Cours accéléré Statistiques
Partie 2: Modèle linéaire
Master 2 Mathématiques et Applications

Christine Keribin

[1]Laboratoire de Mathématiques d'Orsay
Université Paris-Saclay

2024-2025

université PARIS-SACLAY

FACULTÉ DES SCIENCES D'ORSAY

Sommaire

Introduction

- Expliquer ou prédire une variable aléatoire réponse Y en fonction d'une liste de variables explicatives $\mathbf{X} = (X_1, \ldots, X_p)$ observées sur des individus

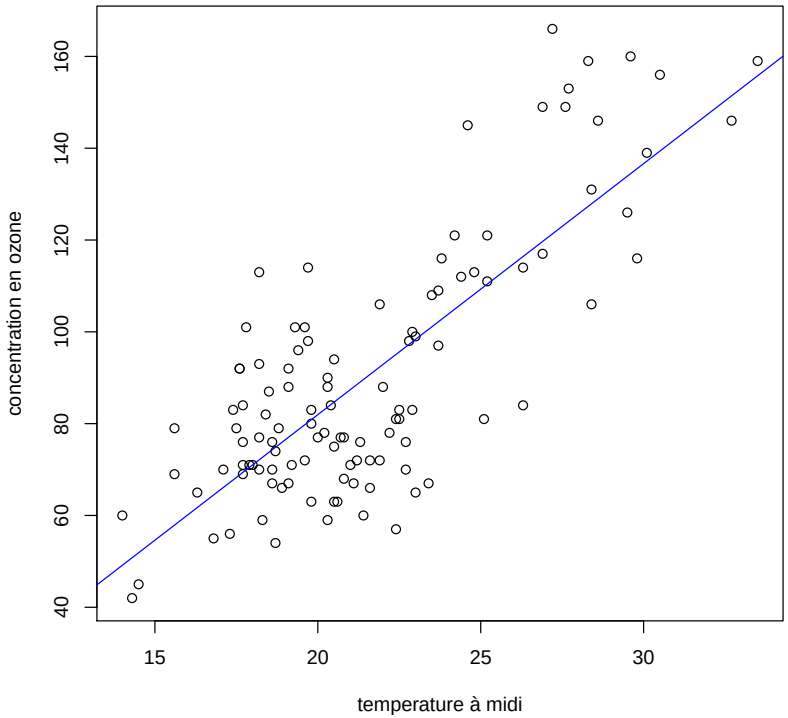
$$Y = f(\mathbf{X})$$

- Modéliser Y par une fonction des variables explicatives $\mathbf{X}$, par exemple, modèle à bruit additif

$$Y_{|\mathbf{X}=x} = m(x) + \varepsilon$$

$m(x) = \mathbb{E}(Y|\mathbf{X} = x)$ est la fonction de régression
  - choix est guidé par la connaissance d'un phénomène physique, biologique,...
  - par l'observation
  - ou non...
- apprentissage supervisé : à partir d'un échantillon étiqueté $((\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n))$, estimer $m(x)$, prédire Y pour une nouvelle condition x ... avec le moins d'erreur

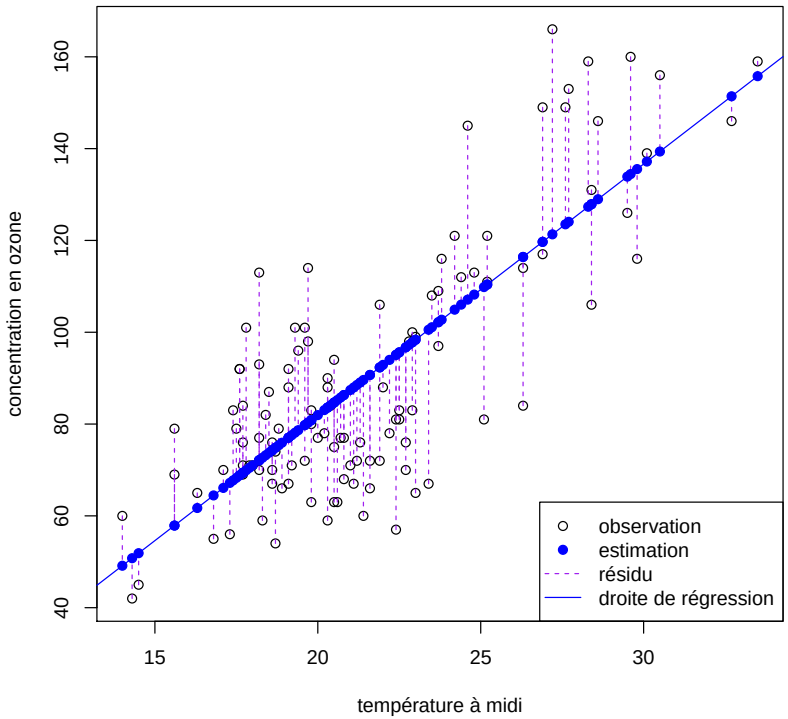
Exemple : régression linéaire (simple) [1]



$$Y_i = \underbrace{\theta_1 + \theta_2 x_i}_{\text{déterministe}} + \underbrace{\varepsilon_i}_{\text{bruit aléatoire}}, \quad \mathbb{E}_{x_i}(\varepsilon_i) = 0$$

1. *http ://www.agrocampus-ouest.fr/math/livreR/*

Exemple : régression linéaire (simple)

