{MV} \sim \frac{\sigma^2}{n-p}\chi^2(n-p) .$$

On a d'après ci-dessus,

$$\hat{\sigma}^2 \sim \frac{\sigma^2}{n-p}\chi^2(n-p) .$$

Formes linéaires en θ Souvent, le statisticien est intéressé par des estimations sur une parties seulement de θ (par exemple, un coefficient précis). Notons que si $a \in \mathbb{R}^p$ alors d'après la Proposition 3.4, $a^T\hat{\theta} \sim \mathcal{N}(a^T\theta, \sigma^2 a^T(X^TX)^{-1}a)$. Cela donne, lorsque l'on connaît σ^2, un intervalle de confiance pour des formes linéaires en θ.

Région de confiance pour des images de θ par des applications linéaires On rappelle qu'on a défini $\hat{\sigma}^2 = \frac{n-p}{n}\hat{\sigma}^2_{MV} \sim \frac{\sigma^2}{n-p}\chi^2(n-p)$ d'après la Proposition 3.4.

Proposition 3.5. Dans le modèle linéaire gaussien identifiable, pour tout $A \in \mathbb{R}^{q\times p}$ avec $q \leq p$ et de rang q, on a

$$F := \frac{1}{q\hat{\sigma}^2}(A\hat{\theta}_{MC} - A\theta)^T[A(X^TX)^{-1}A^T]^{-1}(A\hat{\theta}_{MC} - A\theta) \sim \mathcal{F}(q, n-p)$$

On déduit de la propriété ci-dessus que

Proposition 3.6. Dans le modèle linéaire gaussien identifiable, pour tout $A \in \mathbb{R}^{q\times p}$ avec $q \leq p$ et de rang q, une région de confiance de probabilité de couverture $1-\alpha$ pour $A\theta$ est donnée par :

$$RC_\alpha(A\theta) := \left\{ y \in \mathbb{R}^q,\ \frac{1}{q\hat{\sigma}^2}(A\hat{\theta}_{MC} - y)^T[A(X^TX)^{-1}A^T]^{-1}(A\hat{\theta}_{MC} - y) \leq f^{q,n-p}_{1-\alpha} \right\},$$

où $f^{q,n-p}_{1-\alpha}$ désigne le quantile d'ordre $1-\alpha$ de la loi $\mathcal{F}(q, n-p)$.

4. Tests d'hypothèses classiques dans le modèle linéaire gaussien

Dans toute cette partie, on travaille dans le modèle linéaire gaussien identifiable.

4.1. Test de Student

C'est lorsqu'on veut tester une relation affine des composontes de θ, de la forme $a^T\theta = c$ avec $a \in \mathbb{R}^p$. On cherche à tester

$$\mathcal{H}_0 : a^T\theta = c \quad \text{contre} \quad \mathcal{H}_1 : a^T\theta \neq c$$

On prend comme statistique de test

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

Sous $\mathcal{H}_0$, d'après l'exo-cours 3.4., on a $T \sim \mathcal{T}(n-p)$ On rejette l'hypothèse nulle avec risque α lorsque $|T| > t^{n-p}_{1-\alpha/2}$

Reamrque : si $\mathcal{H}_1$ est plutôt $a^T\theta < c$ on rejettera $\mathcal{H}_0$ si $T < t^{n-p}_\alpha$, pour augmenter la puissance du test.

4.2. Modèle emboités et test de Fisher

Modèle emboités : un exemple On a le modèle classique gaussien $Y = X\theta + \varepsilon$ et on veut tester la nullité des $q > 0$ derniers paramètres du modèle. On note $p_0 = p - q$. Le problème sécrit :

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

En terme de modèle $\theta_{p_0+1} = \ldots = \theta_p = 0$ signifie que le modèle devient $Y = X_0\theta_0 + \varepsilon_0$ avec $X_0 \in \mathbb{R}^{n\times p_0}$, de rang p_0, $\varepsilon_0 \sim \mathcal{N}(0, \sigma^2 I_n)$. On est parti d'un modèle avec $\mathbb{E}[Y] \in \Omega = Im(X)$

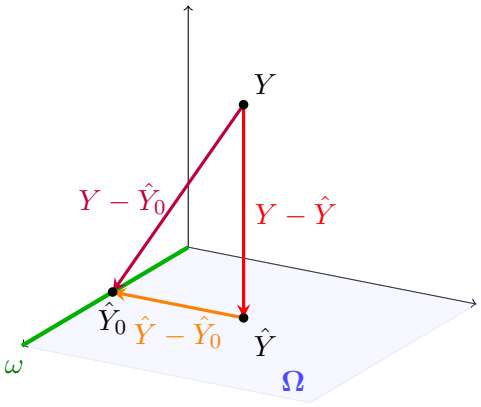


Figure 7 – *Illustration du théorème de Pythagore dans les modèles emboités.*

de dimension p, et sous $\mathcal{H}_0$, $\mathbb{E}[Y] \in \omega$, avec $\omega = Im(X_0)$ un sous espace de Ω de dimension $p_0 < p$.

