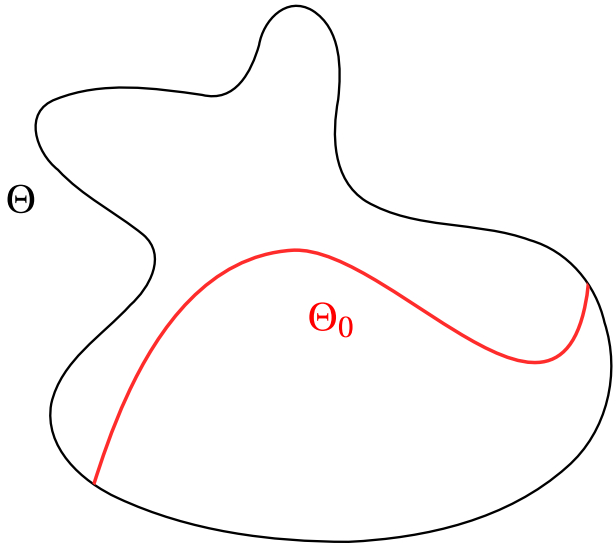
# STATISTIQUES II

HERVÉ PERDRY




# MASTER GEB

# 1. Vraisemblance (modèles à un paramètre)

## 1.1. La vraisemblance

### 1.1.1. Définitions

On considère une loi de probabilité $\mathscr{L}(\theta)$ qui dépend d'un paramètre $\theta$. La loi d'une variable $X \sim \mathscr{L}(\theta)$ sera soit

— une loi discrète, décrite par $\mathbb{P}_\theta(X = k)$ ;
— une loi continue à densité, décrite par $f_\theta(x)$.

L'ensemble des valeurs que $\theta$ peut prendre est « l'espace des paramètres ». On le note souvent $\Theta$.
Si par exemple $X \sim \mathscr{B}(p)$ (une loi de Bernoulli), le paramètre est noté $p$ et prend ses valeurs dans $\Theta = [0,1]$.
Les points 0 et 1 sont particuliers : on dit qu'il sont aux bords de l'espace des paramètres. Les autres points sont dits à l'intérieur de l'espace des paramètres.

La vraisemblance d'une valeur particulière de $\theta$, étant donnée une observation $x$, est définie par
— dans le cas discret, $L(\theta\,;\,x) = \mathbb{P}_\theta(X = x)$ ;
— dans le cas continu, $L(\theta\,;\,x) = f_\theta(x)$ ;

La vraisemblance est donc grosso modo « la probabilité d'observer $x$ » vue comme une fonction de $\theta$.

En fait, on peut prendre une définition un peu plus compliquée, mais qui permet par la suite de simplifier les calculs dans bien des cas :

La vraisemblance d'une valeur particulière de $\theta$, étant donnée une observation $x$, est définie par
— dans le cas discret, $L(\theta\,;\,x) = C(x) \times \mathbb{P}_\theta(x)$ ;
— dans le cas continu, $L(\theta\,;\,x) = C(x) \times f_\theta(x)$ ;
où $C(x)$ ne dépend pas de $\theta$. Dans la première définition on a pris $C(x) = 1$.

Si on suit cette définition la vraisemblance n'est définie qu'à une constante multiplicative près – ici « constante » signifie « dont la valeur ne dépend pas de $\theta$ ».

La vraisemblance de $\theta$, étant données $n$ observations indépendantes $x_1, \dots, x_n$, est définie par

$$L(\theta\,;\,x_1,\dots,x_n) = L(\theta\,;\,x_1) \times \cdots \times L(\theta\,;\,x_n).$$

On définit la log-vraisemblance par $\ell(\theta\,;\,x_1,\dots,x_n) = \log L(\theta\,;\,x_1,\dots,x_n)$ (il s'agit bien du logarithme népérien et non du logarithme décimal !). On remarque que

$$\ell(\theta\,;\,x_1,\dots,x_n) = \ell(\theta\,;\,x_1) + \cdots + \ell(\theta\,;\,x_n). \tag{1.1}$$

En pratique, quand il n'y a pas d'ambiguïté sur les observations $x_1, \dots, x_n$ à partir desquelles la vraisemblance est calculée, on notera simplement $L(\theta)$ et $\ell(\theta)$.

## 1.1.2. Exemples

Tout au long de ce chapitre, nous allons illustrer les notions rencontrées au travers de ces deux exemples « simples ».

### Exemple de la loi de Bernoulli

On joue à pile ou face, avec une pièce qui peut présenter un biais : on ne connaît donc pas à priori $p$, la probabilité de faire face. Le résultat d'un tirage est codé par la variable aléatoire X, avec X = 0 si on a fait pile et X = 1 si on a fait face : la loi de X est donc une variable de Bernoulli de paramètre $p$.

On fait $n$ observations $x_1,\dots,x_n$. La vraisemblance d'une observation $x$ est

$$\mathrm{L}(p\,;x)=\begin{cases}p & \text{si } x=1\\ 1-p & \text{si } x=0\end{cases}$$

La vraisemblance de l'ensemble des observations est donc $\mathrm{L}(p\,;x_1,\dots,x_n)=p^{n_1}(1-p)^{n_0}$, où $n_1$ est le nombre de tirages égaux à 1 dans les observations $x_1,\dots,x_n$ et $n_0$ est le nombre de tirages égaux à 0.

La log-vraisemblance d'une observation est

$$\ell(p\,;x)=\begin{cases}\log p & \text{si } x=1\\ \log(1-p) & \text{si } x=0\end{cases}$$

et la log-vraisemblance de l'ensemble des observations est

$$\ell(p\,;x_1,\dots,x_n)=n_1\log p+n_0\log(1-p)$$

Il est frappant de constater que la vraisemblance a résumé le résultat des $n$ $x_1,\dots,x_n$ à $n_1=x_1+\cdots+x_n$ (le nombre de tirages égaux à 1) et $n_0=n-n_1$ (le nombre de tirages égaux à 0). On conçoit qu'en effet, aucune information n'est contenue dans l'ordre dans lesquelles les tirages indépendants ont été effectués.

Dans la suite nous considérons qu'il y a eu 10 tirages : $x_1=0,x_2=0,x_3=1,x_4=0,x_5=1,x_6=1,x_7=1,x_8=0,x_9=0,x_{10}=0$, soit $n_1=4$ et $n_0=6$.

La figure 1.1 montre comment l'allure de la vraisemblance évolue quand la quantité d'observations prises en compte augmente.

### Exemple de la Gaussienne

On suppose qu'on observe $x_1,\dots,x_n$ tirés dans une loi Gaussienne $\mathcal{N}(\mu,\sigma_0^2)$. La variance $\sigma_0^2$ est supposée connue, on ne s'intéresse qu'à l'espérance $\mu$.

La densité de la loi normale est

$$f_\mu(x)=\frac{1}{\sqrt{2\pi\sigma_0^2}}e^{-\frac{(x-\mu)^2}{2\sigma_0^2}}.$$

Comme $\sigma_0^2$ est connue, on peut prendre $\mathrm{C}(x)=\sqrt{2\pi\sigma_0^2}$ dans la définition de la vraisemblance, et la vraisemblance pour une observation $x$ est

$$\mathrm{L}(\mu\,;x)=e^{-\frac{(x-\mu)^2}{2\sigma_0^2}}.$$

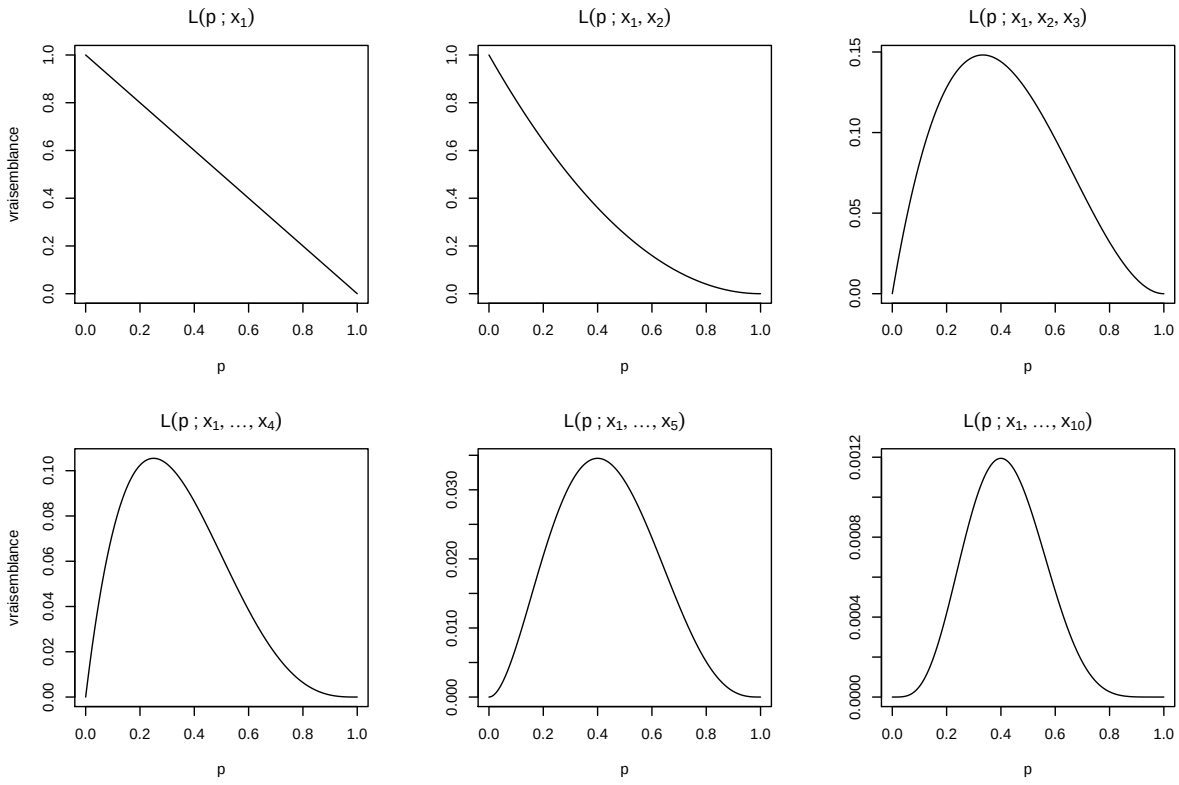



FIGURE 1.1. – Vraisemblances pour la loi de Bernoulli

La log-vraisemblance pour une observation est

$$\begin{aligned}\ell(\mu\,;\,x) &= -\frac{(x-\mu)^2}{2\sigma_0^2}\\ &= -\frac{1}{2\sigma_0^2}\left(x^2 - 2\mu x + \mu^2\right)\end{aligned}$$

On a alors

$$\begin{aligned}\ell(\mu\,;\,x_1,\ldots,x_n) &= \sum_{i=1}^{n}\ell(\mu;x_i)\\ &= -\frac{1}{2\sigma_0^2}\left(\left(x_1^2+\cdots+x_n^2\right) - 2\mu(x_1+\cdots+x_n) + n\mu^2\right)\\ &= -\frac{1}{2\sigma_0^2}\left(s_2 - 2\mu s_1 + n\mu^2\right)\end{aligned}$$

où on a posé $s_2 = x_1^2+\cdots+x_n^2$ et $s_1 = x_1+\cdots+x_n$. On peut en déduire la vraisemblance elle-même, en prenant l'exponentielle de la log-vraisemblance :

$$\mathrm{L}(\mu\,;\,x_1,\ldots,x_n) = \exp\left(-\frac{1}{2\sigma_0^2}\left(s_2 - 2\mu s_1 + n\mu^2\right)\right).$$

On constate cette fois que la vraisemblance a résumé les tirages $x_1,\ldots,x_n$ aux sommes $s_1$ et $s_2$.

Pour notre exemple numérique, nous prendrons $\sigma_0^2 = 1$ et 10 observations $x_1 = 2{,}47, x_2 = 1{,}71, x_3 = 0{,}35, x_4 = 1{,}8, x_5 = -0{,}37, x_6 = 1{,}99, x_7 = 1{,}47, x_8 = 0{,}12, x_9 = -0{,}83, x_{10} = 2{,}77$. On a $s_1 = 11{,}48$ et $s_2 = 27{,}0216$. La figure 1.2 montre comment l'allure de la vraisemblance évolue quand la quantité d'observations prises en compte augmente.

## 1.2. Les dérivées de la log-vraisemblance

Avant de pouvoir entrer vraiment dans le vif du sujet, nous avons besoin de quelques nouvelles définitions. Nous avons vu dans l'exemple de la Gaussienne que la log-vraisemblance est plus facile à manipuler que la vraisemblance ; la bonne nouvelle est qu'elle nous sera également plus utile en pratique.

La figure 1.3 illustre comment, alors que les vraisemblances ont des allures de Gaussienne, les log-vraisemblances ont des airs de parabole. On voit également que la vraisemblance et la log-vraisemblance atteignent leur maximum au même point (ligne verticale pointillée) : c'est parce que le logarithme est une fonction strictement croissante (et donc, si $\mathrm{L}(\theta) < \mathrm{L}(\theta_0)$, alors on a aussi $\ell(\theta) < \ell(\theta_0)$).

### 1.2.1. Score et information observée

Le score $\mathrm{U}(\theta\,;\,x_1,\ldots,x_n)$ est la dérivée de la log-vraisemblance :

$$\mathrm{U}(\theta\,;\,x_1,\ldots,x_n) = \frac{\partial}{\partial\theta}\ell(\theta\,;\,x_1,\ldots,x_n).$$

De la propriété (1.1) on déduit que

$$\mathrm{U}(\theta\,;\,x_1,\ldots,x_n) = \mathrm{U}(\theta\,;\,x_1)+\cdots+\mathrm{U}(\theta\,;\,x_n).$$

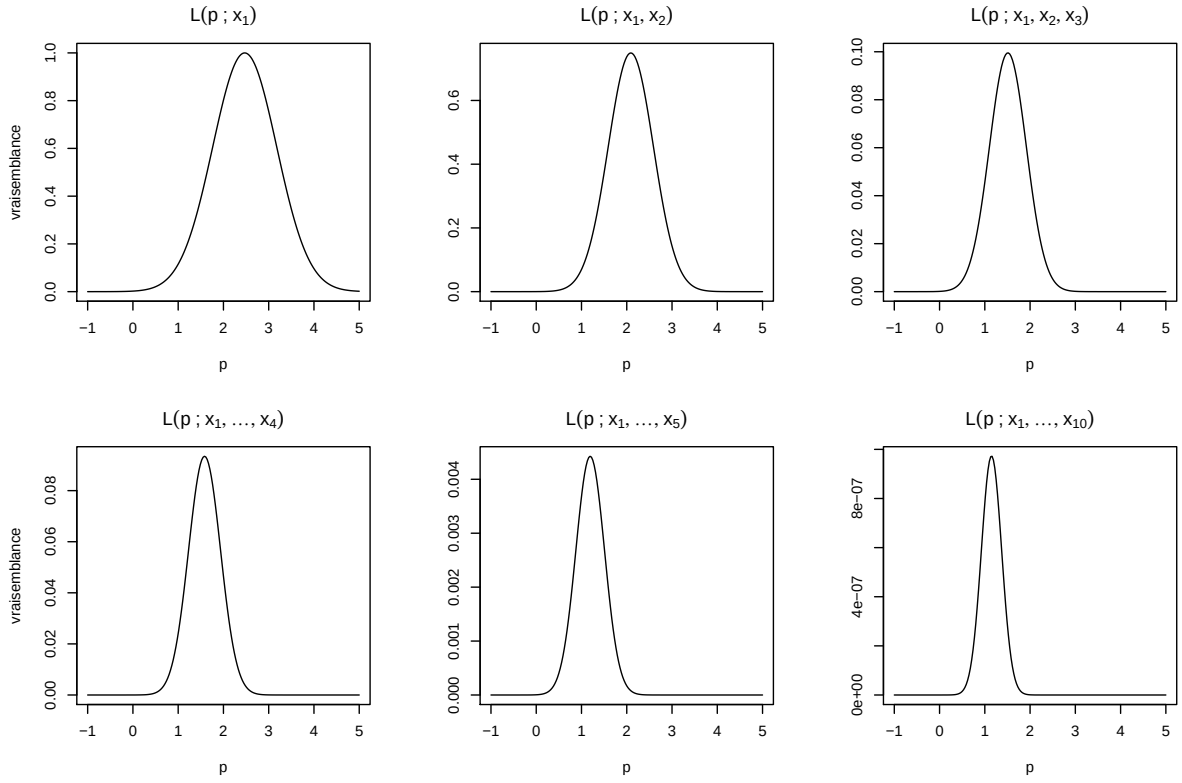



FIGURE 1.2. – Vraisemblances pour la loi normale

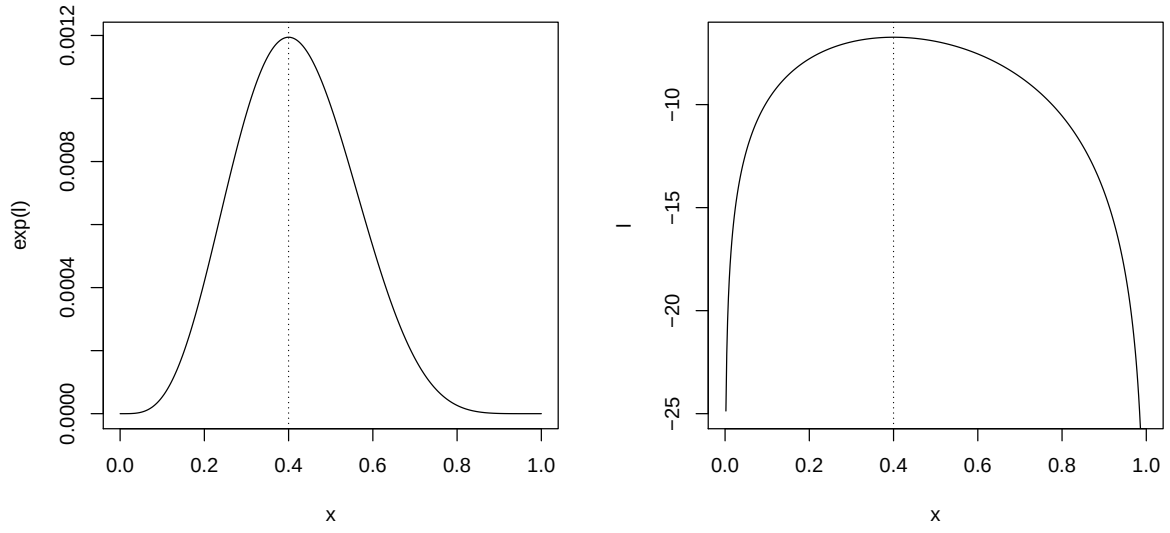



FIGURE 1.3. – Vraisemblance et log-vraisemblance pour la Bernoulli

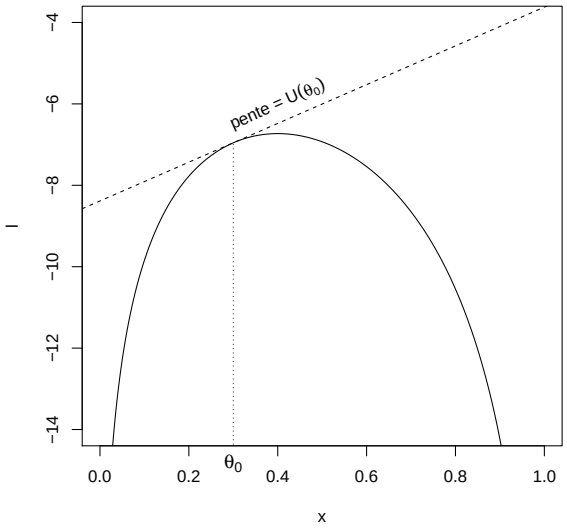



FIGURE 1.4. – Le score est la pente de la log-vraisemblance

Le score $U(\theta_0)$ a une interprétation géométrique simple : c'est la pente de la log-vraisemblance au point d'abscisse $\theta_0$, cf figure 1.4.

L'information observée $J(\theta\,;x_1,\dots,x_n)$ est l'opposé de la dérivée du score :

$$J(\theta\,;x_1,\dots,x_n) = -\frac{\partial}{\partial\theta}U(\theta\,;x_1,\dots,x_n).$$

Remarque : $U(\theta)$ et $J(\theta)$ ne dépendent pas du choix de $C(x)$ dans la définition de la vraisemblance. On a comme précédemment

$$J(\theta\,;x_1,\dots,x_n) = J(\theta\,;x_1)+\cdots+J(\theta\,;x_n).$$

## 1.2.2. Information de Fisher

On considère ici une log-vraisemblance $\ell(\theta\,;x)$, et l'information observée $J(\theta\,;x)$.

Supposons qu'on a $n$ observations $x_1,\dots,x_n$ issues de $n$ tirages dans la loi $\mathscr{L}(\theta_0)$. Les valeurs $J(\theta_0\,;x_1)$, …, $J(\theta_0,x_n)$ sont des valeurs prises au fil des expériences par la variable aléatoire $J(\theta_0\,;X)$ avec $X\sim\mathscr{L}(\theta_0)$. On va considérer l'espérance de cette variable aléatoire.

L'information de Fisher pour une expérience, $I_1(\theta_0)$, est l'espérance de $J(\theta_0\,;X)$ avec $X\sim\mathscr{L}(\theta_0)$.
L'information de Fisher pour $n$ expériences indépendantes, $I_n(\theta_0)$, est l'espérance de $J(\theta_0\,;X_1,\dots,X_n)$ avec $X_1,\dots,X_n\sim\mathscr{L}(\theta_0)$ : elle est égale à $nI_1(\theta_0)$.

Notons que pour une valeur de $\theta_0$ donnée, on peut estimer $I_1(\theta_0)$ par

$$\frac{1}{n}\big(J(\theta_0\,;x_1)+\cdots+J(\theta_0\,;x_n)\big) = \frac{1}{n}J(\theta_0\,;x_1,\dots,x_n);$$

c'est une conséquence de la loi des grands nombres, chaque $J(\theta_0 ; x_i)$ étant tiré dans une loi d'espérance $I_1(\theta_0)$.

Il faut insister sur la différence entre l'information de Fisher et l'information observée :

> L'information de Fisher $nI_1(\theta_0)$ ne dépend que de $\theta_0$, c'est une quantité déterministe ; l'information observée $J(\theta_0)$ dépend de l'échantillon que l'on a observé, c'est une variable aléatoire.

### 1.2.3. Exemples

#### Exemple de la loi de Bernoulli

On calcule le score pour une observation $x$ :

$$U(p ; x) = \begin{cases} \frac{1}{p} & \text{si } x = 1 \\ -\frac{1}{1-p} & \text{si } x = 0 \end{cases}$$

Donc on a

$$U(p ; x_1, \dots, x_n) = \frac{n_1}{p} - \frac{n_0}{1-p}$$

Procédons de même pour l'information observée :

$$J(p ; x) = \begin{cases} \frac{1}{p^2} & \text{si } x = 1 \\ \frac{1}{(1-p)^2} & \text{si } x = 0 \end{cases}$$

On peut donc calculer l'information de Fisher $I_1(p_0)$ en prenant l'espérance de $J(p ; X)$ avec $X \sim \mathscr{B}(p_0)$ :

$$\begin{aligned} I_1(p_0) &= E(J(p_0 ; X)) \\ &= J(p_0 ; 0) \times \mathbb{P}(X = 0) + J(p_0 ; 1) \times \mathbb{P}(X = 1) \\ &= \frac{1}{(1-p_0)^2}(1-p_0) + \frac{1}{p_0^2} p_0 \\ &= \frac{1}{1-p_0} + \frac{1}{p_0} \\ &= \frac{1}{p_0(1-p_0)}. \end{aligned}$$

L'information observée pour l'échantillon $x_1, \dots x_n$ est

$$J(p ; x_1, \dots, x_n) = \frac{n_1}{p^2} + \frac{n_0}{(1-p)^2}.$$

On note que $J(p_0)$, contrairement à $I_1(p_0)$, dépend des observations (à travers $n_1$ et $n_0$).

**Exercice** Vérifier que l'espérance de $J(p_0)$ est bien $nI_1(p_0)$.

#### Exemple de la Gaussienne

La log-vraisemblance pour une observation est $\ell(\mu ; x) = -\frac{1}{2\sigma_0^2}(x-\mu)^2$ (il s'agit là d'une « vraie » parabole !). On calcule donc

$$U(\mu ; x) = \frac{1}{\sigma_0^2}(x - \mu)$$

d'où

$$\begin{aligned} \mathrm{U}(\mu\,;\,x_1,\ldots,x_n) &= \frac{1}{\sigma_0^2}\left(\sum_{i=1}^{n} x_i - n\mu\right) \\ &= \frac{n}{\sigma_0^2}\left(\overline{x}-\mu\right), \end{aligned}$$

où bien sûr $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Passons à l'information observée :

$$\begin{aligned} \mathrm{J}(\mu\,;\,x) &= \frac{1}{\sigma_0^2} \\ \mathrm{J}(\mu\,;\,x_1,\ldots,x_n) &= \frac{n}{\sigma_0^2}. \end{aligned}$$

On a donc $\mathrm{I}_1(\mu) = \frac{1}{\sigma_0^2}$. Ici l'information de Fisher coincide avec son estimation par l'information observée, qui ne dépend que de la taille de l'échantillon et pas de la valeur de $\mu$.

## 1.3. Estimation par maximum de vraisemblance

L'idée est que la vraisemblance $\mathrm{L}(\theta\,;\,x_1,\ldots,x_n)$ mesure à quel point le paramètre $\theta$ rend les observations $x_1,\ldots,x_n$ vraisemblables.

La valeur de $\mathrm{L}(\theta_0)$ en un point donné $\mathrm{L}(\theta_0)$ n'est pas très utile en elle-même (elle dépend d'ailleurs de notre choix de $\mathrm{C}(x)$). Ce sont les valeurs relatives de $\mathrm{L}(\theta_0)$ et $\mathrm{L}(\theta_1)$, plus précisément le rapport $\mathrm{L}(\theta_0)/\mathrm{L}(\theta_1)$, qui permettent de savoir des deux paramètres $\theta_0$ et $\theta_1$ lequel est le plus vraisemblable. Tout ceci motive la définition de l'estimateur du maximum de vraisemblance.

> L'estimateur du maximum de vraisemblance (en abrégé, EMV) de $\theta$ est la quantité $\widehat{\theta}$ qui maximise $\mathrm{L}(\theta\,;\,x_1,\ldots,x_n)$. En notation mathématique on écrit
>
> $$\widehat{\theta} = \operatorname*{argmax}_{\theta} \mathrm{L}(\theta\,;\,x_1,\ldots,x_n).$$

On peut généralement trouver $\widehat{\theta}$ en cherchant les valeurs qui annulent le score $\mathrm{U}(\theta)$.

Pour que ceci soit bien sensé, il faut que les observations $x_1,\ldots,x_n$ aient été tirées selon une loi $\mathscr{L}(\theta_0)$ (la valeur précise de $\theta_0$ n'étant généralement pas connue de l'observateur, qui cherche justement à l'estimer). Il est clair que si l'expérience consistant à tirer $x_1,\ldots,x_n$ et à calculer $\widehat{\theta}$ est répétée plusieurs fois, on n'obtiendra pas à chaque fois la même valeur : $\widehat{\theta}$ est une variable aléatoire. Nous allons dans la suite préciser sa loi.

### 1.3.1. Propriétés de l'estimateur du maximum de vraisemblance

Les figures 1.1 et 1.2 montrent que quand la taille de l'échantillon augmente la vraisemblance prend une forme de cloche de plus en plus resserée autour de son maximum ; on a l'intuition que cela signifie que la précision de l'estimation augmente avec la taille de l'échantillon.

C'est ce qu'exprime le résultat suivant :

> — Si $\theta_0$ est un point intérieur de l'espace des paramètres ;
> — si la fonction $L(\theta) = L(\theta; x_1, \dots, x_n)$ est concave autour de son maximum $\widehat{\theta}$ (c'est-à-dire si $U(\widehat{\theta}) = 0$ et $I(\theta) > 0$)
>
> alors pour $n$ assez grand, la loi de $\widehat{\theta}$ est approximativement normale, d'espérance $\theta_0$ et de variance $(nI_1(\theta_0))^{-1}$.
>
> En pratique, si $\theta_0$ est inconnu, $nI_1(\theta_0)$ pourra être estimé par $nI_1\left(\widehat{\theta}\right)$ ou même par $J\left(\widehat{\theta}; x_1, \dots, x_n\right)$.