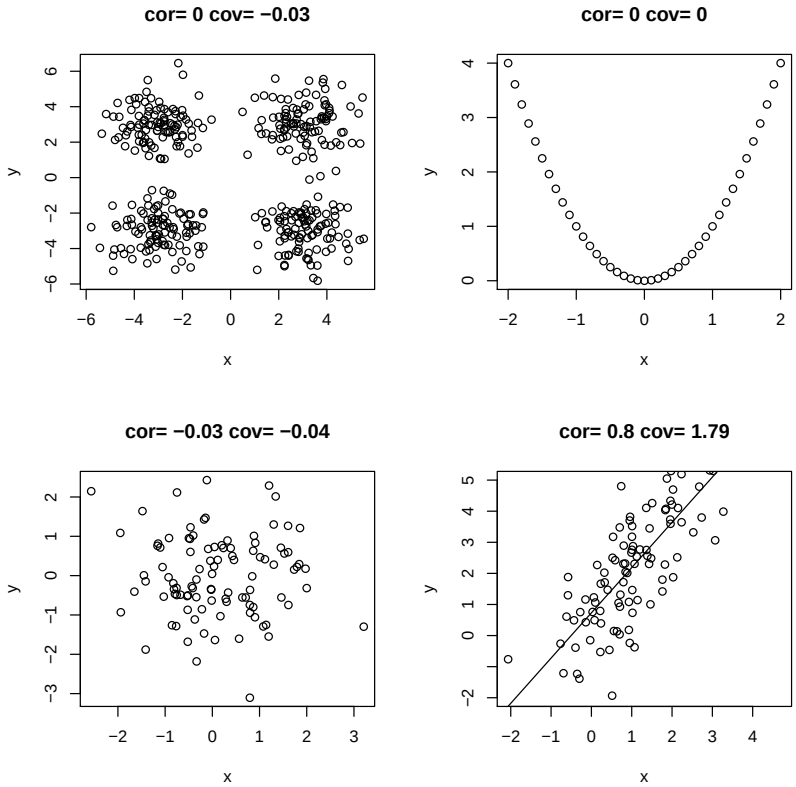


$\mathbb{E}(Y_i|\mathrm{x}_i) = \theta_1 + \theta_2 x_i$ est l'équation de la droite de régression [2] : modélise le comportement moyen pour chaque condition d'expérience $\hookrightarrow$ estimer $\theta = (\theta_1, \theta_2)$

2. "Régression" : Sir Galton (1822-1911)

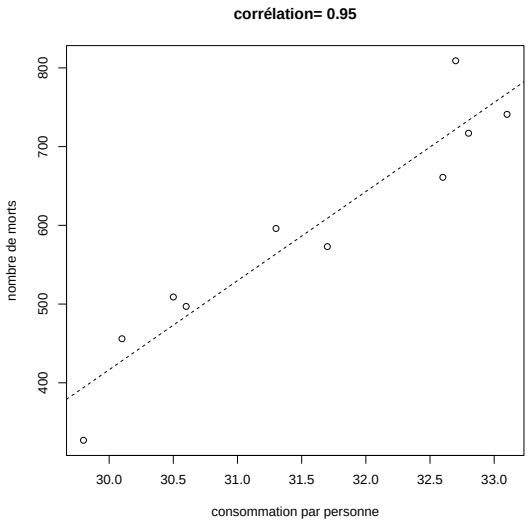
Que cherche-t-on à détecter ?

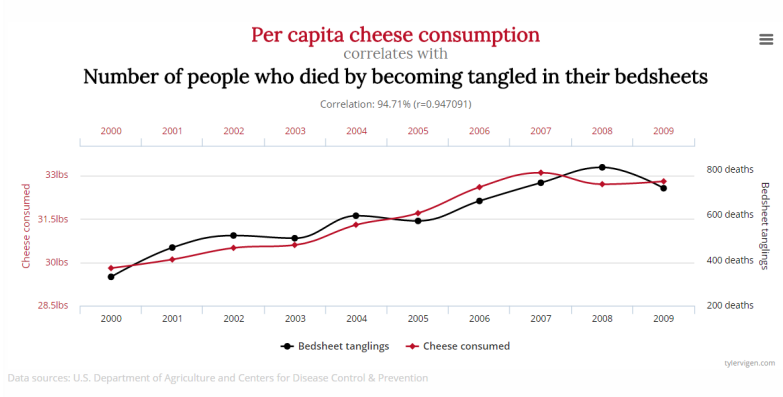
La corrélation entre Y et x... qui n'est que l'expression d'une liaison linéaire



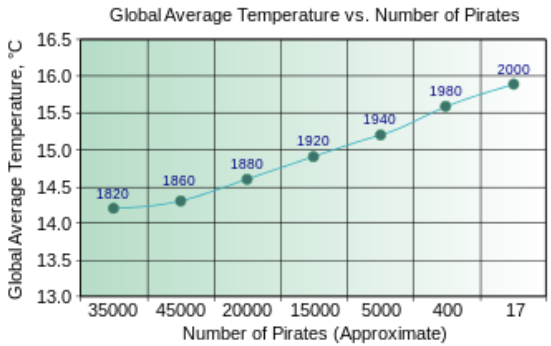
Une corrélation nulle n'indique pas forcément une absence de liaison

Mais attention : corrélation $\neq$ causalité





source : http ://tylervigen.com/spurious-correlations



source : Wikipédia

Apprentissage supervisé

- Données d'apprentissage :
  $\mathcal{D} = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\} \quad (\text{i.i.d.} \sim \mathbf{P})$
- Prédicteur : $f : \mathcal{X} \to \mathcal{Y}$ mesurable
- Fonction de perte ou de coût : $\ell(f(\mathbf{X}), Y)$ mesure avec quelle qualité $f(\mathbf{X})$ "prédit" Y, d'où le risque :

$$\mathcal{R}(f) = \mathbb{E}\ell(Y, f(\mathbf{X}))$$

$\hookrightarrow$ souvent $\ell(f(\mathbf{X}), Y) = \|f(\mathbf{X}) - Y\|^2$ ou $\ell(f(\mathbf{X}), Y) = \mathbf{1}_{Y \neq f(\mathbf{X})}$

But : apprendre une règle pour construire un classifieur / régresseur $\widehat{f} \in \mathcal{F}$ à partir des données d'apprentissage $\mathcal{D}_n$ avec un risque $\mathcal{R}(\widehat{f})$ petit en moyenne.

Fonction de régression

Approcher Y par une fonction $h(X)$ de la variable X.
Risque quadratique :

$$R(h) = \mathbb{E}[(Y - h(X))^2] = \int (y - h(x))^2 dP(x, y).$$

Théorème

La fonction qui minimise le risque quadratique $R(.)$ est $m(X) = \mathbb{E}(Y|X)$, l'espérance de Y conditionnellement à X

$$\forall x, m(x) = \mathbb{E}(Y|X = x).$$

Elle est appelée fonction de régression.

- PB : on ne sait pas la calculer dans le cas général
- Si X est déterministe, on travaille également à $X = x$ fixé... et les développements sont identiques.

$\hookrightarrow$ poser des hypothèses de modélisation.

Bibliographie

J. Pagès. *Statistique générales pour utilisateurs*. Presses Universitaires de Rennes, Rennes, 2005.

J.-M. Azais et J.-M. Bardet. *Le modèle linéaire par l'exemple*. Dunod, Paris, 2005.

P.-A. Cornillon et E. Matzner-Løber. *Régression Théorie et applications*. Springer, Paris, 2007.

P.-A. Cornillon et al. *Statistiques avec R*. Presses Universitaires de Rennes, Rennes, 2008.

