Université Pierre et Marie Curie **IFMA/MPE**

Statistique de base

Ce document rassemble des résultats élémentaires de statistiques. Il reprend une grande partie des notes du cours de Bertrand Michel, et s'appuie sur le livre *Statistique en action* de Rivoirard et Stoltz (Vuibert, 2009).

1 Le Modèle statistique

Etant donnée une certaine expérience aléatoire, le statisticien construit d'abord un modèle statistique censé modéliser cette expérience. L'observation $\boldsymbol{Y}$ est le résultat de cette expérience. Si $\boldsymbol{y}$ est une réalisation de $\boldsymbol{Y}$, on aimerait s'aider de cette information pour en déduire la loi de $\boldsymbol{Y}$. Si nous ne faisons aucune hypothèse sur la loi de $\boldsymbol{Y}$, on dit que le modèle est non-paramétrique. Si on suppose que la loi de $\boldsymbol{Y}$ est de forme connue mais dépend d'un nombre fini de paramètres réels qui sont inconnus, on dira que le modèle est paramétrique.

Soit $(E, \mathcal{E})$ un espace mesurable.

Définition. *On appelle modèle statistique la donnée de* $(E, \mathcal{E}, (\mathrm{P}_\theta)_{\theta\in\boldsymbol{\Theta}})$ *où* $(\mathrm{P}_\theta)_{\theta\in\boldsymbol{\Theta}}$ *désigne une famille de lois de probabilités sur* $(E, \mathcal{E})$. *On notera* E_θ *l'espérance associée à* P_θ *et* V_θ *la variance.*

Si $\Theta \subset \mathbb{R}^d$, on dit que le modèle est *paramétrique*. Sinon le modèle est dit *non-paramétrique*.

Définition. *Une observation* $\boldsymbol{Y}$ *est une variable aléatoire à valeurs dans* $(E, \mathcal{E})$ *dont la loi appartient à la famille de lois* $(\mathrm{P}_\theta)_{\theta\in\boldsymbol{\Theta}}$.

Définition. *Lorsque l'observation* $\boldsymbol{Y}$ *a la forme* $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ *avec* $(Y_i)_{1\leq i\leq n}$ *indépendantes et identiquement distribuées, on parlera d'échantillon. Dans ce cas,* $\mathrm{P}_\theta = \mathrm{p}_\theta^{\otimes \mathrm{n}}$ *où* p_θ *est la loi de* Y_1.

Exemples.
Sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ on considère $\mathrm{P}_\theta = \mathcal{N}(\mu, \sigma^2)$ avec $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$. L'observation $\boldsymbol{Y}$ est la donnée d'une variable Gaussienne $\mathcal{N}(\mu, \sigma^2)$.

Modèle d'un lancer de pièces. On lance une pièce n fois. On a $E = \{0, 1\}^n$, $\mathcal{E}$ la tribu triviale, p_θ est la loi de Bernoulli de paramètre $\theta \in \boldsymbol{\Theta} = [0, 1]$ et $\mathrm{P}_\theta = \mathrm{p}_\theta^{\otimes \mathrm{n}}$. L'observation $\boldsymbol{Y}$ est un échantillon de Bernoulli.

Soit une certaine variable aléatoire sans atomes à valeurs dans $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ dont on cherche à connaître la loi. Dans ce cas, on peut prendre pour $\boldsymbol{\Theta}$ l'ensemble des lois sans atomes ou de manière équivalente l'ensemble des fonctions θ continues croissantes de $\mathbb{R}$ sur $[0, 1]$ avec $\theta(-\infty) = 0$ et $\theta(+\infty) = 1$, puis P_θ est la loi de fonction de répartition θ. Ce modèle est non-paramétrique.

2 Estimateurs

2.1 Premières propriétés

Soit g une fonction de $\mathbf{\Theta}$ dans $\mathbb{R}^p$.

Définition. *Un estimateur $\hat{g}$ de $g(\theta)$ est toute application mesurable en l'observation $\boldsymbol{Y}$ ne dépendant pas de θ.*

On peut donc écrire $\hat{g} = h(\boldsymbol{Y})$ pour une certaine fonction mesurable $h : (E, \mathcal{E}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}^p))$.

Définition.

- *La fonction de biais d'un estimateur $\hat{g}$ est (si elle existe) la fonction $\theta \in \mathbf{\Theta} \to b(\theta) := \mathrm{E}_\theta[\hat{g}] - \theta$.*
- *L'erreur quadratique d'un estimateur $\hat{g}$ est la fonction $\theta \in \mathbf{\Theta} \to R(\theta) := \mathrm{E}_\theta[(\hat{g} - g(\theta))^2]$. Sa décomposition biais-variance s'écrit*

$$R(\theta) = \mathrm{V}_\theta(\hat{g}) + b(\theta)^2$$

Définition. *On dit qu'un estimateur $\hat{g}$ est* **sans biais** *si $\mathrm{E}_\theta[\hat{g}] = g(\theta)$ pour tout $\theta \in \mathbf{\Theta}$.*

La définition d'un estimateur sans biais contient que $\hat{g}$ est P_θ-intégrable pour tout $\theta \in \mathbf{\Theta}$. Un estimateur sans biais donne donc en moyenne la bonne valeur de ce qu'il estime, ce qui est satisfaisant. Cependant, c'est un critère parfois contraignant (dans certains cas un estimateur sans biais n'existe pas, dans d'autres cas un estimateur "naturel" est avec biais). L'erreur quadratique mesure la dispersion des valeurs données par l'estimateur autour de $g(\theta)$, donc sa précision.

Soit un échantillon $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ et $\hat{g}_n = h_n(Y_1, \ldots, Y_n)$ un estimateur de $g(\theta)$. Les propriétés suivantes permettent de connaître l'erreur commise par l'estimation lorsque le nombre de répétitions devient grand.

Définition. *On dit que l'estimateur $\hat{g}_n$ est*

- **asymptotiquement sans biais** *si $\lim_{n\to+\infty} \mathrm{E}_\theta[\hat{g}_n] = g(\theta)$ pour tout $\theta \in \mathbf{\Theta}$.*
- **consistant** *si pour tout $\theta \in \mathbf{\Theta}$, $\hat{g}_n$ converge vers $g(\theta)$ en probabilité.*
- **fortement consistant** *si la convergence a lieu p.s.*
- **asymptotiquement normal** *si, pour tout $\theta \in \mathbf{\Theta}$, $a_n(\hat{g}_n - g(\theta))$ converge en loi vers une loi Gaussienne centrée, pour une certaine suite déterministe a_n (dépendant éventuellement de θ) telle que $\lim_{n\to+\infty} a_n = +\infty$.*

2.2 Estimateurs classiques

On donne ici deux méthodes pour trouver des estimateurs.

2.2.1 Méthode des moments

Soit $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ un échantillon. Supposons que la quantité $g(\theta)$ que l'on cherche à estimer est donnée par $E_\theta[\phi(Y_1)]$ pour une certaine fonction mesurable ϕ (ne dépendant pas de θ) avec $\phi(Y_1)$ P_θ-intégrable pour tout $\theta \in \boldsymbol{\Theta}$. Dans ce cas,

$$\hat{g}_n = \frac{1}{n}\sum_{i=1}^{n} \phi(Y_i).$$

est un estimateur fortement consistant par la loi des grands nombres. Si la variance de $\phi(Y_1)$ sous P_θ est finie pour tout $\theta \in \boldsymbol{\Theta}$, alors $\hat{g}_n$ est asymptotiquement normal par le théorème central limite, avec $a_n = \sqrt{n}$. De plus, si la fonction $\theta \to g(\theta)$ est injective, alors un estimateur de θ par la méthode des moments sera le $\hat{\theta}_n$ tel que $g(\hat{\theta}_n) = \hat{g}_n$.

Exemple. On peut ainsi estimer le moment d'ordre p (s'il existe) de p_θ $\nu_p := \mathrm{E}_\theta[Y_1^p]$ par

$$\hat{\nu}_{p,n} := \frac{1}{n}\sum_{k=1}^{n} Y_k^p.$$

Remarque. Un estimateur sans biais et fortement consistant de la variance est

$$\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{k=1}^{n}(Y_k - \overline{Y}_n)^2.$$

2.2.2 Méthode du maximum de vraisemblance

Dans cette partie, on prend $g(\theta) = \theta$ cad que l'on souhaite estimer θ. Grossièrement, étant donnée notre observation $\boldsymbol{y}$, le principe de la méthode est de trouver la valeur de θ qui maximise la probabilité de voir l'observation $\boldsymbol{y}$.

Revenons à notre modèle statistique $(E, \mathcal{E}, (\mathrm{P}_\theta)_{\theta\in\boldsymbol{\Theta}})$. On se place dans le cas paramétrique $\boldsymbol{\Theta} \subset \mathbb{R}^d$. On suppose que les lois $(\mathrm{P}_\theta, \theta \in \boldsymbol{\Theta})$ du modèle statistique sont dominées par une même mesure μ supposée σ-finie. Le théorème de Radon-Nikodym implique que P_θ a alors une densité que l'on note $L_\theta(\boldsymbol{y})$ par rapport à μ, c'est-à-dire que pour tout ensemble mesurable $A \in \mathcal{E}$, on a $\mathrm{P}_\theta(A) = \int_{\boldsymbol{y}\in A} L_\theta(\boldsymbol{y}) d\mu(\boldsymbol{y})$. Cette densité $L_\theta(\boldsymbol{y})$ est appelée la vraisemblance du modèle.

Définition. *L'estimateur du maximum de vraisemblance noté $\hat{\theta}_{\mathrm{EMV}}$ est, s'il en existe, un θ qui maximise la vraisemblance*

$$\theta \to L_\theta(\boldsymbol{Y}).$$

Dans le cas d'un échantillon $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, on supposera que la loi p_θ de Y_1 (donc aussi des Y_i) a une densité $s_\theta(y_1)$ par rapport à une mesure de référence σ-finie $\mu(dy_1)$. La vraisemblance s'écrira alors

$$L_\theta(\boldsymbol{y}) = L_\theta(y_1, \ldots, y_n) = \prod_{i=1}^{n} s_\theta(y_i)$$

et l'estimateur du maximum de vraisemblance sera le θ qui maximise la fonction $\theta \to L_\theta(\boldsymbol{Y})$.

Sous des hypothèses de régularité, on peut montrer que l'estimateur du maximum de vraisemblance est asymptotiquement normal.

2.3 Estimateur de variance minimale

Supposons que l'on veuille estimer $g(\theta)$. On aimerait comparer plusieurs estimateurs. Une idée naturelle est de comparer l'erreur commise par ces estimateurs.

Définition. *Un estimateur sans biais de variance minimale (ESBVM) est un estimateur $\hat{g}_{ESBVM}$ sans biais qui minimise l'erreur quadratique. En d'autres termes,*

$$\mathrm{V}_\theta(\hat{g}_{ESBVM}) \leq \mathrm{V}_\theta(\hat{g})$$

pour tout $\theta \in \boldsymbol{\Theta}$ et tout $\hat{g}$ estimateur sans biais de $g(\theta)$. S'il existe, un tel estimateur est nécessairement unique.

Remarque. Rien ne nous dit qu'on ne peut pas trouver un estimateur biaisé avec une erreur quadratique plus petite.

La suite de cette partie montre que l'on a une borne inférieure sur la variance des estimateurs sans biais, appelée *borne de Cramer-Rao*. Cette borne inférieure dépend de la quantité d'information apportée par les observations telle que mesurée par *l'information de Fisher*.

On se place désormais dans le cas paramétrique. On a donc $\boldsymbol{\Theta} \subset \mathbb{R}^p$ et supposons pour simplifier que $p = 1$.

Définition. *L'information de Fisher est définie par*

$$I(\theta) := \mathrm{E}_\theta \left[(\partial_\theta \ln L_\theta(\boldsymbol{Y}))^2 \right]$$

lorsque cette quantité existe. Elle peut se réécrire sous certaines hypothèses de régularité

$$I(\theta) = -\mathrm{E}_\theta \left[\partial_\theta^2 \ln L_\theta(\boldsymbol{Y}) \right].$$

Dans le cas d'un échantillon $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, on obtient (sous hypothèses de régularité) que l'information de Fisher du modèle est donnée par

$$I_n(\theta) = nI_1(\theta)$$

avec $I_1(\theta) = \mathrm{E}_\theta \left[(\partial_\theta \ln s_\theta(Y_1))^2 \right]$, qui s'écrit encore $-\mathrm{E}_\theta \left[\partial_\theta^2 \ln s_\theta(Y_1) \right]$.

Intuitivement, l'information de Fisher est une mesure de l'information contenue dans l'observation $\boldsymbol{Y}$. Plus $I(\theta)$ est élevée, meilleure sera l'information et donc plus précis pourront être les estimateurs. Cette heuristique se retrouve dans le théorème suivant qui dit que l'erreur commise par un estimateur sans biais est bornée inférieurement par l'inverse de l'information de Fisher.

Théorème (Borne de Cramer-Rao). *Soit g une fonction $\boldsymbol{\Theta} \to \mathbb{R}$ dérivable, et $\hat{g} = h(\boldsymbol{Y})$ un estimateur sans biais de $g(\theta)$. Sous certaines hypothèses de régularité, on a*

$$\mathrm{V}_\theta(\hat{g}) \geq \frac{(g'(\theta))^2}{I(\theta)}.$$

La borne inférieure de cette inégalité est appelée borne de Cramer-Rao.

Un estimateur sans biais qui atteint la borne de Cramer-Rao est nécessairement de variance minimale. Un tel estimateur est dit *efficace*.

Dans le cas d'un échantillon, si $\hat{g}_n$ est sans biais et vérifie quand $n \to +\infty$, $\mathrm{V}_\theta(\hat{g}_n) \sim \frac{g'(\theta)^2}{I_n(\theta)}$, on dit que l'estimateur est *asymptotiquement efficace*. Sous des hypothèses de régularité, l'estimateur du maximum de vraisemblance est asymptotiquement efficace.

L'existence d'un estimateur efficace est en fait liée à la forme de la vraisemblance du modèle.

Théorème. *(Sous certaines hypothèses de régularité). Un estimateur efficace existe si et seulement si la vraisemblance vérifie*

$$\ln L_\theta(\boldsymbol{y}) = a(\boldsymbol{y})\alpha(\theta) + b(\boldsymbol{y}) + \beta(\theta).$$

Dans ce cas, $g(\theta) = -\frac{\beta'(\theta)}{\alpha'(\theta)}$ admet un estimateur efficace qui est $\hat{g} := a(\boldsymbol{Y})$. C'est l'unique paramètre (à une transformation linéaire près) admettant un estimateur efficace.

Dans le cas d'un échantillon, on peut montrer que l'estimateur du maximum de vraisemblance est asymptotiquement efficace.

Théorème. *(Sous certaines hypothèses de régularité) L'estimateur de maximum de vraisemblance est asymptotiquement efficace.*

3 Intervalles de confiance

3.1 Définition

Soit $(E, \mathcal{E}, (\mathrm{P}_\theta)_{\theta\in\boldsymbol{\Theta}})$ un modèle statistique et $\boldsymbol{Y}$ une observation. On souhaite estimer le paramètre $g(\theta)$.

Définition. *Soit $\alpha \in [0,1]$. On appelle région de confiance au niveau $1-\alpha$ de $g(\theta)$ un ensemble $\hat{\mathcal{C}}$ construit mesurablement par rapport à $\boldsymbol{Y}$, ne dépendant pas de θ, et tel que pour tout $\theta \in \boldsymbol{\Theta}$,*

$$\mathrm{P}_\theta\left(g(\theta) \in \hat{\mathcal{C}}\right) \geq 1-\alpha.$$

Remarque. Dire que $\hat{\mathcal{C}}$ est construit mesurablement signifie que pour tout $\theta \in \boldsymbol{\Theta}$, l'évènement $\{g(\theta) \in \hat{\mathcal{C}}\}$ est mesurable. Si l'inégalité est en fait une égalité dans la définition précédente, on parle de niveau exact.

Remarque. Lorsque $\hat{\mathcal{C}}$ est un intervalle, on parlera plutôt d'intervalle de confiance.