Les deux conditions préliminaires du théorèmes sont vraies dans les cas simples. Retenez surtout que si la vraie valeur $\theta_0$ peut être au bord de l'espace des paramètres, ces propriétés peuvent être fausses.

Attention, ce résultat n'affirme pas que l'espérance de $\widehat{\theta}$ est $\theta_0$ : l'estimateur du maximum de vraisemblance peut être biaisé ; mais quand $n$ est grand, ce biais est négligeable (il est asymptotiquement sans biais).

Notons que la qualité de l'approximation normale est également d'autant meilleure que $n$ est grand. On voit que plus la taille de l'échantillon $n$ est grande, plus faible sera la variance de $\widehat{\theta}$ ; la variance d'un estimateur qui décroît en $1/n$ est un phénomène déjà rencontré en statistiques ! De plus on peut montrer que $\widehat{\theta}$ est (asymptotiquement) l'estimateur de plus petite variance.

### 1.3.2. Intervalle de confiance de Wald

Puisqu'on dispose d'une estimation de la variance de $\widehat{\theta}$ et que sa loi est approximativement normale, on obtient un intervalle de confiance (approximatif) de niveau $1-\alpha$ de la forme

$$\widehat{\theta} \pm z_{1-\alpha/2}\sqrt{\frac{1}{nI_1\left(\widehat{\theta}\right)}} \tag{1.2}$$

où $z_{1-\alpha/2}$ est le quantile $1-\alpha/2$ de la loi normale standard (en particulier, pour un intervalle de confiance de niveau $1-\alpha = 0{,}95$, on a $z_{1-\alpha/2} = z_{0{,}975} = 1{,}96$).

Bien entendu, là encore, plus $n$ est grand plus l'intervalle de confiance est fiable.

### 1.3.3. Exemples

Nous calculons l'estimateur du maximum de vraisemblance pour nos deux exemples.

#### Exemple de la loi de Bernoulli

La vraisemblance est $L(p) = p^{n_1}(1-p)^{n_0}$ et la log-vraisemblance est $\ell(p) = n_1 \log p + n_0 \log(1-p)$. Le score est donc

$$U(p) = \frac{n_1}{p} - \frac{n_0}{1-p}$$

et l'information observée est

$$J(p) = \frac{n_1}{p^2} + \frac{n_0}{(1-p)^2}.$$

L'estimateur du maximum de vraisemblance s'obtient en annulant le score. On a $\mathrm{U}(\widehat{p}) = 0$ pour

$$\frac{n_1}{\widehat{p}} = \frac{n_0}{1-\widehat{p}}$$

d'où on tire

$$\widehat{p} = \frac{n_1}{n_0+n_1} = \frac{n_1}{n}.$$

L'information de Fisher est $n\mathrm{I}_1(p_0) = \frac{n}{p_0(1-p_0)}$ donc la variance de $\widehat{p}$ est approximativement

$$\frac{1}{n\mathrm{I}_1(\widehat{p})} = \frac{\widehat{p}\left(1-\widehat{p}\right)}{n},$$

ce qui ne devrait pas nous surprendre.

On peut également l'approcher par l'inverse de l'information observée en $\widehat{p}$ : en utilisant $n_1 = n\widehat{p}$ et $n_0 = n(1-\widehat{p})$, on a

$$\begin{aligned}\mathrm{J}\left(\widehat{p}\right) &= \frac{n_1}{\widehat{p}^2} + \frac{n_0}{\left(1-\widehat{p}\right)^2} \\ &= \frac{n}{\widehat{p}} + \frac{n}{1-\widehat{p}} \\ &= \frac{n}{\widehat{p}\left(1-\widehat{p}\right)},\end{aligned}$$

et on a l'agréable surprise de retrouver le même estimateur pour la variance de $\widehat{p}$.

Appliquons ceci à notre exemple où $n_0 = 6$ et $n_1 = 4$ : on a $\widehat{p} = 0{,}4$, avec une variance estimée par $\frac{1}{10}0{,}4 \times 0{,}6 = 0{,}024$, d'où un intervalle de confiance à 95%

$$0{,}4 \pm 1{,}96\sqrt{0{,}024} \simeq [0{,}10; 0{,}70].$$

### Exemple de la Gaussienne

On se souvient de la log-vraisemblance $\ell(\mu) = -\frac{1}{2\sigma_0^2}\left(s_2 - 2\mu s_1 + n\mu^2\right)$ et du score

$$\mathrm{U}(\mu) = \frac{1}{\sigma_0^2} n\left(\overline{x} - \mu\right).$$

L'information observée est

$$\mathrm{J}(\mu) = \frac{n}{\sigma_0^2}.$$

L'estimateur du maximum de vraisemblance est donc $\widehat{\mu} = \overline{x} = \frac{1}{n}(x_1 + \cdots + x_n)$ et on estime sa variance par $\frac{\sigma_0^2}{n}$. À nouveau, on retrouve un résultat familier.

Dans notre exemple où $\sigma_0^2 = 1$, $n = 10$ et $s_1 = 11{,}48$, on estime $\mu$ par $\widehat{\mu} = 1{,}148$ ; on a un intervalle de confiance à 95%

$$1{,}148 \pm 1{,}96\sqrt{\frac{1}{10}} \simeq [0{,}53; 1{,}77].$$

## 1.4. Les trois tests

Nous allons voir dans cette section trois façons différentes d'utiliser la vraisemblance pour tester une hypothèse nulle $H_0 : \theta = \theta_0$.

On note dans la suite $\widehat{\theta}$ l'estimateur du maximum de vraisemblance. On se place sous des hypothèses de régularité similaires à celles déjà énoncées pour la loi l'estimateur du maximum de vraisemblance ; en particulier, $\theta_0$ n'est pas au bord de l'espace des paramètres.

### 1.4.1. Test du score

Si l'hypothèse $H_0$ est vraie, la loi de $U(\theta_0)$ est d'espérance nulle et de variance $nI_1(\theta_0)$ ; elle est approximativement normale. On en déduit un test de $H_0$ en comparant

$$z = \frac{U(\theta_0)}{\sqrt{nI_1(\theta_0)}}$$

aux quantiles de la loi normale standard.

On pourra également calculer un degré de significativité $p = \mathbb{P}(|Z| > z)$ pour $Z \sim \mathcal{N}(0,1)$.

### 1.4.2. Test de Wald

Si l'hypothèse $H_0$ est vraie, $\widehat{\theta}$ suit approximativement une loi $\mathcal{N}\left(\theta_0, (nI_1(\theta_0))^{-1}\right)$, et on comparera

$$z = \frac{\widehat{\theta} - \theta_0}{\sqrt{(nI_1(\theta_0))^{-1}}} = \sqrt{nI_1(\theta_0)}\left(\widehat{\theta} - \theta_0\right)$$

aux quantiles de la loi normale standard.

### 1.4.3. Test du rapport de vraisemblance

Si l'hypothèse $H_0$ est vraie, alors $2\ell\left(\widehat{\theta}\right) - 2\ell(\theta_0) = 2\log\left(\frac{L(\widehat{\theta})}{L(\theta_0)}\right)$ suit approximativement une loi $\chi^2(1)$.

### 1.4.4. Interprétation géométrique

Les trois tests peuvent s'interpréter tous trois sur la courbe « quasi-parabolique » de la log-vraisemblance : le test du score rejette $H_0$ quand la pente en $\widehat{\theta}_0$ est trop forte ; le test de Wald quand $\widehat{\theta}$ est trop loin de $\widehat{\theta}_0$ ; et le test du rapport de vraisemblance, quand $\ell(\widehat{\theta})$ est trop loin de $\ell(\widehat{\theta}_0)$ (cf figure 1.5).

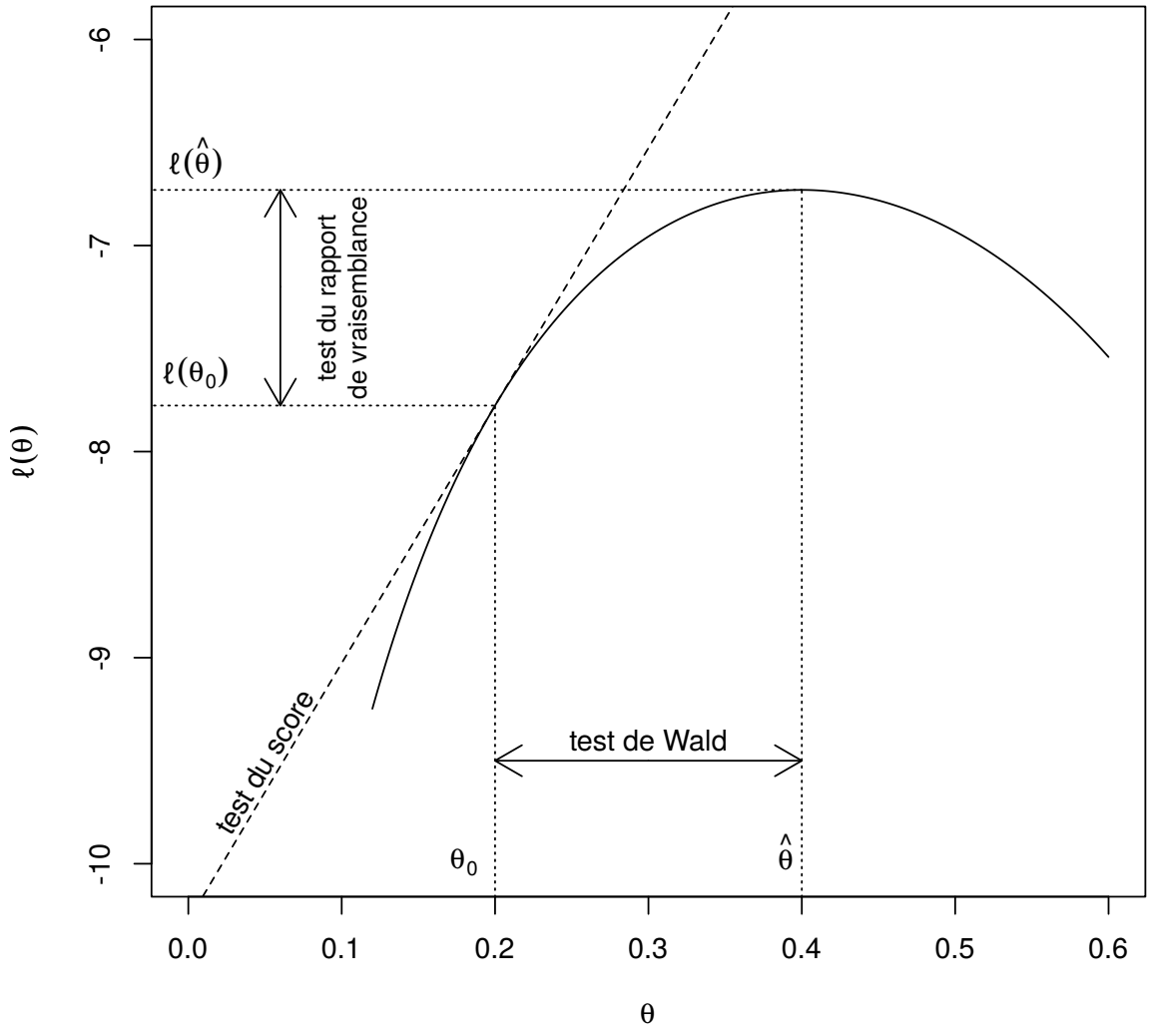



FIGURE 1.5. – les trois tests

## 1.4.5. Exemples

### Exemple de la loi de Bernoulli

Examinons comment tester $H_0 : p = p_0$. Rappelons que $\ell(p) = n_1 \log(p) + n_0 \log(1-p)$, et $U(p_0) = \frac{n_1 - n p_0}{p_0(1-p_0)} = \frac{n}{p_0(1-p_0)}\left(\widehat{p} - p_0\right)$.

La variance de $U(p_0)$ est donnée par l'information de Fisher :

$$nI_1(p_0) = \frac{n}{p_0(1-p_0)}$$

Ce résultat pouvait être retrouvé directement en utilisant nos connaissances sur la binomiale !

Le test du score est

$$\begin{aligned} z &= \frac{U(p_0)}{\sqrt{nI_1(p_0)}} \\ &= \frac{\frac{n}{p_0(1-p_0)}\left(\widehat{p} - p_0\right)}{\sqrt{\frac{n}{p_0(1-p_0)}}} \\ &= \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}. \end{aligned}$$

On retrouve le test classique basé sur l'approximation normale. Le test de Wald est basé sur $\widehat{p} - p_0$ dont la variance est bien $\left(nI_1(p_0)\right)^{-1} = \frac{1}{n} p_0(1-p_0)$. Ici le test de Wald coincide avec le test du score. Le test du rapport de vraisemblance est légèrement différent dans sa forme, cependant on pourrait vérifier qu'en pratique il donne des résultats similaires.

### Exemple de la Gaussienne

Examinons les trois tests de $H_0 : \mu = \mu_0$.

On a

$$\ell(\mu) = -\frac{1}{2\sigma_0^2}\left(s_2 - 2\mu s_1 + n\mu^2\right)$$

et

$$U(\mu_0) = \frac{1}{\sigma_0^2}(s_1 - n\mu_0) = \frac{n}{\sigma_0^2}(\overline{x} - \mu_0).$$

On peut en déduire, d'après les propriétés usuelles de la moyenne empirique, que la variance de $U(\mu_0)$ est

$$\frac{n^2}{\sigma_0^4} \times \frac{\sigma_0^2}{n} = \frac{n}{\sigma_0^2}$$

ce qui est bien la valeur de l'information de Fisher $nI_1(\mu_0)$. Le test du score est donc

$$\begin{aligned} z &= \frac{U(\mu_0)}{\sqrt{nI_1(\mu_0)}} \\ &= \frac{\frac{n}{\sigma_0^2}(\overline{x} - \mu_0)}{\sqrt{\frac{n}{\sigma_0^2}}} \\ &= \frac{\overline{x} - \mu_0}{\sqrt{\sigma_0^2/n}} \end{aligned}$$

et on a retrouvé un test classique. Le test de Wald est basé sur $\widehat{\mu} - \mu_0 = \overline{x} - \mu_0$ dont la variance est bien $\left(n\mathrm{I}_1(\mu_0)\right)^{-1}$. Il coincide avec le test du score. Le test du rapport de vraisemblance est

$$\begin{aligned}
2\ell(\widehat{\mu}) - 2\ell(\mu_0) &= 2\ell\left(\overline{x}\right) - 2\ell(\mu_0) \\
&= -\frac{1}{\sigma_0^2}\left(\left(s_2 - 2\overline{x}s_1 + n\overline{x}^2\right) - \left(s_2 - 2\mu_0 s_1 + n\mu_0^2\right)\right) \\
&= -\frac{1}{\sigma_0^2}\left(\left(s_2 - 2n\left(\overline{x}\right)^2 + n\overline{x}^2\right) - \left(s_2 - 2n\mu_0\overline{x} + n\mu_0^2\right)\right) \\
&= -\frac{1}{\sigma_0^2}\left(-n\left(\overline{x}\right)^2 + 2n\mu_0\overline{x} - n\mu_0^2\right) \\
&= \frac{n}{\sigma_0^2}\left(\overline{x} - \mu_0\right)^2 \\
&= \left(\frac{\overline{x} - \mu_0}{\sqrt{\sigma_0^2/n}}\right)^2 .
\end{aligned}$$

On a retrouvé le carré du test du score. Ici, les trois tests coincident.

# 2. Matrices et vecteurs

## 2.1. Définitions

### 2.1.1. Matrices

Une matrice est un tableau de nombres rectangulaire, écrit entre parenthèses ou entre crochets. On note $\mathbb{R}^{n\times m}$ l'ensemble des matrices à $n$ lignes et $m$ colonnes.

#### Exemples

Voici quelques matrices.

$$\mathrm{A} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} \in \mathbb{R}^{2\times 3} \qquad \mathrm{B} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \in \mathbb{R}^{3\times 3} \qquad \mathrm{C} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} \in \mathbb{R}^{4\times 1}$$

$$\mathrm{D} = \begin{pmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \\ d_{31} & d_{32} \end{pmatrix} \in \mathbb{R}^{3\times 2} \qquad \mathrm{E} = \begin{pmatrix} e_{11} & e_{12} & \dots & e_{1m} \\ e_{21} & e_{22} & \dots & e_{2m} \\ \vdots & \vdots & & \vdots \\ e_{n1} & e_{n2} & \dots & e_{nm} \end{pmatrix} \in \mathbb{R}^{n\times m} \qquad \mathrm{F} = \begin{pmatrix} 1 & 2 & 3 \end{pmatrix} \in \mathbb{R}^{1\times 3}$$

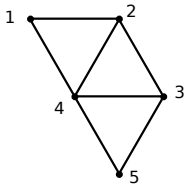
Les éléments d'une matrice $\mathrm{E} \in \mathbb{R}^{n\times m}$ sont souvent notés comme ci-dessus $e_{ij}$ pour $1 \leqslant i \leqslant n$ et $1 \leqslant j \leqslant m$. On note alors également

$$\mathrm{E} = \left(e_{ij}\right)_{\substack{1\leqslant i\leqslant n \\ 1\leqslant j\leqslant m}},$$

ou simplement $\mathrm{E} = (e_{ij})$ en précisant les dimensions par ailleurs.

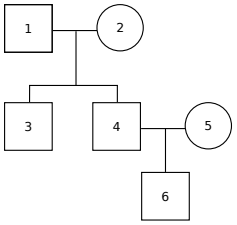
L'élément $e_{ij}$ est l'élement à l'intersection la $i^{\text{e}}$ligne et de la colonne $j$ : le premier indice donne le numéro de la ligne, le second le numéro de la colonne, de la même façon que dans l'écriture $\mathbb{R}^{n\times m}$ $n$ donne le nombre de lignes et $m$ le nombre de colonnes des matrices.

Il est fréquent d'utiliser des matrices pour condenser des informations. Ainsi par exemple, la matrice d'adjacence d'un graphe dont les sommets sont numérotés de 1 à $n$ est la matrice $\mathrm{A} = (a_{ij}) \in \mathbb{R}^{n\times n}$ avec $a_{ij} = 1$ s'il y a une arête entre les sommets $i$ et $j$, 0 sinon.




$$\mathrm{A} = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

Un autre exemple, classique en génétique, est celui de la matrice d'apparentement *(kinship matrix)* de plusieurs individus, qui est la matrice $\Phi = (\phi_{ij})$ avec $\phi_{ij}$ le coefficient d'apparentement des individus $i$ et $j$.




$$\Phi = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{8} \\ 0 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{8} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 & \frac{1}{8} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & 0 & \frac{1}{4} \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{pmatrix}$$

## 2.1.2. Transposée d'une matrice

La transposée d'une matrice $A = (a_{ij}) \in \mathbb{R}^{n\times m}$ est la matrice $A^t$ de $\mathbb{R}^{m\times n}$ de terme $a_{ji}$ ; autrement dit, « on inverse les lignes et les colonnes » ; la $j^e$ ligne de $A^t$ est la $i^e$colonne de A.

### Exemple

Si

$$A = \begin{pmatrix} 1 & 2 & 0 \\ -2 & 1 & 3 \end{pmatrix}$$

alors

$$A^t = \begin{pmatrix} 1 & -2 \\ 2 & 1 \\ 0 & 3 \end{pmatrix}.$$

On a bien sûr $\boxed{\left(A^t\right)^t = A}$.

On rencontre aussi les notations $A'$ ou $A^{\mathrm{T}}$ pour la transposée.

## 2.1.3. Matrices particulières

Les matrices de $\mathbb{R}^{n\times 1}$ sont appelés *vecteurs colonnes*, et celles de $\mathbb{R}^{1\times m}$ sont appelés *vecteurs lignes*.

Dans le texte, n notera souvent les vecteurs lignes $x = (x_1, \ldots, x_n)$. La transposée d'un vecteur ligne étant un vecteur colonne, on peut noter un vecteur colonne par $x = (x_1, \ldots, x_n)^t$ (pour des raisons de commodité typographique avant tout).

On note $\mathrm{diag}(a_1, \ldots, a_n) \in \mathbb{R}^{n\times n}$ la matrice carrée dont tous les éléments sont nuls sauf ceux de la diagonale, égaux à $a_1, a_2, \ldots, a_n$.

$$\mathrm{diag}(a_1, a_2, \ldots, a_n) = \begin{pmatrix} a_1 & 0 & \ldots & 0 \\ 0 & a_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & a_n \end{pmatrix} \in \mathbb{R}^{n\times n}.$$

La *matrice identité* $I_n$ est une matrice diagonale, avec des 1 sur la diagonale : $I_n = \mathrm{diag}(1, 1, \ldots, 1) \in \mathbb{R}^{n\times n}$.

Une matrice est dite *symétrique* si $A = A^t$. Par exemple, la matrice A ci-dessous est symétrique :

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 0 & -1 \\ 3 & -1 & 2 \end{pmatrix}.$$

## 2.2. Arithmétique des matrices

### 2.2.1. Addition

On ne peut additionner deux matrices que si elles ont la même dimension ! Si $A, B \in \mathbb{R}^{n\times m}$ sont deux matrices à $n$ lignes et $m$ colonnes, $A + B \in \mathbb{R}^{n\times m}$ est la matrice de terme $(i, j)$ donné par $a_{ij} + b_{ij}$.

#### Exemples

$$\begin{pmatrix} 1 & 2 & 0 \\ -2 & 1 & 1 \end{pmatrix} + \begin{pmatrix} 1 & 1 & 1 \\ 3 & -1 & 2 \end{pmatrix} = \begin{pmatrix} 1+1 & 2+1 & 0+1 \\ -2+3 & 1+(-1) & 1+2 \end{pmatrix} = \begin{pmatrix} 2 & 3 & 1 \\ 1 & 0 & 3 \end{pmatrix}$$

$$\begin{pmatrix} 2 & 1 \\ -2 & 1 \\ 1 & 0 \end{pmatrix} + \begin{pmatrix} 2 & 0 \\ 0 & 4 \\ 3 & -1 \end{pmatrix} = \begin{pmatrix} 4 & 1 \\ -2 & 5 \\ 4 & -1 \end{pmatrix}$$

L'addition est commutative :

$$\boxed{A + B = B + A}$$

et se comporte bien vis-à-vis de la transposition :

$$\boxed{(A + B)^t = A^t + B^t}$$

### 2.2.2. Multiplication par un nombre

On dit aussi de façon plus jolie « multiplication par un scalaire ». Si $c \in \mathbb{R}$ est un nombre (un scalaire) et $A \in \mathbb{R}^{n\times m}$ est une matrice, alors $c \cdot A \in \mathbb{R}^{n\times m}$ est la matrice de terme $(i, j)$ donné par $ca_{ij}$.

#### Exemples

$$2 \cdot \begin{pmatrix} 1 & 2 & 0 \\ -2 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 2\times 1 & 2\times 2 & 2\times 0 \\ 2\times(-2) & 2\times 1 & 2\times 1 \end{pmatrix} = \begin{pmatrix} 2 & 4 & 0 \\ -4 & 2 & 2 \end{pmatrix}$$

$$3 \cdot \begin{pmatrix} 2 & 1 \\ -2 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 6 & 3 \\ -6 & 3 \\ 3 & 0 \end{pmatrix}$$

## 2.2.3. Multiplication de matrices

### Définition

On ne peut multiplier deux matrices A et B que si le nombre de colonnes de A est égal au nombre de lignes de B : $A \in \mathbb{R}^{n\times m}$ et $B \in \mathbb{R}^{m\times p}$. Le résultat AB est alors dans $\mathbb{R}^{n\times p}$.

La matrice AB a pour terme de coordonnées $i, j$

$$\sum_{k=1}^{m} a_{ik}b_{kj}, \text{ pour } 1 \leqslant i \leqslant n,\ 1 \leqslant j \leqslant p.$$

L'exemple qui suit devrait clarifier cette définition un peu technique.

### Un exemple pas à pas

On veut effectuer la multiplication des matrices A et B suivantes :

$$A = \begin{pmatrix} 1 & 2 & 0 \\ -2 & 1 & 1 \end{pmatrix} \in \mathbb{R}^{2\times 3} \qquad\qquad B = \begin{pmatrix} 2 & 1 \\ -2 & -1 \\ 1 & 0 \end{pmatrix} \in \mathbb{R}^{3\times 2}$$

Pour aider au calcul, on dispose les matrices ainsi, la matrice $2\times 2$ « vide » étant destinée à recevoir le résultat de la multiplication :

$$\begin{array}{cc} & \begin{pmatrix} 2 & 1 \\ -2 & -1 \\ 1 & 0 \end{pmatrix} \\ \begin{pmatrix} 1 & 2 & 0 \\ -2 & 1 & 1 \end{pmatrix} & \begin{pmatrix} \cdot & \cdot \\ \cdot & \cdot \end{pmatrix} \end{array}$$

Pour calculer le premier élément du résultat, on parcourt la première ligne de la matrice A et la première colonne de la matrice B ; on effectue les opérations $1\times 2+2\times(-2)+0\times 1=-2$. On remplit la « case correspondante » :

$$\begin{array}{cc} & \begin{pmatrix} 2 & 1 \\ -2 & -1 \\ 1 & 0 \end{pmatrix} \\ \begin{pmatrix} 1 & 2 & 0 \\ -2 & 1 & 1 \end{pmatrix} & \begin{pmatrix} -2 & \cdot \\ \cdot & \cdot \end{pmatrix} \end{array}$$

On passe à la deuxième case de la première ligne : on parcourt la première ligne de la matrice A et la deuxième colonne de la matrice B, et on calcule $1\times 1+2\times(-1)+0\times 0=-1$.

$$\begin{array}{cc} & \begin{pmatrix} 2 & 1 \\ -2 & -1 \\ 1 & 0 \end{pmatrix} \\ \begin{pmatrix} 1 & 2 & 0 \\ -2 & 1 & 1 \end{pmatrix} & \begin{pmatrix} -2 & -1 \\ \cdot & \cdot \end{pmatrix} \end{array}$$

On continue par l'élement en deuxième ligne, première colonne : on parcourt la deuxième ligne de la matrice A et la première colonne de la matrice B : $-2\times 2+1\times(-2)+1\times 1=-5$.

$$\begin{array}{cc} & \begin{pmatrix} 2 & 1 \\ -2 & -1 \\ 1 & 0 \end{pmatrix} \\ \begin{pmatrix} 1 & 2 & 0 \\ -2 & 1 & 1 \end{pmatrix} & \begin{pmatrix} -2 & -1 \\ -5 & \cdot \end{pmatrix} \end{array}$$

Et enfin, pour la case restante, $-2\times 1+1\times(-1)+1\times 0=-3$.

$$\begin{array}{cc} & \begin{pmatrix} 2 & 1 \\ -2 & -1 \\ 1 & 0 \end{pmatrix} \\ \begin{pmatrix} 1 & 2 & 0 \\ -2 & 1 & 1 \end{pmatrix} & \begin{pmatrix} -2 & -1 \\ -5 & -3 \end{pmatrix} \end{array}$$

### Cas particulier : multiplication de vecteurs

Si $x=(x_1,\dots,x_n)\in\mathbb{R}^{1\times n}$ est un vecteur ligne et $y=(y_1,\dots,y_n)^t\in\mathbb{R}^{n\times 1}$ est un vecteur colonne, alors $xy\in\mathbb{R}$ est un scalaire. On a

$$xy=x_1y_1+\cdots+x_ny_n$$

On rencontre souvent le cas ou $u,v\in\mathbb{R}^{n\times 1}$ sont des vecteurs colonnes, et où on fait le produit $u^tv$. On parle du « produit scalaire » des vecteurs $u$ et $v$.

Dans les cas $n=2$ et $n=3$, il s'agit bien du produit scalaire « usuel ». De façon générale on dira que les vecteurs $u$ et $v$ sont orthogonaux si $u^tv=0$. On pourra également appeler « norme de $u$ » la racine carrée de $u^tu$.

### Associativité et distributivité

L'associativité est une propriété du produit matriciel qui peut sembler évidente (vu ce qu'on sait du produit usuel des scalaires) :

Si on a trois matrices $A\in\mathbb{R}^{n\times m}, B\in\mathbb{R}^{m\times p}, C\in\mathbb{R}^{p\times q}$, alors

$$(AB)C=A(BC)$$

On peut donc écrire sans ambiguïté ABC, sans avoir à indiquer par des parenthèses l'ordre dans lequel on doit effectuer les multiplications.