Sommaire

Régression linéaire : Y est une variable quantitative

à partir de n observations $(\mathrm{x}_1, Y_1), \ldots, (\mathrm{x}_n, Y_n)$

1. Modèle à bruit additif posant une relation de linéarité de l'espérance $\mathbb{E}(Y_i|\mathrm{x}_i)$ en un paramètre θ :

$$\begin{aligned} Y_i &= x_{i1}\theta_1 + \ldots + x_{ip}\theta_p + \varepsilon_i \\ &= \mathrm{x}_i\theta + \varepsilon_i, \end{aligned}$$

2. Les résidus sont centrés : $\mathbb{E}(\varepsilon_i|\mathrm{x}_i) = 0$.
3. La variance de l'erreur est constante : $\mathrm{var}(\varepsilon_i|\mathrm{x}_i) = \sigma^2$.
4. Les résidus sont décorrélés : $\mathrm{cov}(\varepsilon_i, \varepsilon_j|\mathrm{x}_i, \mathrm{x}_j) = 0$ pour $i \neq j$.
5. Eventuellement : ε_i est gaussien

$\hookrightarrow$ x_i est la valeur fixée des variables explicatives pour la i-ème observation. Elle n'est pas aléatoire

$\hookrightarrow$ $m(\mathrm{x}_i) = \mathrm{x}_i\theta$ est appelée fonction de régression

Modèle linéaire gaussien

- Matriciellement : $Y = X\theta + \varepsilon$

$$Y_{n\times 1} = X_{n\times p}\theta_{p\times 1} + \varepsilon_{n\times 1};$$

$$\mathbb{E}(\varepsilon_{n\times 1}) = 0; \quad \text{var}(\varepsilon_{n\times 1}) = \sigma^2 Id_n,$$

X est la matrice du plan d'expérience, concaténation des n vecteurs lignes x_i ou des p variables colonnes $X_j = (x_{1j}, \ldots, x_{nj})'$.
Remarque : En général, $\forall i, x_{i1} = 1$, X_1 est l'intercept.

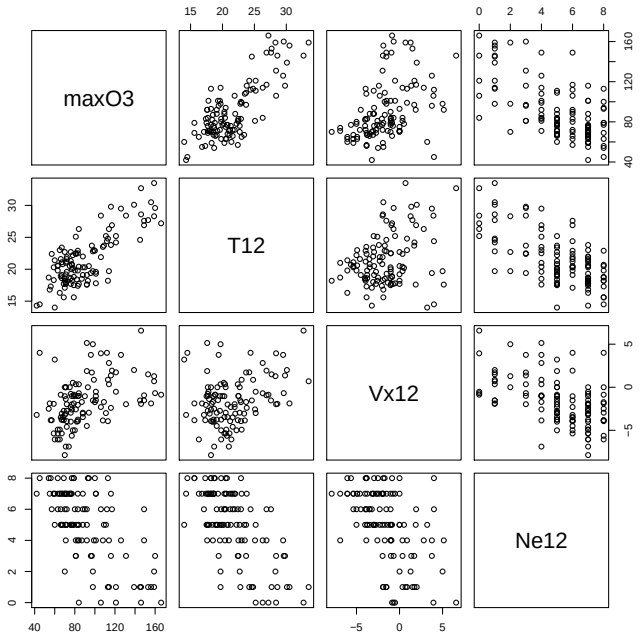
- Modèle linéaire gaussien

$$Y = m + \varepsilon; \quad \varepsilon \sim \mathcal{N}(0, \Sigma); \quad (\Sigma = \sigma^2 Id_n)$$

où $m \in V$, sous espace vectoriel de $\mathbb{R}^n$ de dimension p.

Régression multiple

- les covariables X sont quantitatives
- si $p = 2$ dont un intercept : régression simple
  $m(\mathrm{x}_i) = \theta_1 + \theta_2 x_i$; $\mathrm{x}_i = (1 \;\; x_i)$; $\theta = (\theta_1, \theta_2)'$
- si $p > 2$, régression multiple



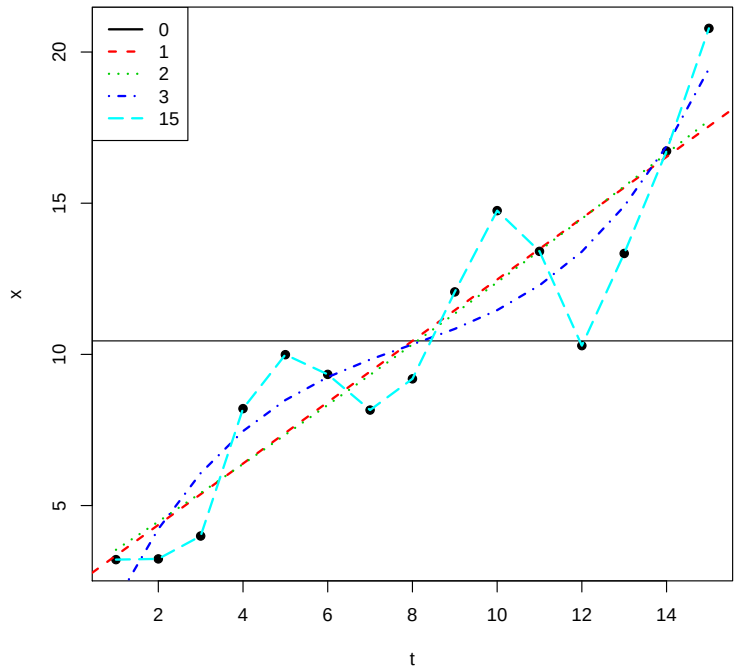
$$\texttt{maxO3} = \theta_1 + \texttt{T12}\ \theta_2 + \texttt{Vx12}\ \theta_3 + \texttt{Ne12}\ \theta_4 + \varepsilon$$

Régression polynomiale

Un cas particulier de régression multiple : régression polynômiale

$$Y_t = \theta_1 + t\theta_2 + t^2\theta_3 + t^3\theta_4 + \varepsilon_t$$

$$X = \begin{pmatrix} 1 & t_1 & \cdots & t_1^{p-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_n & \cdots & t_n^{p-1} \end{pmatrix}$$



Un régression linéaire (en θ) peut permettre de représenter un phénomène non-linéaire (en t)

Identifiabilité

Définition : deux paramètres différents définissent deux lois différentes

Théorème (CNS d'identifiabilité)

La paramétrisation $\mathbb{E}(Y) = X\theta$ est identifiable si et seulement si l'une des propriétés équivalentes suivantes est vérifiée :

- *les colonnes de X sont indépendantes,*
- *X est de rang plein,*
- *X est injective,*
- *$Ker(X) = \{0\}$.*

Définition

On appelle dimension du modèle la dimension de $Im(X)$.
On a : $dim(Im(X)) \leq dim(\theta)$

$\hookrightarrow$ On suppose dans la suite que le modèle est identifiable ou régulier.