Définition. *Dans le cas d'un échantillon $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, on appelle région (resp. intervalle) de confiance asymptotique au niveau $1-\alpha$ de $g(\theta)$ un ensemble (resp. un intervalle) $\hat{\mathcal{C}}_n$ construit mesurablement par rapport à $\boldsymbol{Y}$, ne dépendant pas de θ, et tel que pour tout $\theta \in \boldsymbol{\Theta}$,*

$$\liminf_{n\to+\infty} \mathrm{P}_\theta\left(g(\theta) \in \hat{\mathcal{C}}_n\right) \geq 1-\alpha.$$

3.2 Méthode du pivot

La méthode du pivot consiste à trouver une fonction $f(\boldsymbol{y}, g(\theta))$ mesurable en $\boldsymbol{y} \in E$ dont la loi sous P_θ ne dépend pas de θ. On cherche ensuite a, b tels que $P_\theta(f(\boldsymbol{Y}, g(\theta)) \in [a, b]) \geq 1 - \alpha$. La région de confiance est alors déterminée par $\hat{\mathcal{C}} := \{g(\theta) \,:\, f(\boldsymbol{Y}, g(\theta)) \in [a, b]\}$.

Exemple. On veut estimer la moyenne μ d'une loi Gaussienne $\mathcal{N}(\mu, \sigma^2)$ où σ^2 est connue. Pour cela on a accès à un échantillon $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$. Un estimateur est donné par

$$\hat{\mu}_n = \overline{Y}_n := \frac{1}{n}\left(Y_1 + \ldots + Y_n\right).$$

On sait que $\frac{\hat{\mu}_n - \mu}{\sigma/\sqrt{n}}$ suit une loi Gaussienne $\mathcal{N}(0,1)$. En notant z_β le quantile d'ordre β de la loi Gaussienne centrée réduite, on obtient $\hat{\mathcal{C}}_n := [\hat{\mu}_n \pm \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}]$ pour un intervalle de confiance bilatère de niveau $1-\alpha$. Un intervalle de confiance unilatère serait $\hat{\mathcal{C}}_n :=]-\infty; \hat{\mu}_n + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}]$ (à gauche) ou $[\hat{\mu}_n - \frac{\sigma}{\sqrt{n}} z_{1-\alpha}; +\infty[$ (à droite).

Exemple. Soit un échantillon $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ de v.a. moyenne μ et variance σ^2 toutes deux inconnues. On ne suppose plus que les v.a. suivent une loi Gaussienne. Le théorème central limite dit que $\frac{\hat{\mu}_n - \mu}{\sigma/\sqrt{n}}$ converge en loi vers $\mathcal{N}(0,1)$. On estime σ par son estimateur $\hat{\sigma}_n$. Le lemme de Slutsky entraine que $\frac{\hat{\mu}_n - \mu}{\hat{\sigma}_n/\sqrt{n}}$ converge en loi vers $\mathcal{N}(0,1)$. Cela permet d'obtenir l'intervalle de confiance asymptotique de niveau $1-\alpha$ pour μ:

$$\left[\hat{\mu}_n - \frac{\hat{\sigma}_n}{\sqrt{n}} z_{1-\alpha/2}; \hat{\mu}_n + \frac{\hat{\sigma}_n}{\sqrt{n}} z_{1-\alpha/2}\right].$$

3.3 Utilisation d'une inégalité de probabilité

Supposons que l'on veuille estimer la moyenne $\mu(\theta)$ d'une loi de probabilité dont on sait que la variance est bornée par une constante connue M^2. Pour cela on a accès à un échantillon $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$. Un estimateur sans biais est donné par

$$\hat{\mu}_n = \overline{Y}_n = \frac{1}{n}\left(Y_1 + \ldots + Y_n\right).$$

L'inégalité de Bienaymé-Tchebychev implique que

$$\mathrm{P}_\theta\left(|\hat{\mu}_n - \mu| > t\right) \leq \frac{\mathrm{V}_\theta(\hat{\mu}_n)}{t^2} \leq \frac{M^2}{nt^2}.$$

En choisissant t tel que $\frac{M^2}{nt^2} = \alpha$, on obtient l'intervalle de confiance de niveau $1-\alpha$ de μ suivant

$$\left[\hat{\mu}_n - \frac{M}{\sqrt{\alpha n}}; \hat{\mu}_n + \frac{M}{\sqrt{\alpha n}}\right].$$

Exemple. Intervalle de confiance pour le paramètre d'un échantillon de Bernoulli.

3.4 Intervalle de confiance et réalisation d'un intervalle de confiance

Il est important de garder à l'esprit qu'un intervalle de confiance est une quantité aléatoire. Dans la pratique, c'est-à-dire pour les données étudiées, les bornes de l'intervalle de confiance construit sont la réalisation de l'intervalle de confiance pour les observations. Par exemple, dans le cas de l'échantillon Gaussien i.i.d., on a vu que

$$\left[\hat{\mu}_n \pm \frac{\hat{\sigma}_n t_{0,975}}{\sqrt{n}}\right]$$

est un intervalle de confiance de la moyenne au niveau 95 %. Si $n = 100$, $\hat{\mu}_n = 7,45$ et $\hat{\sigma}_n = 1,21$, la réalisation de cet intervalle vaut $[6,76; 7,24]$. Par abus, on dira parfois que $[6,76; 7,24]$ est l'intervalle de confiance de la moyenne, mais en toute rigueur cela n'a pas de sens : soit la moyenne est dans $[6,76; 7,24]$, soit elle ne l'est pas, la probabilité que la moyenne y soit est donc 0 ou 1!

3.5 Intervalles de confiance simultanés

Supposons que l'on souhaite construire une région de confiance pour un vecteur

$$g(\theta) = (g_1(\theta), g_2(\theta), \ldots, g_k(\theta)).$$

Supposons que l'on dispose pour chaque $g_i(\theta)$ pris séparément d'un intervalle de confiance $\hat{I}_j$ de niveau de confiance $1 - \frac{\alpha}{k}$. Alors $\hat{I}_1 \times \hat{I}_2 \times \ldots \times \hat{I}_k$ est une région de confiance pour le vecteur $g(\theta)$ de niveau $1 - \alpha$ (méthode de Bonferroni). En effet, pour tout j,

$$\mathrm{P}_\theta\left(g_j(\theta) \notin \hat{I}_j\right) \leq \frac{\alpha}{k}.$$

D'où $\mathrm{P}_\theta\left(\bigcup_{j=1}^k \{g_j(\theta) \notin \hat{I}_j\}\right) \leq k\frac{\alpha}{k}$ et donc

$$\mathrm{P}_\theta\left(\bigcap_{j=1}^k \{g_j(\theta) \in \hat{I}_j\}\right) \geq 1 - \alpha.$$

Notons que cette méthode n'est intéressante que pour de petites valeurs de k. Pour de grandes valeurs de k, la région de confiance $\hat{I}_{1,1-\frac{\alpha}{k}} \times \hat{I}_{2,1-\frac{\alpha}{k}} \times \ldots \times \hat{I}_{k,1-\frac{\alpha}{k}}$ a une probabilité beaucoup plus grande que $1 - \alpha$: elle encadre $g(\theta)$ beaucoup trop largement.

4 Tests d'hypothèses

4.1 Introduction

Exemple Une variété de souris présente des cancers spontanés avec un taux de 20%. On aimerait connaître l'efficacité d'un traitement. On fait donc l'expérience sur un échantillon de 200 souris, et on aimerait décider si le traitement a un effet ou non.

On formule donc deux hypothèses:

- $\mathcal{H}_0$: le traitement est sans effet.
- $\mathcal{H}_1$: le traitement a un effet.

On veut choisir entre l'hypothèse $\mathcal{H}_0$ (*hypothèse nulle*) et $\mathcal{H}_1$ (*hypothèse alternative*). Pour cela, on veut effectuer un test, c'est-à-dire que l'on veut se donner une règle pour trancher entre les deux hypothèses en fonction des résultats de l'expérience.

L'étape du choix des hypothèses n'est pas anodine. Les hypothèses nulle et alternative ne sont pas symétriques. On privilégie l'hypothèse $\mathcal{H}_0$: je vais rejeter $\mathcal{H}_0$ si je suis vraiment sûr que $\mathcal{H}_0$ est fausse. Les deux hypothèses ne jouant pas le même rôle, dans la pratique on choisira pour $\mathcal{H}_0$ l'hypothèse "la plus raisonnable", la plus communément admise, l'hypothèse privilégiée par un parti pris subjectif ou encore l'hypothèse la plus simple. Ainsi, dans un même contexte où l'on souhaite confronter deux hypothèses $\mathcal{H}_a$ et $\mathcal{H}_b$, deux tests statistiques définis par $\mathcal{H}_0 = \mathcal{H}_a$, $\mathcal{H}_1 = \mathcal{H}_b$ pour le premier et par $\mathcal{H}_0 = \mathcal{H}_b$, $\mathcal{H}_1 = \mathcal{H}_a$ pour le second, peuvent aboutir à des conclusions différentes. On verra une illustration de ce phénomène dans le cas du test de Student.

Une fois que je rejette $\mathcal{H}_0$, j'accepte alors l'hypothèse $\mathcal{H}_1$. Sinon, j'accepte $\mathcal{H}_0$, ce qui ne veut pas dire que $\mathcal{H}_0$ est nécessairement vraie. Cela voudra juste dire que je n'ai pas assez de certitude pour préférer $\mathcal{H}_1$ à $\mathcal{H}_0$. C'est pourquoi l'on dit qu'un test est *significatif* lorsque la conclusion du test est que l'on rejette $\mathcal{H}_0$.

La démarche.
On suppose que $\mathcal{H}_0$ est vraie. Dans ce cas, le nombre de souris malades suit une loi binômiale $\mathcal{B}(N, p)$ avec $N = 200$ et $p = 0.2$. La proportion $\hat{p}$ de souris malades est une *statistique de test*. C'est une variable aléatoire dont on peut identifier la distribution sous $\mathcal{H}_0$. C'est elle qui va nous dire si $\mathcal{H}_0$ est acceptable ou non. On calcule que la probabilité que $\hat{p}$ (dont on connaît la distribution) se trouve dans l'intervalle $[0.1984, 0.2016]$ est de 95%.

La règle de décision sera donc la suivante: si $\hat{p} \notin [0.1984, 0.2016]$, je rejette $\mathcal{H}_0$ et choisis donc $\mathcal{H}_1$. Sinon je ne rejette pas $\mathcal{H}_0$.

La probabilité de rejeter $\mathcal{H}_0$ alors qu'elle est vraie est, par construction, égale à $\alpha = 5\%$. On appelle α le *risque de première espèce*. C'est ce risque que l'on cherche à minimiser dans un premier temps.

La probabilité β d'accepter $\mathcal{H}_0$ à tort est appelée *risque de deuxième espèce*. La puissance est égale à $1 - \beta$. À α fixé, on cherchera un test le plus puissant possible.

L'hypothèse $\mathcal{H}_1$ influe sur la règle de décision. Supposons que l'on prenne maintenant pour $\mathcal{H}_1$ l'hypothèse que le traitement est bénéfique, cad que la proportion de souris malades diminue avec le traitement. Notre test d'hypothèses est donc:

- $\mathcal{H}_0$ le traitement est sans effet.
- $\mathcal{H}_1$ le traitement a un effet bénéfique.

Il va falloir choisir entre $\mathcal{H}_0$ et $\mathcal{H}_1$ (même dans le cas où ni l'une ni l'autre ne sont vraies!). On pourrait utiliser la règle de décision précédente. Cependant, on sera amené à accepter l'hypothèse $\mathcal{H}_1$ lorsque $\hat{p}$ est grand, ce qui est contre-intuitif: on ne veut pas dire que le traitement est bénéfique si le nombre de souris malades est plus élevé que la normale! On va plutôt rejeter $\mathcal{H}_0$ et donc accepter $\mathcal{H}_1$ si la proportion de souris malades est faible. Cela donnera un test plus puissant. Sous l'hypothèse $\mathcal{H}_0$, la probabilité que $\hat{p} \leq 0.1986$ est de 5%. La règle de décision sera donc: si $\hat{p} \leq 0.1986$ je rejette H_0 (et choisis donc $\mathcal{H}_1$), sinon je choisis $\mathcal{H}_0$. C'est ce qu'on appelle un test *unilatéral*, comparé au précédent qui était un test *bilatéral*.

4.2 Définitions

Soit $(E, \mathcal{E}, (\mathrm{P}_\theta)_{\theta\in\boldsymbol{\Theta}})$ un modèle statistique. Soit $\boldsymbol{\Theta}_0$ et $\boldsymbol{\Theta}_1$ deux sous-ensembles disjoints de $\boldsymbol{\Theta}$. On souhaite tester $\mathcal{H}_0 : \theta \in \boldsymbol{\Theta}_0$ contre $\mathcal{H}_1 : \theta \in \boldsymbol{\Theta}_1$. Lorsque $\boldsymbol{\Theta}_i$ est réduit à un singleton, on parle d'hypothèse simple (par opposition avec hypothèse composite). Dans l'exemple précédent, on avait P_θ loi d'un échantillon de Bernoulli de paramètre $\theta \in \boldsymbol{\Theta} = [0, 1]$, $\boldsymbol{\Theta}_0 = \{0.2\}$ et $\boldsymbol{\Theta}_1 = \boldsymbol{\Theta}\backslash\{0.2\}$ puis $\boldsymbol{\Theta}_1 =]0.2, 1]$.

Définition. *Un test est une application mesurable $\Phi : E \to \{0, 1\}$ associée à la stratégie suivante: pour l'observation $\boldsymbol{y}$, $\mathcal{H}_1$ est acceptée si $\Phi(\boldsymbol{y}) = 1$ et $\mathcal{H}_0$ est acceptée si $\Phi(\boldsymbol{y}) = 0$.*

La région $\mathcal{R} = \{\boldsymbol{y} : \Phi(\boldsymbol{y}) = 1\}$ est la zone de rejet du test. On rejettera ainsi $\mathcal{H}_0$ si $\boldsymbol{Y} \in \mathcal{R}$ et on l'acceptera sinon.

Définition. *L'erreur de première espèce du test Φ est l'application*

$$\begin{aligned} \boldsymbol{\Theta}_0 &\to [0, 1] \\ \theta &\to \mathrm{P}_\theta(\Phi = 1) \end{aligned}$$

Le niveau du test Φ est défini par $\alpha = \sup_{\theta\in\boldsymbol{\Theta}_0} \mathrm{P}_\theta(\Phi = 1)$.

Avoir un niveau α signifie que la probabilité de rejeter à tort l'hypothèse $\mathcal{H}_0$ est inférieure à α.

Définition. *L'erreur de seconde espèce du test Φ est l'application*

$$\begin{aligned} \boldsymbol{\Theta}_1 &\to [0, 1] \\ \theta &\to \mathrm{P}_\theta(\Phi = 0) \end{aligned}$$

La puissance du test Φ est l'application

$$\begin{aligned} \boldsymbol{\Theta}_1 &\to [0, 1] \\ \pi : \theta &\to \mathrm{P}_\theta(\Phi = 1). \end{aligned}$$

La fonction puissance est la probabilité d'accepter $\mathcal{H}_1$ quand celle-ci est vraie. Elle mesure donc la qualité du test. En général, diminuer l'erreur de première espèce se fait au détriment de la puissance. La démarche consiste à se fixer d'abord un niveau du test puis à trouver un test maximisant la puissance.