Modèle emboités : intuition et interprétation géométrique . Sous $\mathcal{H}_0$, $\mathbb{E}[Y]$ appartient à un sous espace ω plus petit que Ω. Lidée est donc de projeter Y sur ce petit sous-espace. Notons $\hat{Y} = \Pi_\Omega(Y)$ et $\hat{Y}_0 = \Pi_\omega(Y)$.

L'idée intuitive est la suivante. Si la projection $\hat{Y}_0$ est "proche" de $\hat{Y}$, cela veut dire qu'avec une dimension plus faible, on explique "presque aussi bien" Y que dans le modèle plus complexe. Autant prendre moins de varaibles explicative : on conserve l'hypothèse nulle. C'est le *prinicipe de parcimonie*.

On compare donc $\|\hat{Y} - \hat{Y}_0\|^2$ à l'erreur sous le modèle standard, i.e. la SCR $\|Y - \hat{Y}\|^2$. Les vecteurs aléatoires $(\hat{Y} - \hat{Y}_0)$, resp. $(Y - \hat{Y})$ vivent dans des sous-espaces de dimension $p - p_0$ resp. $n - p$. On normalise par les dimensions qui sont les degrés de liberté. On obtient la statistique

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

On rejette l'hypothèse nulle si cette quantité est "trop grande".

On a de plus, par Pythagore

$$\underbrace{\|Y - \hat{Y}_0\|^2}_{\text{SCR petit modèle}} = \underbrace{\|\hat{Y} - \hat{Y}_0\|^2}_{\text{erreur supplémentaire entre modèles}} + \underbrace{\|Y - \hat{Y}\|^2}_{\text{SCR grand modèle}}$$

ou "petit" = moins de paramètres = plus faible dimension.

En remarquant que $Y - \hat{Y} = \Pi_{\Omega^\perp}(Y)$ et $\hat{Y}_0 = \Pi_\omega(\hat{Y})$ donc $\hat{Y} - \hat{Y}_0 = \Pi_{\omega^\perp}(\Pi_\Omega(Y)) = \Pi_{\omega^\perp \cap \Omega}(Y)$ et ces deux sous-espaces de $\mathbb{R}^n$, $\Omega^\perp$ et $\omega^\perp \cap \Omega$ sont orthogonaux. Le théorème de Cochran nous donne donc que $\|\hat{Y} - \hat{Y}_0\|^2$ et $\|Y - \hat{Y}\|^2$ sont indépendants, et qu'avec le mêmes notation que ci-dessus, on a que sous $\mathcal{H}_0$,

$$F \sim \mathcal{F}(p - p_0, n - p) = \mathcal{F}(q, n - p) .$$

Une zone de rejet de $\mathcal{H}_0$ de risque α est donc $F > f_{1-\alpha}^{q,n-p}$.

4.3. Cas du sous-modèle constant : le critère du R^2

Un cas particulier pratique est le cas où ω est le sous-modèle constant. Dans ce cas, on fait en fait un test de notre modèle contre le sous-modèle où tous les coefficients sont nuls, sauf le coefficient constant. Ce test apporte une information sur la pertinence globale du modèle, mais elle est très fortement encline à valider le modèle plutôt qu'à le rejeter.

Le modèle ω est donc juste Vect($\mathbf{1}$), et dans ce cas,

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

La variance totale de nos observations est mesurée par $\|Y - \bar{y}\mathbf{1}\|^2$. La variance expliquée par le modèle $\|\hat{Y} - \bar{y}\mathbf{1}\|^2$. La variance résiduelle est $SCR(\hat{\theta}_{MC}) = \|Y - \hat{Y}\|^2$.

On réécrit Pythagore donne (comme avant avec $\omega = \text{Vect}(\mathbf{1})$) mais on réinterprète :

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

Le *coefficient de détermination du* R^2 est défini par

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

On sait cependant qu'une façon plus équitable de comparer ces normes est de normaliser par la dimension des sous-espaces dans lesquels vivent les vecteurs. On a aussi le *coefficient de détermination ajusté* noté R^2_a défini par :

$$R^2_a = 1 - \frac{\|Y - \hat{Y}\|^2/(n-p)}{\|Y - \bar{y}\mathbf{1}\|^2/(n-1)} = 1 - \frac{n-1}{n-p}(1 - R^2) \quad (< R^2)$$

Dans `R`, le coefficient de détermination R^2 est appelé `Multiple R-Squared`, tandis que le coefficient de détermination ajusté R^2_a est appelé `Adjusted R-Squared`.

5. Pratique du modèle linéaire

5.1. Lecture des résultats sur `R`

5.2. Cas d'une variable quantitative

Présenter le modèle dans ce cas, le message quali quanti, puis la notion de référence. Equivalence statistique de deux groupes, non transitivité de cette équivalence (exemple).