Sommaire

Estimateur des Moindres Carrés (EMC)

Recherche un estimateur de θ sous forme d'un minimiseur de la somme des carrés résiduels $SCR(\theta) = ||Y - X\theta||^2$:

$$\widehat{\theta} = \arg\min_{\theta \in \Theta} ||Y - X\theta||^2.$$

On a : $\widehat{\theta} = (X'X)^{-1}X'Y$

Proposition

$\widehat{Y} := \widehat{\mathbb{E}(Y)} = X\widehat{\theta} = H_X Y$ *est la projection orthogonale de* Y *sur* $Im(X)$ *:*

$$H_X = X(X'X)^{-1}X'$$

- H_X : Hat matrix
- $\widehat{Y}$: valeurs ajustées
- $\widehat{\varepsilon} = Y - \widehat{Y}$: résidus (estimés)
- la droite de régression $y = x\widehat{\theta}$ passe par le point moyen $(\bar{Y}, \bar{X}^{(1)}, \ldots, \bar{X}^{(p)})$

Exemple : régression linéaire (simple)

Le modèle s'écrit $Y = \theta_1 \mathbb{1} + \theta_2 X_2 + \varepsilon$, où $\mathbb{1}$ est l'intercept, et $X_2 = (x_1, \ldots, x_n)'$

$$X'X = \begin{pmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix}$$

$$(X'X)^{-1} = \frac{1}{n\sum_i x_i^2 - \sum_i x_i \sum_i x_i} \begin{pmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{pmatrix}$$

soit

$$\widehat{\theta} = (X'X)^{-1}X'Y = \begin{pmatrix} \widehat{\theta}_1 = \bar{Y} - \bar{x}\widehat{\theta}_2 \\ \widehat{\theta}_2 = \frac{(\frac{1}{n}\sum_i x_i Y_i) - \bar{x}\bar{Y}}{(\frac{1}{n}\sum_i x_i^2) - \bar{x}^2} \end{pmatrix}$$

La droite de régression estimée $\widehat{y} = \widehat{\theta}_1 + \widehat{\theta}_2 x$ passe par le point moyen $(\overline{x}, \overline{Y})$.

Propriétés ne nécessitant pas l'hyp. gaussienne

Proposition

L'EMC $\widehat{\theta}$ de θ est

- *sans biais* : $\mathbb{E}(\widehat{\theta}) = \theta$
- *de variance* $\text{var}(\widehat{\theta}) = \sigma^2(X'X)^{-1}$.

De plus, $\widehat{\theta}$ vérifie la propriété de Gauss-Markov :

Théorème (Gauss-Markov)

Parmi les estimateurs linéaires et sans biais de θ, l'EMCO $\widehat{\theta}$ est de variance minimum.

Estimateurs de la variance σ^2 :

- $\widehat{s}^2 = SCR(\widehat{\theta})/n$ est biaisé à distance finie, et asymptotiquement sans biais
- $\widehat{\sigma}^2 = SCR(\widehat{\theta})/(n-p)$ est sans biais

Estimateur du Maximum de Vraisemblance (EMV)

Si la loi de ε est connue (gaussienne), la vraisemblance de l'échantillon est

$$L_n(\theta, \sigma^2; Y) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\left(-\frac{1}{2\sigma^2} \underbrace{||Y - X\theta||^2}_{SCR(\theta)} \right).$$

L'EMV $\widehat{\beta} = (\widehat{\theta}, \widehat{s^2})$ est

$$\widehat{\beta} = \arg \max_{\beta \in \mathbb{R}^p \times \mathbb{R}^{+*}} \log L_n(\beta; Y).$$

D'où

$$\widehat{\theta} = (X'X)^{-1}X'Y.$$

... identique à l'EMC. L'EMV de σ^2 est

$$\widehat{s^2} = \frac{1}{n}||Y - X\widehat{\theta}||^2 = \frac{1}{n}SCR(\widehat{\theta})$$

Sommaire

Cas gaussien

Théorème

L'EMV $\widehat{\beta} = (\widehat{\theta}, \widehat{s}^2)$ du modèle linéaire gaussien

$$Y = X\theta + \varepsilon, \quad \varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_n), \quad \theta \in \mathbb{R}^p, \quad \sigma \in \mathbb{R}^{+*},$$

supposé régulier, suit la loi suivante :

$$\widehat{\theta} \sim \mathcal{N}\left(\theta, \sigma^2 (X'X)^{-1}\right); \quad \widehat{s}^2 = \frac{||Y - X\widehat{\theta}||^2}{n} \sim \frac{\sigma^2}{n}\chi^2(n-p).$$