Exemple. On considère un échantillon $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ de v.a. indépendantes de loi $\mathcal{N}(\theta, \sigma^2)$ avec $\theta \in \mathbb{R}$ et σ^2 connu. On veut tester $\mathcal{H}_0 : \theta = \theta_0$ contre $\mathcal{H}_1 : \theta \neq \theta_0$ où θ_0 est un réel fixé. Prenons le test $\Phi = \mathbf{1}_{\{|\sqrt{n}(\overline{Y}_n - \theta_0)| > z_{1-\alpha/2}\sigma\}}$. Alors le niveau du test est

$$\mathrm{P}_{\theta_0}(\Phi = 1) = \mathrm{P}_{\theta_0}(|\sqrt{n}(\overline{Y}_n - \theta_0)| > z_{1-\alpha/2}\sigma) = \mathrm{P}_{\theta_0}(|Z| > z_{1-\alpha/2}) = \alpha$$

où Z est une variable Gaussienne centré réduite et z_β le quantile β de sa loi. La fonction puissance est donnée par

$$\mathrm{P}_\theta(\Phi = 1) = \mathrm{P}_\theta\left(|\sqrt{n}(\overline{Y}_n - \theta_0)| > z_{1-\alpha/2}\sigma\right) = \mathrm{P}_\theta\left(\left|Z + \frac{\theta - \theta_0}{\sigma\sqrt{n}}\right| > z_{1-\alpha/2}\right)$$

où Z est encore une variable Gaussienne centrée réduite.

Exemple. Si on teste plutôt $\mathcal{H}_0 : \theta = \theta_0$ contre $\mathcal{H}_1 : \theta > \theta_0$, on considèrera le test $\Phi = \mathbf{1}_{\{\sqrt{n}(\overline{Y}_n - \theta_0) > z_{1-\alpha}\sigma\}}$ qui est encore de niveau α. Sa fonction puissance est donnée par

$$\mathrm{P}_\theta(\Phi = 1) = \mathrm{P}_\theta(\sqrt{n}(\overline{Y}_n - \theta_0) > z_{1-\alpha}\sigma) = \mathrm{P}_\theta\left(Z + \frac{\theta - \theta_0}{\sigma\sqrt{n}} > z_{1-\alpha}\right).$$

On peut montrer que ce test (unilatéral) est en effet plus puissant que le test précédent (bilatéral). On verra dans la partie 4.5 qu'il est en fait plus puissant que n'importe quel test de niveau α. De plus, si on teste $\mathcal{H}_0 : \theta \leq \theta_0$ contre $\mathcal{H}_1 : \theta > \theta_0$, le même test $\Phi = \mathbf{1}_{\{\sqrt{n}(\overline{Y}_n - \theta_0) > z_{1-\alpha}\sigma\}}$ reste de niveau α et le plus puissant.

Remarque. Il y a un lien étroit entre intervalle de confiance et tests statistiques. Supposons que l'on veuille tester $\mathcal{H}_0 : \theta = \theta_0$ contre $\mathcal{H}_1 : \theta \neq \theta_0$. Alors il suffit de choisir un intervalle de confiance I pour θ de niveau de confiance $1-\alpha$. On prendra alors $\Phi = \mathbf{1}_{\{\theta_0 \notin I\}}$ comme test de niveau (inférieur à) α.

4.3 Tests asymptotiques

On considère le cas d'un échantillon $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$. On teste encore $\theta \in \Theta_0$ contre $\theta \in \Theta_1$.

Définition. *Un test Φ_n est dit asymptotiquement de niveau α si* $\sup_{\theta \in \boldsymbol{\Theta}_0} \limsup_{n \to +\infty} \mathrm{P}_\theta(\Phi_n = 1) = \alpha$.

Définition. *Un test Φ_n est dit convergent si* $\lim_{n \to +\infty} \mathrm{P}_\theta(\Phi_n = 1) = 1$ *pour tout* $\theta \in \boldsymbol{\Theta}_1$.

Cela veut dire que l'on pourra trancher entre les deux hypothèses avec de plus en plus de certitude si le nombre d'observations est suffisant.

Soit $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ un échantillon de moyenne μ et variance σ^2. On aimerait tester l'égalité de la moyenne à un certain réel μ_0.

$$\mathcal{H}_0 : \mu = \mu_0 \qquad \textit{contre} \qquad \mathcal{H}_1 : \mu \neq \mu_0$$

On utilise la statistique de test (avec $\hat{\sigma}_n$ estimateur de l'écart-type),

$$Z_n = \frac{\overline{Y}_n - \mu_0}{\hat{\sigma}_n / \sqrt{n}}.$$

La variable Z_n converge en loi vers $\mathcal{N}(0,1)$ (par le théorème central limite et le lemme de Slutsky). Donc le test

$$\Phi = \mathbf{1}_{\{|Z_n| > z_{1-\alpha/2}\}}$$

est asymptotiquement de niveau α.

Exercice. Faire un test asymptotique $\mathcal{H}_0 : \theta = 1/2$ contre $\mathcal{H}_1 : \theta \neq 1/2$ dans le cas d'un échantillon de Bernoulli de paramètre $\theta \in [0,1]$ (*Test de proportion*).

4.4 Pratique et interprétation des tests

Plutôt que de donner une réponse définitive telle que "j'accepte $\mathcal{H}_0$" ou telle que " je rejette $\mathcal{H}_0$", il est préfèrable de "mesurer" la facilité avec laquelle $\mathcal{H}_0$ peut être acceptée. Cette mesure est donnée par la *probabilité critique* (aussi appelée *niveau de significativité* ou encore *p-value*) :

$$\begin{aligned} p &:= \max\{\alpha \in [0,1] \,|\, \text{le test accepte } \mathcal{H}_0 \text{ au niveau } \alpha\} \\ &= \max\{\alpha \in [0,1] \,|\, X \notin \mathcal{R}_\alpha\} \\ &= \min\{\alpha \in [0,1] \,|\, \text{le test rejette } \mathcal{H}_0 \text{ au niveau } \alpha\} \\ &= \min\{\alpha \in [0,1] \,|\, X \in \mathcal{R}_\alpha\} \end{aligned}$$

Il s'agit donc du niveau limite, pour lequel la réponse du test change. Notons que la définition donnée ci-dessus sous-entend que la région de rejet $\mathcal{R}_\alpha$ croît avec α, ce qui est naturel. Les logiciels statistiques fournissent comme réponse à un test la p-value correspondante, il appartient à l'utilisateur d'interpréter cette valeur :

- Supposons que la p-value soit de 0.65, on accepte alors $\mathcal{H}_0$ sans problème pour des niveaux standards (par exemple de l'ordre de 0.05).
- Supposons que la p-value d'un test soit de 10^{-2}, on rejette alors $\mathcal{H}_0$ pour des niveaux standards.

De façon générale, accepter $\mathcal{H}_0$ ne constitue pas une validation radicale de l'hypothèse nulle car les tests sont généralement trop conservatifs (ils privilégient $\mathcal{H}_0$). Accepter $\mathcal{H}_0$ signifie uniquement que rien d'anormal n'a été détecté dans les données qui ne contredise l'hypothèse nulle. À l'inverse, on rejette $\mathcal{H}_0$ lorsque les observations (ou une statistique) prennent une valeur extrême sous l'hypothèse de loi imposée par $\mathcal{H}_0$. En fin de compte, la réponse à un test n'est donc claire et franche que lorsque la p-value est très faible ; on dit alors que le résultat du test est *significatif*.

L'étude complète d'un test d'hypothèse se déroule de la façon suivante :

1. Choix des hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$,
2. Détermination d'une statistique adaptée pour le problème posé, choix d'une région de rejet,
3. Étude de la fonction puissance,
4. Calcul de la p-value pour les données observées,
5. Décision finale, interprétation.

4.5 Test du rapport de vraisemblance

Soit $(E, \mathcal{E}, (\mathrm{P}_\theta)_{\theta\in\mathbf{\Theta}})$ le modèle statistique, avec $\boldsymbol{Y}$ l'observation. On rappelle que $L_\theta(\mathbf{y})$ désigne la vraisemblance du modèle. Le test du rapport de vraisemblance s'appuie sur la statistique de test

$$h(\boldsymbol{Y}) = \frac{\sup_{\theta\in\mathbf{\Theta}_1} L_\theta(\boldsymbol{Y})}{\sup_{\theta\in\mathbf{\Theta}_0} L_\theta(\boldsymbol{Y})}.$$

Sous l'hypothèse $\mathcal{H}_1$, le numérateur prendra une valeur importante alors que le dénominateur sera faible: au final la statistique $h(\boldsymbol{Y})$ sera grande. A l'inverse, sous l'hypothèse $\mathcal{H}_0$, la statistique $h(\boldsymbol{Y})$ sera proche de 0. Le test du rapport de vraisemblance s'écrit

$$\Phi := \mathbf{1}_{\{h(\boldsymbol{Y})>k_\alpha\}}$$

où k_α est une constante à fixer selon le seuil α. Le lemme suivant montre que ce test est optimal lorsque les hypothèses sont des hypothèses simples: $\mathcal{H}_0 : \theta = \theta_0$ contre $\mathcal{H}_1 : \theta = \theta_1$ où θ_0 et θ_1 sont 2 éléments de $\mathbf{\Theta}$.

Lemme de Neyman-Pearson. *Pour $\alpha \in]0;1[$, s'il existe k_α tel que le test*

$$\Phi = \mathbf{1}_{\{L_{\theta_1}(\boldsymbol{Y}) > k_\alpha L_{\theta_0}(\boldsymbol{Y})\}}$$

soit de niveau α, alors ce test est le plus puissant des tests de niveau inférieur à α, cad que pour tout test Φ' de niveau inférieur à α, on a $\mathrm{P}_{\theta_1}(\Phi' = 1) \leq \mathrm{P}_{\theta_1}(\Phi = 1)$.

Remarque. Le test de vraisemblance ne s'exprime pas forcément en fonction de la vraisemblance. Il est possible que le test puisse s'écrire en fonction d'une statistique de test S plus simple.

Exemple. Soit $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ un échantillon i.i.d. de loi $\mathcal{N}(\theta, \sigma^2)$ avec θ inconnu et σ^2 connue. On teste $\theta = \theta_0$ contre $\theta = \theta_1$ avec $\theta_0 < \theta_1$. On a $\ln L_\theta(\boldsymbol{y}) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{k=1}^n (y_k - \theta)^2$. Or, pour toute constante c,

$$\begin{aligned} L_{\theta_1}(\boldsymbol{y}) > cL_{\theta_0}(\boldsymbol{y}) &\Leftrightarrow \ln L_{\theta_1}(\boldsymbol{y}) > c_1 + \ln L_{\theta_0}(\boldsymbol{y}) \\ &\Leftrightarrow \sum_{k=1}^n (y_k - \theta_1)^2 > c_2 + \sum_{k=1}^n (y_k - \theta_0)^2 \\ &\Leftrightarrow \sum_{k=1}^n y_k > c_3 \Leftrightarrow \frac{\overline{y}_n - \theta_0}{\sigma/\sqrt{n}} > c_4 \end{aligned}$$

où c_1, c_2, c_3, c_4 sont des constantes ne dépendant pas de $\boldsymbol{y}$ (mais pouvant dépendre de θ_0, θ_1, n, σ^2). On voit donc que le test du rapport de vraisemblance se réécrit

$$\Phi = \mathbf{1}\left\{\frac{\overline{Y}_n - \theta_0}{\sigma/\sqrt{n}} > z_{1-\alpha}\right\}$$

où la constante $c_4 = z_{1-\alpha}$ a été choisie de sorte que Φ soit de niveau α.

Remarque. Le lemme de Neyman-Pearson permet d'avoir un test dans le cas d'hypothèses simples. Dans certains cas, il est possible d'en déduire un test pour des hypothèses composites.

Proposition. *Soit θ_0 fixé et $\mathbf{\Theta}_0$ contenant θ_0. Si le test fourni par le lemme de Neyman-Pearson ne dépend pas de $\theta_1 \in \mathbf{\Theta}_1$ et vérifie $\sup_{\theta \in \mathbf{\Theta}_0} \mathrm{P}_\theta(\Phi = 1) \leq \alpha$, alors c'est aussi un test de $\theta \in \mathbf{\Theta}_0$ contre $\theta \in \mathbf{\Theta}_1$ et il est uniformément plus puissant parmi les tests de niveau inférieur à α (UPP(α)).*

Dire que le test de vraisemblance ne dépend pas de θ_1 signifie que le test de vraisemblance de $\theta = \theta_0$ contre $\theta = \theta_1$ est le même quel que soit $\theta_1 \in \mathbf{\Theta}_1$. Dire que Φ est uniformément plus puissant parmi les tests de niveau inférieur à α siginifie que pour tout test Φ' de niveau inférieur à α de $\theta \in \mathbf{\Theta}_0$ contre $\theta \in \mathbf{\Theta}_1$, on a $\mathrm{P}_{\theta_1}(\Phi' = 1) \leq \mathrm{P}_{\theta_1}(\Phi = 1)$ pour tout $\theta_1 \in \mathbf{\Theta}_1$. Remarquons qu'il n'y a pas de raison qu'un tel test existe forcément.

Exemple. Reprenons le cas de l'échantillon i.i.d. $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ de loi $\mathcal{N}(\theta, \sigma^2)$ avec θ inconnu et σ^2 connue. Soit θ_0 fixé, $\mathbf{\Theta}_0 = \{\theta \leq \theta_0\}$ et $\mathbf{\Theta}_1 = \{\theta > \theta_0\}$. D'après l'exemple précédent le test de vraisemblance s'écrit

$$\Phi = \mathbf{1}\left\{\frac{\overline{Y}_n - \theta_0}{\sigma/\sqrt{n}} > z_{1-\alpha}\right\}.$$

On peut remarquer qu'il ne dépend pas de $\theta_1 > \theta_0$ et on a $\mathrm{P}_\theta(\Phi = 1) \leq \alpha$ pour tout $\theta \leq \theta_0$. On déduit que Φ est $UPP(\alpha)$.

5 Echantillons Gaussiens

On présente dans cette partie les tests statistiques classiques associés à des échantillons Gaussiens.

5.1 Vecteurs Gaussiens

5.1.1 Définition et premières propriétés

Définition. *On appelle vecteur Gaussien de $\mathbb{R}^d$ un vecteur aléatoire $\boldsymbol{Y} = (Y_1, \ldots, Y_d) \in \mathbb{R}^d$ tel que toute combinaison linéraire des composantes suit une loi Gaussienne. On notera $\boldsymbol{m}$ son vecteur moyenne et $\boldsymbol{\Sigma}$ sa matrice de covariance.*

On a donc $\boldsymbol{m} = (\mathrm{E}[Y_1], \ldots, \mathrm{E}[Y_d])'$, $\boldsymbol{\Sigma} = \mathrm{E}[(\boldsymbol{Y} - \boldsymbol{m})(\boldsymbol{Y} - \boldsymbol{m})'] = (Cov(Y_i, Y_j))_{1 \leq i,j \leq d}$.

Exemple. Le vecteur $\boldsymbol{Y} = (Y_1, \ldots, Y_d)$ composé de d variables Gaussiennes i.i.d. est un vecteur Gaussien.

Proposition. *Si A est une matrice $p \times d$ et b un vecteur dans $\mathbb{R}^p$, alors $A\boldsymbol{Y} + b$ est un vecteur Gaussien de $\mathbb{R}^p$ de moyenne $A\boldsymbol{m} + b$ et de matrice de covariance $A\boldsymbol{\Sigma}A'$.*

Proposition. *La fonction caractéristique d'un vecteur Gaussien de $\mathbb{R}^d$ s'écrit*

$$\mathrm{E}\left[e^{i<t,\boldsymbol{Y}>}\right] = e^{i<t,\boldsymbol{m}>} e^{-\frac{1}{2}t'\boldsymbol{\Sigma}t}$$

où $t \in \mathbb{R}^d$. En particulier la loi d'un vecteur Gaussien est caractérisée par sa moyenne et sa variance. On notera $\mathcal{N}(\boldsymbol{m}, \boldsymbol{\Sigma})$ cette loi.

Exemple. La loi $\mathcal{N}(\boldsymbol{m}, \boldsymbol{\Sigma})$ est la loi du vecteur $\boldsymbol{m} + \boldsymbol{Y}$ où $\boldsymbol{Y}$ suit la loi $\mathcal{N}(0, \boldsymbol{\Sigma})$.

Exemple. Notons I_d la matrice identité en dimension d. La loi $\mathcal{N}(0, I_d)$ est celle de $(Y_1, \ldots, Y_d)$ où les $(Y_i)_{1 \leq i \leq d}$ sont des variables Gaussiennes i.i.d. centrées réduites.