La distributivité par rapport à l'addition est également vérifié : si $A\in\mathbb{R}^{n\times m}, B\in\mathbb{R}^{m\times p}, C\in\mathbb{R}^{m\times p}$, alors

$$A(B+C)=AB+AC.$$

Enfin, si $A\in\mathbb{R}^{n\times m}$ et $B\in\mathbb{R}^{m\times p}$, on a $B^t\in\mathbb{R}^{p\times m}$ et $A^t\in\mathbb{R}^{m\times n}$, donc le produit $B^tA^t$ peut être effectué ; on a

$$(AB)^t=B^tA^t.$$

Par contre, on ne peut généralement pas simplifier dans un produit de matrices !

AB = AC n'implique pas B = C.

**Exemple**

$$\begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix}\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 4 & 6 \\ -4 & -6 \end{pmatrix}$$
$$\begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix}\begin{pmatrix} 2 & 3 \\ 2 & 3 \end{pmatrix} = \begin{pmatrix} 4 & 6 \\ -4 & -6 \end{pmatrix}$$

**Exercice**

Montrez que si $A \in \mathbb{R}^{n\times m}$, $AA^t$ est symétrique.

## 2.3. Cas des matrices carrées

Le produit de matrices carrées a ceci de particulier qu'on ne change pas de dimensions : tout se passe dans $\mathbb{R}^{n\times n}$. Si $A, B \in \mathbb{R}^{n\times n}$ sont deux matrices carrées, on peut calculer le produit AB ou bien le produit BA, et tous deux sont dans $\mathbb{R}^{n\times n}$. Cependant

| en général $AB \neq BA$. |
|---|

La matrice identité $I_n$ a la propriété remarquable que

| pour tout $A \in \mathbb{R}^{n\times n}$, $AI_n = I_nA = A$ |
|---|

Enfin, on peut donner un sens à l'écriture $A^k$ pour $k$ entier positif en posant $A^0 = I_n$, $A^1 = A$, $A^2 = A^1 \times A$, $A^3 = A^2 \times A$, etc.

### 2.3.1. Inverse d'une matrice carrée

Soient $A, B \in \mathbb{R}^{n\times n}$. Si on a $AB = I_n$ alors on a $BA = I_n$, et B est la seule matrice à avoir cette propriété. On dit alors que B est l'inverse de A et on note $B = A^{-1}$.

On a bien sûr $\left(A^{-1}\right)^{-1} = A$. On peut donner un sens à $A^{-k}$ (pour $k$ entier positif) en posant $A^{-k} = \left(A^{-1}\right)^k$. Les règles de calculs usuelles sur les puissances s'appliquent : $A^kA^\ell = A^{k+\ell}$ et $\left(A^k\right)^\ell = A^{k\ell}$.

On a également

| $\left(A^{-1}\right)^t = \left(A^t\right)^{-1}$ |
|---|

**Exemple (et exercice)**

On vérifiera que

$$\begin{pmatrix} 2 & 1 & -2 \\ 2 & 2 & 3 \\ 3 & 3 & 4 \end{pmatrix}\begin{pmatrix} 1 & 10 & -7 \\ -1 & -14 & 10 \\ 0 & 3 & -2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Mais attention : certaines matrices n'ont pas d'inverse ! L'objet de la section suivante est de donner un critère pour l'existence d'un inverse.

#### Exercice

Si $A, B \in \mathbb{R}^{n\times n}$ on toutes les deux un inverse, montrer que

$$\boxed{(AB)^{-1} = B^{-1}A^{-1}.}$$

### 2.3.2. Déterminant

Le déterminant d'une matrice $2\times 2$ est défini par

$$\det\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = a_{11}a_{22} - a_{21}a_{12}.$$

Le déterminant d'une matrice $3\times 3$ est défini par

$$\det\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = a_{11} \times \det\begin{pmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{pmatrix} - a_{21} \times \det\begin{pmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{pmatrix} + a_{31} \times \det\begin{pmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{pmatrix}.$$

Les matrices $2\times 2$ à droite du signe égal sont obtenues en supprimant la première colonne de la matrice $3\times 3$, et une de ses lignes.

Le déterminant d'une matrice $4\times 4$ est défini par

$$\begin{aligned} \det\begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix} &= a_{11} \times \det\begin{pmatrix} a_{22} & a_{23} & a_{24} \\ a_{32} & a_{33} & a_{34} \\ a_{42} & a_{43} & a_{44} \end{pmatrix} - a_{21} \times \det\begin{pmatrix} a_{12} & a_{13} & a_{14} \\ a_{32} & a_{33} & a_{34} \\ a_{42} & a_{43} & a_{44} \end{pmatrix} \\ &\quad + a_{31} \times \det\begin{pmatrix} a_{12} & a_{13} & a_{14} \\ a_{22} & a_{23} & a_{24} \\ a_{42} & a_{43} & a_{44} \end{pmatrix} - a_{41} \times \det\begin{pmatrix} a_{12} & a_{13} & a_{14} \\ a_{22} & a_{23} & a_{24} \\ a_{32} & a_{33} & a_{34} \end{pmatrix} \end{aligned}$$

...et ainsi de suite.

#### Exemples

$$\det\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = 1\times 4 - 3\times 2 = -2$$

$$\begin{aligned} \det\begin{pmatrix} 5 & 4 & 2 \\ 0 & 0 & -2 \\ 1 & 1 & 1 \end{pmatrix} &= 5 \times \det\begin{pmatrix} 0 & -2 \\ 1 & 1 \end{pmatrix} - 0 \times \det\begin{pmatrix} 4 & 2 \\ 1 & 1 \end{pmatrix} + 1 \times \det\begin{pmatrix} 4 & 2 \\ 0 & -2 \end{pmatrix} \\ &= 5\times 2 + 1 \times (-8) \\ &= 2 \end{aligned}$$

Un des intérêts du déterminant réside dans la propriété suivante :

Soit $A \in \mathbb{R}^{n\times n}$. La matrice A admet un inverse $A^{-1}$ si, et seulement si, $\det A \neq 0$.

En pratique, si un calcul numérique demande l'inversion d'une matrice, un déterminant non nul mais très petit peut poser problème... (l'ordinateur gromellera peut-être un message d'erreur à propos d'une matrice « numériquement singulière »).

On a également d'autres propriétés remarquables dont notamment

Soient $A, B \in \mathbb{R}^{n\times n}$. On a $\det(A^t) = \det(A)$, $\det(AB) = \det(A)\det(B)$.
Si $x \in \mathbb{R}$, $\det(xA) = x^n \det(A)$.

### 2.3.3. Trace d'une matrice

La notion suivante est utile également.

La trace d'une matrice carrée est la somme des éléments sur la diagonale.

#### Exemple

$$\mathrm{Tr}\begin{pmatrix} 5 & 4 & 2 \\ 0 & 0 & -2 \\ 1 & 1 & 1 \end{pmatrix} = 5 + 0 + 1 = 6$$

On a les propriétés suivantes :

Si $A, B \in \mathbb{R}^{n\times n}$, $\mathrm{Tr}(A+B) = \mathrm{Tr}(A) + \mathrm{Tr}(B)$.
Si $x \in \mathbb{R}$, $\mathrm{Tr}(xA) = x\,\mathrm{Tr}(A)$.
Si $A \in \mathbb{R}^{n\times n}$, $\mathrm{Tr}(A^t) = \mathrm{Tr}(A)$.
Si $A \in \mathbb{R}^{n\times m}$ et $B \in \mathbb{R}^{m\times n}$, alors $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$.

#### Exemple

Avec

$$A = \begin{pmatrix} 5 & 4 & 2 \\ 0 & 0 & -2 \end{pmatrix}$$
$$B = \begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 2 & 0 \end{pmatrix}$$

on calcule

$$\mathrm{AB} = \begin{pmatrix} 9 & 9 \\ -4 & 0 \end{pmatrix}$$
$$\mathrm{A}^t\mathrm{B} = \begin{pmatrix} 5 & 4 & 0 \\ 0 & 0 & -2 \\ 10 & 8 & 4 \end{pmatrix}$$

qui ont toutes deux leur trace égale à 9. Notons que $\mathrm{B}^t\mathrm{A} = \left(\mathrm{A}^t\mathrm{B}\right)^t$ et $\mathrm{BA}^t = \left(\mathrm{AB}^t\right)^t$, et leur trace est bien égale à 9 également.

### 2.3.4. Valeurs et vecteurs propres

Soit $\mathrm{A} \in \mathbb{R}^{n\times n}$ et $u \in \mathbb{R}^{n\times 1}$ un vecteur non nul. Si le vecteur $\mathrm{A}u$ est proportionnel à $u$, $\mathrm{A}u = \lambda u$, pour $\lambda \in \mathbb{R}$, on dit que $\lambda$ est une valeur propre de A, et que $u$ est un vecteur propre.

Nous laissons de côté les méthodes utilisées pour trouver valeurs et vecteurs propres. Nous verrons un peu plus loin quelques cas particuliers.

## 2.4. Formes quadratiques

### 2.4.1. Définitions

Une forme quadratique est une fonction du type $f(x) = x^t\mathrm{A}x$ avec A une matrice symétrique $\mathrm{A} \in \mathbb{R}^{n\times n}$ et $x \in \mathbb{R}^{n\times 1}$.

#### Exemples

L'exemple le plus simple est $\mathrm{A} = \mathrm{I}_n$, par exemple avec $n = 2$ :

$$\begin{aligned} x'\mathrm{I}_2 x &= \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \\ &= \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \\ &= x_1^2 + x_2^2 + x_3^2 \\ &= ||x||^2 \end{aligned}$$

Voici un exemple un peu moins particulier :

$$\begin{aligned} x'\mathrm{A}x &= \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 2 \\ 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \\ &= \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \begin{pmatrix} x_1 + x_3 \\ x_2 + 2x_3 \\ x_1 + 2x_2 + 3x_3 \end{pmatrix} \\ &= x_1(x_1 + x_3) + x_2(x_2 + 2x_3) + x_3(x_1 + 2x_2 + 3x_3) \\ &= x_1^2 + x_2^2 + 3x_3^2 + 2x_1x_3 + 4x_2x_3 \end{aligned}$$

On voit que les termes sur la diagonale de A donnent les coefficients des $x_i^2$ alors que les termes hors-diagonale donnent, après multiplication par deux, ceux des $x_i x_j$.

De manière générale, si $u \in \mathbb{R}^{n\times 1}$ est un vecteur, on a $(u^t x)^2 = (u^t x)^t \times (u^t x) = x^t(uu^t)x$, et donc $(u^t x)^2$ est une forme quadratique. Ceci fournit d'autres exemples :

$$(x_1 + x_2 + x_3)^2 = \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

$$(x_1 - x_2)^2 = \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

Si pour tout $x$ non nul, on a $x^t\mathrm{A}x > 0$, on dira que A est définie positive. Si on a simplement $x^t\mathrm{A}x \geqslant 0$ on dira que A est positive (ou parfois, semi-définie positive).
Si $(-\mathrm{A})$ est (définie) positive, on dira que A est (définie) négative.

#### Exemples

La matrice $\mathrm{I}_n$ est définie positive. La matrice

$$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

n'est ni positive, ni négative (pourquoi ?).

Si $\mathrm{B} \in \mathbb{R}^{n\times m}$, alors $\mathrm{A} = \mathrm{B}\mathrm{B}^t$ est une matrice positive (pourquoi ?).

### 2.4.2. Valeurs et vecteurs propres des matrices symétriques

Si $\mathrm{A} \in \mathbb{R}^{n\times n}$ est symétrique, alors il y a $n$ vecteurs propres deux à deux orthogonaux $u_1, \ldots, u_n$ (c'est-à-dire que $u_i^t u_j = 0$ si $i \neq j$), associés à des valeurs propres $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$.

On peut imposer que la norme des $u_i$ soit égale à 1, c'est-à-dire que $u_i^t u_i = 1$ pour tout $i$. Dans ce cas, si on note U la matrice $n \times n$ dont les colonnes sont $u_1, \ldots, u_n$ et $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, on a

$$\mathrm{A} = \mathrm{U}\Lambda\mathrm{U}^t.$$

Connaître les valeurs propres de A permet de vérifier si A est (définie) positive.

Si $\mathrm{A} \in \mathbb{R}^{n\times n}$ est symétrique, et si $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ sont ses valeurs propres, alors A est définie positive si, et seulement si, $\lambda_i > 0$ pour tout $i$ ; A est positive si, et seulement si, $\lambda_i \geqslant 0$ pour tout $i$.

## 2.5. Manipulation de matrices en R

### 2.5.1. Création de matrices, opérations simples

On peut créer une matrice comme ceci :

```
> A <- matrix( c(1,2,3,4,5,6), nrow=3)
> A
     [,1] [,2]
[1,]    1    4
[2,]    2    5
[3,]    3    6
```

Pour que les éléments donnés servent à remplir la matrice ligne par ligne et non colonne par colonne, on utilise

```
> matrix( c(1,2,3,4,5,6), nrow= 3, byrow=TRUE)
     [,1] [,2]
[1,]    1    2
[2,]    3    4
[3,]    5    6
```

Il y a une fonction `diag()` pour créer des matrices diagonales :

```
> diag( c(4,6,8) )
     [,1] [,2] [,3]
[1,]    4    0    0
[2,]    0    6    0
[3,]    0    0    8
```

On peut extraire des vecteurs lignes et colonnes ainsi :

```
> A[1,]
[1] 1 4
> A[,2]
[1] 4 5 6
```

Pour que le résultat reste sous forme matricielle, on peut ajouter `drop=FALSE` :

```
> A[1,,drop=FALSE]
     [,1] [,2]
[1,]    1    4
> A[,2,drop=FALSE]
     [,1]
[1,]    4
[2,]    5
[3,]    6
```

La transposée s'obtient avec la fonction `t()` :

```
> A
     [,1] [,2]
[1,]    1    4
[2,]    2    5
[3,]    3    6
> t(A)
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
```

La somme de matrices est facile à obtenir :

```
> B <- matrix( c(-1,0, 0,2, -1,-3), nrow=3)
> B
     [,1] [,2]
[1,]   -1    2
[2,]    0   -1
[3,]    0   -3
> A+B
     [,1] [,2]
[1,]    0    6
[2,]    2    4
[3,]    3    3
```

### 2.5.2. Produits

Il y a une subtilité pour le produit matriciel, il s'obtient avec l'opérateur spécial `%*%` :

```
> A
     [,1] [,2]
[1,]    1    4
[2,]    2    5
[3,]    3    6
> t(B)
     [,1] [,2] [,3]
[1,]   -1    0    0
[2,]    2   -1   -3
> A %*% t(B)
     [,1] [,2] [,3]
[1,]    7   -4  -12
[2,]    8   -5  -15
[3,]    9   -6  -18
```

Les vecteurs « simples », créés avec `c()`, seront considérés comme des vecteurs lignes ou colonnes selon les circonstances :

```
> A <- matrix( c(1,2,3, 2,0,1, 3,1,4), nrow=3)
> A
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    2    0    1
[3,]    3    1    4
> u <- c(0,1,2)
> A %*% u
     [,1]
[1,]    8
[2,]    2
[3,]    9
> u %*% A
     [,1] [,2] [,3]
[1,]    8    2    9
```

Du coup la forme quadratique $u^t \mathrm{A} u$ se calcule simplement comme ceci :

```
> u %*% A %*% u
     [,1]
[1,]   20
```

### 2.5.3. Inverse, déterminant, trace

L'inverse d'une matrice s'obtient avec la fonction `solve()` :

```
> C <- matrix( c(2,2,3, 1,2,3, -2,3,4), nrow=3)
> C
     [,1] [,2] [,3]
[1,]    2    1   -2
[2,]    2    2    3
[3,]    3    3    4
> solve(C)
     [,1] [,2] [,3]
[1,]    1   10   -7
[2,]   -1  -14   10
[3,]    0    3   -2
> C %*% solve(C)
     [,1]          [,2]          [,3]
[1,]    1 -8.881784e-16  0.000000e+00
[2,]    0  1.000000e+00 -8.881784e-16
[3,]    0  1.776357e-15  1.000000e+00
```

Autre exemple :

```
> D <- matrix( c(5,0,1, 4,0,1, 2,-2,1), nrow=3)
> D
     [,1] [,2] [,3]
[1,]    5    4    2
[2,]    0    0   -2
[3,]    1    1    1
> solve(D)
     [,1] [,2] [,3]
[1,]    1 -1.0   -4
[2,]   -1  1.5    5
[3,]    0 -0.5    0
```

et le déterminant avec `det()` :

```
> det(C)
[1] -1
> det(D)
[1] 2
```

La trace peut s'obtenir ainsi :

```
> sum(diag(C))
[1] 8
```

### 2.5.4. Valeurs propres

Les vecteurs propres et les valeurs propres d'une matrice s'obtiennent avec `eigen` (allemand pour « propre » ; en anglais, les vecteurs et valeurs propres sont appelés *eigenvectors* et *eigenvalues*).

```
> A <- matrix( c(1,6,-1, 2,-1,-2, 1,0,-1), nrow=3)
> x <- eigen(A)
> x
```

```
eigen() decomposition
$values
[1] -4.000000e+00  3.000000e+00  9.616736e-17

$vectors
           [,1]       [,2]       [,3]
[1,]  0.4082483 -0.4850713 -0.0696733
[2,] -0.8164966 -0.7276069 -0.4180398
[3,] -0.4082483  0.4850713  0.9057529
```

Les valeurs propres sont dans `x$values`, et les vecteurs propres dans les colonnes de `x$vectors`.

```
> u <- x$vectors[,1]
> u
[1]  0.4082483 -0.8164966 -0.4082483
> A %*% u
          [,1]
[1,] -1.632993
[2,]  3.265986
[3,]  1.632993
> x$values[1] * u
[1] -1.632993  3.265986  1.632993
```

Considérons le cas d'une matrice symétrique :

```
> A <- matrix(c(5, 2, 4, 2, 3, 2, 4, 2, 1), nrow=3)
> A
     [,1] [,2] [,3]
[1,]    5    2    4
[2,]    2    3    2
[3,]    4    2    1
> eigenA <- eigen(A)
> eigenA
eigen() decomposition
$values
[1]  8.790213  1.794590 -1.584803

$vectors
           [,1]        [,2]       [,3]
[1,] -0.7525758  0.45794385 -0.4731987
[2,] -0.4317041 -0.88573564 -0.1705987
[3,] -0.4972536  0.07589338  0.8642795
```

Cette matrice n'est ni positive, ni négative. On peut vérifier qu'on a $A = U\Lambda U^t$ :

```
> U <- eigenA$vectors
> Lambda <- diag(eigenA$values)
> U %*% Lambda %*% t(U)
     [,1] [,2] [,3]
[1,]    5    2    4
[2,]    2    3    2
[3,]    4    2    1
```

# 3. Fonctions à plusieurs variables

## 3.1. Fonctions d'un vecteur

On peut voir une fonction à plusieurs variables comme une fonction d'un vecteur, ainsi

$$f(a,b) = 1 + a^2 + b^4 + a(b-1)$$

peut être vue comme une fonction du vecteur $x = (a,b)^t$. On peut la noter également

$$f(x_1, x_2) = 1 + x_1^2 + x_2^4 + x_1(x_2 - 1)$$

pour renforcer ce point de vue.

## 3.2. Représentation des fonctions à deux variables

On ne peut plus faire des graphes comme pour les fonctions à une variable... pour les fonctions à deux variables $f(x,y)$, on pourrait faire une représentation en trois dimensions, en mettant le plan $(x,y)$ horizontal (le « plancher ») et en dessinant la surface $z = f(x,y)$.

Pour représenter cette surface sur la feuille de papier, on a 2 possibilités : d'une part, les courbes de niveau (avec ou sans « heat map »), d'autre part le dessin en perspective – qui est souvent peu lisible ; voir figure 3.1.

## 3.3. Dérivées partielles ; gradient, hessienne

La dérivée partielle de $f(a,b) = 1 + a^2 + b^4 + a(b-1)$ par rapport à $a$ s'obtient en traitant $b$ comme une constante, et en dérivant la variable $a$ :

$$\frac{\partial}{\partial a} f(a,b) = 2a + b - 1$$

et de la même façon

$$\frac{\partial}{\partial b} f(a,b) = 4b^3 + a$$

Le gradient de $f(a,b)$ en $(a,b)$ est le vecteur $\underset{(a,b)}{\operatorname{grad}} f(a,b) = \left(\frac{\partial}{\partial a} f(a,b), \frac{\partial}{\partial b} f(a,b)\right)^t$. De façon générale, si $x$ est un vecteur de dimension $n$, le gradient de $f(x)$ en $x$ est

$$\underset{x}{\operatorname{grad}} f(x) = \begin{pmatrix} \frac{\partial}{\partial x_1} f(x) \\ \vdots \\ \frac{\partial}{\partial x_n} f(x) \end{pmatrix}.$$

Dans le cas de notre exemple on a

$$\underset{(a,b)}{\operatorname{grad}} f(a,b) = \begin{pmatrix} 2a + b - 1 \\ 4b^3 + a \end{pmatrix}.$$

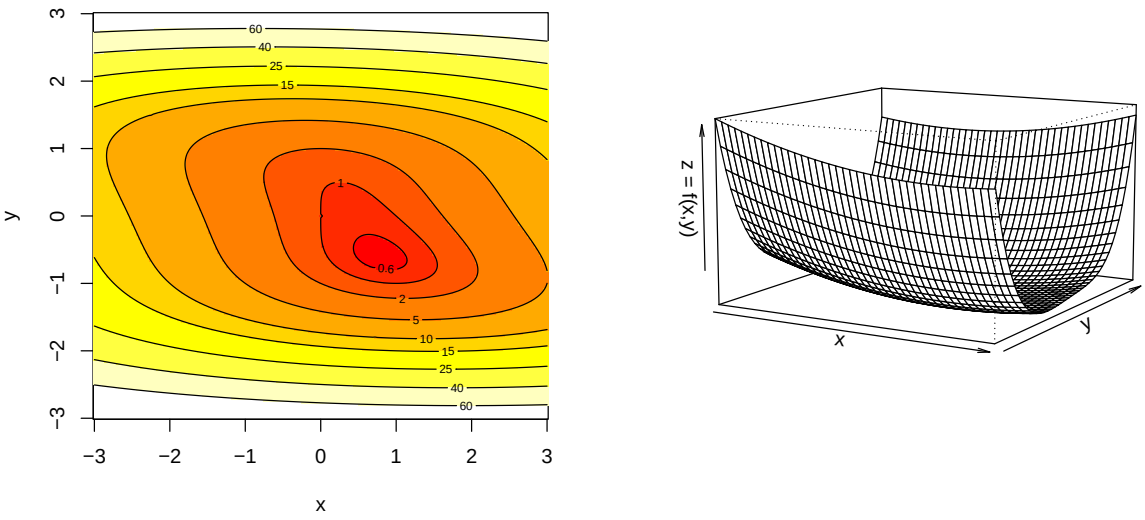



FIGURE 3.1. – Deux représentations de $f(x,y) = 1 + x^2 + y^4 + x(y-1)$

On peut évaluer ce gradient, par exemple en $a = 1$ et $b = 2$ :

$$\operatorname*{grad}_{(a,b)} f(1,2) = \begin{pmatrix} 3 \\ 33 \end{pmatrix}.$$

Passons maintenant à la dérivée seconde de $f$ :

La hessienne de $f(a,b)$ en $(a,b)$ est la matrice (symétrique) des dérivées secondes de $f$ :

$$\mathrm{H}_{(a,b)} f(a,b) = \begin{pmatrix} \frac{\partial^2}{\partial a^2} f(a,b) & \frac{\partial^2}{\partial a \partial b} f(a,b) \\ \frac{\partial^2}{\partial b \partial a} f(a,b) & \frac{\partial^2}{\partial b^2} f(a,b) \end{pmatrix}$$

De façon générale, si $x$ est un vecteur de dimension $n$, la hessienne de $f(x)$ en $x$ est la matrice $n \times n$ symétrique

$$\mathrm{H}_x f(x) = \begin{pmatrix} \frac{\partial^2}{\partial x_1^2} f(x) & \frac{\partial^2}{\partial x_1 \partial x_2} f(x) & \dots & \frac{\partial^2}{\partial x_1 \partial x_n} f(x) \\ \frac{\partial^2}{\partial x_2 \partial x_1} f(x) & \frac{\partial^2}{\partial x_2^2} f(x) & \dots & \frac{\partial^2}{\partial x_2 \partial x_n} f(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_n \partial x_1} f(x) & \frac{\partial^2}{\partial x_n \partial x_2} f(x) & \dots & \frac{\partial^2}{\partial x_n^2} f(x) \end{pmatrix}$$

Retournons à notre exemple :

$$\frac{\partial^2}{\partial a^2} f(a,b) = 2$$
$$\frac{\partial^2}{\partial a \partial b} f(a,b) = 1$$
$$\frac{\partial^2}{\partial b^2} f(a,b) = 12b^2$$

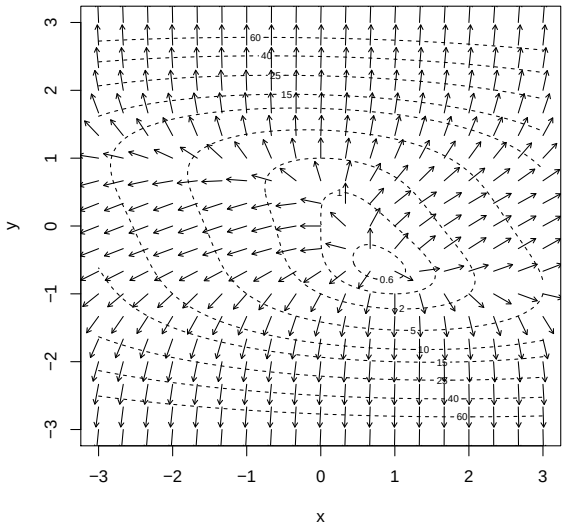



FIGURE 3.2. – Représentation de la direction pointée par le gradient.
Elle est perpendiculaire aux courbes de niveau.

D'où

$$\mathrm{H}_{(a,b)} f(a,b) = \begin{pmatrix} 2 & 1 \\ 1 & 12b^2 \end{pmatrix}.$$

## 3.4. Interprétation géométrique du gradient

Prenons le cas de la fonction à deux variables, et imaginons un petit bonhomme posé sur la surface $z = f(x,y)$, au point de coordonnées au « plancher » $(x_0, y_0)$ et d'« altitude » $f(x_0,y_0)$. Le gradient $\underset{(x,y)}{\mathrm{grad}}\, f(x_0,y_0)$ est un vecteur de dimension 2 qui peut être interprété comme pointant dans une direction : c'est la direction de la plus grande pente ascendante – elle est donc perpendiculaire aux courbes de niveau ! D'autre part, plus la norme de ce vecteur est grande, plus la pente est forte. Si le bonhomme est au fond d'un creux ou au sommet d'une bosse, le gradient sera nul.

La figure 3.2 représente la direction du gradient pour notre exemple $f(x,y) = 1 + x^2 + y^4 + x(y-1)$.

## 3.5. Minimum et maximum

En un point $x_0$ où $f$ atteint un maximum ou un minimum (local), on a $\underset{x}{\mathrm{grad}}\, f(x_0) = 0$.

On peut donc chercher les minima et maxima parmi les points (dits points critiques de $f$) qui annulent le gradient (de même qu'en dimension 1 on cherche les points qui annulent la dérivée).