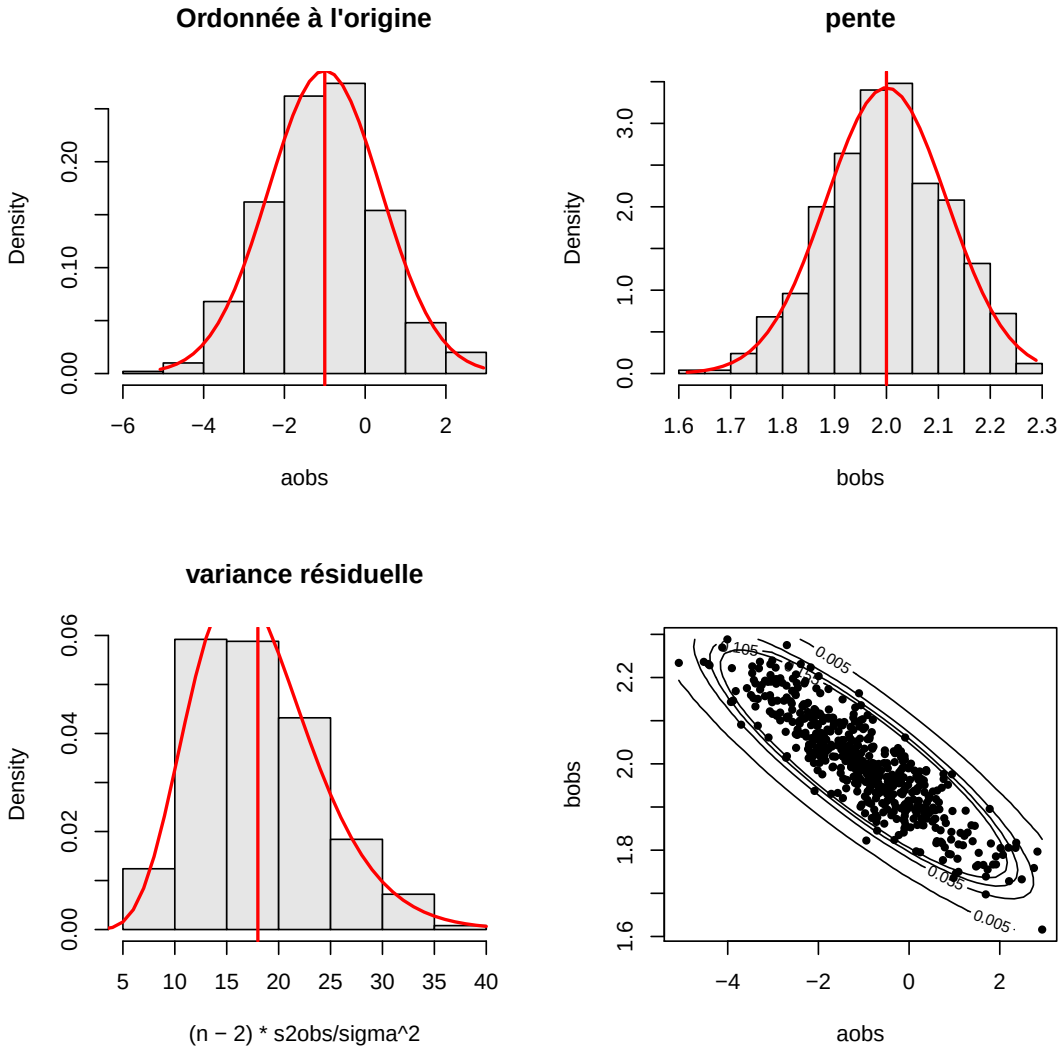
De plus, $\widehat{\theta}$ et $\widehat{s}^2$ sont indépendants

Csq : Soit A une matrice $q \times p$ de rang $q \leq p$

$$A\widehat{\theta} \sim \mathcal{N}_q(A\theta, \sigma^2 A(X'X)^{-1}A')$$

Illustration simulée

500 échantillons (Y, X) de taille $n = 30$, avec $x = 1 : 30$ et $m(x) = -1 + 2x$



Cas gaussien

Quand ε est gaussien, on utilise plutôt l'estimateur non biaisé de la variance $\widehat{\sigma}^2 = ||Y - X\widehat{\theta}||^2/(n-p)$

Résultats

- $\widehat{\theta}_{MV} = \widehat{\theta}_{MC}$ et $\widehat{\sigma}^2$ sont indépendants
- $\widehat{\theta} \sim \mathcal{N}(\theta, \sigma^2(X'X)^{-1})$, $(n-p)\widehat{\sigma}^2/\sigma^2 \sim \chi^2(n-p)$
- pour toute matrice L non nulle de dimension $(1 \times p)$: $T_n = (L(X'X)^{-1}L')^{-1/2}(L\widehat{\theta} - L\theta)/\widehat{\sigma} \sim \mathcal{T}(n-p)$
- pour toute matrice A de dimension $r \times p$ et de rang $r \leq p$: $F_n = (A(\widehat{\theta} - \theta))'[A(X'X)^{-1}A']^{-1}A(\widehat{\theta} - \theta)/(r\widehat{\sigma}^2) \sim \mathcal{F}(r, n-p)$
- si $r = 1$, $F_n = T_n^2$

Intervalle de confiance d'une espérance

Théorème

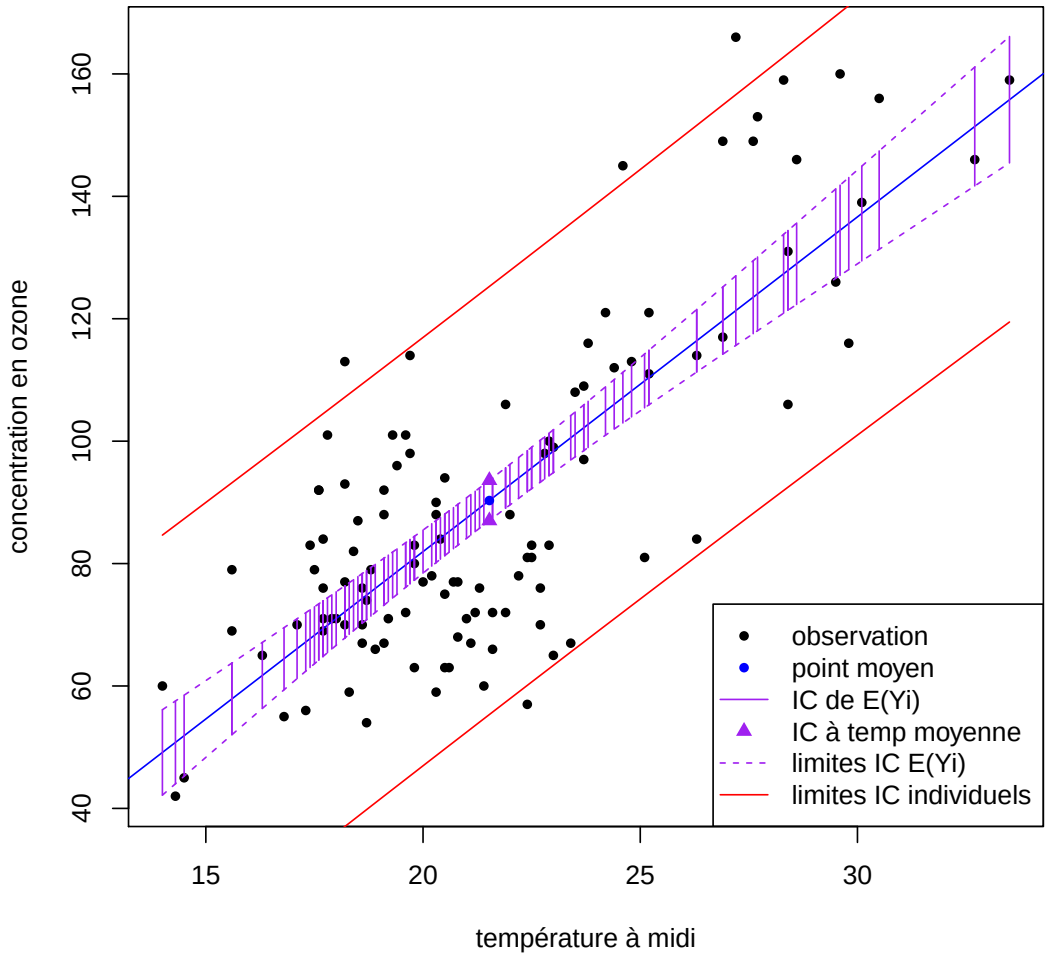
Soit L une matrice de dimension $1 \times p$. Un intervalle de confiance de niveau $1-\alpha$ d'une forme linéaire de $L\theta$ est donné par

$$\left[L\widehat{\theta} \pm q_{1-\alpha/2}^{\mathcal{T}_{n-p}} \widehat{\sigma} \sqrt{L(X'X)^{-1}L'} \right],$$

où $q_{1-\alpha/2}^{\mathcal{T}_{n-p}}$ est le quantile d'ordre $1-\alpha/2$ de la loi de Student à $n-p$ degrés de liberté.