Proposition. *Lorsque $\boldsymbol{\Sigma}$ est inversible, la densité du vecteur Gaussien $\mathcal{N}(\boldsymbol{m}, \boldsymbol{\Sigma})$ en $\boldsymbol{x} \in \mathbb{R}^d$ s'écrit*

$$\frac{1}{(2\pi)^{d/2}} \frac{1}{\sqrt{\det(\boldsymbol{\Sigma})}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{m})}.$$

Dans le cas où $\boldsymbol{\Sigma}$ n'est pas inversible, la loi de $\boldsymbol{Y}$ est portée par un sous-espace vectoriel strict de $\mathbb{R}^d$.

Proposition. *Soient X et Y deux variables aléatoires réelles. Si (X, Y) est un vecteur Gaussien, alors X et Y sont indépendantes si et seulement si $Cov(X, Y) = 0$.*

Théorème central limite. *Soit $(\mathbf{X}_i)_{i \geq 1}$ une suite i.i.d. de vecteurs aléatoires de $\mathbb{R}^d$ de moyenne nulle et de matrice de covariance finie $\boldsymbol{\Sigma}$. Alors $\frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{X}_i$ converge en loi vers $\mathcal{N}(0, \boldsymbol{\Sigma})$.*

5.1.2 Loi du khi-deux, loi de Student, loi de Fisher

On définit ici trois lois qui vont revenir fréquemment dans le cadre des échantillons Gaussiens.

Définition. *On appelle loi du χ^2 à d degrés de libertés de paramètre de décentrage $||\boldsymbol{m}||^2$, notée $\chi^2(d, ||\boldsymbol{m}||^2)$ (ou simplement $\chi^2(d)$ si $\boldsymbol{m} = 0$) la loi de la variable $||\boldsymbol{Y}||^2$ où $\boldsymbol{Y}$ est de loi $\mathcal{N}(\boldsymbol{m}, I_d)$.*

Remarque. Dans la définition, $||\boldsymbol{x}|| = \sqrt{x_1^2 + \ldots + x_d^2}$ représente la norme euclidienne du vecteur $\boldsymbol{x} = (x_1, \ldots, x_d)' \in \mathbb{R}^d$.

Définition. *On appelle loi de Student à d degrés de libertés de paramètre de décentrage $\mu \in \mathbb{R}$, notée $\mathcal{T}(d, \mu)$ (et $\mathcal{T}(d)$ dans le cas $\mu = 0$) la loi de la variable aléatoire*

$$\frac{Z}{\sqrt{U/d}}$$

où U, Z sont indépendantes, U suit une loi $\chi^2(d)$ et Z suit une loi $\mathcal{N}(\mu, 1)$.

Définition. *Soient Z_1 et Z_2 deux variables aléatoires indépendantes de loi respectivement $\chi^2(d_1, \lambda)$ et $\chi^2(d_2)$. On appelle loi de Fisher à d_1 et d_2 degrés de liberté et de paramètre de décentrage λ la loi de la variable aléatoire*

$$F = \frac{Z_1/d_1}{Z_2/d_2}.$$

On note $\mathcal{F}(d_1, d_2, \lambda)$ cette loi, et plus simplement $\mathcal{F}(d_1, d_2)$ si $\lambda = 0$ (loi de Fisher à d_1 et d_2 degrés de liberté).

5.1.3 Le théorème de Cochran

Voici un théorème fondamental pour les échantillons Gaussiens.

Théorème de Cochran. *Soit $\boldsymbol{Y}$ de loi $\mathcal{N}(\boldsymbol{m}, I_d)$. Soit $E_1 \oplus \ldots \oplus E_\ell$ une décomposition de $\mathbb{R}^d$ en ℓ espaces vectoriels orthogonaux de dimension respective $d_1, \ldots, d_\ell$. Alors les projections orthogonales $\Pi_{E_1}\boldsymbol{Y}$, $\ldots$, $\Pi_{E_\ell}\boldsymbol{Y}$ sont indépendantes et la loi de $||\Pi_{E_k}\boldsymbol{Y}||^2$ est $\chi^2(d_k, ||\Pi_{E_k}\boldsymbol{m}||^2)$.*

Corollaire. *Soit $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ un échantillon de n variables Gaussiennes i.i.d. de loi $\mathcal{N}(\mu, \sigma^2)$. Alors les variables $\overline{Y}_n$ et $\frac{1}{\sigma^2}\sum_{i=1}^n (Y_i - \overline{Y}_n)^2$ sont indépendantes, de loi respective $\mathcal{N}(\mu, \sigma^2/n)$ et $\chi^2(n-1)$. Ainsi,*

$$\frac{\overline{Y}_n - \mu}{\hat{\sigma}_n/\sqrt{n}} \sim \mathcal{T}(n-1)$$

où $\hat{\sigma}_n^2 := \frac{1}{n-1}\sum_{i=1}^n (Y_i - \overline{Y}_n)^2$.

5.2 Intervalles de confiance

On déduit des résultats précédents des intervalles de confiance pour la moyenne et la variance d'un échantillon Gaussien. Soit donc $(Y_1, \ldots, Y_n)$ des variables aléatoires Gaussiennes $\mathcal{N}(\mu, \sigma^2)$ indépendantes.

Moyenne. *On a déjà vu comment construire un intervalle de confiance lorsque σ^2 est connue. Dans le cas où σ^2 est inconnue, un intervalle de confiance (bilatéral) au niveau $1-\alpha$ de μ est donné par*

$$\left[\overline{Y}_n - \frac{\hat{\sigma}_n}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}}, \overline{Y}_n + \frac{\hat{\sigma}_n}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}}\right]$$

où $t_{d,\beta}$ est le quantile β de la loi de Student $\mathcal{T}(d)$.

Variance. *Un intervalle de confiance (bilatéral) au niveau $1-\alpha$ de σ^2 est donné par*

$$\left[\frac{(n-1)\hat{\sigma}_n^2}{c_{n-1,1-\alpha/2}}, \frac{(n-1)\hat{\sigma}_n^2}{c_{n-1,\alpha/2}}\right]$$

où $c_{d,\beta}$ est le quantile β de la loi khi-deux $\chi^2(d)$.

Exercice. Donner les intervalles de confiance unilatéraux à droite.

5.3 Test de comparaison à une référence

Soit une série de variables aléatoires Gaussiennes indépendantes $(Y_1, \ldots, Y_n)$ de loi $\mathcal{N}(\mu, \sigma^2)$.

Moyenne. *On veut tester (test bilatéral)*

$$\mathcal{H}_0 : \mu = \mu_0 \quad \textit{contre} \quad \mathcal{H}_1 : \mu \neq \mu_0$$

pour un certain nombre μ_0. On utilise que sous $\mathcal{H}_0$, la variable

$$T = \frac{\overline{Y}_n - \mu_0}{\hat{\sigma}_n/\sqrt{n}} \sim \mathcal{T}(n-1).$$

La variable T est notre statistique de test. Pour obtenir un test de risque α, on rejettera $\mathcal{H}_0$ si

$$|T| > t_{n-1,1-\alpha/2}.$$

- *Dans le cas où $\mathcal{H}_1 : \mu > \mu_0$, on rejette $\mathcal{H}_0$ si $T > t_{n-1,1-\alpha}$.*
- *Dans le cas où $\mathcal{H}_1 : \mu < \mu_0$, on rejette $\mathcal{H}_0$ si $T < t_{n-1,\alpha}$.*

Pro-OGM et anti-OGM. Cet exemple (inventé) a pour objectif d'illustrer le caractère non équitable des tests d'hypothèses déjà souligné plus haut. La protéine P435 permet de mesurer le taux de mutation cellulaire dans le pancréas. Un échantillon de $n = 20$ souris est nourri avec du maïs OGM H76. Il est communément admis que pour un taux supérieur à 80 unités de P435, le risque de cancer est élevé. On modélise ce taux de P435 chez une souris nourrie au maïs H76 par une variable aléatoire notée Y, et on suppose que $Y \sim \mathcal{N}(\mu, \sigma^2)$ où μ et σ^2 sont inconnus. Sur l'échantillon des 20 souris, les mesures donnent $\overline{Y} = 79.5$ et $\hat{\sigma}_n^2 = 2.13$. Soit la statistique de Student $T_n := \sqrt{n}\frac{\overline{Y}_n - 80}{\hat{\sigma}_n} = 1.53$.

- Les militants anti-OGM effectuent le test de $\mathcal{H}_0 : \mu > 80$ contre $\mathcal{H}_1 : \mu \leq 80$ au niveau 5%. Il s'agit du test

  $$\Phi_{\text{anti}}(X) := \mathbf{1}\{T_n < t_{n-1,0.05}\}$$

  où $t_{n-1,0.05} \approx -1.73$. Les militants anti-OGM gardent donc l'hypothèse $\mu > 80$.

- Les industriels pro-OGM effectuent le test de $\mathcal{H}_0 : \mu \leq 80$ contre $\mathcal{H}_1 : \mu > 80$ au niveau 5%. Il s'agit du test
$$\Phi_{\text{pro}}(X) := \mathbf{1}\{T_n > t_{n-1,0.95}\}$$
où $t_{n-1,0.95} \approx 1.73$. Les pro-OGM gardent donc l'hypothèse $\mu \leq 80$.

La faible puissance du test de Student pour $n = 20$ explique cette situation paradoxale. On peut calculer par exemple $\pi(\mu = 0.85) = 0.12$ et $\pi(\mu = 0.9) = 0.25$; le test est très conservatif !

Variance. *On veut tester (test bilatéral)*
$$\mathcal{H}_0 : \sigma^2 = \sigma_0^2 \text{ contre } \mathcal{H}_1 : \sigma^2 \neq \sigma_0^2$$
où σ_0^2 est un certain nombre positif. On se rappelle que la variable $\frac{(n-1)\hat{\sigma}_n^2}{\sigma^2} \sim \chi^2(n-1)$. Ainsi sous $\mathcal{H}_0$, on aura
$$U = \frac{(n-1)\hat{\sigma}_n^2}{\sigma_0^2} \sim \chi^2(n-1).$$
La variable U est notre statistique de test. Pour obtenir un test de risque α, on rejettera $\mathcal{H}_0$ si
$$U \notin [c_{n-1,\alpha/2}, c_{n-1,1-\alpha/2}].$$
- *Dans le cas où $\mathcal{H}_1 : \sigma^2 > \sigma_0^2$, on rejette $\mathcal{H}_0$ si $U > c_{n-1}(1-\alpha)$.*
- *Dans le cas où $\mathcal{H}_1 : \sigma^2 < \sigma_0^2$, on rejette $\mathcal{H}_0$ si $U < c_{n-1}(\alpha)$.*

Remarque. On aurait pu déduire ces tests des intervalles de confiance de la partie précédente.

Vocabulaire. On appelle communément le test sur la moyenne test de Student ou t-test (en anglais).

5.4 Test d'homogénéité de deux échantillons Gaussiens

On considère deux échantillons Gaussiens indépendants $\boldsymbol{X} = (X_1, \ldots, X_n)$ et $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$. On aimerait comparer les caractéristiques des deux échantillons.

5.4.1 Comparaison des variances

On suppose que $\boldsymbol{X}$ et $\boldsymbol{Y}$ sont indépendants. Les variables $(X_i)_i$ sont i.i.d. de loi $\mathcal{N}(\mu_X, \sigma_X^2)$ et les variables $(Y_i)_i$ de loi $\mathcal{N}(\mu_Y, \sigma_Y^2)$.

On aimerait tester $\mathcal{H}_0 : \sigma_X = \sigma_Y$ contre $\mathcal{H}_1 : \sigma_X \neq \sigma_Y$.

Sous $\mathcal{H}_0$, on remarque que
$$F = \frac{\hat{\sigma}_X^2}{\hat{\sigma}_Y^2} \sim \frac{\chi^2(n-1)/(n-1)}{\chi^2(m-1)/(m-1)} = Fisher(n-1, m-1).$$
On rejette donc $\mathcal{H}_0$ si $F < f_{n-1,m-1}(\alpha/2)$ ou $F > f_{n-1,m-1}(1-\alpha/2)$.

- Dans le cas où $\mathcal{H}_1 : \sigma_X > \sigma_Y$, on rejette $\mathcal{H}_0$ si $F > f_{n-1,m-1}(1-\alpha)$.
- Dans le cas où $\mathcal{H}_1 : \sigma_X < \sigma_Y$, on rejette $\mathcal{H}_0$ si $F < f_{n-1,m-1}(\alpha)$.

Remarque. Ce test est appelé test de Fisher.

5.5 Comparaison des moyennes

Echantillons indépendants. On suppose que $\boldsymbol{X} = (X_1, \ldots, X_n)$ et $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$ sont indépendants et ont **même variance** $\sigma_X = \sigma_Y = \sigma$ (il existe encore un test dans le cas $\sigma_X \neq \sigma_Y$ appelé test de Welch). Les variables $(X_i)_i$ sont i.i.d. de loi $\mathcal{N}(\mu_X, \sigma^2)$ et les variables $(Y_i)_i$ i.i.d. de loi $\mathcal{N}(\mu_Y, \sigma^2)$.

On teste (test bilatéral)

$$\mathcal{H}_0 : \mu_X = \mu_Y \text{ contre } \mathcal{H}_1 : \mu_X \neq \mu_Y$$

Sous $\mathcal{H}_0$,

$$T = \frac{\overline{X}_n - \overline{Y}_m}{\hat{\sigma}_{intra}\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim \mathcal{T}(n + m - 2)$$

avec $\hat{\sigma}^2_{intra} = \frac{1}{n+m-2}\left(\sum_{k=1}^n (X_k - \overline{X}_n)^2 + \sum_{k=1}^m (Y_k - \overline{Y}_m)^2\right)$.

On rejette $\mathcal{H}_0$ si $|T| > t_{n+m-2}(1 - \alpha/2)$.

- Dans le cas où $\mathcal{H}_1 : \mu_X > \mu_Y$, on rejette $\mathcal{H}_0$ si $T > t_{n+m-2}(1 - \alpha)$.
- Dans le cas où $\mathcal{H}_1 : \mu_X < \mu_Y$, on rejette $\mathcal{H}_0$ si $T < t_{n+m-2}(\alpha)$.

Echantillons appariés. On suppose cette fois-ci que les échantillons $\boldsymbol{X}$ et $\boldsymbol{Y}$ ne sont plus indépendants, mais qu'ils proviennent de la même population. On supposera donc que les couples (X_1, Y_1), $\ldots$, (X_n, Y_n) sont des vecteurs Gaussiens indépendants, chaque vecteur (X_i, Y_i) étant de moyenne $\boldsymbol{m}_i$ et de variance $\boldsymbol{\Sigma}$. Le vecteur moyenne $\boldsymbol{m}_i = (\mu_{X,i}, \mu_{Y,i})$ peut dépendre de i, mais la matrice de covariance $\boldsymbol{\Sigma}$ est identique pour tous les couples.

On aimerait tester $\mathcal{H}_0 : \mu_{X,i} = \mu_{Y,i}$ pour tout i contre $\mathcal{H}_1$: il existe i tel que $\mu_{X,i} \neq \mu_{Y,i}$.

Cela revient à faire un test de comparaison de la moyenne des variables $\Delta_i := X_i - Y_i$ à zéro. On appelle ce test le test de Student apparié. La statistique de test est

$$T = \frac{\overline{\Delta}_n}{\hat{\sigma}_n/\sqrt{n}}$$

avec $\hat{\sigma}^2_n = \frac{1}{n-1}\sum_{k=1}^n (\Delta_k - \overline{\Delta}_n)^2$. On compare ensuite T aux quantiles de la loi de Student $\mathcal{T}(n-1)$.

6 Tests d'adéquation

Il s'agit de tester l'adéquation d'un échantillon à une loi de référence, voire à une famille de lois de référence (par exemple les lois exponentielles). Le test de Kolmogorov repose sur une comparaison des fonctions de répartition et permet de répondre à cette question. Son caractère non paramétrique est appréciable si la loi de référence n'est pas standard, mais la puissance de ce test est faible pour des échantillons de petite taille. Une alternative est le test du χ^2, quitte à regrouper les observations en sous-catégories. Pour tester la normalité d'un échantillon, c'est-à-dire tester s'il suit une loi gaussienne, on préfèrera le test de Shapiro-Wilk, en particulier pour des échantillons de petite taille. On finira par présenter le test d'indépendance du χ^2 qui peut se voir comme le test d'adéquation de la loi d'un couple de variables aléatoires à une loi produit.