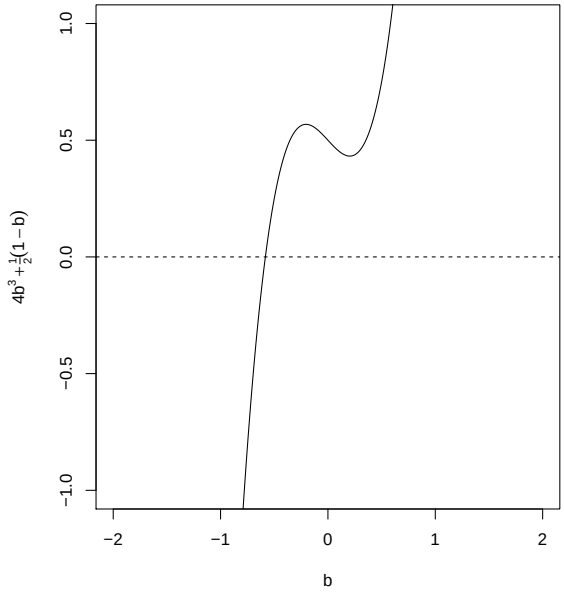



FIGURE 3.3. – Graphe de $4b^3 + \frac{1}{2}(1-b)$

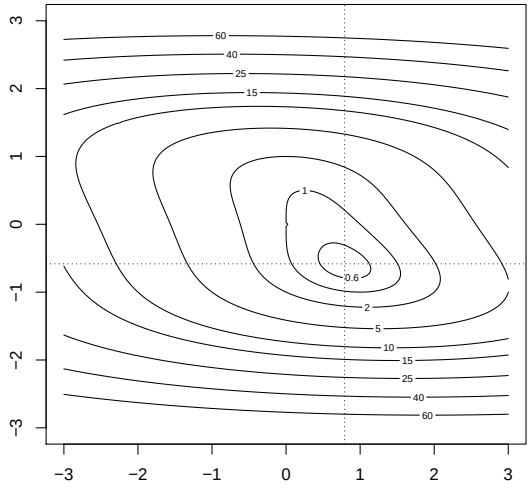



FIGURE 3.4. – Position du minimum de $f(x,y)$

Retournons à notre exemple : $\underset{(a,b)}{\text{grad}}\, f(a,b) = 0$ implique

$$\begin{cases} 2a+b-1=0 \\ 4b^3+a=0 \end{cases}$$

d'où

$$\begin{cases} a = \dfrac{1}{2}(1-b) \\ 4b^3+\dfrac{1}{2}(1-b)=0 \end{cases}$$

Nous traçons le graphe de $4b^3 + \frac{1}{2}(1-b)$ (cf figure 3.3) afin de chercher où cette quantité s'annule. Il n'y a qu'un seul point $b$ où $4b^3+\frac{1}{2}(1-b)=0$, c'est $b \simeq -0{,}583$ ; on a ensuite $a = (1-b)/2 = 0{,}791$, d'où un seul point critique : $(0{,}791,\ -0{,}583)$. Ici c'est un minimum de $f$.

**Remarque :** On peut déterminer si un point critique $x_0$ est un maximum ou un minimum (local) de $f$ en regardant la hessienne $\mathrm{H}_x f(x_0)$ : en particulier, si elle est définie positive, c'est un minimum, et si son opposée est définie positive c'est un maximum.

## 3.6. Fonctions quadratiques

Les fonctions de la forme suivante sont particulièrement intéressantes :

$$\varphi(x) = c + v^t x + x^t \mathrm{A} x \tag{3.1}$$

avec

- $x \in \mathbb{R}^{n\times 1}$ un vecteur colonne,
- $c \in \mathbb{R}$, un nombre,
- $v \in \mathbb{R}^{n\times 1}$ un vecteur colonne,
- $\mathrm{A} \in \mathbb{R}^{n\times n}$ une matrice symétrique.

Le terme $c$ s'appelle « terme constant », $(v^t x)$ « terme linéaire », et $(x^t \mathrm{A} x)$ est un « terme quadratique ».

**Exemples** Commençons par un terme linéaire, avec $v = (3,4,-2)^t$ :

$$u^t x = \begin{pmatrix} 3 & 4 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = 3x_1 + 4x_2 - 2x_3$$

Examinons maintenant un terme quadratique :

$$\begin{aligned} x^t \mathrm{A} x &= \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \begin{pmatrix} 1 & 2 & 0 \\ 2 & 3 & -5 \\ 0 & -5 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \\ &= \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \begin{pmatrix} x_1 + 2x_2 \\ 2x_1 + 3x_2 - 5x_3 \\ -5x_2 + 4x_3 \end{pmatrix} \\ &= (x_1^2 + 2x_1 x_2) + (2x_1 x_2 + 3x_2^2 - 5x_2 x_3) + (-5x_2 x_3 + 4x_3^2) \\ &= x_1^2 + 3x_2^2 + 4x_3^2 + 4x_1 x_2 - 10x_2 x_3 \end{aligned}$$

On voit que les termes sur la diagonale de C donnent les coefficients des $x_i^2$ alors que les termes hors-diagonale donnent, après multiplication par deux, ceux des $x_i x_j$.

## 3.6.1. Gradient et hessienne

Le gradient et la hessienne de $\varphi(x) = c + v^t x + x^t \mathrm{A} x$ sont

$$\begin{aligned} \operatorname*{grad}_x \varphi(x) &= v + 2\mathrm{A}x \\ \mathrm{H}_x \varphi(x) &= 2\mathrm{A}. \end{aligned}$$

Dans le cas où A est inversible, on en déduit que $\varphi$ n'a qu'un point critique, $x = -\frac{1}{2}\mathrm{A}^{-1} v$.

## 3.6.2. Lignes de niveau d'une fonction quadratique

On se place ici en dimension $n = 2$.

### Matrices définies positives ou négatives

Si la matrice A est définie positive (pour tout $x$ non nul, $x^t \mathrm{A} x > 0$), les lignes de niveau de $\varphi$ sont des ellipses. C'est également le cas si (A) est définie négative.

Le centre des ellipses est l'unique point critique de $\varphi$. La direction des axes est donnée par les vecteurs propres $u_1$ et $u_2$ de A. Les valeurs propres $\lambda_1, \lambda_2$ déterminent les longueurs des « demi-grand axe » et « demi petit-axe », qui sont proportionnelles à $1/\sqrt{\lambda_i}$. La figure 3.6 montre des lignes de niveau pour

$$\begin{aligned} f(x) &= 2 + \begin{pmatrix} 3 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} 3 & -1 \\ -1 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ &= 2 + 3x_1 - x_2 + 3x_1^2 - 2x_1 x_2 + 3x_2^2. \end{aligned}$$

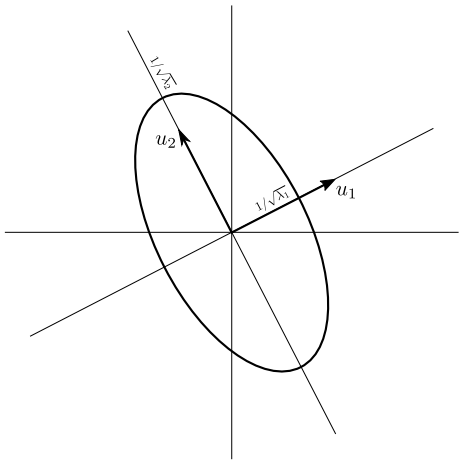



FIGURE 3.5. – Lien entre les axes d'une ellipses et valeurs et vecteurs propres

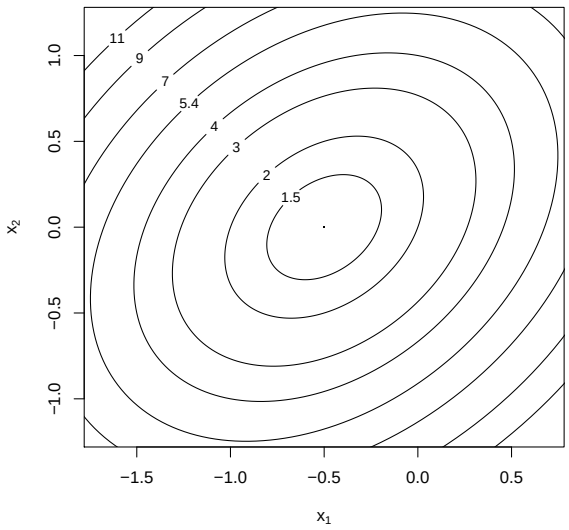



FIGURE 3.6. – Lignes de niveau elliptiques

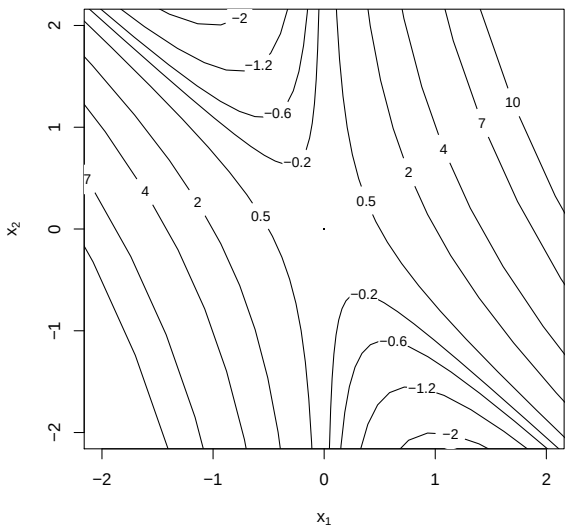



FIGURE 3.7. – Lignes de niveau hyperboliques

#### Le cas hyperbolique

Dans le cas où les deux valeurs propres de A n'ont pas le même signe, les lignes de niveau sont des hyperboles ; on a un « point selle » ou un « col ». La figure 3.7 illustre ce cas pour

$$f(x) = \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

### 3.6.3. Approximation d'une fonction quelconque par une fonction quadratique

Si $f(x)$ est une fonction de $x$ (un vecteur de dimension $n$), alors pour $x$ proche d'un point $x_0$ on a les approximations suivantes, de qualité croissante :

$$\begin{aligned} &f(x) \simeq f(x_0) && \text{(ordre 0)} \\ &f(x) \simeq f(x_0) + (\operatorname{grad}_x f(x_0))^t (x - x_0) && \text{(ordre 1)} \\ &f(x) \simeq f(x_0) + (\operatorname{grad}_x f(x_0))^t (x - x_0) + \frac{1}{2}(x - x_0)^t \left(\mathrm{H}_x f(x_0)\right)(x - x_0) && \text{(ordre 2)} \end{aligned}$$

Considérons le cas de notre exemple $f(x_1, x_2) = 1 + x_1^2 + x_2^4 + x_1(x_2 - 1)$. Pour $x$ proche du point $x_0 = (0,0)$, on obtient l'approximation suivante, illustrée à la figure 3.8 :

$$\begin{aligned} f(x) &\simeq 1 + \begin{pmatrix} -1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \tfrac{1}{2} \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ &= 1 - x_1 + \frac{1}{2}(2x_1^2 + 2x_1 x_2) \end{aligned}$$

Note : avec cet exemple, si on s'éloigne un peu de (0,0) la qualité de l'approximation se dégrade franchement : la forme quadratique n'est pas définie positive (c'est celle qu'on a utilisée pour la figure 3.7) alors que $f$ admet un minimum (cf figure 3.4).

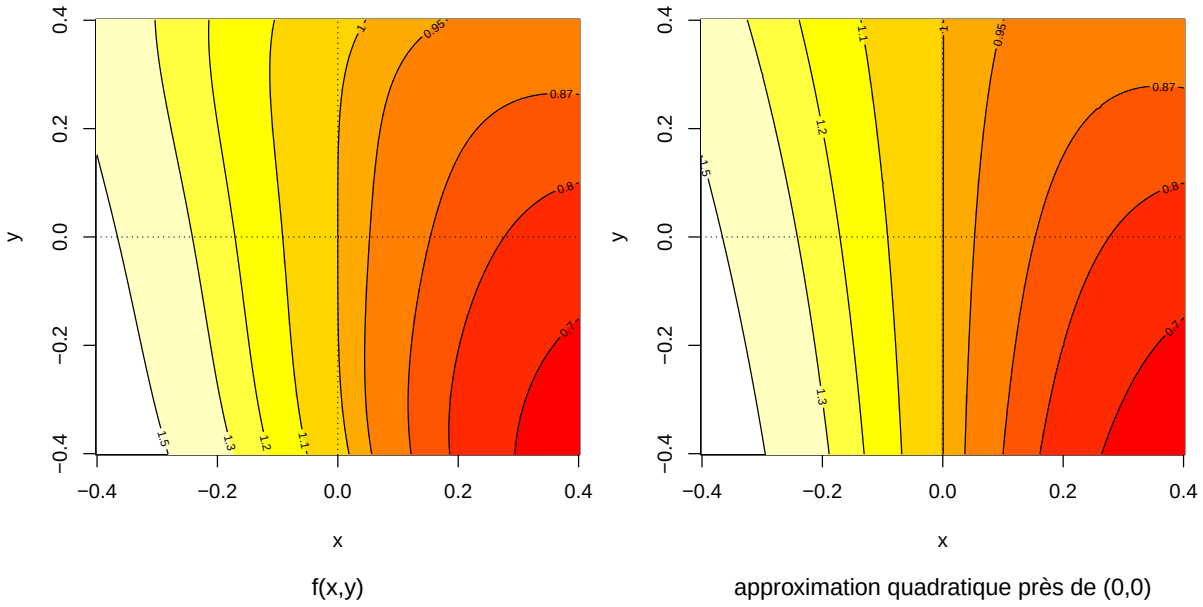



FIGURE 3.8. – Courbes de niveau de $f(x,y)$ et de son approximation quadratique près de $(0,0)$

# 4. Vecteurs aléatoires

## 4.1. Définitions

Un vecteur aléatoire de dimension $n$ est un vecteur $X = (X_1, \ldots, X_n)^t$ où les $X_i$ sont des variables aléatoires. Son espérance est le vecteur $E(X) = (E(X_1), \ldots, E(X_n))^t$.

Sa matrice de variance est la matrice symétrique $n \times n$

$$\mathrm{var}(X) = \begin{pmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1,X_2) & \cdots & \mathrm{cov}(X_1,X_n) \\ \mathrm{cov}(X_2,X_1) & \mathrm{var}(X_2) & \cdots & \mathrm{cov}(X_2,X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(X_n,X_1) & \mathrm{cov}(X_n,X_2) & \cdots & \mathrm{var}(X_n) \end{pmatrix}$$

Si $X = (X_1, \ldots, X_n)^t$ et $Y = (Y_1, \ldots, Y_m)^t$ sont des vecteurs aléatoires, leur matrice de covariance est la matrice $n \times m$

$$\mathrm{cov}(X,Y) = \begin{pmatrix} \mathrm{cov}(X_1,Y_1) & \mathrm{cov}(X_1,Y_2) & \cdots & \mathrm{cov}(X_1,Y_m) \\ \mathrm{cov}(X_2,Y_1) & \mathrm{cov}(X_2,Y_2) & \cdots & \mathrm{cov}(X_2,Y_m) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(X_n,Y_1) & \mathrm{cov}(X_n,X_2) & \cdots & \mathrm{cov}(X_n,Y_m) \end{pmatrix}$$

Attention, on n'a pas cov(X,Y) = cov(Y,X), mais bien $\mathrm{cov}(X,Y) = \mathrm{cov}(Y,X)^t$. Modulo quelques aménagements, les propriétés usuelles de l'espérance et de la variance restent vraies avec ces définitions :

Si X et Y sont des vecteurs de même dimension,
- $E(X+Y) = E(X) + E(Y)$ ;
- $\mathrm{var}(X+Y) = \mathrm{var}(X) + \mathrm{cov}(X,Y) + \mathrm{cov}(Y,X) + \mathrm{var}(Y)$ ;
- si X et Y sont indépendants, $\mathrm{cov}(X,Y) = \mathrm{cov}(Y,X) = 0$.

### 4.1.1. La loi multinomiale

Rappelons qu'une expérience de Bernoulli est une expérience aléatoire à deux issues possibles, appelées usuellement « succès » et « échec ». Quand on répète $n$ expériences de Bernoulli indépendantes ayant une probabilité de succès $p$, le nombre total de succès suit une loi binomiale $\mathcal{B}in(n,p)$.

La loi multinomiale est une généralisation au cas où l'expérience a plus de deux issues possibles, disons $d$ issues numérotées de 1 à $d$ ; ces issues se réalisent avec probabilités respectives $p_1, \ldots, p_d$ (on a donc $p_1 + \cdots + p_d = 1$).

On réalise $n$ telles expériences et on compte le nombre d'issues observées dans chaque catégorie : soient $X_1, \ldots, X_d$ le nombre d'observations dans les catégories $1, \ldots, d$. La loi du vecteur $(X_1, \ldots, X_d)^t$ est une loi multinomiale de paramètre $n$ et $p = (p_1, \ldots, p_d)^t$.

$X \sim \mathcal{M}ult(n,p)$

Chaque $X_i$ suit une loi $\mathcal{B}in(n, p_i)$, mais les $X_i$ ne sont pas indépendantes (leur somme étant contrainte à être égale à $n$). L'espérance et la variance de X sont (preuve laissée en exercice) :

$$\boxed{\begin{aligned} &E(X) = np = (np_1, \dots, np_d) \\ &\mathrm{var}(X) = n \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_d \\ -p_2p_1 & p_2(1-p_2) & \cdots & -p_2p_d \\ \vdots & \vdots & \ddots & \vdots \\ -p_dp_1 & -p_dp_2 & \cdots & p_d(1-p_d) \end{pmatrix} \end{aligned}}$$

Autrement dit, $\mathrm{var}(X_i) = np_i(1-p_i)$ et $\mathrm{cov}(X_i, X_j) = -np_ip_j$ quand $i \neq j$.

Pour finir la loi de $X_i$ est donnée par

$$\mathbb{P}(X = x) = \mathbb{P}(X_1 = x_1, \dots, X_d = x_d) = \binom{n}{x_1\ x_2\ \dots\ x_d} p_1^{x_1} \cdots p_d^{x_d},$$

où le coefficient multinomial est

$$\binom{n}{x_1\ x_2\ \dots\ x_d} = \frac{n!}{x_1!x_2!\cdots x_d!}$$

pour $x_1 + \cdots + x_d = n$.

#### Exemple (et exercice)

On considère un locus di-allélique A/a, et on suppose que les génotypes sont dans les proportions de Hardy-Weinberg. On observe les génotypes de $n$ individus : il y en $X_{AA}$ (resp. $X_{Aa}$, $X_{aa}$) qui sont AA (resp. Aa, aa).

Le vecteur $X \sim (X_{AA}, X_{Aa}, X_{aa})$ suit une loi multinomiale $\mathcal{M}ult\left(n, \left(p^2, 2pq, q^2\right)\right)$.

On pose $Y = aX_{AA} + bX_{Aa} + cX_{aa}$. Calculez $E(Y)$ et $\mathrm{var}(Y)$.

### 4.1.2. Multiplication d'un vecteur aléatoire par une matrice

Si $X \in \mathbb{R}^{n\times 1}$ est un vecteur colonne aléatoire et $A \in \mathbb{R}^{m\times n}$ est une matrice, alors AX est un vecteur aléatoire (de dimension $m$). On a

$$\boxed{\begin{aligned} E(AX) &= A\,E(X) \\ \mathrm{var}(AX) &= A\,\mathrm{var}(X)\,A^t \end{aligned}}$$

Si $B \in \mathbb{R}^{d\times n}$, alors BX est un vecteur aléatoire de dimension $d$. On a

$$\boxed{\mathrm{cov}(AX, BX) = A\,\mathrm{var}(X)\,B^t.}$$

En particulier, si $u \in \mathbb{R}^{n\times 1}$ est un vecteur colonne, alors $u^tX = u_1X_1 + \cdots u_nX_n$ est une variable aléatoire et $u^t\,\mathrm{var}(X)\,u$ est la variance de $u^tX$. On en déduit que $u^t\,\mathrm{var}(X)\,u \geqslant 0$, et ceci pour n'importe quel vecteur $u$ : $\mathrm{var}(X)$ est une matrice positive.

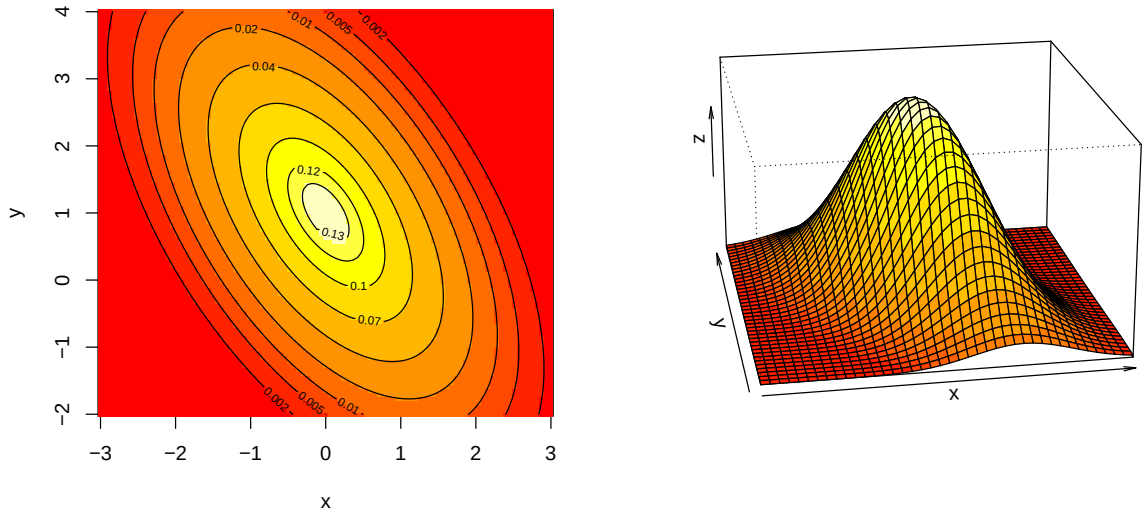



FIGURE 4.1. – Densité d'une loi normale avec $\mu = (0,1)^t$ et $\Omega = \begin{pmatrix} 1 & -0{,}8 \\ -0{,}8 & 2 \end{pmatrix}$.

Soit $A \in \mathbb{R}^{n\times n}$ une matrice symétrique. Si pour tout vecteur $u \in \mathbb{R}^{n\times 1}$, on a $u^t A u \geqslant 0$, la matrice A est dite *positive*. Les matrices de variance d'un vecteur aléatoire sont toujours positives.

Si pour tout vecteur $u \in \mathbb{R}^{n\times 1}$, on a $u^t A u > 0$, la matrice A est dite *définie positive*.

Si il existe un vecteur $u$ tel que $u^t A u = 0$, on dit que la matrice var(X) est *singulière*. Dans le cas où A = var(X) est une matrice de variance, on a $\mathrm{var}(u^t X) = u^t A u = 0$, ce qui signifie que $u^t X$ est une constante.

Les matrices symétriques définies positives sont toujours inversibles ; les matrices singulières ne le sont pas.

**Exemple** Si $X = (X_1, \ldots, X_d)^t$ suit une loi multinomiale ... alors $X_1 + \cdots + X_d = n$ est une constante ; en posant $u = (1, \ldots, 1)^t$, on a $u^t X = n$ et $\mathrm{var}(u^t X) = u^t \,\mathrm{var}(X)\, u = 0$. La matrice var(X) est donc singulière.

## 4.2. La loi normale multivariée

### 4.2.1. Densité

La loi normale multivariée est une loi continue à densité. Soit $\Omega \in \mathbb{R}^{n\times n}$ une matrice définie positive et soit $\mu = (\mu_1, \ldots, \mu_n)^t \in \mathbb{R}^{n\times 1}$. La densité de la loi normale $\mathcal{N}(\mu, \Omega)$ est

$$f(x) = \frac{1}{\sqrt{(2\pi)^n \det(\Omega)}} \exp\left(-\frac{1}{2}(x-\mu)^t \Omega^{-1} (x-\mu)\right)$$

pour $x \in \mathbb{R}^{n\times 1}$.

Si $X \sim \mathcal{N}(\mu, \Omega)$, alors $E(X) = \mu$ et $\mathrm{var}(X) = \Omega$. La figure 4.1 montre l'aspect typique d'une telle densité, dans le cas de la dimension 2 : les courbes de niveaux sont des ellipses, la densité s'écrase très vite quand on s'éloigne de l'espérance.

### 4.2.2. Lois marginales

Si $X \in \mathcal{N}(\mu, \Omega)$, avec $\Omega = (\omega_{ij})$, alors

$$X_i \sim \mathcal{N}(\mu_i, \omega_{ii})$$

La covariance de $X_i$ et $X_j$ est $\text{cov}(X_i, X_j) = \omega_{ij}$. Si $\omega_{ij} = 0$ alors $X_i$ et $X_j$ sont indépendantes.

**Remarque :** On sait que si U et V sont indépendantes, alors cov(U,V) = 0 ; la réciproque n'est pas vraie en général ! Le cas du vecteur gaussien est très particulier.

Plus généralement, on a les résultats suivants :

Soit X un vecteur normal d'espérance $\mu \in \mathbb{R}^{n\times 1}$ et de matrice de variance $\Omega \in \mathbb{R}^{n\times n}$ : $X \sim \mathcal{N}(\mu, \Omega)$.

- si $A \in \mathbb{R}^{d\times n}$, AX est un vecteur normal d'espérance $A\mu$ et de variance $A\,\text{var}(X)\,A^t$.
- si $A \in \mathbb{R}^{d\times n}$ et $B \in \mathbb{R}^{c\times n}$, (AX, BX) est un vecteur gaussien, et la covariance de AX et BX est $\text{cov}(AX, BX) = A\,\text{var}(X)\,B^t$.
- si $u = (u_1, \ldots, u_n) \in \mathbb{R}^{n\times 1}$, $u^t X = u_1 X^1 + \cdots + u_n X_n$ est une variable normale d'espérance $u^t\mu$ et de variance $u^t\,\text{var}(X)\,u$.
- si $u, v \in \mathbb{R}^{n\times 1}$, alors $(u^t X, v^t X)$ est un vecteur gaussien ; la covariance de $u^t X$ et $v^t X$ est $\text{cov}(u^t X, v^t X) = u^t\,\text{var}(X)\,v$. En particulier, si $u^t\,\text{var}(X)\,v = 0$, $u^t X$ et $v^t X$ sont indépendantes.

Quand $A\,\text{var}(X)\,A^t$ est singulière (n'est pas définie positive) on parle de *distribution normale dégénérée*. L'exemple suivant doit permettre de comprendre ce qui se passe dans ce cas.

#### Exemple de distribution normale dégénérée

En dimension $n = 2$, on prend $\mu = (0,0)^t$ et $\Omega = I_2$ (la matrice identité) ; et $X = (X_1, X_2)^t$ un vecteur gaussien $X \sim \mathcal{N}(\mu = 0, \Omega = I_2)$. La matrice

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$$

est singulière. La figure 4.2 montre 100 valeurs de tirées selon la loi de X, $\mathcal{N}(\mu, \Omega)$, et les 100 valeurs correspondantes de Y = AX.

En notant $Y = (Y_1, Y_2)^t$ on vérifie que $Y_2 = 2Y_1$ (à faire en exercice !), ce qui explique que les valeurs prises par Y sont dans une droite. Quelles sont les lois marginales de $Y_1$ et $Y_2$ ?

#### Factorisation de Cholesky

Posons le problème informatique suivant : générer des tirages d'un vecteur gaussien de loi $\mathcal{N}(0, \Omega)$, où $\Omega$ est une matrice définie positive ; on suppose qu'on sait générer des tirages d'une variable gaussienne $\mathcal{N}(0,1)$.

Dans le cas où $\Omega = I_n$, le problème est facile, les composantes d'un vecteur $X \sim \mathcal{N}(0, I_n)$ étant indépendantes ; il suffit donc de tirer ses composantes $X_1, \ldots, X_n$ une à une. Dans le cas général, on utilisera un vecteur $(X_1, \ldots, X_n) \sim \mathcal{N}(0, I_n)$ qu'on multipliera (à gauche) par une matrice A « bien choisie » : la variance de AX étant $AA^t$, il faut $AA^t = \Omega$.

Une telle matrice est fournie par la factorisation de Cholesky.

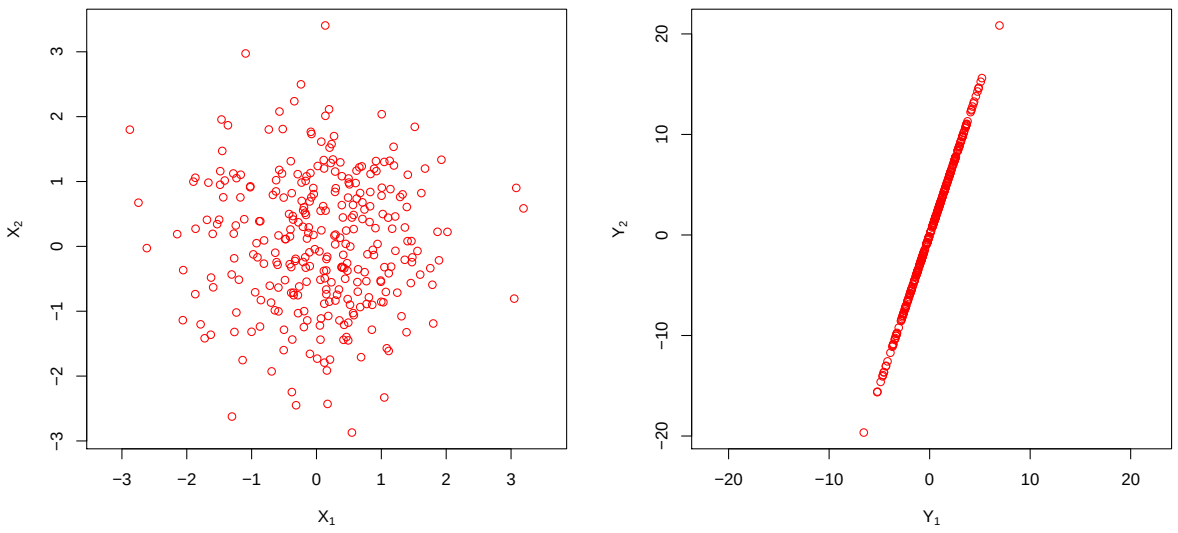



FIGURE 4.2. – Le vecteur $Y = AX$ suit une distribution normale dégénérée.