Applications : IC d'une composante de θ, IC de $\mathbb{E}(Y|x_0)$

IC d'une valeur moyenne ou d'une prédiction individuelle



Intervalle de confiance d'une prévision

A partir de n observations ($\hookrightarrow \widehat{\theta}$), prédire une nouvelle observation individuelle indépendante sous la condition d'expérience $\mathrm{x}_{n+1} = (x_{n+1,1}, \ldots, x_{n+1,p})$

- modèle : $Y_{n+1} = \mathrm{x}_{n+1}\theta + \varepsilon_{n+1}$,
- observation prédite : $\widehat{Y}^p_{n+1} = \mathrm{x}_{n+1}\widehat{\theta}$
- erreur de prévision

$$\widehat{\varepsilon}^p_{n+1} = Y_{n+1} - \widehat{Y}^p_{n+1},$$

$\mathbb{E}(\widehat{\varepsilon}^p_{n+1}) = 0$, $\mathrm{var}(\widehat{\varepsilon}^p_{n+1}) = \sigma^2(\mathrm{x}_{n+1}(X'X)^{-1}\mathrm{x}'_{n+1} + 1)$.

- IC de prévision de niveau $1-\alpha$:

$$\left[\mathrm{x}_{n+1}\widehat{\theta} \pm q^{\mathcal{T}_{n-p}}_{1-\alpha/2}\widehat{\sigma}\sqrt{\mathrm{x}_{n+1}(X'X)^{-1}\mathrm{x}'_{n+1} + 1}\right].$$

Région de confiance simultanée (Bonferroni)

Soit θ_1 et θ_2 deux composantes de θ.
On cherche une région de confiance de la forme
$IC(\theta_1) \times IC(\theta_2)$

$$\mathbb{P}\left((\theta_1, \theta_2) \notin RC_\alpha(\theta_1, \theta_2)\right)$$

$$\leq \mathbb{P}(\theta_1 \notin IC_{1-\alpha/2}(\theta_1)) + \mathbb{P}(\theta_2 \notin IC_{1-\alpha/2}(\theta_2)) \leq \alpha.$$

K intervalles de confiance de Bonferroni de risque α/K forment une région de confiance de risque simultané α.

Région de confiance simultanée (autre méthode)

Théorème

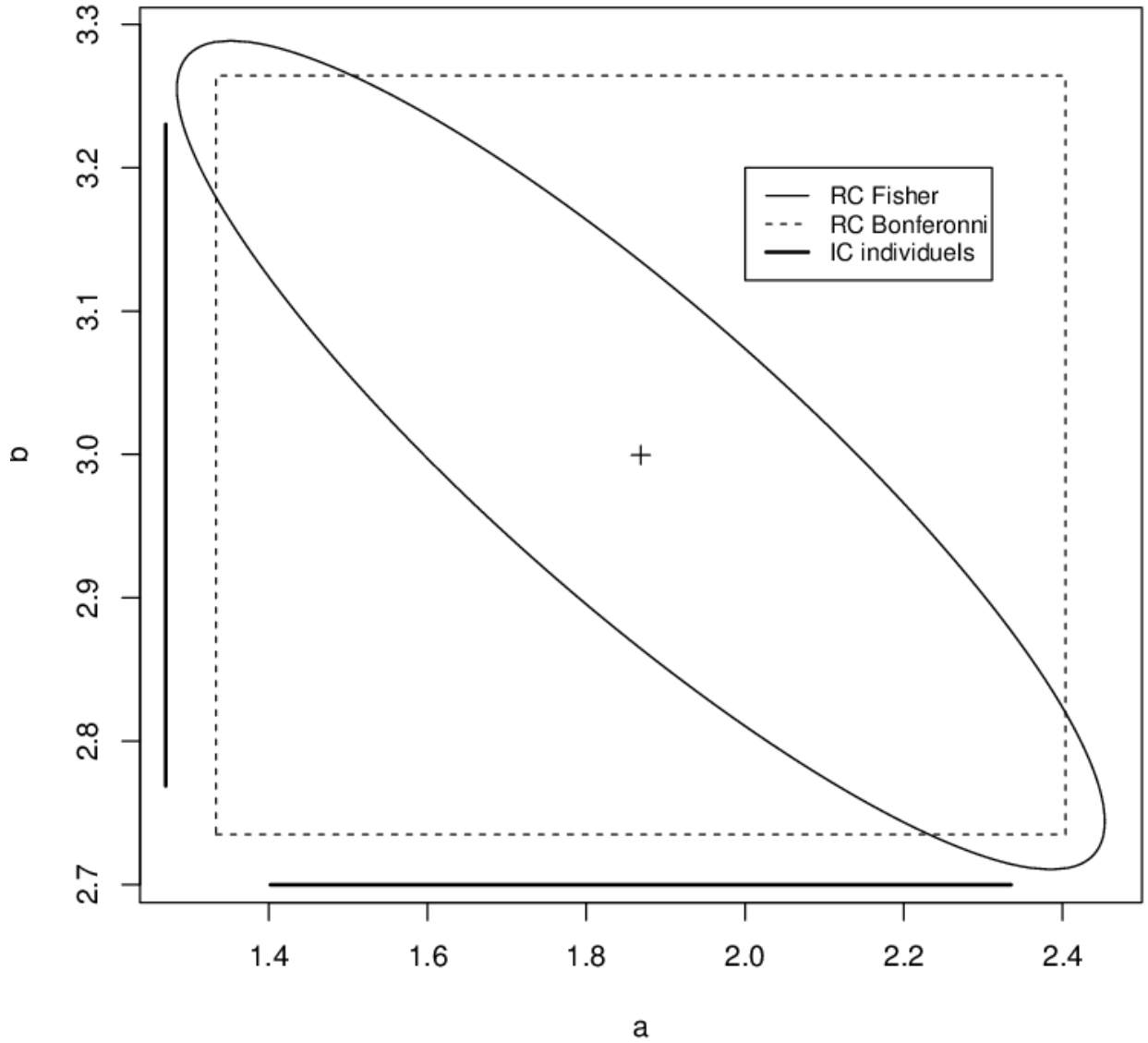
Soit A une matrice de dimension $r \times p$ et de rang r. Une région de confiance de niveau $1-\alpha$ de $A\theta$ est donnée par