6.1 Test de Kolmogorov

Fonction de répartition empirique.

Soit $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ un échantillon de variables i.i.d. de loi commune ayant pour fonction de répartition F.

Définition. *On appelle fonction de répartition empirique la fonction*

$$F_n(x) := \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{\{Y_k \leq x\}}.$$

La fonction de répartition empirique est un estimateur de F.

Théorème de Glivenko-Cantelli. *On a* $||F_n - F||_\infty \to 0$ *quand* $n \to +\infty$ *p.s.*

Statistique de test.

On aimerait savoir si la loi commune de l'échantillon $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ est égale à une certaine loi donnée que l'on suppose sans atome . Soit F^{ref} la fonction de répartition de cette loi, qui est donc continue.

On veut tester $\mathcal{H}_0 : F = F^{ref}$ contre $\mathcal{H}_1 : F \neq F^{ref}$. Soit la statistique de test

$$H_n := ||F_n - F||_\infty$$

On peut montrer que H_n a sous $\mathcal{H}_0$ même loi que

$$\sup_{q \in [0,1]} \Big| \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{\{U_k \leq q\}} - q \Big|$$

où les variables $(U_k)_{k \leq n}$ sont des variables i.i.d. de loi uniforme sur $[0, 1]$.

En particulier, la statistique a la même loi quelle que soit F^{ref} (continue). On note $h_{n,1-\alpha}$ les quantiles de cette loi. Le test consistera à comparer la statistique H_n au quantile $h_{n,1-\alpha}$.

6.2 Test d'adéquation du χ^2

On a une série de données $Y_1, \ldots, Y_n$. On pense pouvoir modéliser chaque donnée Y_i par une certaine loi (par exemple loi Gaussienne, de Poisson, etc). On appellera cette loi la loi de référence, et une variable aléatoire suivant la loi de référence une variable aléatoire de référence. On aimerait un test pour valider cette hypothèse $\mathcal{H}_0$.

On choisit d classes de valeurs, où d est fixé. On regarde le nombre de données Y_i qui tombent dans la i-ième classe, que l'on note n_i. On note p_i^{ref} la probabilité théorique que la variable aléatoire de référence prenne une valeur dans cette classe. Le nombre théorique de données dans la classe i doit donc être proche de np_i^{ref} si notre hypothèse est vraie.

Si p_i est la probabilité que Y_1 appartienne à la classe i, on veut tester $\mathcal{H}_0 : p_i = p_i^{ref}$ pour tout $1 \leq i \leq d$ contre $\mathcal{H}_1 : p_i \neq p_i^{ref}$ pour un certain i.

Soit la statistique de Pearson

$$D_n = \sum_{i=1}^{d} \frac{(n_i - np_i^{ref})^2}{np_i^{ref}}.$$

Proposition.

- *Sous $\mathcal{H}_0$, la statistique de Pearson converge en loi vers $\chi^2(d-1)$ quand $n \to +\infty$.*
- *Sous $\mathcal{H}_1$, la statistique de Pearson converge vers $+\infty$ p.s.*

En pratique on utilisera le test pour $n \geq 30$ et $n_i \geq 5$ pour tout i.

On rejettera donc $\mathcal{H}_0$ si $D_n \geq c_{d-1,1-\alpha}$.

Si la loi théorique dépend de m paramètres inconnus (par exemple $\mathcal{N}(\mu, \sigma^2)$ dépend de deux paramètres si on ne connaît ni μ ni σ), on estime d'abord chaque paramètre (dans notre exemple: $\overline{Y}_n$ et $\hat{\sigma}_n^2$). On note alors $\hat{p}_i^{ref}$ la probabilité théorique de se trouver dans la classe i (dans notre exemple: probabilité que $\mathcal{N}(\overline{Y}_n, \hat{\sigma}_n^2)$ se trouve dans la classe i), et on calcule

$$D_n = \sum_{i=1}^{d} \frac{(n_i - n\hat{p}_i^{ref})^2}{n\hat{p}_i^{ref}}.$$

qui converge en loi (sous des conditions idoines) vers la loi $\chi^2(d-1-m)$ quand $n \to +\infty$. On rejette $\mathcal{H}_0$ si $D_n \geq c_{d-1-m,1-\alpha}$.

6.3 Test d'indépendance du χ^2

Le test d'indépendance du χ^2 est utilisé pour tester si deux variables aléatoires sont indépendantes. Supposons que l'on observe un n-échantillon bivarié $X = (Y, Z) = ((Y_1, Z_1) \ldots, (Y_n, Z_n))$. On note ν la loi du couple (Y_1, Z_1), μ celle de Y_1 et λ celle de Z_1. Le test d'indépendance des lois μ et λ correspond au test de $\mathcal{H}_0 : \nu = \mu \otimes \lambda$ contre $\mathcal{H}_1 : \nu \neq \mu \otimes \lambda$.

Le test décrit plus bas suppose que les variables aléatoires en jeu sont qualitatives. Quitte à effectuer des regroupements de variables, on suppose que les variables aléatoires Y_i et Z_i prennent presque sûrement des valeurs dans $\{y_1, \ldots, y_\ell\}$ (ensemble des modalités de Y) et dans $\{z_1, \ldots, z_m\}$ (ensemble des modalités de Z).

On considère donc un tableau croisé. Dans la case (i, j), on fait apparaître le nombre $n_{i,j}$ de $k \leq n$ tels que $Y_k = y_i$ et $Z_k = z_j$:

On va tester l'hypothèse que les deux variables Y et Z sont indépendantes. Notons $p_i = \mathrm{P}(Y = y_i)$ et $q_j = \mathrm{P}(Z = z_j)$. Si les deux variables étaient indépendantes, on devrait avoir $\mathrm{P}(Y = y_i, Z = z_j) = p_i q_j$.

$Y \backslash Z$	z_1	z_2	z_3	...	z_m	Total
y_1	$n_{1,1}$	$n_{1,2}$	$n_{1,3}$	...	$n_{1,m}$	n_{1+}
...	...	...	...	...	...	...
y_ℓ	$n_{\ell,1}$	$n_{\ell,2}$	$n_{\ell,3}$	...	$n_{\ell,m}$	$n_{\ell+}$
Total	n_{+1}	n_{+2}	n_{+3}	...	n_{+m}	n

On voudrait donc faire un test de χ^2 en comparant $n_{i,j}/n$ à $p_i q_j$. Il s'écrirait

$$\sum_{i,j} \frac{(n_{i,j} - np_iq_j)^2}{np_iq_j} \sim \chi^2(m\ell - 1).$$

Cependant on ne connaît ni p_i ni q_j. On va donc les estimer par

$$\hat{p}_i = \frac{n_{i,+}}{n} \qquad \hat{q}_j = \frac{n_{+,j}}{n}$$

où $n_{i+} = \sum_j n_{i,j}$ et $n_{+j} = \sum_i n_{i,j}$. Avec cette approximation, on perd $(m-1) + (\ell - 1)$ degrés de libertés.

En résumé, on teste

$$\mathcal{H}_0 : \text{ les variables } Y \text{ et } Z \text{ sont indépendantes}$$

contre

$$\mathcal{H}_1 : \text{ elles ne sont pas indépendantes}$$

La statistique de test est

$$D_n = \sum_{i,j} \frac{(n_{i,j} - \frac{n_{i+}n_{+j}}{n})^2}{\frac{n_{i+}n_{+j}}{n}}$$

qui, sous $\mathcal{H}_0$, converge en loi quand $n \to +\infty$ vers la loi $\chi^2((m-1)(\ell-1))$. On rejette $\mathcal{H}_0$ si $D_n > c_{(m-1)(\ell-1),1-\alpha}$.

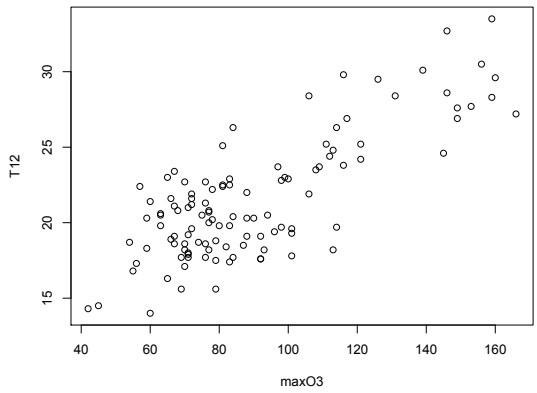
7 Régression linéaire

7.1 Deux exemples introductifs

7.1.1 Le modèle de la régression linéaire simple

Pour introduire le modèle, on étudie un jeu de données portant sur des niveaux de pollution enregistrés sur 112 journées de l'été 2001 à Rennes [1]. On dispose des deux variables maxO3 pour le maximum journalier de la concentration en ozone et T12 pour la température relevée à 12 h. Le graphique ci-dessous représente le nuage des points pour ces deux variables.

[1] exemple tiré de l'ouvrage *Statistiques avec R, P.A. Cornillon et al., PUR 2008*



Les points semblent s'aligner sur une droite que l'on souhaiterait déterminer.

De façon générale, soient deux variables numériques $\boldsymbol{x}$ et $\boldsymbol{y}$ dont on observe n couples (x_i, y_i). Si la corrélation linéaire empirique

$$\mathrm{corr}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y})}{\sqrt{\mathrm{V}(\boldsymbol{x})\mathrm{V}(\boldsymbol{y})}}$$

entre les deux variable est importante (proche de 1 ou -1), il existe alors une liaison linéaire importante entre les deux variables. Dans ce cas, d'après Cauchy-Schwartz, les variables recentrées sont presque colinéaires et le nuage des points (x_i, y_i) est positionné au voisinage d'une droite de pente a et d'ordonnée à l'origine b.

Remarque. Attention : une corrélation linéaire importante ne signifie pas qu'il existe une relation de causalité entre $\boldsymbol{x}$ et $\boldsymbol{y}$!

Le *modèle de régression linéaire simple (gaussien)* permet de modéliser une relation linéaire entre deux variables numériques $\boldsymbol{x}$ et $\boldsymbol{Y}$:

$$Y_i = ax_i + b + \varepsilon_i \quad i = 1 \dots n$$

où

- Y est la variable dite *dépendante* (à expliquer),
- $\boldsymbol{x}$ est une variable explicative supposée déterministe (design fixe),
- $\boldsymbol{\varepsilon}$ est le vecteur des erreurs, ces erreurs sont supposées indépendantes et de même loi $\mathcal{N}(0, \sigma^2)$,
- $\boldsymbol{\theta} = (b, a) \in \mathbb{R}^2$ et $\sigma > 0$ sont les paramètres du modèle.

Il est possible d'écrire ce modèle sous la matricielle suivante:

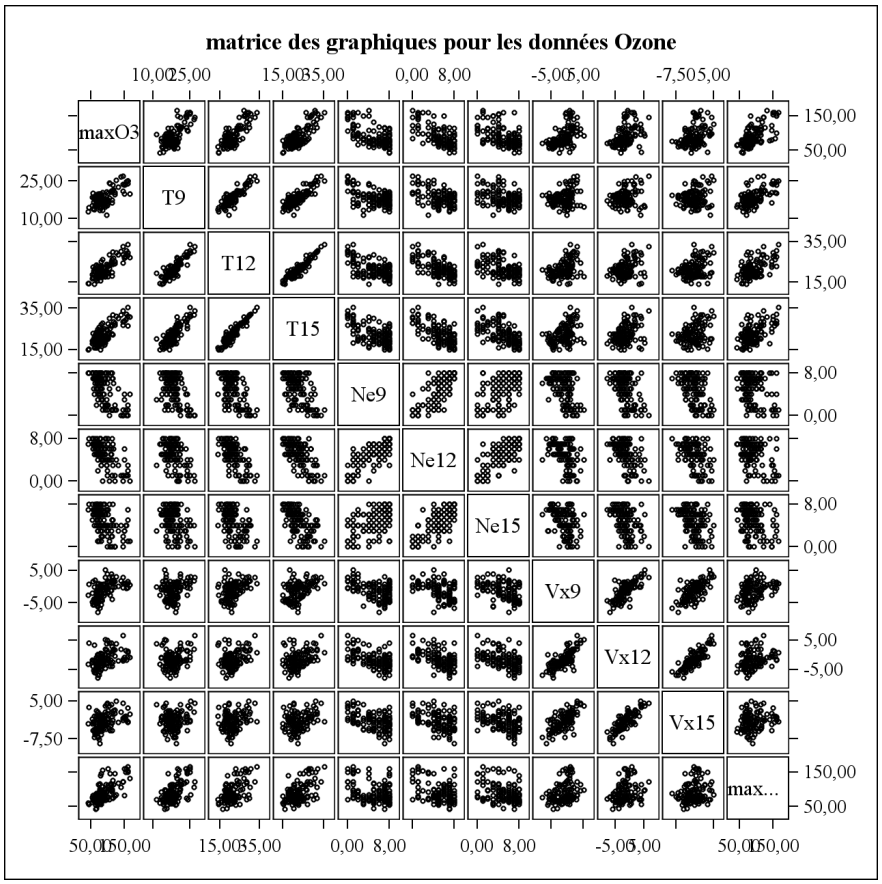
$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}}_{\mathbf{M}} \underbrace{\begin{bmatrix} b \\ a \end{bmatrix}}_{\boldsymbol{\theta}} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}}_{\boldsymbol{\varepsilon}} = \mathbf{M}\boldsymbol{\theta} + \boldsymbol{\varepsilon}.$$

7.1.2 Le modèle de la régression linéaire multiple

Dans le même contexte que celui décrit pour la régression simple, on souhaite maintenant expliquer la quantité d'ozone maxO3 par les mesures suivantes :

- T9, T12, T15: température à 9h, 12h, 15h;
- Ne9, N12, N15 : nébulosité à 9h, 12h, 15h;
- Vx9, Vx12, VX15: vitesse du vent sur un axe Est-Ouest à 9h, 12h, 15h;
- MaxO3v: ozone mesurée la veille.

La graphique ci-dessous représente la "matrice" des nuages de points pour tous les croisements possibles de deux variables :



Ce problème peut être modélisé par un *modèle de régression linéaire multiple*. Disposant de n observations indépendantes Y_i d'une variable Y numérique, on souhaite expliquer Y par $p-1$ variables numériques $\boldsymbol{x}^j$ supposées déterministes. Dans le modèle de régression multiple, on suppose que l'espérance de Y est une fonction linéaire des variables explicatives $\boldsymbol{x}^j$:

$$Y_i = a_0 + a_1 x_i^1 + \cdots + a_{p-1} x_i^{p-1} + \varepsilon_i, \quad i = 1 \dots n \tag{1}$$

où

- Y est la variable dite *dépendante* (à expliquer),

- $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^{p-1}$ sont les variables explicatives ou régresseurs, supposées déterministes (design fixe), la i-ème observation est le vecteur $\boldsymbol{x}_i = (1, x_i^1, \ldots, x_i^{p-1})'$ (il sera en effet plus pratique de donner la valeur 1 à la première coordonnée dans la suite, afin de prendre en compte le terme constant).
- $\boldsymbol{\varepsilon}$ est le vecteur des erreurs, ces erreurs sont supposées indépendantes et de même loi $\mathcal{N}(0, \sigma^2)$.
- $\boldsymbol{\theta} = (a_0, \ldots, a_{p-1}) \in \mathbb{R}^p$ et $\sigma > 0$ sont les paramètres du modèle.