Si $\Omega$ est une matrice définie positive, il existe une matrice triangulaire inférieure C telle que $CC^t = \Omega$. Cette écriture s'appelle la factorisation de Cholesky de $\Omega$.
La matrice C est inversible et on a $\Omega^{-1} = \left(C^{-1}\right)^t C^{-1}$.

### 4.2.3. Forme quadratique d'un vecteur gaussien

On a le résultat suivant.

Soit $X \sim \mathcal{N}(0, I_n)$. Si A est une matrice symétrique de valeurs propres $\lambda_1, \ldots, \lambda_n$, alors $X^t AX$ est une somme de $n$ variables $\chi^2(1)$ indépendantes pondérées par $\lambda_1, \ldots, \lambda_n$.

$$X^t AX \sim \lambda_1 \chi^2(1) + \cdots + \lambda_n \chi^2(1)$$

En particulier, avec $A = I_n$, on retrouve $X^t X \sim \chi^2(n)$.

Dans le cas général, si $X \sim \mathcal{N}(0, \Omega)$, on utilise la factorisation de Cholesky $\Omega = CC^t$ : X s'écrit $X = CY$ avec $Y \sim \mathcal{N}(0, I_n)$. On a alors $X^t AX = Y^t \left(C^t AC\right) Y$ et on est ramené au cas précédent. Ce sont alors les valeurs propres de $C^t AC$ qui apparaissent dans la combinaison de variables de $\chi^2$.

On en déduit le cas particulier suivant :

Soit $X \in \mathbb{R}^n$ de loi $\mathcal{N}(0, \Omega)$. Alors $X^t \Omega^{-1} X$ suit une loi $\chi^2(n)$.

# 5. Vraisemblance (lois multiparamétriques)

## 5.1. La vraisemblance et ses dérivées

### 5.1.1. La vraisemblance

Ici nous considérons le cas d'une loi de probabilité $\mathscr{L}(\theta)$ où $\theta$ est un vecteur de $\mathbb{R}^d$. L'espace des paramètres $\Theta$ est donc une partie de $\mathbb{R}^d$ ; on parle d'un espace de dimension $d$.

La vraisemblance de $\theta$ pour une observation $x$ est définie comme précédemment par

$$\mathrm{L}(\theta;x) = \begin{cases} \mathrm{C}(x) \times \mathbb{P}_\theta(x) & \text{pour une loi discrète} \\ \mathrm{C}(x) \times f_\theta(x) & \text{pour une loi à densité.} \end{cases}$$

Pour $n$ observations indépendantes $x_1, \dots, x_n$ on a

$$\mathrm{L}(\theta;x_1,\dots,x_n) = \mathrm{L}(\theta;x_1) \times \cdots \times \mathrm{L}(\theta;x_n).$$

On pose $\ell(\theta;x_1,\dots,x_n) = \log \mathrm{L}(\theta;x_1,\dots,x_n)$ et on a

$$\boxed{\ell(\theta;x_1,\dots,x_n) = \ell(\theta;x_1) + \cdots + \ell(\theta;x_n).} \tag{5.1}$$

### 5.1.2. Le score, information observée, information de Fisher

Le score est maintenant le gradient de $\ell(\theta;x_1,\dots,x_n)$, donc un vecteur de dimension $d$ :

$$\boxed{\mathrm{U}(\theta;x_1,\dots,x_n) = \operatorname*{grad}_\theta \ell(\theta;x_1,\dots,x_n)}$$

Et l'information observée est l'opposée de la matrice $n \times n$ des dérivées secondes de $\ell(\theta)$ :

$$\boxed{\mathrm{J}(\theta;x_1,\dots,x_n) = -\mathrm{H}_\theta \ell(\theta;x_1,\dots,x_n)}$$

L'information de Fisher $\mathrm{I}_1(\theta)$ est l'espérance de $\mathrm{J}(\theta;\mathrm{X})$ avec $\mathrm{X} \sim \mathscr{L}(\theta)$.

## 5.2. Estimation et tests

### 5.2.1. Estimateur du maximum de vraisemblance

L'estimateur du maximum de vraisemblance (EMV) est

$$\widehat{\theta} = \operatorname*{argmax}_\theta \mathrm{L}(\theta;x_1,\dots,x_n).$$

Il peut être déterminé en cherchant $\widehat{\theta}$ qui annule le score $U(\widehat{\theta})$.

Sous des hypothèses similaires à celle que nous avons vues dans le cas d'un seul paramètre, quand $n$ est grand la loi de l'estimateur du maximum de vraisemblance s'approche d'une loi normale d'espérance $\theta_0$ et de matrice de variance $(nI_1(\theta_0))^{-1}$.
On peut en pratique approcher $nI_1(\theta_0)$ (l'information de Fisher en $\theta_0$, qu'on ne connaît pas) par $nI_1(\widehat{\theta})$ (l'information de Fisher en $\widehat{\theta}$), ou même par $J(\widehat{\theta}; x_1, \ldots, x_n)$ (l'information observée en $\widehat{\theta}$).

### 5.2.2. Les trois tests

On veut tester $H_0 : \theta = \theta_0$. Si $H_0$ est vraie :

— **test du score :** la loi de $U(\theta_0)$ est approximativement normale, centrée et de matrice de variance $(nI_1(\theta_0))$. On en déduit (cf chapitre sur la loi normale multivariée) que

$$U(\theta_0)^t \times (nI_1(\theta_0))^{-1} \times U(\theta_0)$$

suit approximativement une loi $\chi^2(d)$ ;

— **test de Wald :** la loi de $\widehat{\theta} - \theta_0$ est approximativement normale, centrée et de matrice de covariance $(nI_1(\theta_0))^{-1}$. On en déduit que

$$\left(\widehat{\theta} - \theta_0\right)^t \times (nI_1(\theta_0)) \times \left(\widehat{\theta} - \theta_0\right)$$

suit approximativement une loi $\chi^2(d)$ ;

— **test du rapport de vraisemblance :** la quantité $2\ell\left(\widehat{\theta}\right) - 2\ell(\theta_0)$ suit approximativement une loi $\chi^2(d)$.

## 5.3. Exemples

### 5.3.1. La gaussienne

On considère $n$ observations $x_1, \ldots, x_n$ tirées dans une gaussienne $\mathcal{N}(\mu, \sigma^2)$. La densité est

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

On prend comme paramètre la variance $\sigma^2$ et non l'écart-type $\sigma$. Pour faciliter la lecture des calculs, on pose $v = \sigma^2$.

La log-vraisemblance pour une observation $x$ est

$$\ell(v, \mu; x) = -\frac{1}{2}\log v - \frac{1}{2v}(x-\mu)^2.$$

Calculons le score ; on a

$$\frac{\partial}{\partial v}\ell(v,\mu;x) = -\frac{1}{2v} + \frac{1}{2v^2}(x-\mu)^2$$
$$\frac{\partial}{\partial \mu}\ell(v,\mu;x) = \frac{1}{v}(x-\mu)$$

d'où le score

$$\mathrm{U}(v,\mu;x) = \begin{pmatrix} -\frac{1}{2v} + \frac{1}{2v^2}(x-\mu)^2 \\ \frac{1}{v}(x-\mu) \end{pmatrix},$$

$$\mathrm{U}(v,\mu;x_1,\dots,x_n) = \begin{pmatrix} -\frac{n}{2v} + \frac{1}{2v^2}\sum_{i=1}^n (x_i-\mu)^2 \\ \frac{1}{v}\left(\sum_{i=1}^n x_i - n\mu\right) \end{pmatrix}.$$

Passons à l'information observée. On calcule les dérivées secondes :

$$\frac{\partial^2}{\partial v^2}\ell(v,\mu) = \frac{1}{2v^2} - \frac{1}{v^3}(x-\mu)^2$$
$$\frac{\partial^2}{\partial v \partial \mu}\ell(v,\mu) = -\frac{1}{v^2}(x-\mu)$$
$$\frac{\partial^2}{\partial \mu^2}\ell(v,\mu) = -\frac{1}{v}$$

On a donc

$$\mathrm{J}(v,\mu;x) = \begin{pmatrix} -\frac{1}{2v^2} + \frac{1}{v^3}(x-\mu)^2 & \frac{1}{v^2}(x-\mu) \\ \frac{1}{v^2}(x-\mu) & \frac{1}{v} \end{pmatrix}$$

et

$$\mathrm{J}(v,\mu;x_1,\dots,x_n) = \begin{pmatrix} -\frac{n}{2v^2} + \frac{1}{v^3}\sum_i (x_i-\mu)^2 & \frac{1}{v^2}\sum_i (x_i-\mu) \\ \frac{1}{v^2}\sum_i (x_i-\mu) & \frac{n}{v} \end{pmatrix}.$$

L'information de Fisher pour une observation est facile à calculer : si $\mathrm{X} \sim \mathcal{N}(\mu,\sigma^2)$, on a $\mathrm{E}(\mathrm{X}-\mu) = 0$ et $\mathrm{E}\left((\mathrm{X}-\mu)^2\right) = \sigma^2 = v$, et donc

$$\begin{aligned} \mathrm{I}_1(v,\mu) &= \mathrm{E}\left(\mathrm{J}(v,\mu;\mathrm{X})\right) \\ &= \begin{pmatrix} -\frac{1}{2v^2} + \frac{1}{v^3}\times v & 0 \\ 0 & \frac{1}{v} \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{2v^2} & 0 \\ 0 & \frac{1}{v} \end{pmatrix} \end{aligned}$$

On trouve l'estimateur du maximum de vraisemblance en annulant le score :

$$\widehat{v} = \frac{1}{n}\sum_{i=1}^n (x_i - \widehat{\mu})^2$$
$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^n x_i$$

La variance asymptotique de $\widehat{v}$ est $\frac{2v^2}{n}$ et celle de $\widehat{\mu}$ est $\frac{v}{n}$.

### 5.3.2. Un modèle pour les données cas/témoins

Cet exemple va nous permettre de prendre conscience du fait que les tests proposés plus haut ne suffisent pas toujours.

On a étudié en exercice le modèle suivant :
- un locus di-allélique A/$a$ est à l'équilibre de Hardy-Weinberg dans la population générale ;
- la pénétrance des génotypes AA, A$a$, et $aa$ est, respectivement, $f_0, r f_0$, $r^2 f_0$ ;
- on dispose d'un échantillon de témoins considéré comme tiré dans la population générale, et d'un échantillon de cas.

Les fréquences génotypiques dans les deux échantillons sont alors donnés par la table 5.1, où on a posé $q = 1-p$, $p' = \frac{p}{p+qr}$ et $q' = 1-p'$.

| Génotype | AA | Aa | aa |
|---|---|---|---|
| Témoins | $p^2$ | $2pq$ | $q^2$ |
| Cas | $p'^2$ | $2p'q'$ | $q'^2$ |

TABLE 5.1. – Fréquences génotypiques chez les cas et les témoins

La table 5.2 donne les effectifs génotypiques observés.

| Génotype | AA | Aa | aa | |
|---|---|---|---|---|
| Témoins | 20 | 48 | 32 | 100 |
| Cas | 24 | 73 | 23 | 120 |

TABLE 5.2. – Effectifs génotypiques

Travailler avec les paramètres $p$ et $p'$ est beaucoup plus confortable qu'avec les paramètres $p$ et $r$. On écrit la log-vraisemblance :

$$\begin{aligned}\ell(p,p') &= 20\log\left(p^2\right) + 48\log\left(2p(1-p)\right) + 32\log\left((1-p)^2\right) + 24\log\left(p'^2\right) + 73\log\left(2p'(1-p')\right) + 23\log\left((1-p')^2\right)\\ &= (2\times 20+48)\log p + (48+2\times 32)\log(1-p) + (2\times 24+73)\log p' + (73+2\times 23)\log\left(1-p'\right) + \text{cste}\\ &= 88\log p + 112\log(1-p) + 121\log p' + 119\log\left(1-p'\right) + \text{cste}.\end{aligned}$$

Il est alors facile de calculer le maximum de vraisemblance :

$$\widehat{p} = \frac{88}{200} = 0{,}440$$

$$\widehat{p'} = \frac{121}{240} = 0{,}504$$

Les trois tests décrits ci-dessus permettent de tester une hypothèse du type $H_0 : p = p_0, p' = p'_0$ Cependant ça n'a pas un grand intérêt pratique ! On veut pouvoir tester l'hypothèse $H_0 : p' = p$.

La section suivante va donner une façon de le faire.

## 5.4. Comparaison de modèles emboîtés

Rappelons que le paramètre $\theta = (\theta_1, \ldots, \theta_d)^t$ prend ses valeurs dans un espace de dimension $d$, $\Theta \subseteq \mathbb{R}^d$. On va donner une procédure pour tester une hypothèse $H_0$ qui impose $c$ contraintes à $\theta_1, \ldots, \theta_d$, par exemple

$\theta_1 = 0$, ou $\theta_1 = \theta_2$... Ces contraintes ont pour effet de restreindre les valeurs prises par $\theta$ à un sous-espace $\Theta_0 \subset \Theta$ de dimension $d - c$ (figure 5.1). L'hypothèse nulle peut alors s'écrire $\theta \in \Theta_0$.

On considère l'estimateur du maximum de vraisemblance « sous $H_0$ » $\widehat{\theta}_0 \in \Theta_0$ qui est la valeur de $\theta \in \Theta_0$ qui maximise $L(\theta)$ :

$$\widehat{\theta}_0 = \underset{\theta \in \Theta_0}{\operatorname{argmax}} L(\theta)$$

Sous $H_0 : \theta \in \Theta_0$, la statistique

$$2\ell(\widehat{\theta}) - 2\ell(\widehat{\theta}_0)$$

suit une loi $\chi^2(c)$.
Le test construit sur cette statistique s'appelle naturellement le test du rapport de vraisemblances.

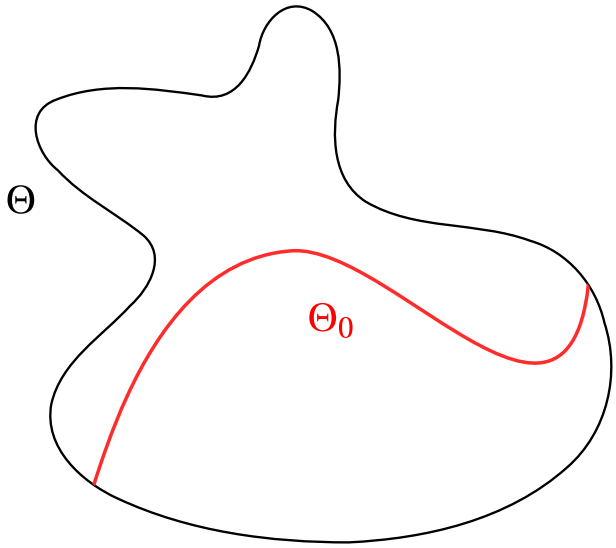



FIGURE 5.1. – L'espace des paramètres $\Theta$
et le sous-espace $\Theta_0$ correspondant à $H_0$

Il est également possible de construire un test du score et un test de Wald pour une hypothèse de cette forme, mais nous ne les présenterons pas ici.

Nous pouvons reprendre le fil de notre exemple afin d'illustrer l'usage de ce théorème.

### 5.4.1. Suite de l'exemple des données cas/témoins

La figure 5.2 montre l'espace des paramètres $\Theta = \{(p,p') ; p \in [0,1],\ p' \in [0,1]\}$, qui est de dimension deux ; le sous-espace $\Theta_0$ (en rouge) défini par la contrainte $p' = p$ est de dimension 1.

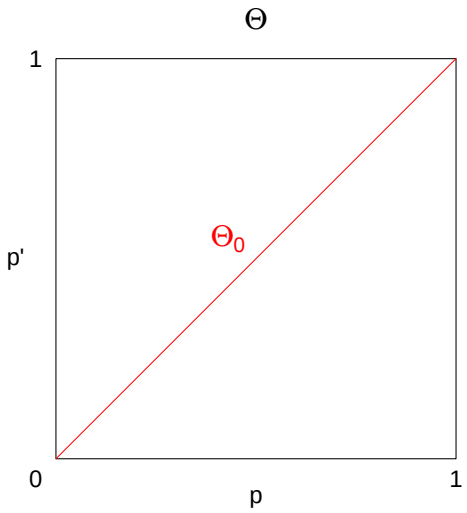



FIGURE 5.2. – Les espaces de paramètres sous $H_0$ et sous $H_1$

On maximise la vraisemblance sous $H_0$, en prenant $p' = p$ :

$$\ell(p, p' = p) = 88 \log p + 112 \log(1-p) + 121 \log p + 119 \log(1-p)$$

$$= 209 \log p + 231 \log(1-p).$$

Son maximum est atteint pour $\widehat{p} = \frac{209}{440} = 0{,}475$.

Le maximum de vraisemblance sous $H_1$ est donc $\ell(0{,}440, 0{,}504) = -303{,}533$ et sous $H_0$, $\ell(0{,}475, 0{,}475) = -304{,}435$. La statistique de test

$$2\ell(0{,}440,\ 0{,}504) - 2\ell(0{,}475,\ 0{,}475) = 1{,}804.$$

On compare 1,804 aux quantiles de la loi $\chi^2(1)$ : le test n'est pas significatif.

On pourrait vérifier que le test obtenu est équivalent au test d'Armitage.

# 6. Régression linéaire simple

## 6.1. Motivation et exemple

Le problème général est le suivant : on dispose d'une variable d'intérêt $y$ et d'une variable $x$ (souvent appelée « covariable ») dont la valeur de $y$ est supposée dépendre, selon une relation linéaire

$$y = \alpha + \beta x + \varepsilon,$$

où $\varepsilon$ est un terme d'erreur : une variable aléatoire d'espérance nulle (ou « centrée »).

On a les résultats de $n$ expériences : les valeurs $y_1, \ldots, y_n$, pour des valeurs $x_1, \ldots, x_n$ de la covariable ; on veut estimer les valeurs de $\alpha$ et de $\beta$, ainsi que la variance des erreurs résiduelles (la variance de $\varepsilon$).

On pourra par exemple examiner la façon dont la pression artérielle systolique dépend de l'âge. On dispose de 40 mesures (extraites du jeu de données `diabetes` du package R `faraway`).

| TA | age | TA | age | TA | age |
|---|---|---|---|---|---|
| 160 | 65 | 111 | 28 | 136 | 41 |
| 136 | 37 | 136 | 50 | 115 | 57 |
| 130 | 28 | 110 | 31 | 150 | 83 |
| 170 | 79 | 130 | 68 | 132 | 32 |
| 158 | 26 | 138 | 36 | 160 | 53 |
| 118 | 19 | 131 | 63 | 141 | 58 |
| 139 | 53 | 159 | 50 | 150 | 41 |
| 150 | 48 | 150 | 59 | 138 | 34 |
| 160 | 63 | 124 | 23 | 112 | 21 |
| 110 | 23 | 110 | 36 | 170 | 71 |
| 130 | 64 | 180 | 43 | 122 | 31 |
| 130 | 44 | 132 | 60 | 141 | 43 |
| 111 | 48 | 140 | 56 | 136 | 55 |
| 150 | 49 | | | | |

TABLE 6.1. – Tension artérielle systolique (TA) et âge de 40 sujets

La table 6.1 contient les 40 mesures. La figure 6.1 montre le nuage de points correspondant. La table 6.2 récapitule les données.

| $n$ | $\sum_i x_i$ | $\sum_i x_i^2$ | $\sum_i x_i y_i$ | $\sum_i y_i$ | $\sum_i y_i^2$ |
|---|---|---|---|---|---|
| 40 | 1869,00 | 97699,00 | 263290,00 | 5506,00 | 771004,00 |

TABLE 6.2. – Récapitulatif des données

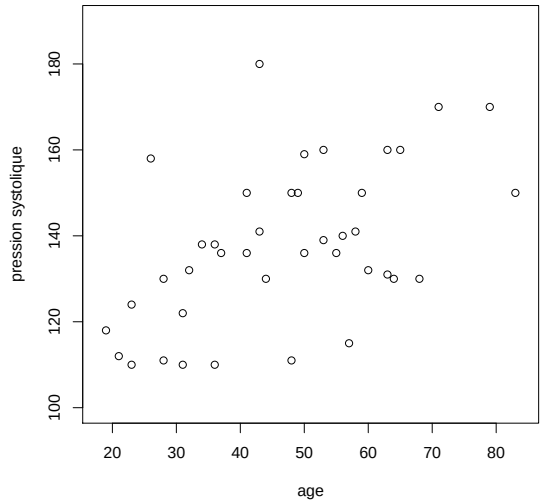



FIGURE 6.1. – La pression artérielle et l'âge pour 40 sujets

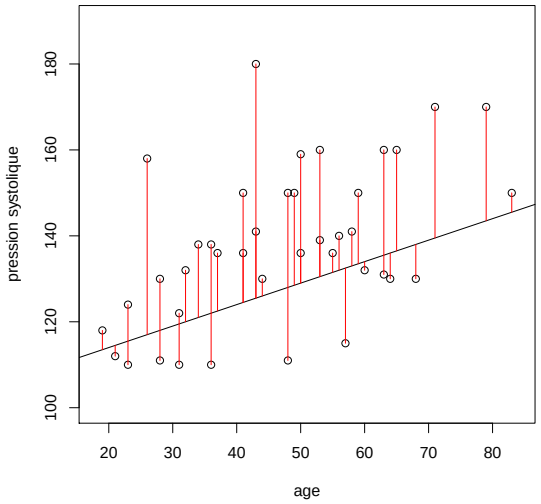



FIGURE 6.2. – Erreur pour le modèle $y = 104{,}00 + 0{,}50 \times x + \varepsilon$

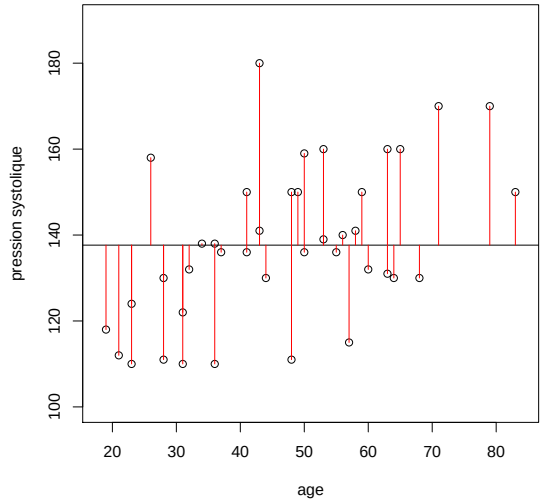



FIGURE 6.3. – Erreur pour le modèle $y = \overline{y} + \varepsilon = 136{,}00 + \varepsilon$

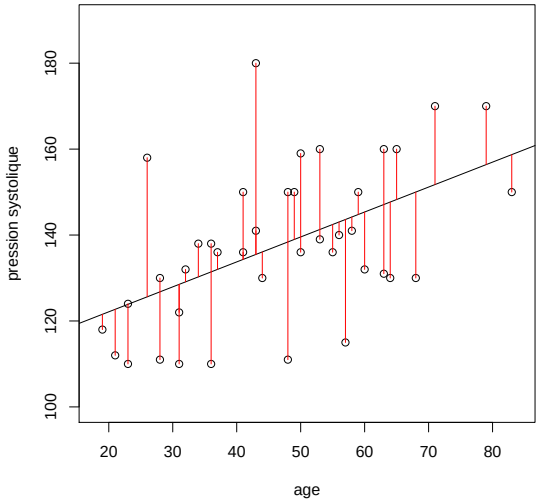



FIGURE 6.4. – Plus petite erreur possible : modèle $y = 110{,}52 + 0{,}58x + \varepsilon$

On va chercher à estimer $\alpha$ et $\beta$ qui donnent « la meilleure prédiction » des valeurs observées $y_1, \ldots, y_n$. Pour cela il faut choisir une façon de mesurer la qualité d'une prédiction ; le choix classique est la somme des carrés des écarts entre les valeurs prédites $\alpha + \beta x_i$ et les valeurs observées $y_i$ :

$$S(\alpha,\beta) = \sum_{i=1}^{n} (\alpha + \beta x_i - y_i)^2.$$

Plus $S(\alpha,\beta)$ est petit, mieux les valeurs observées sont prédites. On va donc estimer $\alpha$ et $\beta$ par $\widehat{\alpha}$ et $\widehat{\beta}$ tels que $S(\widehat{\alpha},\widehat{\beta})$ soit le plus petit possibble.

La figure 6.2 illustre la façon dont les erreurs sont mesurées : les longueurs des barres rouges donnent les erreurs (absolues) de prédiction, ce sont leurs carrés qui sont sommés. La figure 6.4 correspond aux valeurs $\widehat{\alpha}$ et $\widehat{\beta}$ de $\alpha$ et $\beta$ qui minimisent cette somme de carrés.

On parle pour $\widehat{\alpha}$ et $\widehat{\beta}$ d'*estimateurs des moindres carrés*. Nous verrons qu'en fait, si on suppose que le terme d'erreur $\varepsilon$ est pris dans une loi normale, ce sont également les estimateurs du maximum de vraisemblance. La théorie de la vraisemblance fournit des tests asymptotiques (c'est-à-dire portant sur des grands échantillons), cependant il est possible de fournir des tests exacts pour les petits échantillons, exactement comme on le fait avec le test $t$ pour comparer les moyennes de deux échantillons supposés pris dans des lois normales. C'est cette théorie qui est développée dans ce chapitre.

La figure 6.3 correspond à l'erreur commise en ignorant la covariable $x$ : la valeur de $\alpha$ qui minimise $\sum_u (y_i - \alpha)^2$ est $\alpha = \overline{y}$.
On testera l'hypothèse $H_0 : \beta = 0$, c'est-à-dire l'hypothèse selon laquelle la valeur de $x$ n'influence pas la valeur de $y$, en comparant les valeurs des erreurs commises dans ces deux modèles ; c'est-à-dire en comparant la somme des carrés des longueurs des barres rouges de la figure 6.3 à celles de la figure 6.4.

## 6.2. Les moindres carrés

On considère donc le modèle $y = \alpha + \beta x + \varepsilon$, où $\varepsilon$ est une variable aléatoire d'espérance nulle qui représente un « bruit » ou une « erreur ».

On a des observations numérotées par $i = 1, \ldots, n$ : $(y_1, x_1), \ldots, (y_i, x_i)$.

### 6.2.1. Estimation des paramètres

On va estimer $\alpha$ et $\beta$ en minimisant les carrés des écarts entre la *valeur prédite* $\widehat{y_i} = \widehat{\alpha} + \widehat{\beta} x_i$ et la valeur observée $y_i$ :

$$S(\alpha,\beta) = \sum_{i=1}^{n} (\alpha + \beta x_i - y_i)^2.$$

On calcule le gradient de S :

$$\frac{\partial}{\partial \alpha} S(\alpha,\beta) = 2 \sum_{i=1}^{n} (\alpha + \beta x_i - y_i)$$
$$\frac{\partial}{\partial \beta} S(\alpha,\beta) = 2 \sum_{i=1}^{n} x_i (\alpha + \beta x_i - y_i)$$

et on cherche $\widehat{\alpha},\widehat{\beta}$ qui annulent le gradient. On obtient le système d'équations

$$\begin{cases} n\widehat{\alpha} + \widehat{\beta} \sum_i x_i = \sum_i y_i \\ \widehat{\alpha} \sum_i x_i + \widehat{\beta} \sum_i x_i^2 = \sum_i x_i y_i \end{cases}$$

La première équation permet d'obtenir

$$\boxed{\widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x}}$$

où $\overline{x} = \frac{1}{n}\sum_i x_i$ et $\overline{y} = \frac{1}{n}\sum_i y_i$.