$$RC_\alpha(A\theta) =$$

$$\{u \in \mathbb{R}^p, \frac{1}{r\widehat{\sigma}^2}(A\widehat{\theta} - Au)'[A(X'X)^{-1}A']^{-1}(A\widehat{\theta} - Au) \leq f_{r,n-p,1-\alpha}\}$$

où $f_{r,n-p,1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi de $\mathcal{F}(r, n-p)$.

Comparaison de régions de confiance



Sommaire

Test de Student

Test d'une relation affine des composantes du paramètre $L\theta = c$, avec L matrice ligne de dimension p.

- test bilatéral : $H_0 : L\theta = c$ contre $H_1 : L\theta \neq c$
- statistique de test

$$T = \frac{L\widehat{\theta} - c}{\widehat{\sigma}\sqrt{L(X'X)^{-1}L'}} \sim_{H_0} \mathcal{T}(n-p),$$

- région de rejet du test bilatère de risque α

$$\mathcal{R}_\alpha = \{|T| > t_{n-p,1-\alpha/2}\}.$$

- utilisé pour tester la significativité (non-nullité) d'une composante de θ

Remarque : test unilatéral : $H_0 : L\theta = c$ contre $H_1 : L\theta < c$ de région de rejet à gauche :

$$\mathcal{R}_\alpha = \{T < t_{n-p,\alpha}\}.$$

Exemple : $Y = \mu + \beta_1 V1 + \beta_2 V2 + \beta_3 V3 + \varepsilon$

```
> out=lm(Y~V1+V2+V3, data=df); summary(out)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -20.3129    51.8246  -0.392    0.701        H0       H1
V1           -3.2241     3.6100  -0.893    0.386      ~V2+V3 vs ~V1+V2+V3
V2            0.2334     0.2768   0.843    0.412      ~V1+V3 vs ~V1+V2+V3
V3           20.3895     1.2662  16.103  7.1e-11 ***  ~V1+V2 vs ~V1+V2+V3

Residual standard error: 4.163 on 15 degrees of freedom
F-statistic: 584.6 on 3 and 15 DF,  p-value: 9.386e-16
```

Test de Wald avec loi exacte

Modèle emboîté $\omega \subset \Omega$ défini par des hypothèses affines : $A\theta = c$ contre $A\theta \neq c$, A matrice $r \times p$ de rang r, $dim(\omega) = p - r = q$

$$W_{exact} = W/r = W = \frac{1}{r\widehat{\sigma}^2}(A\widehat{\theta}-c)'[A(X'X)^{-1}A']^{-1}(A\widehat{\theta}-c) \sim_{(H_0)} \mathcal{F}(r, n-p)$$

- tester la significativité globale de la régression : tous les coeff. sont nuls sauf celui de l'intercept : $(H_0) : A\theta = 0$ contre $(H_0) : A\theta \neq 0$

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

```
F-statistic: 584.6 on 3 and 15 DF,  p-value: 9.386e-16
```

- tester des modèles emboîtés.

Petite interprétation géométrique

Soit ω un sev de $\Omega \in \mathbb{R}^n$, $dim(\omega) = q < dim(\Omega) = p$.

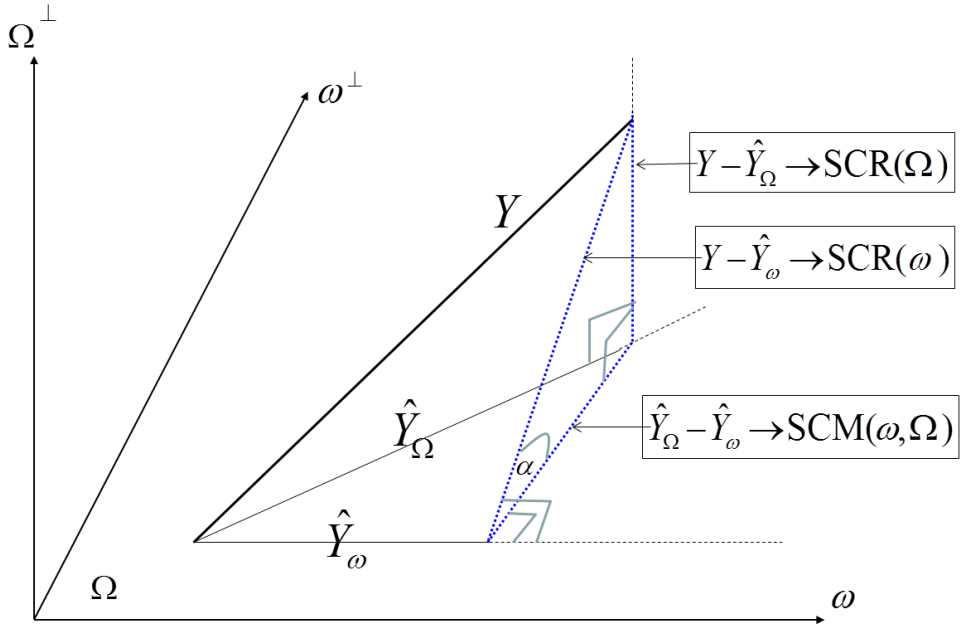
On note $\widehat{Y}_\Omega = H_\Omega(Y)$. On a :

$$||Y - \widehat{Y}_\omega||^2 = ||\widehat{Y}_\Omega - \widehat{Y}_\omega||^2 + ||Y - \widehat{Y}_\Omega||^2$$

$$\text{SCR}(\omega) = \text{SCM}(\omega,\Omega) + \text{SCR}(\Omega).$$

$$F = \frac{(\text{SCR}(\omega) - \text{SCR}(\Omega))/(p-q)}{\text{SCR}(\Omega)/(n-p)} = \frac{\text{SCM}(\omega,\Omega)/(p-q)}{\text{SCR}(\Omega)/(n-p)}$$

$$\sim_\omega \mathcal{F}(p-q, n-p).$$



Test de Fisher

Théorème (Test de Fisher d'un sous-modèle linéaire)

Soit le modèle linéaire gaussien $Y = m + \varepsilon, m \in \Omega$, $\varepsilon \sim \mathcal{N}(0, I_n)$, *de dimension* $dim(\Omega) < n$. *Soit* ω *un sev de* Ω *tel que* $dim(\omega) = q < p$. *Le test de Fisher*

$$H_0 : m \in \omega \text{ contre } H_1 : m \in \Omega \backslash \omega.$$

de région de rejet

$$\mathcal{R}_\alpha = \{F > f_{p-q,n-p,1-\alpha}\},$$

où $f_{p-q,n-p,1-\alpha}$ *est le quantile de la loi de Fisher* $\mathcal{F}(p-q, n-p)$, *est de niveau* α.