Il est possible d'écrire ce modèle sous la forme matricielle suivante :

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & x_1^1 & \ldots & x_1^{p-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n^1 & \ldots & x_n^{p-1} \end{bmatrix}}_{\mathbf{M}=[\mathbf{e},\boldsymbol{x}^1,\ldots,\boldsymbol{x}^{p-1}]} \underbrace{\begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_{p-1} \end{bmatrix}}_{\boldsymbol{\theta}} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}}_{\boldsymbol{\varepsilon}} = \mathbf{M}\boldsymbol{\theta} + \boldsymbol{\varepsilon}.$$

7.2 Estimation

7.2.1 Le modèle linéaire Gaussien

Définition. *Soit $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ vecteur de $\mathbb{R}^n$. On dit que $\boldsymbol{Y}$ suit un modèle linéaire Gaussien si*

$$\boldsymbol{Y} = \boldsymbol{m} + \boldsymbol{\varepsilon}$$

où $\boldsymbol{m}$ appartient à un sous-espace vectoriel $\mathcal{V}$ (de dimension notée p) de $\mathbb{R}^n$, et $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$. Les paramètres inconnus sont $\boldsymbol{m}$ et σ^2.

On peut détailler les hypothèses du modèle linéaire Gaussien ainsi:

- (H1) les erreurs sont centrées : $\mathrm{E}\,\boldsymbol{\varepsilon} = 0$,
- (H2) homoscédasticité : $\forall i \in \{1, \ldots, n\}$, $\mathrm{V}(\varepsilon_i) = \sigma^2$,
- (H3) les erreurs ε_i sont indépendantes.
- (H4) normalité : $\forall i \in \{1, \ldots, n\}$, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Lorsqu'on suppose simplement que les 3 premières hypothèses sont vérifiées, on dit seulement que le modèle est linéaire.

Dans la proposition suivante, on note $\Pi_{\mathcal{V}}$ la projection orthogonale sur l'espace vectoriel $\mathcal{V}$.

Proposition.

- *Les estimateurs de maximum de vraisemblance de $\mathbf{m}$ et σ^2 sont donnés par*

$$\hat{\mathbf{m}} = \Pi_{\mathcal{V}} Y \qquad \hat{\sigma}^2 = \frac{1}{n} ||Y - \Pi_{\mathcal{V}}||^2$$

  où $\Pi_{\mathcal{V}}$ est la projection orthogonale sur $\mathcal{V}$.

- *L'estimateur $\hat{\mathbf{m}}$ est un vecteur Gaussien de loi $\mathcal{N}(\mathbf{m}, \sigma^2 \Pi_{\mathcal{V}})$.*

- *La loi de* $\frac{n}{\sigma^2}s^2$ *est* $\chi^2(n-p)$*. L'estimateur* s^2 *est donc biaisé. On préfèrera donc l'estimateur sans biais*

$$\hat{\sigma}^2 := \frac{1}{n-p}||\boldsymbol{Y} - \Pi_{\mathcal{V}}\boldsymbol{Y}||^2.$$

  La loi de $\frac{n-p}{\sigma^2}\hat{\sigma}^2$ *est* $\chi^2(n-p)$*.*

- *Les estimateurs* $\hat{\mathbf{m}}$ *et* $\hat{\sigma}$ *sont indépendants.*

Exemple. Si $\mathcal{V}$ est l'espace vectoriel engendré par le vecteur $\boldsymbol{e}' = (1, \ldots, 1)'$, alors les variables $(Y_i)_{i\leq n}$ sont i.i.d. de moyenne et variance inconnus. On appellera ce modèle le modèle i.i.d.

Exemple. En prenant pour $\mathcal{V}$ l'espace vectoriel image d'une matrice $\mathbf{M}$ connue de taille $n \times \tilde{p}$ de rang p, on trouve le modèle de la régression linéaire multiple

$$\boldsymbol{Y} = \mathbf{M}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \tag{2}$$

où

- $\boldsymbol{Y} = (Y_1, \ldots, Y_n)' \in \mathbb{R}^n$ est le vecteur des observations,
- $\boldsymbol{M}$ est une matrice connue de taille $n \times \tilde{p}$ de rang p
- $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{\tilde{p}})'$ est le vecteur des paramètres (en général inconnu),
- $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ est le vecteur aléatoire des erreurs.

Lorsque $\tilde{p} = p$, on dit que le modèle est *régulier*. Il y a alors une bijection entre les vecteurs $\boldsymbol{\theta}$ et $\mathbf{m} \in \mathcal{V}$. Dans le cas contraire, le modèle est dit *singulier*: il faut alors imposer des conditions supplémentaires sur $\boldsymbol{\theta}$ pour pouvoir l'identifier.

Ce modèle permet d'expliquer et prédire le phénomène Y à l'aide d'une représentation simple (linéaire) de l'espérance de Y. On note que plus l'espace vectoriel $\mathcal{V} = \mathrm{Im}(\mathbf{M})$ est grand, plus le modèle aura une grande capacité à expliquer et prédire le phénomène. L'hypothèse H1 signifie qu'il n'y a pas d'information dans la moyenne des erreurs. L'hypothèse H3 signifie qu'il y a eu un échantillonnage indépendant pour obtenir les données, ou bien qu'un processus physique fournit les données de façon indépendante. On prendra garde aux contextes où le temps intervient.

Proposition. *Supposons le modèle régulier.*

- *L'estimateur du maximum de vraisemblance de* $\boldsymbol{\theta}$ *est donné par*

$$\hat{\boldsymbol{\theta}} = (\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}'Y$$

  de loi $\mathcal{N}(\boldsymbol{\theta}, \sigma^2(\mathbf{M}'\mathbf{M})^{-1})$*.*

- *On peut réécrire*

$$\hat{\sigma}^2 = \frac{1}{n-p}||\boldsymbol{Y} - \mathbf{M}\hat{\boldsymbol{\theta}}||^2.$$

  La loi de $\frac{n-p}{\sigma^2}\hat{\sigma}^2$ *est* $\chi^2(n-p)$*.*

- *Les estimateurs* $\hat{\boldsymbol{\theta}}$ *et* $\hat{\sigma}^2$ *sont indépendants.*

Définition. *On appelle le vecteur* $\hat{\boldsymbol{\varepsilon}} := \boldsymbol{Y} - \mathbf{M}\hat{\boldsymbol{\theta}}$ *vecteur des résidus.*

Remarque. L'estimateur $\hat{\boldsymbol{\theta}}$ est aussi l'estimateur des moindres carrés: C'est le $\boldsymbol{\theta}$ qui minimise la norme L^2 des résidus (On a $\mathbf{M}\hat{\boldsymbol{\theta}} = \hat{\mathbf{m}} = \Pi_{\mathcal{V}}\boldsymbol{Y}$).

Remarque. On peut montrer que $\hat{\boldsymbol{\theta}}$ est de variance minimale parmi les estimateurs sans biais de $\boldsymbol{\theta}$.

7.3 Exemple: le modèle ANOVA

Modèle ANOVA à un facteur.
Supposons que l'on veuille déceler une différence de salaires entre les hommes et les femmes. On veut donc trouver une relation entre une variable numérique, le salaire, et une variable catégorielle, le sexe. C'est un modèle à un facteur qui peut s'écrire

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

où Y_{1j} $(1 \leq j \leq n_1)$ est le salaire du j-ième individu pris dans le groupe homme et Y_{2j} $(1 \leq j \leq n_2)$ est le salaire du j-ième individu pris dans le groupe femme. En règle générale, le modèle ANOVA à un facteur s'écrit

$$Y_{ij} = \mu_i + \varepsilon_{ij},\ i = 1, \ldots, p,\ j = 1, \ldots, n_i$$

où p représente le nombre de niveaux du facteur, et n_i le nombre de données dans le groupe i. C'est un modèle linéaire Gaussien de la forme

$$\boldsymbol{Y} = \mathbf{m} + \boldsymbol{\varepsilon}$$

avec

$$\begin{aligned}
\boldsymbol{Y} &= (\underbrace{Y_{11}, \ldots, Y_{1n_1}}_{\text{groupe } 1}, \ldots, \underbrace{Y_{p1}, \ldots, Y_{pn_p}}_{\text{groupe } p})' \\
\mathbf{m} &= (\underbrace{\mu_1, \ldots, \mu_1}_{\text{groupe } 1}, \ldots, \underbrace{\mu_p, \ldots, \mu_p}_{\text{groupe } p})' \\
\boldsymbol{\varepsilon} &= (\varepsilon_{11}, \ldots, \varepsilon_{1n_1}, \ldots, \varepsilon_{p1}, \ldots, \varepsilon_{pn_p})'.
\end{aligned}$$

On peut réécrire $\mathbf{m} = \mathbf{M}\boldsymbol{\theta}$ avec $\boldsymbol{\theta} = (\mu_1, \mu_2, \ldots, \mu_p)'$ et $\mathbf{M} = [C_1\, C_2\, \ldots, C_p]$ avec C_i le vecteur colonne de terme j égal à 1 si $n_1 + \ldots + n_{i-1} < j \leq n_1 + \ldots + n_i$.

On obtient que

$$\hat{\mu}_i = \bar{Y}_{i\cdot} := \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$$

et

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \hat{\mu}_i)^2$$

avec $n := \sum_i n_i$. On a $(\boldsymbol{M}'\boldsymbol{M})_{ij} = n_i$ si $i = j$ et 0 sinon. Le théorème précédent implique que si on a $\frac{n_i}{n} \to p_i > 0$ pour tout i, alors les estimateurs sont consistants et asymptotiquement normaux.

Modèle ANOVA à deux facteurs Supposons que l'on veuille expliquer la production d'une parcelle de terrain par un effet du terrain et un effet de l'engrais utilisé. On a donc

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \qquad i = 1, \dots I \ ; \ j = 1, \dots, J \ ; \ k = 1, \dots, n_{ij}.$$

avec

- Y_{ijk} : production de la parcelle k du terrain i avec l'engrais j,
- μ : effet moyen,
- α : effet du terrain,
- β : effet de l'engrais.

On ajoute les contraintes $\sum_{i=1,\dots,I} \alpha_i = 0$ et $\sum_{j=1,\dots,J} \beta_j = 0$ pour avoir un modèle identifiable (on a ainsi unicité de la solution $\boldsymbol{M}\hat{\boldsymbol{\theta}} = \hat{\mathbf{m}}$). On obtient que $\hat{\mu} = \bar{Y}$, $\hat{\alpha}_i = \bar{Y}_{i\cdot} - \bar{Y}$ et $\hat{\beta}_j = \bar{Y}_{j\cdot} - \bar{Y}$. Si on a $\frac{n_{ij}}{n} \to p_{ij} > 0$, les estimateurs sont consistants et asymptotiquement normaux.

7.4 Intervalles de confiance dans le modèle de régression

7.4.1 Intervalle de confiance et test de Student pour les paramètres

On a $\hat{\theta}_i \sim \mathcal{N}(\theta_i, \sigma^2[(\boldsymbol{M}'\boldsymbol{M})^{-1}]_{ii})$. On en déduit que la variable

$$\frac{\hat{\theta}_i - \theta_i}{\hat{\sigma}\sqrt{[(\boldsymbol{M}'\boldsymbol{M})^{-1}]_{ii}}}$$

suit une loi de Student $\mathcal{T}(n-p)$. Ce qui donne l'intervalle de confiance pour θ_i

$$\hat{\theta}_i \pm \hat{\sigma} t_{n-p,1-\alpha/2}\sqrt{[(\boldsymbol{M}'\boldsymbol{M})^{-1}]_{ii}}.$$

Pour savoir si $\theta_i \neq 0$, on regarde si 0 est contenu dans l'intervalle. Cela revient à calculer la statistique de test

$$T_i = \frac{\hat{\theta}_i}{\hat{\sigma}\sqrt{[(\boldsymbol{M}'\boldsymbol{M})^{-1}]_{ii}}}$$

et à la comparer aux quantiles de la loi de Student à $n-p$ degrés de liberté.

Dans la pratique, il n'est pas recommandé d'utiliser les tests de Student définis ci-dessus pour choisir directement un sous-groupe de régresseurs dans le modèle de régression multiple car, lorsque les prédicteurs sont corrélés entre eux, ces tests sont beaucoup trop conservatifs (voir plus loin pour les méthodes de sélection de variable).

Relation linéaire entre les paramètres. Si on veut tester plus généralement $\mathcal{H}_0 : \mathbf{c}'\boldsymbol{\theta} = a$ contre $\mathcal{H}_1 : \mathbf{c}'\boldsymbol{\theta} \neq a$ pour $\mathbf{c} \in \mathbb{R}^p$ fixé. Il suffit d'utiliser que

$$\mathbf{c}'\hat{\boldsymbol{\theta}} \sim \mathcal{N}(\mathbf{c}'\boldsymbol{\theta}, \sigma^2\mathbf{c}'(\boldsymbol{M}'\boldsymbol{M})^{-1}\mathbf{c}).$$

Ainsi la statistique de test

$$T = \frac{\mathbf{c}'\hat{\boldsymbol{\theta}} - a}{\hat{\sigma}\sqrt{\mathbf{c}'(\boldsymbol{M}'\boldsymbol{M})^{-1}\mathbf{c}}}$$

suit une loi de Sudent $\mathcal{T}(n-p)$ sous $\mathcal{H}_0$. Il reste à comparer T aux quantiles correspondants.

Remarque. C'est aussi un test de sous-hypothèse linéaire; il est équivalent d'utiliser un F-test (cf la suite).

7.4.2 Intervalle de confiance de la régression

On aimerait avoir un intervalle de confiance pour la vraie valeur de $\mathbf{c}'\boldsymbol{\theta}$ pour un $\mathbf{c} \in \mathbb{R}^p$. C'est donc un intervalle de confiance sur la moyenne de Y_i quand le vecteur des observations (à savoir $(1, x_i^1, \ldots, x_i^{p-1})'$ dans (1)) est égal à $\mathbf{c}$. En utilisant l'analyse du paragraphe précédent, un intervalle de confiance au niveau $1-\alpha$ pour $\mathbf{c}'\boldsymbol{\theta}$ est donné par

$$\mathbf{c}'\hat{\boldsymbol{\theta}} \pm t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{\mathbf{c}'(\boldsymbol{M}'\boldsymbol{M})^{-1}\mathbf{c}}.$$

7.4.3 Intervalle de prédiction

Soit $\mathbf{c} \in \mathbb{R}^p$. On effectue N mesures de Y_i au point $\mathbf{c}$, et on prend leur moyenne empirique. On aimerait donner un intervalle auquel doit appartenir le résultat. Cet intervalle est donné par

$$\mathbf{c}'\hat{\boldsymbol{\theta}} \pm t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{\mathbf{c}'(\boldsymbol{M}'\boldsymbol{M})^{-1}\mathbf{c} + 1/N}$$

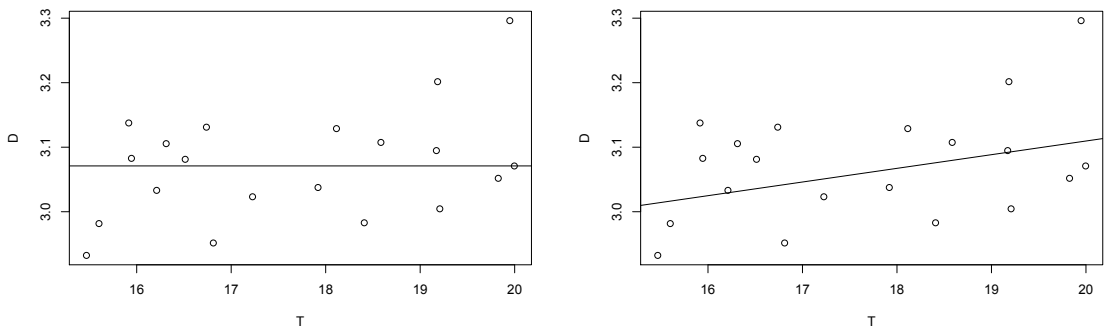
7.5 Test de sous-hypothèse linéaire: le F-test

Définition. *On dit que le modèle linéaire* (ω) *dans* $\mathbb{R}^n$ *est un sous-modèle linéaire de* (Ω) *dans* $\mathbb{R}^n$ *si et seulement si* $Im(\boldsymbol{M}_\omega) \subset Im(\boldsymbol{M}_\Omega)$. Supposons que l'on observe un vecteur $\boldsymbol{Y}$ dans le modèle linéaire régulier :

$$(\Omega) \ : \ \boldsymbol{Y} = \boldsymbol{M}_\Omega\, \boldsymbol{\theta}_\Omega + \boldsymbol{\varepsilon}$$

avec $\mathrm{rg}(\boldsymbol{M}_\Omega) = p$. On considère un sous-modèle linéaire régulier (ω) de (Ω), défini par la matrice $\boldsymbol{M}_\omega$. On a donc $\mathrm{Im}(\boldsymbol{M}_\omega) \subset \mathrm{Im}(\boldsymbol{M}_\Omega)$ et on suppose que $\mathrm{rg}(\boldsymbol{M}_\omega) = q < p$. On souhaite déterminer si le modèle (ω) est suffisamment complexe pour expliquer les données. Il s'agit pour cela de tester si $E\boldsymbol{y}$ appartient à $\mathrm{Im}(\boldsymbol{M}_\omega)$ ou non. Ce contexte englobe notamment la comparaion de deux modèles de régression linéaire multiple, lorsque l'un des deux modèles contient tous les régresseurs de l'autre. Dans ce cas particulier la matrice $\boldsymbol{M}_\omega$ est composée de colonnes extraites de la matrice $\boldsymbol{M}_\Omega$.