En remplaçant $\alpha$ par cette valeur dans la deuxième équation on obtient

$$\left(\overline{y} - \widehat{\beta}\overline{x}\right)\sum_i x_i + \widehat{\beta}\sum_i x_i^2 = \sum_i x_i y_i$$

d'où

$$\boxed{\widehat{\beta} = \frac{\sum_i x_i y_i - \left(\sum_i x_i\right)\overline{y}}{\sum_i x_i^2 - \left(\sum_i x_i\right)\overline{x}}}$$

**Remarque 1 : écritures centrées** On a également les écritures « centrées » suivantes pour les termes qui apparaissent dans l'écriture de $\widehat{\beta}$ :

$$\sum_i x_i y_i - \sum_i x_i \overline{y} = \sum_i \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)$$
$$\sum_i x_i^2 - \sum_i x_i \overline{x} = \sum_i \left(x_i - \overline{x}\right)^2.$$

On obtient alors l'écriture suivante pour $\widehat{\beta}$ :

$$\boxed{\widehat{\beta} = \frac{\sum_i \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sum_i \left(x_i - \overline{x}\right)^2}}$$

**Remarque 2 : une heuristique** On peut également arriver à l'expression de $\widehat{\beta}$ par l'« heuristique » suivante : si $y = \alpha + \beta x + \varepsilon$, alors

$$\mathrm{cov}(y,x) = \beta\,\mathrm{var}(x)$$

d'où

$$\beta = \frac{\mathrm{cov}(x,y)}{\mathrm{var}(x)}$$

et, en remplaçant $\mathrm{cov}(x,y)$ et $\mathrm{var}(x)$ par leurs estimateurs, on retrouve l'expression précendente sous la forme

$$\widehat{\beta} = \frac{\frac{1}{n-1}\sum_i \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\frac{1}{n-1}\sum_i \left(x_i - \overline{x}\right)^2}.$$

L'expression $\widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x}$ se retrouve de la même façon écrivant

$$\mathrm{E}(y) = \alpha + \beta\mathrm{E}(x)$$

(on suppose que les erreurs sont centrées).

Tout ceci n'est *pas* totalement rigoureux car on n'a pas supposé que $x$ était une variable aléatoire (au contraire, on suppose généralement que les valeurs $x_1, \ldots, x_n$ sont fixées par l'expérimentateur).

**Remarque 3 : loi du vecteur des $n$ observations** Le vecteur $\mathrm{Y} = (y_1, \ldots, y_n)^t$ des $n$ observations suit une loi normale multivariée d'espérance $\mathrm{X}\beta$ où $\mathrm{X} = (x_1, \ldots, x_n)^t$ est le vecteur des covariables, et de variance $\sigma^2\mathrm{I}_n$. C'est un moyen d'écrire le modèle de façon très compacte :

$$\mathrm{Y} \sim \mathcal{N}\left(\mathrm{X}\beta, \sigma^2\mathrm{I}_n\right).$$

### 6.2.2. Erreur résiduelle

On note l'estimation des termes d'erreur

$$\widehat{\varepsilon}_i = y_i - (\widehat{\alpha} + \widehat{\beta} x_i) = y_i - \widehat{y}_i.$$

On appelle les $\widehat{\varepsilon}_i$ les *résidus de la régression.*

On définit alors la somme des carrés résiduels (SCR) du modèle, qui est $S\left(\widehat{\alpha}, \widehat{\beta}\right)$ :

$$\boxed{\text{SCR} = S\left(\widehat{\alpha}, \widehat{\beta}\right) = \sum_i \widehat{\varepsilon_i}^2 = \sum_i (y_i - \widehat{y}_i)^2.}$$

En pratique, on peut la calculer en utilisant la « formule décentrée » suivante :

$$\boxed{\text{SCR} = \sum_i y_i^2 - \left(\widehat{\alpha} \sum_i y_i + \widehat{\beta} \sum_i x_i y_i\right).}$$

### 6.2.3. Estimation de la variance des termes d'erreurs

On suppose que les erreurs $\varepsilon_i$ sont tirées dans une loi de variance inconnue $\sigma^2$. On a un estimateur sans biais de $\sigma^2$ en prenant

$$\begin{aligned}\widehat{\sigma^2} &= \frac{1}{n-2} \sum_{i=1}^{n} \left(y_i - \widehat{y}_i\right)^2 \\ &= \frac{1}{n-2} \text{SCR}.\end{aligned}$$

On peut expliquer le terme en $\frac{1}{n-2}$ à la place du $\frac{1}{n}$ qu'on aurait pu attendre par le fait que l'estimation de $\alpha$ et de $\beta$ « retire deux degrés de liberté ». L'inconvénient de cette explication nue est qu'elle tient souvent plus du mantra ou de la formule magique que d'autre chose.

### 6.2.4. Récapitulation

Résumons les résultats énoncés :

On obtient les estimateurs des moindres carrés par

$$\widehat{\beta} = \frac{\sum_i x_i y_i - \left(\sum_i x_i\right)\overline{y}}{\sum_i x_i^2 - \left(\sum_i x_i\right)\overline{x}} = \frac{\sum_i \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sum_i \left(x_i - \overline{x}\right)^2}.$$

et

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x}.$$

Les résidus de la régression sont $\widehat{\varepsilon}_i = y_i - \widehat{y}_i$ où $\widehat{y}_i = \alpha + \widehat{\beta}x_i$. La somme des carrés résiduels est définie par

$$\text{SCR} = \text{S}\left(\widehat{\alpha}, \widehat{\beta}\right) = \sum_i \widehat{\varepsilon}_i{}^2 = \sum_i (y_i - \widehat{y}_i)^2.$$

On peut la calculer avec la formule

$$\text{SCR} = \sum_i y_i^2 - \left(\widehat{\alpha}\sum_i y_i + \widehat{\beta}\sum_i x_i y_i\right).$$

On pose

$$\widehat{\sigma^2} = \frac{1}{n-2}\text{SCR}.$$

$\widehat{\sigma^2}$ est un estimateur sans biais de $\sigma^2$ .

### 6.2.5. Retour à notre exemple

Calculons d'abord $\widehat{\alpha}$ et $\widehat{\beta}$ à partir des données de la table 6.2.

On a $\overline{y} = \frac{1}{40}5506{,}00 = 136{,}00$ et $\overline{x} = \frac{1}{40}1869{,}00 = \bar{x}$. On calcule donc

$$\widehat{\beta} = \frac{263290{,}00 - 1869{,}00 \times 136{,}00}{97699{,}00 - 1869{,}00 \times 46{,}725} = 0{,}58$$

puis

$$\widehat{\alpha} = 136{,}00 - 0{,}58 \times 46{,}725 = 110{,}52.$$

On peut calculer ensuite les carrés résiduels (table 6.3) et la somme de leurs carrés SCR = 9605,86 (en pratique on utilise en fait la formule décentrée donnée plus haut !) ; on en déduit $\widehat{\sigma^2} = \frac{1}{40-2}9605{,}86 = 252{,}786$.

| $x_i$ | $y_i$ | $\widehat{y}_i = \widehat{\alpha} + \widehat{\beta}x$ | $\widehat{\varepsilon}_i$ |
|---|---|---|---|
| 65 | 160 | 148,26 | 11,74 |
| 28 | 111 | 126,78 | -15,78 |
| 41 | 136 | 134,33 | 1,67 |
| 37 | 136 | 132,00 | 4,00 |
| 50 | 136 | 139,55 | -3,55 |
| ⋮ | ⋮ | ⋮ | ⋮ |

TABLE 6.3. – Carrés résiduels

## 6.3. Lien avec le maximum de vraisemblance : le cas Gaussien

On fait ici et pour le reste du chapitre une hypothèse supplémentaire : l'erreur aléatoire suit une loi de Gauss ; les erreurs des observations sont indépendantes et de même variance.

$$\boxed{\varepsilon \sim \mathcal{N}(0,\sigma^2)}$$

Le modèle a donc trois paramètres à estimer : $\alpha,\beta$ et $\sigma^2$.

La log-vraisemblance d'une observation $(y_i,x_i)$ est $\ell(\alpha,\beta,\sigma^2|y_i,x_i) = -\frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y_i-\alpha-\beta x_i)^2$, et la log-vraisemblance de l'ensemble des observations s'obtient en prenant la somme :

$$\begin{aligned}\ell(\alpha,\beta,\sigma^2) &= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2 \\ &= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\mathrm{S}(\alpha,\beta).\end{aligned}$$

On voit qu'indépendament de la valeur de $\sigma^2$, la log-vraisemblance est maximale quand $\mathrm{S}(\alpha,\beta)$ est minimale. Les estimateurs $\widehat{\alpha}$ et $\widehat{\beta}$ calculés plus haut sont les estimateurs du maximum de vraisemblance.

On peut vérifier ensuite que l'estimateur du maximum de vraisemblance pour $\sigma^2$ est

$$\widehat{\sigma^2}_{\mathrm{MV}} = \frac{1}{n}\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2,$$

mais il est (légèrement) biaisé. Nous lui préfèrerons l'estimateur $\widehat{\sigma}^2$ introduit plus haut.

Il est enfin intéressant de noter que la valeur de la vraisemblance à son maximum est

$$\boxed{\ell\left(\widehat{\alpha},\widehat{\beta},\widehat{\sigma^2}_{\mathrm{MV}}\right) = \frac{n}{2}\log\left(\widehat{\sigma^2}_{\mathrm{MV}}\right) - \frac{n}{2}}$$

La théorie générale de la vraisemblance permet d'obtenir des lois et des tests *asymptotiques* pour le modèle linéaire. Cependant il est possible d'obtenir des lois exactes, valides également quand les échantillons sont petits.

## 6.4. Loi des estimateurs dans le cas Gaussien

— Le vecteur $\left(\widehat{\alpha},\widehat{\beta}\right)^t$ suit une loi normale d'espérance $(\alpha,\beta)^t$ et de matrice de covariance

$$\frac{\sigma^2}{\sum_i\left(x_i-\overline{x}\right)^2}\begin{pmatrix}\frac{1}{n}\sum_i x_i^2 & -\overline{x}\\ -\overline{x} & 1\end{pmatrix} = \sigma^2\begin{pmatrix}v_{11} & v_{12}\\ v_{21} & v_{22}\end{pmatrix}.$$

— L'estimateur de la variance $\widehat{\sigma^2}$ suit une loi $\frac{\sigma^2}{n-2}\chi^2(n-2)$.

— Le vecteur des coefficients $(\widehat{\alpha},\widehat{\beta})^t$ et l'estimateur de la variance $\widehat{\sigma^2}$ sont indépendants.

En particulier, on a $\widehat{\alpha}\sim\mathcal{N}(\alpha,\sigma^2 v_{11})$ et $\widehat{\beta}\sim\mathcal{N}(\beta,\sigma^2 v_{22})$. Si la valeur de $\sigma^2$ est connue, on peut en déduire des intervalles de confiance de niveau $1-a$ pour $\alpha$ et $\beta$ :

$$\widehat{\alpha} \pm z_{1-a/2}\sqrt{\sigma^2 v_{11}}$$
$$\widehat{\beta} \pm z_{1-a/2}\sqrt{\sigma^2 v_{22}}$$

où $z_{1-a/2}$ est un quantile de loi normale standard (typiquement pour $a = 0{,}05$ on a $z_{1-a/2} = 1{,}96$).

La valeur de $\sigma^2$ n'étant généralement pas connue, il faut la remplacer par son estimation ; on doit alors utiliser des quantiles de la loi $t(n-2)$ pour obtenir l'intervalle de confiance.

— La loi de $\left(\frac{\widehat{\alpha}-\alpha}{\sqrt{\widehat{\sigma^2} v_{11}}}\right)$ est une loi $t(n-2)$ ;
— la loi de $\left(\frac{\widehat{\beta}-\beta}{\sqrt{\widehat{\sigma^2} v_{22}}}\right)$ est une loi $t(n-2)$ ;
— on a des intervalles de confiance de niveau $1-a$ pour $\alpha$ et $\beta$ :

$$\widehat{\alpha} \pm t_{1-a/2}^{n-2} \sqrt{\widehat{\sigma^2} v_{11}}$$
$$\widehat{\beta} \pm t_{1-a/2}^{n-2} \sqrt{\widehat{\sigma^2} v_{22}}$$

#### Retour à notre exemple

La matrice de variance est estimée par

$$\frac{\widehat{\sigma^2}}{\sum_i (x_i - \overline{x})^2} \begin{pmatrix} \frac{1}{n}\sum_i x_i^2 & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix} = \frac{252{,}786}{97699{,}00 - 1869{,}00 \times 46{,}725} \begin{pmatrix} \frac{1}{40}97699{,}00 & -46{,}725 \\ -46{,}725 & 1 \end{pmatrix}$$
$$= \begin{pmatrix} 59{,}54 & -1{,}139 \\ -1{,}139 & 0{,}024 \end{pmatrix}.$$

(On a utilisé la formule décentrée pour calculer $\sum_i (x_i - \overline{x})^2$). On a donc les intervalles de confiance suivants :

$$110{,}52 \pm 2{,}024\sqrt{59{,}54} = [94{,}89\,;\,126{,}14]$$
$$0{,}58 \pm 2{,}024\sqrt{0{,}024} = [0{,}26\,;\,0{,}90]$$

## 6.5. Prévision

Une fois les paramètres estimés, on peut les utiliser pour prévoir comment va se comporter le système pour une valeur donnée de $x$.

— pour une valeur donnée du prédicteur $x$, comment calculer un intervalle de confiance sur la valeur moyenne de la réponse $y$ ?
— pour une valeur donnée du prédicteur $x$, comment calculer un intervalle de confiance sur la réponse $y$ qui va être observée ?

La première question demande un intervalle de confiance sur $E(y) = \alpha + \beta x$ (qui est naturellement estimé par $\widehat{\alpha} + \widehat{\beta} x$), la seconde sur $\alpha + \beta x + \varepsilon$ (l'incertitude sur la réponse $y$ étant due à la fois au terme d'erreur $\varepsilon$ et à l'incertitude sur les valeurs de $\alpha$ et $\beta$).

Calculons d'abord la variance de $\widehat{\alpha} + \widehat{\beta} x$. C'est

$$\begin{aligned} \operatorname{var}\left(\widehat{\alpha} + \widehat{\beta} x\right) &= \operatorname{var}(\widehat{\alpha}) + 2x \operatorname{cov}\left(\widehat{\alpha}, \widehat{\beta}\right) + x^2 \operatorname{var}\left(\widehat{\beta}\right) \\ &= \sigma^2 v_{11} + 2x\sigma^2 v_{12} + x^2 \sigma^2 v_{22} \end{aligned}$$

Cette écriture pourrait suffire pour toutes les applications pratiques. Cependant une autre façon de l'écrire

nous permet de comprendre la façon dont se comporte ce terme :

$$\begin{aligned}
\operatorname{var}\left(\widehat{\alpha}+\widehat{\beta}x\right) &= \frac{\sigma^2}{\sum_i \left(x_i-\overline{x}\right)^2}\left(\frac{1}{n}\sum_i x_i^2 - 2x\overline{x} + x^2\right) \\
&= \frac{\sigma^2}{\sum_i \left(x_i-\overline{x}\right)^2}\left(\frac{1}{n}\left(\sum_i x_i^2 - \sum_i x_i\overline{x}\right) + \frac{1}{n}\sum_i x_i\overline{x} - 2x\overline{x} + x^2\right) \\
&= \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2 - 2x\overline{x} + x^2}{\sum_i \left(x_i-\overline{x}\right)^2}\right) \\
&= \sigma^2\left(\frac{1}{n} + \frac{\left(\overline{x}-x\right)^2}{\sum_i \left(x_i-\overline{x}\right)^2}\right)
\end{aligned}$$

On voit que la variance de $\widehat{\alpha}+\widehat{\beta}x$ croît avec $\left(\overline{x}-x\right)^2$ : plus on s'éloigne de la moyenne des $x_i$, moins précise est l'estimation de $\alpha+\beta x$.

La variance de $\widehat{\alpha}+\widehat{\beta}x+\varepsilon$ est maintenant facile à calculer, $\varepsilon$ étant un terme d'erreur indépendant des erreurs $\varepsilon_1,\ldots,\varepsilon_n$, et donc des estimateurs $\widehat{\alpha}$ et $\widehat{\beta}$ :

$$\begin{aligned}
\operatorname{var}\left(\widehat{\alpha}+\widehat{\beta}x+\varepsilon\right) &= \operatorname{var}\left(\widehat{\alpha}+\widehat{\beta}x\right) + \operatorname{var}(\varepsilon) \\
&= \sigma^2\left(1 + \frac{1}{n} + \frac{\left(\overline{x}-x\right)^2}{\sum_i \left(x_i-\overline{x}\right)^2}\right)
\end{aligned}$$

Et bien sûr, quand on remplace $\sigma^2$ par son estimateur $\widehat{\sigma^2}$, on a besoin des quantiles de la loi $t$.

On a l'intervalle de confiance de niveau $1-a$ suivant sur $\mathrm{E}(y)=\alpha+\beta x$ :

$$\widehat{\alpha}+\widehat{\beta}x \pm t_{1-a/2}^{n-2}\widehat{\sigma}\sqrt{\frac{1}{n} + \frac{\left(\overline{x}-x\right)^2}{\sum_i \left(x_i-\overline{x}\right)^2}}$$

Et l'intervalle de confiance suivant sur la valeur de $y$ qui va être observée :

$$\widehat{\alpha}+\widehat{\beta}x \pm t_{1-a/2}^{n-2}\widehat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{\left(\overline{x}-x\right)^2}{\sum_i \left(x_i-\overline{x}\right)^2}}$$

Dans le cas de notre exemple, ces intervalles de confiance sont illustrés par la figure 6.5.

## 6.6. Tester $\alpha=\beta=0$

On peut tester l'hypothèse nulle $\mathrm{H}_0 : \alpha=\beta=0$ ; c'est-à-dire, non seulement la covariable n'a pas d'effet, mais $y$ est centrée. Ceci est rarement utile sauf cas très particulier !

On pose $\mathrm{SCE}=\sum_i \widehat{y}_i^2$, la somme des carrés (des écarts de $y$ à 0) expliqués (on dirait mieux « prédits ») par le « modèle nul ». La statistique de test est

$$\mathrm{F} = \frac{\mathrm{SCE}/2}{\mathrm{SCR}/(n-2)} \sim \mathrm{F}(2, n-2).$$

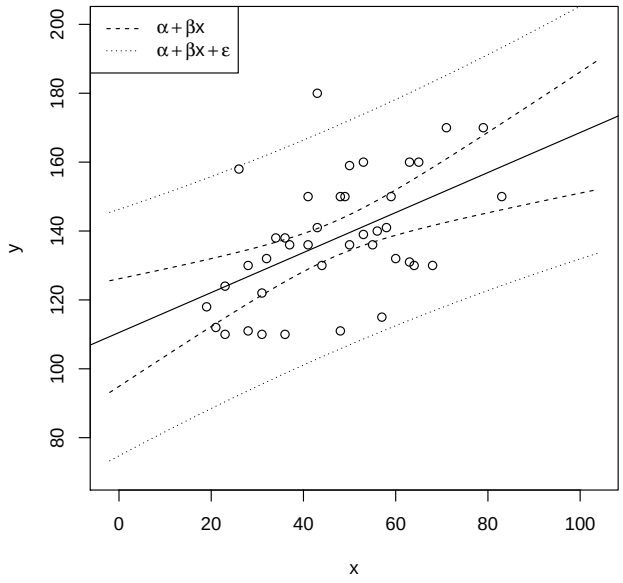



FIGURE 6.5. – Intervalles de confiances sur $\mathrm{E}(y) = \alpha + \beta x$ et sur $y = \alpha + \beta x + \varepsilon$

On peut également l'écrire

$$\boxed{\mathrm{F} = \frac{(\mathrm{SCT} - \mathrm{SCR})/2}{\mathrm{SCR}/(n-2)} \sim \mathrm{F}(2, n-2),}$$

où SCT est la somme des carrés totaux : $\mathrm{SCT} = \sum_i y_i^2$. On a en effet $\mathrm{SCT} = \mathrm{SCE} + \mathrm{SCR}$. On peut interpréter SCT comme la somme des carrés résiduels dans le modèle où $\alpha = \beta = 0$. On fait un test unilatéral à droite, car ce sont les petites valeurs de SCR (et donc, les grandes valeurs de F) qui plaident contre $\mathrm{H}_0$.

## 6.7. Tester $\beta = 0$

En général, on n'a pas de raison de penser que $y$ est centrée, et on veut simplement tester l'effet de la covariable $x$. L'hypothèse nulle est donc $\beta = 0$.

On peut simplement faire un test $t$, en utilisant la statistique

$$t = \frac{\widehat{\beta}}{\widehat{\sigma}\sqrt{v_{22}}} \sim t(n-2).$$

De façon équivalente, on peut définir $\mathrm{SCR}_0$ : la somme des carrés résiduels du « modèle nul », le modèle où $\beta = 0$, par

$$\boxed{\mathrm{SCR}_0 = \sum_i (y_i - \overline{y})^2.}$$

En effet, on estime $\alpha$ par $\widehat{\alpha} = \overline{y}$ et les résidus dans ce modèle sont $(y_i - \overline{y})$.

La statistique de test est alors

$$\boxed{\mathrm{F} = \frac{\mathrm{SCR}_0 - \mathrm{SCR}}{\mathrm{SCR}/(n-2)} \sim \mathrm{F}(1, n-2).}$$

Notons qu'on peut utiliser la formule décentrée classique pour le calcul de $\mathrm{SCR}_0$ :

$$\boxed{\mathrm{SCR}_0 = \sum_i y_i^2 - n\left(\overline{y}\right)^2 = \sum_i y_i^2 - \frac{1}{n}\left(\sum_i y_i\right)^2.}$$

#### Retour à notre exemple

On calcule $\mathrm{SCR}_0$ :

$$\mathrm{SCR}_0 = 771004{,}00 - (5506{,}00)^2/40 = 13103{,}10.$$

Puis on calcule F :

$$\mathrm{F} = \frac{13103{,}10 - 9605{,}86}{9605{,}86/(40-2)} = 13{,}83$$

## 6.8. Coefficient de détermination

Il est fréquent de rapporter, dans les résultats de la régression, le *coefficient de détermination* $\mathrm{R}^2$, qui est défini ainsi :

$$\boxed{\mathrm{R}^2 = \frac{\mathrm{SCR}_0 - \mathrm{SCR}}{\mathrm{SCR}_0} = 1 - \frac{\mathrm{SCR}}{\mathrm{SCR}_0}.}$$

Le coefficient de détermination ajusté est défini en divisant chaque somme de carrés résiduels par son degré de liberté :

$$\boxed{\mathrm{R}^2_{\mathrm{aj}} = 1 - \frac{\mathrm{SCR}/(n-2)}{\mathrm{SCR}_0/(n-1)}.}$$

#### Retour à notre exemple

On a déjà calculé $\mathrm{SCR} = 9605{,}86$ et $\mathrm{SCR}_0 = 13103{,}10$. On a donc

$$\mathrm{R}^2 = 1 - \frac{9605{,}86}{13103{,}10} = 0{,}2669$$

et

$$\mathrm{R}^2_{\mathrm{aj}} = 1 - \frac{9605{,}86/(40-2)}{13103{,}10/(40-1)} = 0{,}2476.$$

## 6.9. La régression avec R

On peut bien sûr – et c'est un exercice recommandable – réaliser la régression en menant pas à pas les calculs comme ci-dessus. Cependant R offre des fonctions qui font la régression toutes seules..

Ci-dessous, les vecteurs `y` et `x` contiennent les données de la table 6.1. On retrouve dans le résultat toutes les valeurs que nous avons calculées au fil du traitement de cet exemple.

```
> reg <- lm(y ~ x)
> summary(reg)
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-28.617 -11.373  -1.083  11.174  44.513

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 110.5154     7.7162   14.32  < 2e-16 ***
x             0.5807     0.1561    3.72 0.000643 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.9 on 38 degrees of freedom
Multiple R-squared:  0.2669,        Adjusted R-squared:  0.2476
F-statistic: 13.83 on 1 and 38 DF,  p-value: 0.0006428
```

## 6.10. Remarque : asymétrie entre $y$ et $x$

Une remarque finale est que les variables $y$ et $x$ n'ont pas ici des rôles symétriques : la droite de régression ne sera pas la même selon qu'on cherche à minimiser l'erreur de prédiction (mesurée par la somme des carrés des erreurs) de $y$ par $x$ ou celle de $x$ par $y$. En effet, dans le premier cas, on s'intéresse aux carrés des barres verticales rouges représentées figure 6.6, dans le second cas aux carrés des barres horizontales vertes (même figure, panel de droite).

La différence entre ces deux façons de faire de la régression peut être illustrée par notre exemple : les deux droites de régressions sont représentées figure 6.7, en rouge la régression qui prédit $y$ par $x$, en vert celle qui prédit $x$ par $y$.

Si on explore la relation entre deux variables qui jouent des rôles symétriques : par exemple, la taille de deux frères ; il faut garder cette particularité à l'esprit si on choisit la régression linéaire comme méthode d'analyse.

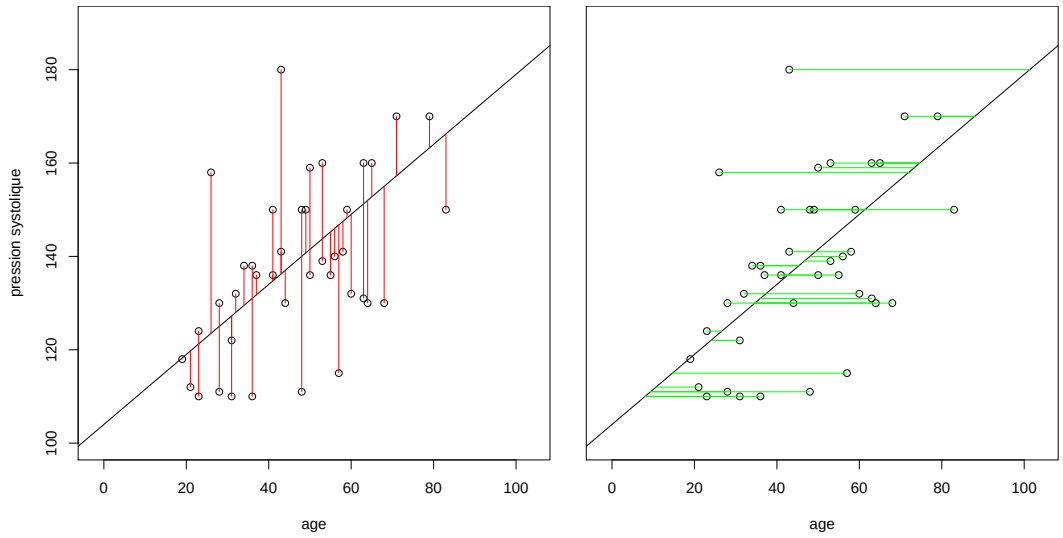



FIGURE 6.6. – Asymétrie entre $x$ et $y$ :
la mesure des erreurs dans les deux cas

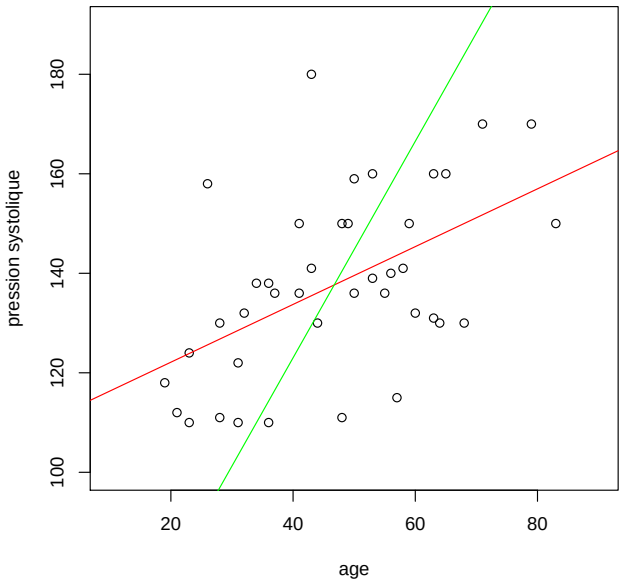



FIGURE 6.7. – Les deux droites de régression

# 7. Régression linéaire multiple

Dans ce chapitre nous nous plaçons d'emblée dans le cas Gaussien. Les mêmes arguments que dans le chapitre précédent permettent de montrer que l'estimateur des moindres carrés coïncide avec l'estimateur du maximum de vraisemblance : nous ne les répéterons pas.

## 7.1. Modèle

On suppose qu'une variable aléatoire $y$ gaussienne a son espérance déterminée par des paramètres (déterministes) $x_1, \dots, x_p$, selon le modèle

$$y = \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

où $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

On dispose de $n$ observations $y_1, \dots, y_n$ pour estimer $\beta_1, \dots, \beta_p$ et pour faire des tests sur ces valeurs. Écrivons d'emblée le modèle sous forme matricielle :

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

ou

$$\mathrm{Y} = \mathrm{X}\beta + \varepsilon,$$

avec $\mathrm{Y} = (y_1, \dots, y_n)^t$, $\mathrm{X} \in \mathbb{R}^{n \times p}$ la matrice des $(x_{ij})$, $\beta = (\beta_1, \dots, \beta_p)^t$ et $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^t$ est un vecteur gaussien centré de variance $\sigma^2 \mathrm{I}_n$.