Application : tableau d'analyse de la variance

$(H_0) : \omega$ contre $(H_1) : \Omega \setminus \omega$ tq. $\omega \subset \Omega$, $dim(\omega) = q$, $dim(\Omega) = p$

Source	Res.Df	RSS	Df	Sum of Sq	F	Prob>F
ω	$n-q$	$SCR(\omega)$				
Ω	$n-p$	$SCR(\Omega)$	$p-q$	$SCM(\omega,\Omega)$	$f_{obs} = \frac{SCM/(p-q)}{SCR(\Omega)/(n-p)}$	p-value

```
> anova(lm(Y~1,data=df), lm(Y~.,data=df))
Model 1: Y ~ 1
Model 2: Y ~ V1+V2+V3   #équivalent à 1+V1+V2+V3

  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1     18 30654
2     15   260  3     30394 584.55 9.386e-16 ***

#F-statistic: 584.6 on 3 and 15 DF,  p-value: 9.386e-16
```

Tableau d'analyse de la variance

Une autre utilisation de la fonction anova

```
> anova(lm(Y~.,data=df))
Response: Y
          Df  Sum Sq Mean Sq F value    Pr(>F)
V1         1 18252.3 18252.3 1053.11 2.603e-15 ***
V2         1  7647.5  7647.5  441.24 1.548e-12 ***
V3         1  4494.2  4494.2  259.31 7.101e-11 ***
Residuals 15   260.0    17.3
```

(H_0)	(H_1)
iid: Y~1	Y~V1
Y~V1	Y~V1+V2
Y~V1+V2	Y~V1+V2+V3

Exemple (suite)

```
> anova(lm(Y~.,data=df))
Response: Y
          Df  Sum Sq Mean Sq F value    Pr(>F)
V1         1 18252.3 18252.3 1053.11 2.603e-15 ***
V2         1  7647.5  7647.5  441.24 1.548e-12 ***
V3         1  4494.2  4494.2  259.31 7.101e-11 ***
Residuals 15   260.0    17.3
```

```
> anova(lm(Y~V3+V2+V1,data=df))
Response: y
          Df  Sum Sq Mean Sq   F value Pr(>F)
V3         1 30377.9 30377.9 1752.7292 <2e-16 ***
V2         1     2.3     2.3    0.1349 0.7186
V1         1    13.8    13.8    0.7976 0.3859
Residuals 15   260.0    17.3
```

Attention à l'ordre des termes quand les variables explicatives ne sont pas orthogonales

Sommaire

Validation

- L'étape de validation fait partie intégrante de la démarche statistique
- Elle peut remettre en cause le modèle initialement choisi, donner des idées de modifications pour repartir pour un tour...
- Outils : tests, critères, étude graphique
  $\hookrightarrow$ étude des résidus

Critère du R^2

Si ω est le modèle iid (constant) :

- Coefficient de détermination

$$R^2 = \frac{SCM}{SCT} = \frac{||\widehat{Y} - \bar{y}\mathbb{1}||^2}{||Y - \bar{y}\mathbb{1}||^2} = \cos^2 \alpha$$

  Part de la variabilité expliquée par le modèle sur la variabilité totale

- Indication de la qualité d'ajustement, mais pas forcément de la qualité de prévision
- Pour compte de la dimension de $Im(X)$: $\hookrightarrow$ R^2 ajusté :

$$R^2_a = 1 - \frac{n-1}{n-p}\frac{||\widehat{\epsilon}||^2}{||Y - \bar{y}\mathbb{1}||^2}$$

```
Multiple R-squared:  0.9915,Adjusted R-squared:  0.9898
```

Tests utiles

- Significativité globale : test de Fisher du modèle iid contre le modèle d'étude.
- Significativité d'un régresseur : $\hookrightarrow$ des stratégies de sélection de variables.
- Adéquation : comparer le modèle d'étude Ω à un surmodèle Ω_S peu contestable qui contient Ω
- Hypothèse gaussienne : test de Kolmogorov-Smirnov ou test de Shapiro-Wilks.

Propriété des résidus

Proposition

Dans le modèle linéaire gaussien, les résidus $\widehat{\varepsilon} = Y - \widehat{Y}$ *possèdent les propriétés suivantes :*

- *centrés :* $\mathbb{E}(\widehat{\varepsilon}) = 0$.
- *hétéroscédastiques :* $\text{var}(\widehat{\varepsilon}) = \sigma^2(I - H)$.
- *décorrélés avec les valeurs estimées :* $\text{cov}(\widehat{Y}, \widehat{\varepsilon}) = 0$.
- *Si* $\mathbb{1} \in Im(X)$, *les résidus estimés sont linéairement dépendants*.

Différents résidus

- Les résidus normalisés éliminent l'hétéroscédasticité

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

- résidus standardisés, parfois également appelés studentisés : de variance 1

$$t_i = \frac{\widehat{\varepsilon}_i}{\widehat{\sigma}\sqrt{1-h_{ii}}}.$$

- Les résidus studentisés par validation croisée sont définis par

$$t_i^* = \frac{\widehat{\varepsilon}_i}{\widehat{\sigma}_{(i)}\sqrt{1-h_{ii}}}.$$

où $\widehat{\sigma}_{(i)}$ est calculé dans un échantillon privé de l'observation i. Dans le cas gaussien, $t_i^* \sim \mathcal{T}(n-1-p)$.
$\hookrightarrow$ points aberrants

Visualisations

Différents types de graphiques permettent une étude visuelle :

- valeurs estimées en fonction de valeurs observées,
- résidus en fonction des valeurs estimées,
- résidus en fonction d'une valeur de covariable,
- résidus de l'observation $i+1$ en fonction du résidu de l'observation i,
- graphe quantile/quantile des résidus.

Exemple de diagnostics visuels