Exemple pour la régression simple : Dans l'industrie : on mesure un échantillon de pièces métalliques de très petites tailles (mm) soumises à des températures variables. Y-a-t-il un effet de la température sur la taille des pièces métalliques ? Il s'agit en fait de tester si l'on peut modéliser la relation moyenne entre taille et température par une droite horizontale (de pente nulle).



(ω) : échantillon iid (à gauche), (Ω) régression linéaire simple (à droite).

On adopte les notation suivantes :

- Prédictions :
$$\hat{\mathbf{Y}}_\Omega := \Pi_{\mathrm{Im}(\boldsymbol{M}_\Omega)}(\boldsymbol{Y}) \quad \text{et} \quad \hat{\mathbf{Y}}_\omega := \Pi_{\mathrm{Im}(\boldsymbol{M}_\omega)}(\boldsymbol{Y}),$$
- Somme des carrés résiduels:
$$SC_\Omega = \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}_\Omega\|^2 = \|\Pi_{\mathrm{Im}(\boldsymbol{M}_\Omega)}(\boldsymbol{Y})\|^2 \quad \text{et} \quad SC_\omega = \|\boldsymbol{Y} - \hat{\mathbf{Y}}_\omega\|^2 = \|\Pi_{\mathrm{Im}(\boldsymbol{M}_\omega)}(\boldsymbol{Y})\|^2,$$
- l'espace vectoriel G est le supplémentaire orthogonal de $\mathrm{Im}(\boldsymbol{M}_\omega)$ dans $\mathrm{Im}(\boldsymbol{M}_\Omega)$:
$$\mathrm{Im}(\boldsymbol{M}_\Omega) = \mathrm{Im}(\boldsymbol{M}_\omega) \overset{\perp}{\oplus} G.$$

Proposition. *La statistique*
$$F_{\omega|\Omega} := \frac{\left\|\hat{\mathbf{Y}}_\Omega - \hat{\mathbf{Y}}_\omega\right\|^2 /(p-q)}{\left\|\boldsymbol{Y} - \hat{\mathbf{Y}}_\Omega\right\|^2 /(n-p)}$$
vérifie
$$F_{\omega|\Omega} = \frac{\|\Pi_G(\boldsymbol{Y})\|^2 /(p-q)}{\left\|\boldsymbol{Y} - \hat{\mathbf{Y}}_\Omega\right\|^2 /(n-p)} = \frac{(SC_\omega - SC_\Omega)/(p-q)}{SC_\Omega/(n-p)}.$$
De plus, on a
$$F_{\omega|\Omega} \sim \mathcal{F}\left(p-q, n-p, \sigma^{-2}\|\Pi_G \mathrm{E}(\boldsymbol{Y})\|^2\right),$$
et donc, si $\mathrm{E}\boldsymbol{Y} \in Im(\boldsymbol{M}_\omega)$,
$$F_{\omega|\Omega} \sim \mathcal{F}(p-q, n-p).$$

Ce résultat permet de construire un test sur les modèles linéaires emboîtés, que l'on note $(\omega|\Omega)$. Plus précisément on considère les deux hypothèses
$$(H_0) \ : \ \mathrm{E}\boldsymbol{Y} \in \mathrm{Im}(\boldsymbol{M}_\omega) \quad \text{et} \quad (H_1) \ : \ \mathrm{E}\boldsymbol{Y} \in \mathrm{Im}(\boldsymbol{M}_\Omega) \setminus \mathrm{Im}(\boldsymbol{M}_\omega).$$

Corollaire . *Pour* $\alpha \in (0,1)$, *soit* $f_{p-q,n-p,1-\alpha}$ *le quantile* $1-\alpha$ *de la loi de Fisher* $\mathcal{F}(p-q, n-p)$. *Le test* $\mathbf{1}_{\{F_{\omega|\Omega} \geq f_{p-q,n-p,1-\alpha}\}}$ *est un test de* $(\omega|\Omega)$ *de niveau* α, *il s'agit du F-test des modèles emboîtés.*

Le numérateur de la statistique de test $F_{\omega|\Omega}$ mesure la distance entre $\Pi_{\mathrm{Im}(\boldsymbol{M}_\Omega)}(\boldsymbol{Y})$ et $\Pi_{\mathrm{Im}(\boldsymbol{M}_\omega)}(\boldsymbol{Y})$, il est donc naturel de rejeter les situations où $F_{\omega|\Omega}$ prend de grandes valeurs. En revanche, rejeter les petites valeurs de $F_{\omega|\Omega}$ n'est pas pertinent car cela reviendrait à rejeter les situations où $\Pi_{\mathrm{Im}(\boldsymbol{M}_\Omega)}(\boldsymbol{Y}) \approx \Pi_{\mathrm{Im}(\boldsymbol{M}_\omega)}(\boldsymbol{Y})$ c'est-à-dire les situations où $\mathrm{E}\boldsymbol{Y}$ est "proche" de l'espace $\mathrm{Im}(\boldsymbol{M}_\omega)$.

Dans les logiciels de statistique, la réalisation d'un F-test est accompagnée d'un tableau de l'analyse de la variance, ou encore tableau de décomposition des carrés. Ce tableau prend la forme suivante :

	Deg Lib	SC	Carré Moyen	F-value	Pr > F
Effet testé	p-q	$SC_\omega - SC_\Omega$	$(SC_\omega - SC_\Omega)/(p-q)$	$\frac{(SC_\omega - SC_\Omega)/(p-q)}{SC_\Omega/(n-p)}$	p-val. du F-test
Résidus	n -p	SC_Ω	$SC_\Omega/(n-p)$		

La colonne des Deg Lib (degrés de libertés) renseigne les degrés de libertés des lois χ^2 que suivent les sommes de carrés.

7.6 Sorties R

Sorties R pour régression simple

Nous reprenons le jeu de données portant sur les niveaux de pollution enregistrés sur 112 journées de l'été 2001 à Rennes. Nous étudions ici les sorties R de la procédure `lm` de R pour l'ajustement du modèle de régression simple de maxO3 par T12.

```
Call:
lm(formula = maxO3 ~ T12)

Residuals:
    Min      1Q  Median      3Q     Max
-38.079 -12.735   0.257  11.003  44.671

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -27.4196     9.0335  -3.035    0.003 **
T12           5.4687     0.4125  13.258   <2e-16 ***
---

Residual standard error: 17.57 on 110 degrees of freedom
Multiple R-squared:  0.6151,        Adjusted R-squared:  0.6116
F-statistic: 175.8 on 1 and 110 DF,  p-value: < 2.2e-16
```

Ces sorties R fournissent notamment des informations à propos de l'estimation des deux paramètres a (sur la ligne T12) et b (sur la ligne Intercept). La colonne `Standard Error` donne une estimation de l'écart type des deux estimateurs $\hat{a}$ et $\hat{b}$. Par exemple, on sait que la variance de $\hat{b}$ vaut $\frac{\sigma^2}{n\mathbb{V}(\boldsymbol{x})}$, quantité qui s'estime par $\frac{\hat{\sigma}^2}{n\mathbb{V}(\boldsymbol{x})}$. La sortie donne aussi les résultats des tests de $a = 0$ et de $b = 0$ (tests de Student).

L'estimateur $\hat{\sigma}$ de σ est renseigné par le champ `Residual standard error`. Les degrés de libertés de $\hat{\sigma}^2$ sont aussi renseignés, ici $n - 2$ pour ce modèle de régression simple.

Soit (Ω) le modèle de régression simple et (ω) le modèle échantillon i.i.d. $\mathcal{N}(\mu, \sigma^2)$, c'est-à-dire le sous-modèle de (Ω) de pente nulle $b = 0$. La statistique F donnée dans la sortie correspond au test des modèles emboîtés $(\omega|\Omega)$, qui est identique au test de la pente. On peut vérifier que la valeur de la statistique de Student t indiquée sur la ligne de T12 vérifie $t^2 = F$. La sortie ci-dessous est le tableau d'analyse de la variance pour ce modèle de régression linéaire simple.

```
Analysis of Variance Table

Response: maxO3
           Df Sum Sq Mean Sq F value    Pr(>F)
T12         1  54244   54244  175.76 < 2.2e-16 ***
Residuals 110  33948     309
```

Sorties R pour la régression multiple

Les sorties ci-dessous ont été obtenues avec la procédure `lm` de R pour les données Ozone en effectuant l'ajustement du modèle de régression multiple de maxO3 par T9, T12, T15, Ne9, N12, N15, Vx9, Vx12, VX15 et MaxO3v. La dimension du modèle est ici égale à $p = 11$.

```
Call:
```

```
lm(formula = maxO3 ~ T9 + T12 + T15 + Ne9 + Ne12 + Ne15 + Vx9 +
    Vx12 + Vx15 + maxO3v)

Residuals:
    Min      1Q  Median      3Q     Max
-53.566  -8.727  -0.403   7.599  39.458

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.24442   13.47190   0.909   0.3656
T9          -0.01901    1.12515  -0.017   0.9866
T12          2.22115    1.43294   1.550   0.1243
T15          0.55853    1.14464   0.488   0.6266
Ne9         -2.18909    0.93824  -2.333   0.0216 *
Ne12        -0.42102    1.36766  -0.308   0.7588
Ne15         0.18373    1.00279   0.183   0.8550
Vx9          0.94791    0.91228   1.039   0.3013
Vx12         0.03120    1.05523   0.030   0.9765
Vx15         0.41859    0.91568   0.457   0.6486
maxO3v       0.35198    0.06289   5.597 1.88e-07 ***

Residual standard error: 14.36 on 101 degrees of freedom
Multiple R-squared:  0.7638,        Adjusted R-squared:  0.7405
F-statistic: 32.67 on 10 and 101 DF,  p-value: < 2.2e-16
```

Ces sorties contiennent tout d'abord des informations à propos de l'estimation des paramètres a_j. La colonne `Standard Error` donne une estimation de l'écart type des estimateurs $\hat{a}_j$. On a vu que la variance de l'estimateur $\hat{a}_j$ vaut $\sigma^2 w_{jj}$ et on estime cette quantité par $\hat{\sigma}^2 w_{jj}$. La sortie donne aussi le résultat de chacun des tests de $a_j = 0$ contre $a_j \neq 0$ (test de Student). Le fait que la majorité des p-values soient élevées s'explique par une corrélation importante des régresseurs entre eux. Il n'est pas recommandé d'utiliser ces tests pour retirer des paquets de variables. En effet l'effet d'une variable n'est testé ici qu'individuellement, c'est-à-dire en conservant tous les autres régresseurs.

L'estimateur $\hat{\sigma}$ de σ est encore renseigné par le champ `Residual standard error`. Les degrés de libertés de $\hat{\sigma}^2$ sont aussi renseignés, ici $n - 11$ pour ce modèle de régression multiple.

Le F-test fourni à la fin des sorties est le "F-test global" pour ce modèle de régression multiple: soit (Ω) le modèle de régression multiple défini ci-dessus, et (ω) le modèle échantillon i.i.d. $\mathcal{N}(\mu, \sigma^2)$, c'est-à-dire le sous-modèle de (Ω) pour lequel aucune des variables n'est retenue. La statistique F donnée dans la sortie correspond au test des modèles emboîtés $(\omega|\Omega)$.

Pour évaluer la pertinence d'inclure ou non chacune des variables, en respectant l'ordre de déclaration de la commande `lm`, nous pouvons considérer l'analyse de la variance des différentes variables :

```
Analysis of Variance Table

Response: maxO3
          Df Sum Sq Mean Sq  F value    Pr(>F)
T9         1  43138   43138 209.1954 < 2.2e-16 ***
T12        1  11125   11125  53.9484  5.42e-11 ***
T15        1    876     876   4.2467 0.0418957 *
Ne9        1   3244    3244  15.7313 0.0001366 ***
```

```
Ne12         1    232     232   1.1260 0.2911615
Ne15         1      5       5   0.0253 0.8740175
Vx9          1   2217    2217  10.7503 0.0014307 **
Vx12         1      1       1   0.0050 0.9437342
Vx15         1     67      67   0.3251 0.5698239
maxO3v       1   6460    6460  31.3251  1.88e-07 ***
Residuals  101  20827     206
```

Chaque ligne correspond à un F-test particulier mais comme vous pouvez le vérifier, les valeurs des statistiques de test ne correspondent pas aux carrés des valeurs des statistiques de Student, comme c'est le cas pour la régression simple. Dans ce tableau, les variables sont incluses dans le modèle de régression dans l'ordre de déclaration de la commande `lm` (d'abord T9, puis T15 etc.) et les F-test proposés comparent deux modèles successifs. Soient $(\Omega[T9])$ et $(\Omega[T9, T12])$ les modèles de régression linéaire incluant respectivement T9, et T9 et T12. Le F-test sur la ligne de T12 est en fait défini par

$$\frac{\left(SC_{\Omega[T9]} - SC_{\Omega[T9,T12]}\right)/(3-2)}{SC_\Omega/(n-p)}$$

où l'on rappelle que (Ω) est le modèle de régression multiple complet (avec toutes les variables).

Exercice : Donner un résultat analogue à la Proposition 7.5 pour cette statistique de test et montrer que cette statistique de test permet effectivement de tester l'effet de la variable T12.

Remarque.Les résultats du tableau de l'analyse de la variance dépendent fortement de l'ordre selon lequel les variables sont déclarées dans la commande `lm`.

7.7 Régions de confiance pour le vecteur des paramètres

Nous avons donné précédemment des intervalles de confiance pour les paramètres pris de façon individuelle. Il est aussi possible de construire un ellipsoïde de confiance pour le vecteur θ des paramètres d'un modèle de régression.

Proposition. *Pour un modèle de régression multiple régulier à $p-1$ prédicateurs, on considère la statistique*

$$\tilde{F} := \frac{\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)' (\boldsymbol{M}'\boldsymbol{M}) \left(\hat{\boldsymbol{\theta}} - \theta\right)/p}{\hat{\sigma}^2}.$$

Alors $\tilde{F} \sim \mathcal{F}(p, n-p)$ et l'ellipsoïde

$$\left\{\boldsymbol{u} \in \mathbb{R}^p \mid \left(\hat{\boldsymbol{\theta}} - \boldsymbol{u}\right)' \boldsymbol{M}'\boldsymbol{M} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{u}\right) \leq p\hat{\sigma}^2 f_{p,n-p,1-\alpha}\right\}$$

est une région de confiance pour $\boldsymbol{\theta}$ de niveau $1-\alpha$.

En pratique, ces ellipsoïdes de confiance ne sont pas faciles à manipuler. La méthode de Bonferroni permet la construction des zones de confiance à l'aide de pavés de confiance pour $\boldsymbol{\theta}$ en s'appuyant sur les intervalles de confiance des θ_j.

Proposition. *Soit $\alpha > 0$, et pour tout $j \in \{0, \ldots p-1\}$ soit $\hat{I}_j$ un intervalle de confiance pour le coefficient a_j de niveau $1 - \frac{\alpha}{p}$. Alors $\hat{I}_0 \times \ldots \times \hat{I}_{p-1}$ est une région de confiance pour le vecteur $\boldsymbol{\theta}$ de niveau $1-\alpha$.*

Remarque. Ces pavés sont certes plus faciles à définir et à représenter que l'ellipsoïde de confiance mais en réalité le pavé a une aire plus grande que l'ellipsoïde, ceci d'autant plus que p est grand.