Ainsi, Y est un vecteur gaussien d'espérance $\mathrm{X}\beta$ et de variance $\sigma^2 \mathrm{I}_n$ :

$$\mathrm{Y} \sim \mathcal{N}\left(\mathrm{X}\beta, \sigma^2 \mathrm{I}_n\right).$$

Le problème posé est d'estimer les paramètres du modèle : $\beta = (\beta_1, \dots, \beta_p)^t$ et $\sigma^2$, et de réaliser des tests sur certains d'entre eux.

Une hypothèse importante nécessaire à l'estimation des paramètres $\beta_1, \dots, \beta_p$ est qu'aucune variable ne puisse s'obtenir par somme pondérée des autres ; en effet, si par exemple $x_3 = 2x_1 + x_2$, alors

$$\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 = (\beta_1 + 2c) x_1 + (\beta_2 + c) x_2 + (\beta_3 - c) x_3$$

pour n'importe quelle valeur de $c$ : on ne peut donc pas trouver une estimation unique des $\beta_i$. La solution est d'imposer par exemple $\beta_3 = 0$, c'est-à-dire de ne pas prendre en compte la variable $x_3$. Il faut également supposer que $n \geqslant p$ : on ne peut pas estimer plus de paramètres qu'on a d'observations !

Notons que les valeurs $x_{ij}$ des covariables sont considérées comme fixées par l'expérimentateur. On peut supposer qu'on a toujours $x_{i1} = 1$ ce qui a pour effet d'inclure un terme constant (en anglais, *intercept*) dans le modèle.

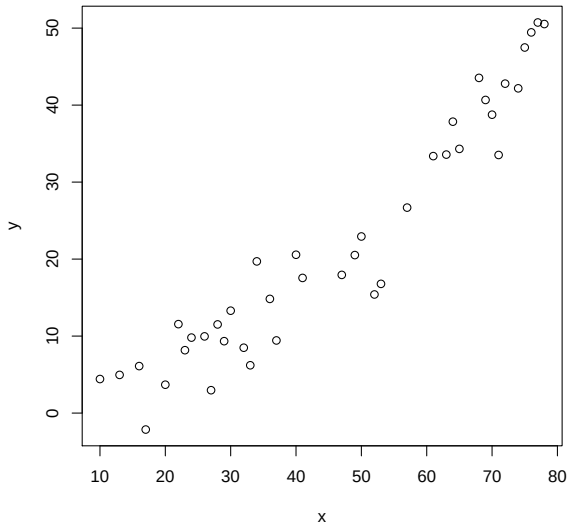



FIGURE 7.1. – Nuage de points présentant une tendance convexe

#### Cas particulier : régression polynomiale

Si on a un nuage de points comme sur la figure 7.1, il peut être intéressant de considérer un modèle $y = \alpha + \beta x + \gamma x^2$ afin que la valeur prédite puisse suivre la tendance observée (figure 7.2). C'est un cas particulier de la régression multiple, il suffit de prendre $x_1 = 1$, $x_2 = x$ et $x_3 = x^2$.

### 7.1.1. Exemple

Dans notre exemple du chapitre précédent, on a une autre variable disponible : le poids. En notant $y$ la pression artérielle, et $x_2$ l'âge, $x_3$ le poids, on va considérer le modèle

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3.$$

La variable $x_1$ est prise constante égale à 1.

La table 7.1 donne l'ensemble des données utilisées.

Pour simplifier sa présention, le traitement de l'exemple sera développé avec R. Les données sont rassemblées dans une matrice `x` dans laquelle on n'a pas oublié de placer une colonne de 1 pour la variable constante $X_1 = 1$. Cette matrice est de dimensions $40 \times 3$ ; nous en montrons ci-après les 10 premières lignes à l'aide de la fonction `head`.

```
> dim(x)
[1] 40  3
> head(x,5)
    un age weight
100  1  65    197
101  1  28    200
102  1  41    154
103  1  37    203
104  1  50    180
```

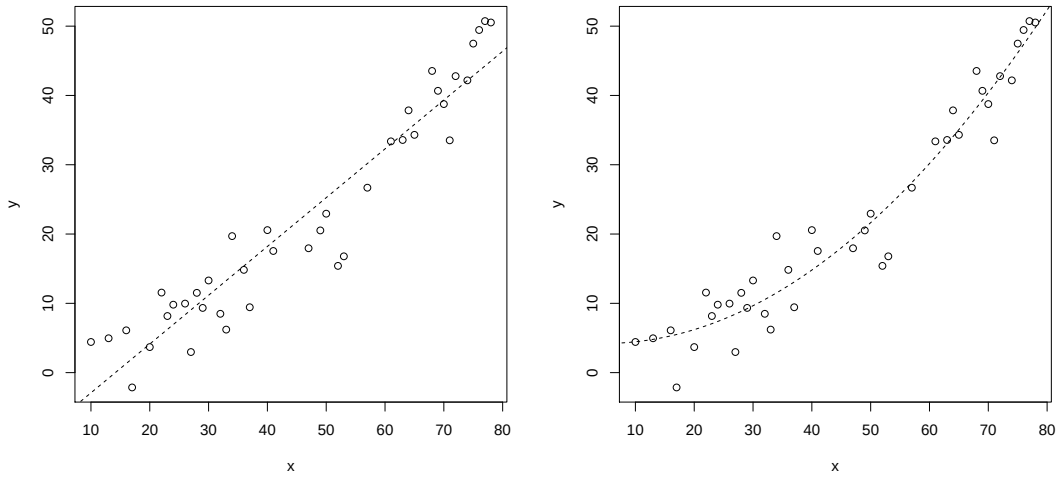



FIGURE 7.2. – Régressions $y = \alpha + \beta x$ et $y = \alpha + \beta x + \gamma x^2$

| TA | âge | poids | TA | âge | poids | TA | âge | poids |
|---|---|---|---|---|---|---|---|---|
| 160 | 65 | 197 | 111 | 28 | 200 | 136 | 41 | 154 |
| 136 | 37 | 203 | 136 | 50 | 180 | 115 | 57 | 150 |
| 130 | 28 | 204 | 110 | 31 | 200 | 150 | 83 | 125 |
| 170 | 79 | 165 | 130 | 68 | 170 | 132 | 32 | 212 |
| 158 | 26 | 227 | 138 | 36 | 150 | 160 | 53 | 174 |
| 118 | 19 | 119 | 131 | 63 | 175 | 141 | 58 | 230 |
| 139 | 53 | 158 | 159 | 50 | 263 | 150 | 41 | 156 |
| 150 | 48 | 120 | 150 | 59 | 172 | 138 | 34 | 170 |
| 160 | 63 | 158 | 124 | 23 | 164 | 112 | 21 | 169 |
| 110 | 23 | 235 | 110 | 36 | 125 | 170 | 71 | 244 |
| 130 | 64 | 225 | 180 | 43 | 140 | 122 | 31 | 227 |
| 130 | 44 | 160 | 132 | 60 | 167 | 141 | 43 | 325 |
| 111 | 48 | 121 | 140 | 56 | 151 | 136 | 55 | 223 |
| 150 | 49 | 266 | | | | | | |

TABLE 7.1. – Tension artérielle systolique (TA), âge et poids de 40 sujets

## 7.2. Les moindres carrés

### 7.2.1. Estimation du vecteur de paramètres β

Comme dans le cas de la régression simple on va estimer les β en minimisant la somme des carrés des écarts entre les valeurs prédites $\sum_{j=1}^{p} x_{ij}\beta_j$ et les valeurs observées $y_i$, ou, en notant $x_i = (x_{i1},\dots,x_{ip})$, $x_i\beta$ :

$$S(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2$$

On utilise la réécriture matricielle d'une somme de carrés :
— soit $Y = (y_1,\dots,y_n)^t \in \mathbb{R}^{n\times 1}$ (vecteur colonne) ;
— soit $X = (x_{ij}) \in \mathbb{R}^{n\times p}$ la matrice à $n$ lignes et $p$ colonnes dont les lignes sont les $x_i$ ;
alors le vecteur $X\beta$ est le vecteur colonne $(x_1\beta,\dots,x_n\beta)^t \in \mathbb{R}^{n\times 1}$, et on a

$$\begin{aligned} S(\beta) &= (Y - X\beta)^t (Y - Y\beta) \\ &= (Y^t - \beta^t X^t)(Y - X\beta) \\ &= Y^tY - \beta^t X^t Y - Y^t X\beta + \beta^t X^t X\beta \\ &= Y^tY - 2(Y^tX)\beta + \beta^t (X^tX)\beta. \end{aligned}$$

Le gradient de $S(\beta)$ est donc

$$\operatorname*{grad}_{\beta} S(\beta) = -2X^tY + 2(X^tX)\beta.$$

On en déduit l'estimateur des moindres carrés,

$$\boxed{\widehat{\beta} = (X^tX)^{-1}X^tY.}$$

On remarque qu'il faut que la matrice $(X^tX)$ soit inversible pour calculer $\widehat{\beta}$. On peut montrer que c'est le cas dès que $n \geqslant p$, et si aucune variable ne peut s'obtenir comme somme pondérée des autres ; ce sont les hypothèses dont la nécessité a déjà été remarquée plus haut.

**Remarque : une heuristique** On peut proposer une heuristique simple qui permet de retrouver l'estimateur de β. En espérance on a

$$Y = X\beta,$$

d'où

$$X^tY = (X^tX)\beta,$$

puis

$$\beta = (X^tX)^{-1}X^tY.$$

Notons en outre que $\frac{1}{n-1}(X^tX) \in \mathbb{R}^{p\times p}$ est une estimation de la matrice de variance des variables $x_1,\dots,x_p$, et $\frac{1}{n-1}X^tY$ est une estimation du vecteur de covariances covariances $(\operatorname{cov}(x_1,y),\dots,\operatorname{cov}(x_p,y))^t$ ; l'expression de $\widehat{\beta}$ correspond à la relation

$$\operatorname{cov}(x,y) = \operatorname{var}(x)\beta.$$

### 7.2.2. Erreur résiduelle

Le vecteur des valeurs $\widehat{y}_i = \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_p x_{ip}$ s'obtient simplement par un produit matriciel : $\widehat{Y} = X\widehat{\beta}$.

Comme dans la régression simple, on note les résidus de régression $\widehat{\varepsilon}_i = y_i - \widehat{y}_i$, et la somme des carrés résiduels est

$$\mathrm{SCR} = \mathrm{S}(\widehat{\beta}) = \sum_i \widehat{\varepsilon}_i^2 = \sum_i (y_i - \widehat{y}_i)^2$$

On a un estimateur sans biais de $\sigma^2$ en prenant

$$\begin{aligned}\widehat{\sigma^2} &= \frac{1}{n-p}\sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2 \\ &= \frac{1}{n-p}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2.\end{aligned}$$

Récapitulons :

L'estimateur des moindres carrés est

$$\widehat{\beta} = (\mathrm{X}^t\mathrm{X})^{-1}\mathrm{X}^t\mathrm{Y}.$$

On pose $\widehat{\mathrm{Y}} = \mathrm{X}\widehat{\beta}$. Les résidus de la régression sont les $\widehat{\varepsilon}_i = y_i - \widehat{y}_i$ et la somme des carrés résiduels est

$$\mathrm{SCR} = \sum_i \widehat{\varepsilon}_i^2 = \sum_i (y_i - \widehat{y}_i)^2.$$

On estime $\sigma^2$ par

$$\widehat{\sigma^2} = \frac{1}{n-p}\mathrm{SCR}.$$

### 7.2.3. Application à notre exemple

On calcule donc $\widehat{\beta}$ avec la formule

$$\widehat{\beta} = (\mathrm{X}^t\mathrm{X})^{-1}\mathrm{X}^t\mathrm{Y}$$

```
> beta <- solve(t(x) %*% x) %*% t(x) %*% y
> beta
               [,1]
un     98.78015873
age     0.59124097
weight  0.06099326
```

On peut calculer $\widehat{\mathrm{Y}}$ et la SCR :

```
> haty <- x %*% beta
> SCR <- sum( (y-haty)**2 )
> SCR
[1] 9307.557
```

et en déduire une estimation de $\sigma^2$ !

```
> n <- dim(x)[1]
> p <- dim(x)[2]
> n
[1] 40
> p
[1] 3
> sigma2 <- SCR/(n-p)
> sigma2
[1] 251.5556
> sqrt(sigma2)
[1] 15.8605
```

## 7.3. Loi des estimateurs

La loi de $\widehat{\beta}$ est facile à déduire du fait qu'un vecteur de $n$ observations $Y = (Y_1, \ldots, Y_n)$ suit une loi $\mathcal{N}(X\beta, \sigma^2 I_n)$.

- Le vecteur $\widehat{\beta}$ suit une loi normale d'espérance $\beta$ et de variance
$$\sigma^2 (X^t X)^{-1} = \sigma^2 V.$$
- L'estimateur de la variance $\widehat{\sigma^2} = \frac{1}{n-p} SCR$ suit une loi $\frac{\sigma^2}{n-p}\chi^2(n-p)$.
- Le vecteur $\widehat{\beta}$ et $\widehat{\sigma^2}$ sont indépendants.

On peut en déduire, comme au chapitre précédent, des intervalles de confiance sur chacun des $\beta_i$ :

$$\widehat{\beta}_i \pm t_{1-a/2}^{n-p} \sqrt{\widehat{\sigma^2} v_{ii}}.$$

De même, tester une hypothèse du type $\beta_i = 0$ par un test $t$ est immédiat.

On peut également construire des intervalles de confiance de prédiction, c'est-à-dire sur

$$x\beta = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

et

$$x\beta + \varepsilon = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon,$$

étant donné une valeur du vecteur (ligne) de variables de prédiction $x = (x_1, \ldots, x_p)$. La variance de $x\beta$ est $\sigma^2 \left(x V x^t\right)$, d'où les intervalles de confiance respectifs

$$x\beta \pm t_{1-a/2}^{n-p} \sqrt{\widehat{\sigma^2} x V x^t}$$

et

$$x\beta \pm t_{1-a/2}^{n-p} \sqrt{\widehat{\sigma^2}\left(1 + x V x^t\right)}.$$

### 7.3.1. Application à notre exemple

On estime la matrice de variance du vecteurs de nos coefficients par

```
> variance <- sigma2 * solve(t(x) %*% x)
> variance
                un           age        weight
un     175.3843997 -1.2374826194 -0.6036025726
age     -1.2374826  0.0243512487  0.0005406592
weight  -0.6036026  0.0005406592  0.0031371862
```

on a en particulier sur la diagonale les variances de chacun d'entre eux. Leurs écart-types sont

```
> sqrt( diag(variance) )
         un         age      weight
13.24327753  0.15604887  0.05601059
```

### 7.3.2. Un élément de preuve

On sait que Y suit une loi normale multivariée $\mathcal{N}(X\beta, \sigma^2 I_n)$. On en déduit que $\widehat{\beta} = (X^tX)^{-1}X^tY$ est normal, d'espérance

$$\begin{aligned} E\left(\widehat{\beta}\right) &= (X^tX)^{-1}X^tE(Y) \\ &= (X^tX)^{-1}X^tX\beta \\ &= \beta \end{aligned}$$

et de variance

$$\begin{aligned} \text{var}(Y) &= \left((X^tX)^{-1}X^t\right)\text{var}(Y)\left((X^tX)^{-1}X^t\right)^{-1} \\ &= \left((X^tX)^{-1}X^t\right) \times \sigma^2 I_n \times \left((X^tX)^{-1}X^t\right)^t \\ &= \left((X^tX)^{-1}X^t\right) \times \sigma^2 I_n \times \left(X(X^tX)^{-1}\right)^t \\ &= \sigma^2 (X^tX)^{-1}X^tX(X^tX)^{-1} \\ &= \sigma^2 (X^tX)^{-1}. \end{aligned}$$

Pour la loi de SCR, se reporter à l'annexe.

## 7.4. Test de $H_0 : \beta = 0$

On a l'égalité suivante :

$$\sum_i y_i^2 = \sum_i \widehat{\epsilon_i}^2 + \sum_i \widehat{y_i}^2,$$

qu'on écrit souvent SCT = SCR + SCE, somme des carrés totaux = somme des carrés résiduels + somme des carrés expliqués (par les variables $x_1, \ldots, x_n$).

On obtient un test de $H_0 : \beta_0$ en utilisant la statistique F :

$$F = \frac{(SCT - SCR)/p}{SCR/(n-p)} = \frac{SCE/p}{SCR/(n-p)} \sim F(p, n-p).$$

On réalise un test unilatéral à droite, les grandes valeurs de SCE plaidant contre l'hypothèse de la nullité de $\beta$ (qui entraine celle de $x\beta$).

Ce test est généralement peu utile en pratique (sauf cas particulier), car on teste la nullité de tous les $\beta$ d'un coup : on veut en général inclure (au moins) l'*intercept* dans le modèle nul.

## 7.5. Comparaison de modèles emboîtés

On suppose ici qu'on a $p = q + r$ variables $x_1, \ldots, x_p$. On veut tester l'hypothèse nulle suivante :

$$H_0 : \beta_{q+1} = \cdots = \beta_p = 0$$

Pour cela on va comparer les résultats de la régression de $y$ sur les $q$ premières variables $x_1, \ldots, x_q$ à ceux de la régression de $y$ sur l'ensemble des $p$ variables $x_1, \ldots, x_p$.

On note $X_A$ la matrice $n \times q$ dont les $q$ colonnes correspondent aux $q$ premières variables, et X la matrice de l'ensemble des variables :

$$X_A = \begin{pmatrix} x_{11} & \dots & x_{1q} \\ x_{21} & \dots & x_{2q} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{nq} \end{pmatrix} \qquad X = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ x_{21} & \dots & x_{2p} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix}$$

— On fait la régression dans le modèle à $q$ variables : on estime $(\beta_1,\dots,\beta_q)^t$ par $\widehat{\beta}_A = (X_A^t X_A)^{-1} X_A^t Y$. On calcule ensuite $SCR_0$, la somme des carrés résiduels de ce modèle, en calculant tout d'abord $\widehat{Y}_A = X_A \widehat{\beta}_A$, puis

$$SCR_0 = \sum_i (y_i - \widehat{y_{Ai}})^2.$$

— On fait d'autre part la régression dans le modèle à $p = q + r$ variables : on estime $(\beta_1,\dots,\beta_p)^t$ par $\widehat{\beta} = (X^t X)^{-1} X^t Y$ ; on calcule $\widehat{Y} = X\widehat{\beta}$ puis

$$SCR = \sum_i (y_i - \widehat{y_i})^2.$$

Le test repose sur la statistique suivante :

$$F = \frac{(SCR_0 - SCR)/r}{SCR/(n-p)} \tag{7.1}$$

qui suit, sous $H_0 : \beta_1 = \dots = \beta_q = 0$, une loi $F(r, n-p)$. On obtient un test de $H_0$ en réalisant un test unilatéral à droite : les grandes valeurs de $SCR_0 - SCR$ plaident contre l'hypothèse que $\widehat{Y}_A$ estime bien l'espérance de Y.

## 7.6. Cas particulier : modèle avec un terme constant

Dans le cas d'un modèle avec un terme constant ou *intercept*, on a posé $x_{i1} = 1$ pour tout $i$ ; le modèle est donc

$$y = \beta_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

et il est alors fréquent de tester l'hypothèse nulle $H_0 : \beta_2 = \dots = \beta_p = 0$.

Dans ce cas la régression sur la première variable $x_1 = 1$ mène à estimer $\beta_1$ par $\overline{y}$ et la somme des carrés résiduels de ce modèle est

$$SCR_0 = \sum_i (y_i - \overline{y})^2 \sim \chi^2(n-1).$$

La statistique de test est

$$F = \frac{(SCR_0 - SCR)/(p-1)}{SCR/(n-p)} \sim F(p-1, n-p). \tag{7.2}$$

### 7.6.1. Application à notre exemple

On teste $H_0 : \beta_2 = \beta_3 = 0$ avec nos données. On calcule $SCR_0$ :

```
> SCR0 <- sum( (y-mean(y))**2 )
> SCR0
[1] 13103.1
```

```
> F <- (SCR0 - SCR)/SCR*(n-p)/(p-1)
> F
[1] 7.544143
```

## 7.7. Coefficient de détermination

### 7.7.1. Définition

— Si on a inclut un terme constant dans le modèle, on pose $\mathrm{SCR}_0 = \sum_i (y_i - \overline{y})^2$ et

$$\mathrm{R}^2 = 1 - \frac{\mathrm{SCR}}{\mathrm{SCR}_0}$$

— Sinon, on pose $\mathrm{SCT} = \sum_i y_i^2$, et

$$\mathrm{R}^2 = 1 - \frac{\mathrm{SCR}}{\mathrm{SCT}}$$

Il y a aussi la version ajustée où on divise par les degrés de liberté, pour obtenir

$$\mathrm{R}^2 = 1 - \frac{\mathrm{SCR}/(n-p)}{\mathrm{SCR}_0/(n-1)}$$

ou

$$\mathrm{R}^2 = 1 - \frac{\mathrm{SCR}/(n-p)}{\mathrm{SCT}/n}$$

selon le cas.

### 7.7.2. Application à notre exemple

```
> R <- 1 - SCR/SCR0
> Ra <- 1 - SCR/SCR0*(n-1)/(n-p)
> R
[1] 0.2896675
> Ra
[1] 0.2512712
```

### 7.7.3. Lien avec les tests de modèles emboîtés

#### Modèle avec un terme constant

On a

$$\mathrm{R}^2 = 1 - \frac{\mathrm{SCR}}{\mathrm{SCR}_0} = \frac{\mathrm{SCR}_0 - \mathrm{SCR}}{\mathrm{SCR}_0}$$

(avec $\mathrm{SCR}_0 = \sum_i (y_i - \overline{y})^2$ : la SCR du modèle avec l'intercept seul), et donc

$$\frac{\mathrm{R}^2}{1-\mathrm{R}^2} = \frac{\mathrm{SCR}_0 - \mathrm{SCR}}{\mathrm{SCR}},$$

d'où on tire que la statistique F de l'équation (7.2) peut s'écrire

$$\mathrm{F} = \frac{n-p}{p-1} \times \frac{\mathrm{R}^2}{1-\mathrm{R}^2}.$$

### Modèles emboîtés

Enfin, si on veut comparer deux modèles emboîtés, le premier avec $q$ variables et le second avec $p = q + r$ variables, en notant $R^2_A$ le coefficient de détermination du petit modèle et $R^2$ celui du grand modèle, la statistique 7.1 s'écrit :

$$F = \frac{n-p}{r} \times \frac{R^2 - R^2_A}{1 - R^2}.$$

## 7.8. Régression avec R

Nous avons montré comment mener tous les calculs avec des opérations matricielles. La fonction `lm()` fait tout ça pour nous :

```
> reg <- lm(bp ~ age + weight)
> summary(reg)
Call:
lm(formula = bp ~ age + weight)

Residuals:
    Min      1Q  Median      3Q     Max
-26.630 -10.238  -0.012   9.283  47.257

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 98.78016   13.24328   7.459    7e-09 ***
age          0.59124    0.15605   3.789  0.00054 ***
weight       0.06099    0.05601   1.089  0.28321
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.86 on 37 degrees of freedom
Multiple R-squared:  0.2897,        Adjusted R-squared:  0.2513
F-statistic: 7.544 on 2 and 37 DF,  p-value: 0.001787
```

# A. L'atroce vérité sur le modèle linéaire

La légende veut que quiconque présente le matériel contenu dans cet appendice dans un master de Santé Publique est maudit sur treize générations. Aussi, il ne sera pas présenté en cours et n'est inclu dans ce document qu'à des fins ésotériques.

L'idée est la suivante : il faut considérer, dans le modèle linéaire, le vecteur $Y = (Y_1, \dots, Y_n)$ comme un vecteur gaussien d'espérance $X\beta$ (X étant une matrice dont les $n$ lignes sont remplies par les vecteurs à $p$ composantes $(x_{i1}, \dots, x_{ip})$) et de variance $\sigma^2 I_n$. Quand on adopte ce point de vue, tout se ramène à des considérations géométriques relativement simples... à ceci près qu'il s'agit de géométrie dans un espace de dimension $n$.

## A.1. Géométrie des vecteurs

En complément des notions matricielles introduites au chapitre 2, un peu de vocabulaire géométrique s'impose.

Si $u, v \in \mathbb{R}^{n\times 1}$ sont deux vecteurs colonnes, on appelle $u^t v = u_1 v_1 + \cdots u_n v_n$ le produit scalaire de $u$ et $v$. On a $u^t v = v^t u$.

On dit que $u, v$ sont orthogonaux si $u^t v = 0$.

La norme de $u$ est définie par $||u||^2 = u^t u = u_1^2 + \cdots + u_n^2$.

Il est facile de vérifier que $||u+v||^2 = ||u||^2 + 2u^t v + ||v||^2$. On en déduit le célèbre

**Théorème de Pythagore**
Si $u, v \in \mathbb{R}^{n\times 1}$ sont orthogonaux, alors $||u+v||^2 = ||u||^2 + ||v||^2$.

Une remarque simple mais utile :

Soit deux matrices $A, B \in \mathbb{R}^{n\times m}$. Supposons que les colonnes de A sont orthogonales aux colonnes de B ; autrement dit, si $A^t B = 0$ (la matrice nulle de $\mathbb{R}^{n\times m}$). Alors, pour tous vecteurs $u, v \in \mathbb{R}^{m\times 1}$, le vecteur $Au \in \mathbb{R}^{n\times 1}$ est orthogonal au vecteur $Bv \in \mathbb{R}^{n\times 1}$.

### A.1.1. Base orthonormale

Une famille de $n$ vecteurs colonnes $u_1, \dots, u_n \in \mathbb{R}^{n\times 1}$ tels que
— $||u_1||^2 = \cdots = ||u_n||^2 = 1$
— $u_i^t u_j = 0$ si $i \neq j$ : ils sont orthogonaux entre eux
est appelée une base orthonormale.
Si $x = (x_1, \dots, x_n)^t \in \mathbb{R}^{n\times 1}$ alors on peut décomposer $x$ en

$$x = (u_1^t x)u_1 + \cdots + (u_n^t x)u_n.$$

### A.1.2. Théorème spectral

Comme son nom l'indique, ce théorème fait peur.

**Théorème spectral (Weierstrass, 1858)** Si $\mathrm{A} \in \mathbb{R}^{n\times n}$ est une matrice symétrique, c.-à-d. $\mathrm{A}^t = \mathrm{A}$, il existe $n$ vecteurs $u_1, \dots, u_n \in \mathbb{R}^{n\times 1}$ tels que
— $\mathrm{A}u_1 = \lambda_1 u_1, \dots, \mathrm{A}u_n = \lambda_n u_n$ : ce sont des vecteurs propres.
— $u_1, \dots, u_n$ est une base orthonormale.
On parle de base orthornormale de vecteurs propres.

#### Conséquence pour les formes quadratiques

Une conséquence du théorème spectral est que toutes les formes quadratiques se décomposent en somme de carrés du type $(u^t x)^2$ avec $u \in \mathbb{R}^{n\times 1}$ de norme $||u|| = 1$.

Soit $\mathrm{A} \in \mathbb{R}^{n\times n}$ est une matrice symétrique, et $u_1, \dots, u_n$ une base orthonormale de vecteurs propres.
La forme quadratique $x^t \mathrm{A} x$ peut alors se décomposer en

$$x^t \mathrm{A} x = \lambda_1 (u_1^t x)^2 + \cdots + \lambda_n (u_n^t x)^2.$$

### A.1.3. Image d'une matrice

On peut interpréter le produit d'un vecteur par une matrice A comme une somme pondérée des colonnes de A. Soit $\mathrm{A} \in \mathbb{R}^{n\times m}$. On note $\mathrm{A}_1, \dots, \mathrm{A}_m \in \mathbb{R}^{n\times 1}$ les colonnes de A. Soit $x = (x_1, \dots, x_m)^t \in \mathbb{R}^{m\times 1}$. On a

$$\mathrm{A}x = x_1 \mathrm{A}_1 + \cdots + x_m \mathrm{A}_m.$$

On note I(A) l'ensemble des vecteurs colonnes de la forme $x_1\mathrm{A}_1 + \cdots + x_m\mathrm{A}_m$. D'après ce qui précède, I(A) est l'ensemble des vecteurs $\mathrm{A}x$ avec $x \in \mathbb{R}^{m\times 1}$.