7.8 Analyse des résidus

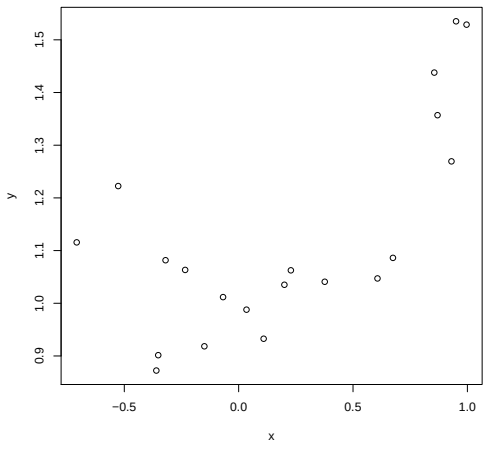
Définition On appelle vecteur des résidus le vecteur $\hat{\boldsymbol{\varepsilon}} := \boldsymbol{Y} - \hat{\boldsymbol{Y}}$.

Les hypothèses du modèle linéaire (Gaussien) portent surtout sur le vecteur $\boldsymbol{\varepsilon}$. Ce vecteur est en pratique inconnu, alors que le vecteur des résidus $\hat{\boldsymbol{\varepsilon}}$ est disponible. L'analyse des résidus a pour objectif de valider les hypothèses du modèle linéaire, en s'appuyant sur des propriétés devant être vérifiées par $\hat{\boldsymbol{\varepsilon}}$ dans ce contexte.

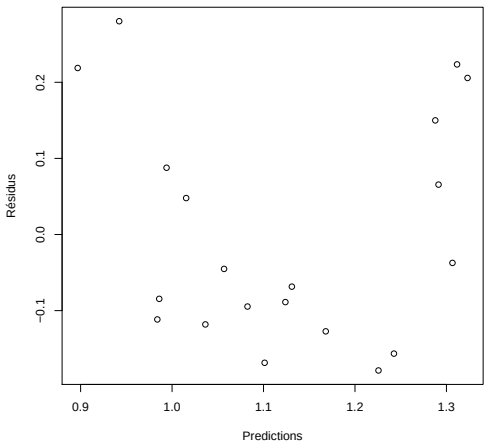
7.8.1 Validation du modèle à l'aide des résidus

L'analyse des résidus consiste à s'appuyer sur de multiples représentations graphiques $(x_i, \hat{\epsilon}_i)$, $(y_i, \hat{\epsilon}_i)$, $(\hat{y}_i, \hat{\epsilon}_i)$, $(i, \hat{\epsilon}_i)$, (i, t_i) ... pour détecter d'éventuelles entorses aux hypothèses du modèle linéaire.

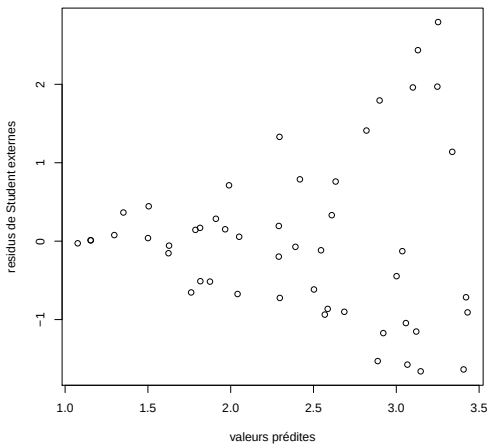
Linéarité. On cherche d'éventuels défauts de linéarité dans les nuages (x_i, y_i).



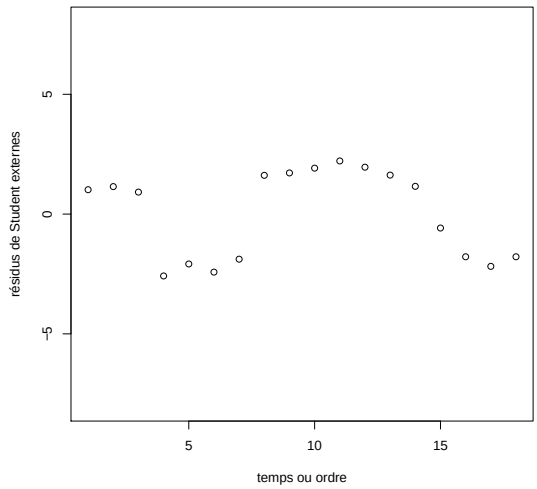
Problème d'adéquation au modèle. Puisque $\hat{\boldsymbol{\varepsilon}}$ et $\hat{\mathbf{Y}}$ sont indépendants, il ne doit pas y avoir de relation franche entre $\hat{\mathbf{Y}}$ et $\hat{\boldsymbol{\varepsilon}}$. Dans le cas contraire, il faudra améliorer le modèle.



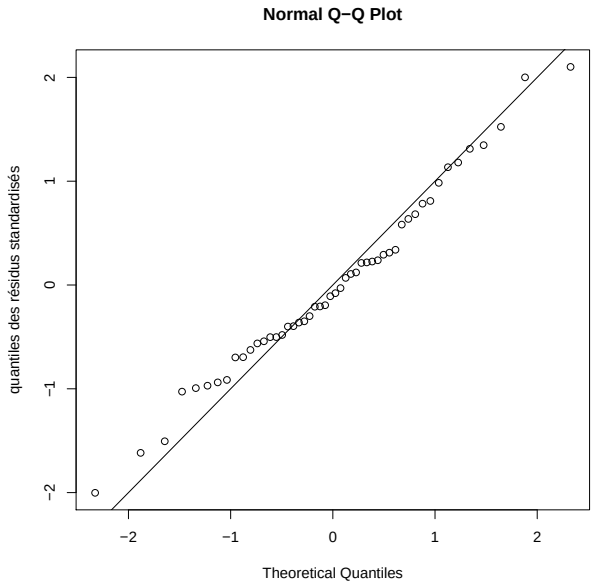
Hétéroscédasticité. Si l'hypothèse H2 n'est pas vérifiée, on observe des structures suspectes dans le nuage $(\hat{Y}_i, r_i)$ ou dans le nuage $(\hat{Y}_i, t_i)$, par exemple en forme d'entonnoir lorsque le bruit augmente avec l'un des prédicteurs. Dans ce cas on pourra effectuer une transformation des données (par exemple passage au log) pour se ramener à une situation où l'hétéroscédasticité est plus faible.



Corrélation. D'après H3, les ε_i sont indépendants. En pratique, on vérifie que les résidus sont peu corrélés. Si la dimension temporelle des données ne peut être ignorée, le modèle linéaire est à proscrire et on pourra par exemple utiliser des modèles de séries chronologiques.



Normalité. Il s'agit (éventuellement) de contrôler la normalité de $\boldsymbol{\varepsilon}$ ou des résidus de Student. On gardera à l'esprit que la validation de l'hypothèse H4 est moins fondamentale si la taille de l'échantillon est importante (en pratique plus de 50 observations). Ce controle est généralement effectué graphiquement à l'aide d'un QQ-plot des résidus de Student :



- en abscisses : les quantiles théoriques de la loi normale centrée réduite,
- en ordonnées : les valeurs ordonnées des résidus de Student (internes ou externes) = quantiles empiriques des résidus de Student.

De façon générale, on appelle QQ plot un graphique croisant les quantiles de deux distributions de probabilités.

7.8.2 Valeurs aberrantes

Notons $\Pi_{Im(\boldsymbol{M})} = \boldsymbol{M}(\boldsymbol{M}'\boldsymbol{M})^{-1}\boldsymbol{M}' = (h_{ij})_{1\leq i,j\leq n}$ la matrice de projection orthogonale sur $Im(\boldsymbol{M})$.

Proposition. *Sous les hypothèses H1-H4 on a :*

- $\hat{\boldsymbol{\varepsilon}} \sim \mathcal{N}\left(0, \sigma^2(\mathbf{I}_n - \Pi_{Im(\boldsymbol{M})})\right)$,
- $\hat{\varepsilon}_i \sim \mathcal{N}\left(0, \sigma^2(1-h_{ii})\right)$,
- $\hat{\boldsymbol{\varepsilon}}$ *et* $\hat{\boldsymbol{Y}}$ *sont indépendants.*

Dans la pratique, on considère aussi d'autres versions des résidus. Puisque les résidus n'ont pas tous la même variance, on les rend comparables en définissant les résidus de Student internes (ou résidus standardisés) :

$$r_i := \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}.$$

Cependant cette standardisation n'aboutit pas en réalité à une statistique de Student, on considère donc aussi les résidus de Student externe :

$$t_i := \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}$$

où $\hat{\sigma}^2_{(i)}$ est le $\hat{\sigma}^2$ calculé sans l'observation i. Sous H4, on a alors que t_i suit une loi de Student $\mathcal{T}(n-p-1)$. On peut alors tester la présence d'observations aberrantes (en utilisant la méthode de Bonferroni pour construire un test multiple). En pratique, on se contente souvent de vérifier que les résidus studentisés sont ou non dans l'intervalle $[-2, 2]$. On peut de plus montrer que

$$t_i = r_i\sqrt{\frac{n-p+1}{n-p-r_i^2}}$$

ce qui facilite leur calcul.

Il faut retenir qu'une valeur aberrante indique si une variable expliquer y_i a une valeur mal expliquée par le modèle.

7.9 Phénomène de levier et influence

On s'intéresse ici à la sensibilité des estimateurs aux observations : il s'agit de repérer les observations i telles qu'une faible variation des valeurs Y_i et $\boldsymbol{x}_i = (x_i^1, \ldots, x_i^p)'$ induit une variation importante des prédictions. On rappelle que $\hat{\boldsymbol{Y}} = \Pi_{Im(\boldsymbol{M})}\boldsymbol{Y}$, où $\Pi_{Im(\boldsymbol{M})}$ est la matrice de projection, aussi appelée *hat matrix*:

$$\Pi_{Im(\boldsymbol{M})} = \boldsymbol{M}(\boldsymbol{M}'\boldsymbol{M})^{-1}\boldsymbol{M}' = (h_{ij})_{1\leq i,j\leq n}\,.$$

Proposition. *Les h_{ij} vérifient :*

1. $\sum_i h_{ii} = p$
2. $\sum_{ij} h_{i,j}^2 = p$
3. $h_{ii}(1-h_{ii}) = \sum_{i\neq j} h_{ij}$

4. $0 \leq h_{ii} \leq 1$ *et* $-0.5 \leq h_{ij} \leq 0.5$ *si* $i \neq j$.

5. *Si* $h_{ii} = 1$ *alors* $h_{ij} = 0$ *pour tout* $j \neq i$.

6. *Si* $h_{ii} = 0$ *alors* $h_{ij} = 0$ *pour tout* $j \neq i$.

***x*-outlier.** Dans l'expression de la prévision de l'observation i

$$\hat{Y}_i = \left(\Pi_{Im(\boldsymbol{M})}\boldsymbol{Y}\right)_i = \sum_{k=1}^{n} h_{ik}Y_k,$$

l'observation Y_i intervient avec un poids h_{ii}, où h_{ii} est le i-ème coefficient de la diagonale de $\Pi_{Im(\boldsymbol{M})}$. Ce coefficient mesure donc la capacité de l'observation i à influencer la valeur de sa prédiction, et donc à modifier l'estimation des paramètres du modèle. Notons que d'après la proposition précédente, un h_{ii} élevé signifie $h_{ii} \approx 1$. Dans le cas de la régression simple, on peut calculer les h_{ii} explicitement :

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{k=1}^{n}(x_k - \overline{x})^2}.$$

Le coefficient h_{ii} est donc élevé lorsque que x_i est loin du coeur du nuage $\overline{x}$. L'importance du coefficient h_{ii} peut se voir aussi dans la formule suivante:

$$y_i - \hat{y}_i = (1 - h_{ii})(y_i - \hat{y}_{i,(i)})$$

où $\hat{y}_{i,(i)}$ représente la valeur prédite par la régression en x_i lorsqu'on enlève l'observation (x_i, y_i). En d'autres termes, lorsque h_{ii} est élevé (donc proche de 1), la régression tend à se rapprocher du point (x_i, y_i) lorsque la i-ième observation est incluse. On dira qu'un point (x_i, y_i) est un point levier si $h_{ii} > 2p/n$ (Hoagin & Welsch 1978) ou $h_{ii} > 0.5$ (Huber 1981).

Il faut retenir qu'un point levier (ou x-outlier) est un point correspondant des variables explicatives x_i qui se trouvent éloignées du centre de gravité $\bar{x}$.

D de Cook. Un x outlier ne perturbe pas nécessairement l'estimation du modèle, il en a simplement la capacité. Pour illustrer le problème, on a représenté sur le graphique ci-dessous la droite de régression simple ajustée sur une vingtaine d'observations, d'abord sans utiliser les deux points 1 et 2 puis en utilisant l'un ou l'autre de ces deux points. Seul le point 1 perturbe fortement l'estimation de la droite. Pour identifier les observations qui perturbent l'estimation du modèle, on utilise la statistique du D de Cook définie par

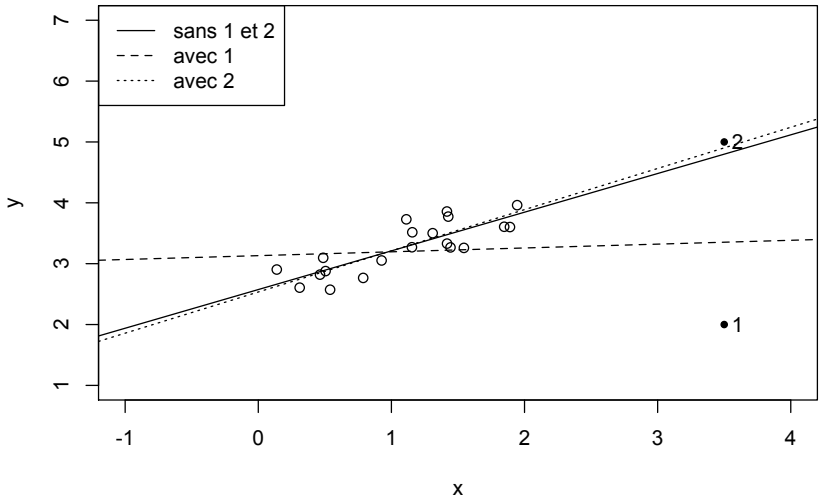
$$D_i := \frac{\sum_{k=1\ldots n}(\hat{Y}_{(i)k} - \hat{Y}_k)^2}{(p+1)\hat{\sigma}^2}$$

où $Y_{(i)k}$ est la prédiction de la k-ième observation effectuée sans utiliser l'observation i. Une observation telle que le D_i est important est appelé un point influent.

Proposition. *On a*

$$D_i = \frac{h_{ii}}{(p+1)(1-h_{ii})} r_i^2$$

où r_i *est le résidu standardisé (interne) de l'observation* i.



On appelle *y-outlier* une observation i telle que Y_i est située loin de sa valeur prédite $\hat{Y}_i$. Dans ce cas, les résidus de Student interne et externe de cette observation seront plus élevés (en valeur absolue) que ceux des autres observations. La proposition précédente montre qu'une observation est influente (D de Cook important) si et seulement si elle est à la fois un *x-outlier* ($h_{ii} \approx 1$) et un *y-outlier* (les magnitudes des résidus internes r_i et externes t_i sont élevées). Si l'on considère que celles-ci correspondent à des données aberrantes, on pourra retirer de la base de données les observations dont les D_i sont trop importants et estimer de nouveau le modèle régression. On pourra considèrer qu'une observation est :

- un x-outlier si $h_{ii} \geq \frac{2(p+1)}{n}$,
- un y-outlier si $|r_i| \geq 2$,
- un point influent si $D_i \geq \frac{4}{n}$.

Cependant dans la pratique il est difficile de retirer tous les points influents, on pourra donc comparer les D_i entre eux et traiter le cas des observations les plus influentes.