### A.1.4. Matrices de projections orthogonales

Soit $\mathrm{A} \in \mathbb{R}^{n\times n}$ une matrice symétrique telle que $\mathrm{A}^2 = \mathrm{A}$. On dit qu'une telle matrice est une matrice de projection orthogonale.

On peut représenter l'action d'une telle matrice sur un vecteur $x \in \mathbb{R}^{n\times 1}$ par un dessin comme celui de la figure A.1. Rappelons que I(A) est l'image de A : c'est l'ensemble des vecteurs de la forme $x_1\mathrm{A}_1 + \cdots + x_n\mathrm{A}_n$ où les $\mathrm{A}_i$ sont les colonnes de A. On a bien sûr $\mathrm{A}x \in \mathrm{I}(\mathrm{A})$ pour n'importe quelle matrice A. Ce qui est très particulier ici c'est que si $x$ est dans I(A), alors $\mathrm{A}x = x$ ; ceci découle de ce que $\mathrm{A}^2 = \mathrm{A}$ ; et de plus, pour tout vecteur $x$, on a $\mathrm{A}x$ orthogonal à $(\mathrm{I}-\mathrm{A})x = x - \mathrm{A}x$ ; ceci découle de $\mathrm{A}^t(\mathrm{I}-\mathrm{A}) = \mathrm{A}-\mathrm{A}^2 = 0$. On en déduit, par une application du théorème de Pythagore, que

$$||x||^2 = ||\mathrm{A}x||^2 + ||x - \mathrm{A}x||^2.$$

On a plus généralement $(\mathrm{I}-\mathrm{A})x$ orthogonal à tous les vecteurs de I(A). Bien que tout ceci se passe en dimension $n$ avec $n$ potentiellement grand, des dessins en perspective comme celui-ci réussissent à bien rendre compte de la situation.

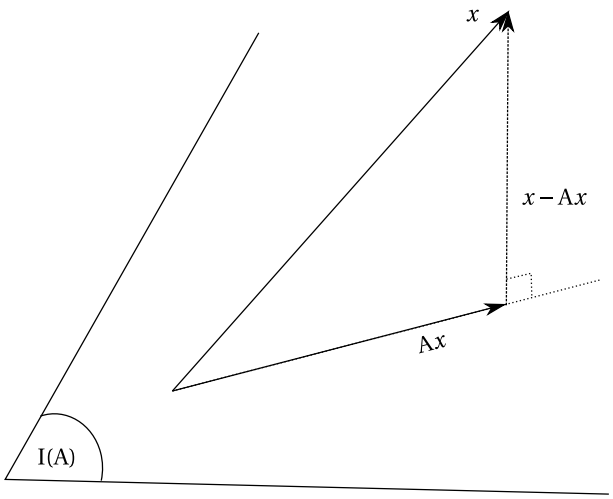



FIGURE A.1. – L'action d'une projection orthogonale

Une autre propriété notable de A est que $||\mathrm{A}x||^2$ coincide avec la forme quadratique $x^t\mathrm{A}x$ :

$$||\mathrm{A}x||^2 = (\mathrm{A}x)^t(\mathrm{A}x) = x^t\mathrm{A}^t\mathrm{A}x = x^t\mathrm{A}^2x = x^t\mathrm{A}x.$$

On a enfin le résultat suivant :

Soit A une matrice de projection orthogonale. Alors il existe $n = r + s$ vecteurs colonnes $u_1, \ldots, u_n \in \mathbb{R}^{n\times 1}$ tels que
- $\mathrm{A}u_1 = u_1, \ldots, \mathrm{A}u_r = u_r$ et $\mathrm{A}u_{r+1} = 0, \ldots, \mathrm{A}u_s = 0$
- $||u_1||^2 = \cdots = ||u_n||^2 = 1$
- $u_i^t u_j = 0$ si $i \neq j$

Le nombre $r$ de vecteurs propres associés à la valeur propre 1 est appelé le rang de A. On a $\mathrm{Tr}(\mathrm{A}) = r$.

Les vecteurs $u_1, \ldots, u_r$ sont bien sûr dans I(A), et $u_{r+1}, \ldots, u_s$ sont orthogonaux à I(A).

Si $\mathrm{A}u = \lambda u$ avec $u$ un vecteur non nul, alors on a $\mathrm{A}\times(\mathrm{A}u) = \mathrm{A}\times(\lambda u)$ d'où $\mathrm{A}u = \lambda\mathrm{A}u$ puis $\lambda u = \lambda^2 u$ ; comme $u$ est non nul on a $\lambda = \lambda^2$ d'où $\lambda = 0$ ou 1 : les seuls valeurs propres possibles pour A sont 0 et 1.

On applique le théorème spectral à A, qui est symétrique, et on en déduit le résultat (à part le résultat final sur la trace, que nous admettons).

Notons que les vecteurs propres de A sont aussi vecteurs propres de $(\mathrm{I}-\mathrm{A})$, les $r$ premiers avec valeur propre 0, et les $s=n-r$ suivant avec valeur propre 1, donc le rang de $\mathrm{I}-\mathrm{A}$ est $n-r$. Ceci peut également se déduire de $\mathrm{Tr}(\mathrm{I}_n-\mathrm{A})=\mathrm{T}r(\mathrm{I}_n)-\mathrm{Tr}(\mathrm{A})=n-\mathrm{Tr}(\mathrm{A})$.

## A.2. Complément sur les formes quadratiques d'un vecteur Gaussien

### A.2.1. Preuve du résultat sur la distribution de $\mathrm{X}^t\mathrm{AX}$

Soit $\mathrm{A}\in\mathbb{R}^{n\times n}$ une matrice symétrique.

On applique le théorème spectral (section A.1.2) à A pour obtenir des vecteurs propres orthogonaux $u_1,\dots,u_n$ avec $\mathrm{A}u_1=\lambda_1 u_1,\dots,\mathrm{A}u_r=\lambda_n u_n$, et tous les $u_i$ on norme 1 : $||u_1||^2=\cdots=||u_n||^2$.

La forme quadratique $x^t\mathrm{A}x$ se décompose alors en

$$x^t\mathrm{A}x=\lambda_1(u_1^t x)^2+\cdots+\lambda_n(u_n^t x)^2$$

On considère un vecteur aléatoire $\mathrm{X}\sim\mathcal{N}(0,\mathrm{I}_n)$, et la forme quadratique $\mathrm{X}^t\mathrm{AX}$, décomposée comme ci-dessus. Chaque $u_i^t\mathrm{X}$ est une variable normale centrée, de variance $u_i^t u_i=||u_i||^2=1$. On a donc $(u_i^t\mathrm{X})^2\sim\chi^2(1)$.

D'autre part l'orthogonalité des $u_i$ entraîne que $\mathrm{cov}(u_i^t\mathrm{X},u_j^t\mathrm{X})=u_i^t u_j=0$ pour $i\neq j$ : les $u_i\mathrm{X}$ sont indépendantes deux à deux ; les variables $(u_i^t\mathrm{X})^2$ le sont donc également.

On a donc bien montré que $\mathrm{X}^t\mathrm{AX}$ est une somme pondérée de $n$ variables $\chi^2(1)$ indépendantes,

$$\boxed{\mathrm{X}^t\mathrm{AX}\sim\lambda_1\chi^2(1)+\cdots+\lambda_n\chi^2(n).}$$

### A.2.2. Le théorème de Cochran

Cette section fait référence à la section A.1.4 sur les projections orthogonales.

Considérons une matrice de projection orthogonale $\mathrm{A}\in\mathbb{R}^{n\times n}$. Pour tout $x$, on écrit $x=\mathrm{A}x+(x-\mathrm{A}x)$ et ces deux vecteurs sont orthogonaux, d'où

$$||x||^2=||\mathrm{A}x||^2+||x-\mathrm{A}x||^2.$$

On considère un vecteur aléatoire $\mathrm{X}\sim\mathcal{N}(0,\mathrm{I}_n)$, et les vecteurs $\mathrm{AX}$ et $(\mathrm{I}-\mathrm{A})\mathrm{X}$. Ils sont gaussiens et d'espérance nulle ; la matrice de variance de $\mathrm{AX}$ est $\mathrm{A}\mathrm{I}_n\mathrm{A}^t=\mathrm{A}^2=\mathrm{A}$, celle de $(\mathrm{I}-\mathrm{A})\mathrm{X}$ est de la même façon $\mathrm{I}-\mathrm{A}$. La matrice de covariance $\mathrm{cov}(\mathrm{AX},(\mathrm{I}-\mathrm{A})\mathrm{X})$ est $\mathrm{A}(\mathrm{I}-\mathrm{A})^t=\mathrm{A}-\mathrm{A}^2=0$, donc ces vecteurs sont indépendants.

On a de plus

$$||\mathrm{X}||^2=||\mathrm{AX}||^2+||(\mathrm{I}-\mathrm{A})\mathrm{X}||^2$$

Nous nous intéressons maintenant à la loi de $||\mathrm{AX}||^2$ et $||(\mathrm{I}-\mathrm{A})\mathrm{X}||^2$ : rappelons que A a $r$ valeurs propres égales à 1 et $n-r$ égales à 0. D'après le résultat précédent,

$$||\mathrm{AX}||^2\sim\chi^2(1)+\cdots+\chi^2(1)\quad(r\text{ termes})$$

et donc $||\mathrm{AX}||^2\sim\chi^2(r)$. De même on a $||(\mathrm{I}-\mathrm{A})\mathrm{X}||^2\sim\chi^2(n-r)$. Si on récapitule, on obtient le théorème suivant.

Soit $X \sim \mathcal{N}(0, I_n)$. Si $A = A^t$ et $A^2 = A$, (c.-à-d. si A est une matrice de projection orthogonale), les vecteurs gaussiens $AX \sim \mathcal{N}(0, A)$ et $(I-A)X \sim \mathcal{N}(0, I-A)$ sont indépendants. On a

$$||X||^2 = ||AX||^2 + ||(I-A)X||^2$$

et $||X||^2 \sim \chi^2(n)$, $||AX||^2 \sim \chi^2(r)$, $||(I-A)X||^2 \sim \chi^2(n-r)$, où $r$ est le rang de A. Les variables $||AX||^2$ et $||(I-A)X||^2$ sont indépendantes.

Ce résultat peut se généraliser ainsi :

Soit $X \sim \mathcal{N}(0, I_n)$.
Si on a une décomposition de la matrice $I_n$ en

$$I_n = A_1 + \cdots + A_k$$

où les $A_i$ sont des matrices de projections orthogonales de rangs $r_1, \ldots, r_k$, alors les vecteurs gaussiens $A_iX \sim \mathcal{N}(0, A_i)$ sont indépendants. On a

$$||X||^2 = ||A_1X||^2 + \cdots + ||A_rX||^2$$

et $||X||^2 \sim \chi^2(n)$, $||A_1X||^2 \sim \chi^2(r_1), \ldots, ||A_rX||^2 \sim \chi^2(r_k)$. On a également $n = r_1 + \cdots + r_k$.

## A.3. Le modèle linéaire

On note $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^t$ le vecteur colonne des termes d'erreur. Le vecteur $\varepsilon$ est tiré dans une loi $\mathcal{N}(0, \sigma^2 I_n)$ où $I_n$ est la matrice identité de $\mathbb{R}^{n\times n}$.

### A.3.1. Loi de $\widehat{\beta}$

Nous avons montré dans le le chapitre sur le modèle linéaire que, sachant que Y suit une loi normale multivariée $\mathcal{N}(X^\beta, \sigma^2 I_n)$, on en déduit que la loi de $\widehat{\beta} = (X^tX)^{-1}X^tY$ est normale, d'espérance

$$E(\widehat{\beta}) = \beta$$

et de variance

$$\text{var}(Y) = \left((X^tX)^{-1}X^t\right)\text{var}(Y)\left((X^tX)^{-1}X^t\right)^{-1} = \sigma^2(X^tX)^{-1}.$$

### A.3.2. Indépendance de $\widehat{\beta}$ et $\widehat{\sigma^2}$

Introduisons tout d'abord une notation utile. La valeur estimée $\widehat{Y}$ pour $E(Y) = X\beta$ est

$$\begin{aligned}\widehat{Y} &= X\widehat{\beta} \\ &= X(X^tX)^{-1}X^t y \\ &= HY\end{aligned}$$

où $H = X(X^tX)^{-1}X^t$ est la « hat matriX », la « matrice chapeau », puisqu'elle « met un chapeau sur Y ».

On vérifie par des calculs élémentaires que la matrice H a la propriétés suivantes :

$$\begin{aligned} \mathrm{H}^t &= \mathrm{H} \\ \mathrm{H}^2 &= \mathrm{H} \\ \mathrm{HX} &= \mathrm{X} \end{aligned}$$

D'autre part, le vecteur $\widehat{\varepsilon} = \mathrm{Y} - \widehat{\mathrm{Y}}$ des résidus s'écrit de deux façons différentes avec la matrice H :

$$\boxed{\widehat{\varepsilon} = (\mathrm{I} - \mathrm{H})\, y = (\mathrm{I} - \mathrm{H})\varepsilon.}$$

C'est la matrice $(\mathrm{I} - \mathrm{H})$ qui « met un chapeau sur $\varepsilon$ ».

Pour montrer l'égalité encadrée ci-dessus on écrit simplement

$$\begin{aligned} \widehat{\varepsilon} &= \mathrm{Y} - \widehat{\mathrm{Y}} \\ &= \mathrm{Y} - \mathrm{HY} \\ &= (\mathrm{I}_n - \mathrm{H})\mathrm{Y} \end{aligned}$$

On remplace ensuite Y par $\mathrm{X}\beta + \varepsilon$ et on développe le produit tranquillement :

$$\begin{aligned} \widehat{\varepsilon} &= (\mathrm{I}_n - \mathrm{H})(\mathrm{X}\beta + \varepsilon) \\ &= (\mathrm{I}_n - \mathrm{H})\mathrm{X}\beta + (\mathrm{I}_n - \mathrm{H})\varepsilon \\ &= (\mathrm{X}\beta - \mathrm{HX}\beta) + (\mathrm{I}_n - \mathrm{H})\varepsilon \\ &= (\mathrm{X}\beta - \mathrm{X}\beta) + (\mathrm{I}_n - \mathrm{H})\varepsilon \\ &= (\mathrm{I}_n - \mathrm{H})\varepsilon \end{aligned}$$

La matrice de variance de $\widehat{\varepsilon}$ est donc

$$\begin{aligned} (\mathrm{I}_n - \mathrm{H}) \times \sigma^2 \mathrm{I}_n \times (\mathrm{I}_n - \mathrm{H})^t &= \sigma^2 (\mathrm{I}_n - \mathrm{H}) \times (\mathrm{I}_n - \mathrm{H}) \\ &= \sigma^2 (\mathrm{I}_n - \mathrm{H}) \end{aligned}$$

Et à partir de $\widehat{\beta} = \beta + (\mathrm{X}^t\mathrm{X})^{-1}\mathrm{X}^t\varepsilon$ et $\widehat{\varepsilon} = (\mathrm{I}_n - \mathrm{H})\varepsilon$, on calcule la covariance de $\widehat{\varepsilon}$ et $\widehat{\beta}$ :

$$\begin{aligned} (\mathrm{I}_n - \mathrm{H}) \times \sigma^2 \mathrm{I}_n \times \left((\mathrm{X}^t\mathrm{X})^{-1}\mathrm{X}^t\right)^t &= \sigma^2 (\mathrm{I}_n - \mathrm{H}) \left(\mathrm{X}(\mathrm{X}^t\mathrm{X})^{-1}\right) \\ &= \sigma^2 \left(\mathrm{X}(\mathrm{X}^t\mathrm{X})^{-1} - \mathrm{HX}(\mathrm{X}^t\mathrm{X})^{-1}\right) \\ &= \sigma^2 \left(\mathrm{X}(\mathrm{X}^t\mathrm{X})^{-1} - \mathrm{X}(\mathrm{X}^t\mathrm{X})^{-1}\right) \\ &= 0 \end{aligned}$$

Comme on est dans le cadre gaussien, on a indépendance de $\widehat{\varepsilon}$ et $\widehat{\beta}$. On en déduit l'indépendance de $\widehat{\sigma^2} = \frac{1}{n-p}||\widehat{\varepsilon}||^2$ et $\widehat{\beta}$.

### Loi de $\widehat{\sigma^2}$

On a mentionné plus haut que $\mathrm{H}^2 = \mathrm{H}$ : c'est une matrice de projection orthogonale. C'est la matrice de projection sur le sous-espace de $\mathbb{R}^n$ engendré par les colonnes de X, soit l'ensemble des valeurs $\mathrm{X}u$ possibles pour $u \in \mathbb{R}^{p\times 1}$. Nous admettrons qu'une conséquence de l'hypothèse selon laquelle les colonnes de X sont indépendantes est que le rang de H est égal à $p$.

On peut utiliser le théorème de Cochran. Les matrices H et $I-H$ permettent de décomposer le vecteur gaussien $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ ainsi :

$$\begin{aligned}\varepsilon &= H\varepsilon + (I_n - H)\varepsilon \\ &= (\varepsilon - \widehat{\varepsilon}) + \widehat{\varepsilon}\end{aligned}$$

où $(\varepsilon - \widehat{\varepsilon})$ et $\widehat{\varepsilon}$ sont des vecteurs orthogonaux. On a donc

$$||\varepsilon||^2 = ||\varepsilon - \widehat{\varepsilon}||^2 + ||\widehat{\varepsilon}||^2.$$

Le rang des $I_n - H$ est $n-p$, et le théorème de Cochran entraîne que $\widehat{\sigma^2} \sim \frac{\sigma^2}{n-p}\chi^2(n-p)$.

### Test de $H_0 : \beta = 0$

D'autre part on a également une décomposition de $y$ en

$$\begin{aligned}Y &= (I-H)Y + HY \\ &= (Y - \widehat{Y}) + \widehat{Y} \\ &= \widehat{\varepsilon} + \widehat{Y},\end{aligned}$$

ces deux vecteurs étant orthogonaux. On a alors

$$||Y||^2 = ||\widehat{\varepsilon}||^2 + ||\widehat{Y}||^2,$$

qu'on écrit souvent SCT = SCR + SCE, somme des carrés totaux = somme des carrés résiduels + somme des carrés expliqués (par les variables $x_1, \ldots, x_n$).

Sous l'hypothèse nulle $\beta = 0$, on a $Y = \varepsilon$ et on en déduit $||\widehat{Y}||^2 = ||\varepsilon - \widehat{\varepsilon}||^2 \sim \sigma^2\chi^2(p)$, indépendant de $||\widehat{\varepsilon}||^2 \sim \sigma^2\chi^2(n-p)$. Sous l'hypothèse alternative, $||\widehat{Y}||^2$ va tendre à avoir des valeurs plus grandes. On obtient un test de $H_0$ en utilisant la statistique F :

$$F = \frac{||\widehat{Y}||^2/p}{||\widehat{\varepsilon}||^2/(n-p)} = \frac{SCE/p}{SCR/(n-p)} \sim F(p, n-p).$$

On réalise un test unilatéral à droite, les grandes valeurs de $||\widehat{Y}||^2$ plaidant contre l'hypothèse de la nullité de $\beta$ (qui entraine celle de $x\beta$).

Ce test est généralement peu utile en pratique (sauf cas particulier), car on teste la nullité de tous les $\beta$ d'un coup : on veut en général inclure l'*intercept* dans le modèle nul.

#### Interprétation géométrique

La figure A.2 montre une interprétation géométrique de ce que nous venons de voir : $\widehat{Y}$ est la projection orthogonale de Y sur l'espace $I(x)$ engendré par les colonnes de $x$, c'est-à-dire les vecteurs de la forme $xu$ ; c'est le vecteur de cette forme le plus proche de Y. L'orthogonalité de $\widehat{Y}$ et de $\widehat{\varepsilon}$, comme celle de $\widehat{\varepsilon}$ et $\varepsilon - \widehat{\varepsilon}$, sont visibles sur le dessin.

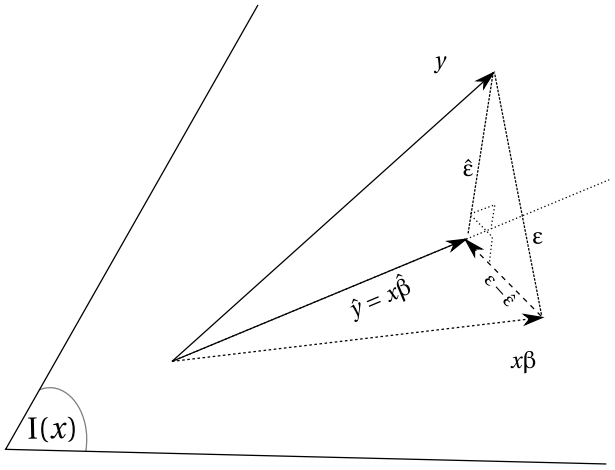



FIGURE A.2. – Interprétation géométrique : projection du vecteur Y sur I($x$)

## A.4. Comparaison de modèles emboîtés

Rappelons ce qu'on a vu au chapitre sur la régression linéaire. On a $p = q + r$ variables $x_1, \ldots, x_p$. On veut tester l'hypothèse nulle suivante :

$$\mathrm{H}_0 : \beta_{q+1} = \cdots = \beta_p = 0$$

On compare les résultats de la régression de Y sur les $q$ premières variables $x_1, \ldots, x_q$ à ceux de la régression de Y sur les $p$ variables $x_1, \ldots, x_p$. On note $\mathrm{X_A}$ la matrice $n \times q$ dont les $q$ colonnes correspondent aux $q$ premières variables, et X la matrice de l'ensemble des variables :

$$\mathrm{X_A} = \begin{pmatrix} x_{11} & \ldots & x_{1q} \\ x_{21} & \ldots & x_{2q} \\ \vdots & & \vdots \\ x_{n1} & \ldots & x_{nq} \end{pmatrix} \qquad \mathrm{X} = \begin{pmatrix} x_{11} & \ldots & x_{1p} \\ x_{21} & \ldots & x_{2p} \\ \vdots & & \vdots \\ x_{n1} & \ldots & x_{np} \end{pmatrix}$$

et on pose $\mathrm{H_A} = \mathrm{X_A}(\mathrm{X_A^t X_A})^{-1}\mathrm{X_A^t}$ et $\mathrm{H} = \mathrm{X}(\mathrm{X}^t\mathrm{X})^{-1}\mathrm{X}^t$, de sorte que

— l'espérance de Y estimée sous $\mathrm{H}_0$ est $\widehat{\mathrm{Y}}_\mathrm{A} = \mathrm{H_A Y}$, et le vecteur des résidus est $\widehat{\varepsilon}_\mathrm{A} = (\mathrm{I} - \mathrm{H_A})\varepsilon$

— l'espérance de Y estimée sous $\mathrm{H}_1$ est $\widehat{\mathrm{Y}} = \mathrm{HY}$, et le vecteur des résidus est $\widehat{\varepsilon} = (\mathrm{I} - \mathrm{H})\varepsilon$

Le test proposé repose sur la statistique

$$\mathrm{F} = \frac{\left(||\mathrm{Y} - \widehat{\mathrm{Y}}_\mathrm{A}||^2 - ||\mathrm{Y} - \widehat{\mathrm{Y}}||^2\right)/r}{||\mathrm{Y} - \widehat{\mathrm{Y}}||^2/(n-p)} \tag{A.1}$$

qui suit, sous $\mathrm{H}_0 : \beta_1 = \cdots = \beta_q = 0$, une loi $\mathrm{F}(r, n-p)$.

#### Interprétation géométrique et preuve

Tout découle une nouvelle fois du théorème de Cochran. Si on suppose que $\mathrm{H}_0$ est vrai, on a $(\varepsilon - \widehat{\varepsilon}_\mathrm{A})$ orthogonal à $\widehat{\varepsilon}_\mathrm{A}$. On a donc

$$\varepsilon = (\varepsilon - \widehat{\varepsilon}_\mathrm{A}) + \widehat{\varepsilon}_\mathrm{A}$$
$$||\varepsilon||^2 = ||\varepsilon - \widehat{\varepsilon}_\mathrm{A}||^2 + ||\widehat{\varepsilon}_\mathrm{A}||^2.$$

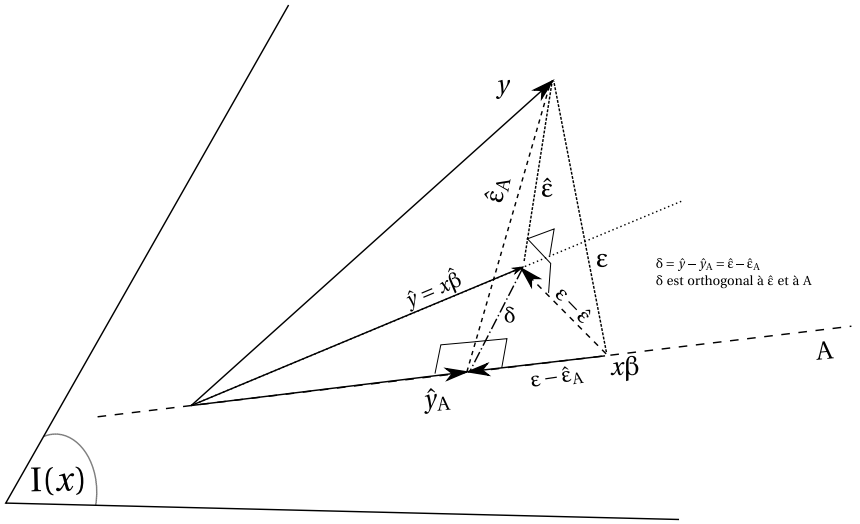



FIGURE A.3. – Interprétation géométrique : projection du vecteur Y sur I(X) et sur A

D'autre part, $\widehat{\varepsilon}$ est orthogonal à toutes les colonnes de X, donc à $X\widehat{\beta} = \widehat{Y}$ et à $X\widehat{\beta}_A = \widehat{Y}_A$. On en déduit que $\widehat{\varepsilon}_A - \widehat{\varepsilon} = \widehat{Y}_A - \widehat{Y}$ est orthogonal à $\widehat{\varepsilon}$. On a donc

$$\widehat{\varepsilon}_A = (\widehat{\varepsilon}_A - \widehat{\varepsilon}) + \widehat{\varepsilon}$$
$$||\widehat{\varepsilon}_A||^2 = ||\widehat{\varepsilon}_A - \widehat{\varepsilon}||^2 + ||\widehat{\varepsilon}||^2.$$

On a donc décomposé $\varepsilon$ en trois vecteurs orthogonaux deux à deux $(\varepsilon - \widehat{\varepsilon}_A)$, $(\widehat{\varepsilon}_A - \widehat{\varepsilon})$ et $\widehat{\varepsilon}$ :

$$\varepsilon = (\varepsilon - \widehat{\varepsilon}_A) + (\widehat{\varepsilon}_A - \widehat{\varepsilon}) + \widehat{\varepsilon}$$
$$||\varepsilon||^2 = ||\varepsilon - \widehat{\varepsilon}_A||^2 + ||\widehat{\varepsilon}_A - \widehat{\varepsilon}||^2 + ||\widehat{\varepsilon}||^2,$$

et $(\varepsilon - \widehat{\varepsilon}_A)$ a $q$ degrés de libertés, $(\widehat{\varepsilon}_A - \widehat{\varepsilon})$ a $r$ degrés de libertés, et $\widehat{\varepsilon}$ en a $n - p$ ; on en déduit (théorème de Cochran)

— $||\varepsilon - \widehat{\varepsilon}_A||^2 \sim \sigma^2\chi^2(q)$,
— $||\widehat{\varepsilon}_A - \widehat{\varepsilon}||^2 \sim \sigma^2\chi^2(r)$,
— $||\widehat{\varepsilon}||^2 \sim \sigma^2\chi^2(n-p)$,

et ces trois variables aléatoires sont indépendantes.

On a d'autre part $\widehat{\varepsilon} = Y - \widehat{Y}$ et $\widehat{\varepsilon}_A - \widehat{\varepsilon} = \widehat{Y}_A - \widehat{Y}$, d'où

$$||Y - \widehat{Y}_A||^2 - ||Y - \widehat{Y}||^2 = ||\widehat{\varepsilon}||^2 - ||\widehat{\varepsilon}_A||^2 = ||\widehat{\varepsilon}_A - \widehat{\varepsilon}||^2$$

et la statistique F s'écrit finalement

$$F = \frac{||\widehat{\varepsilon}_A - \widehat{\varepsilon}||^2 / r}{||\widehat{\varepsilon}||^2 / (n-p)}$$

et c'est bien une variable de loi $F(r, n-p)$.