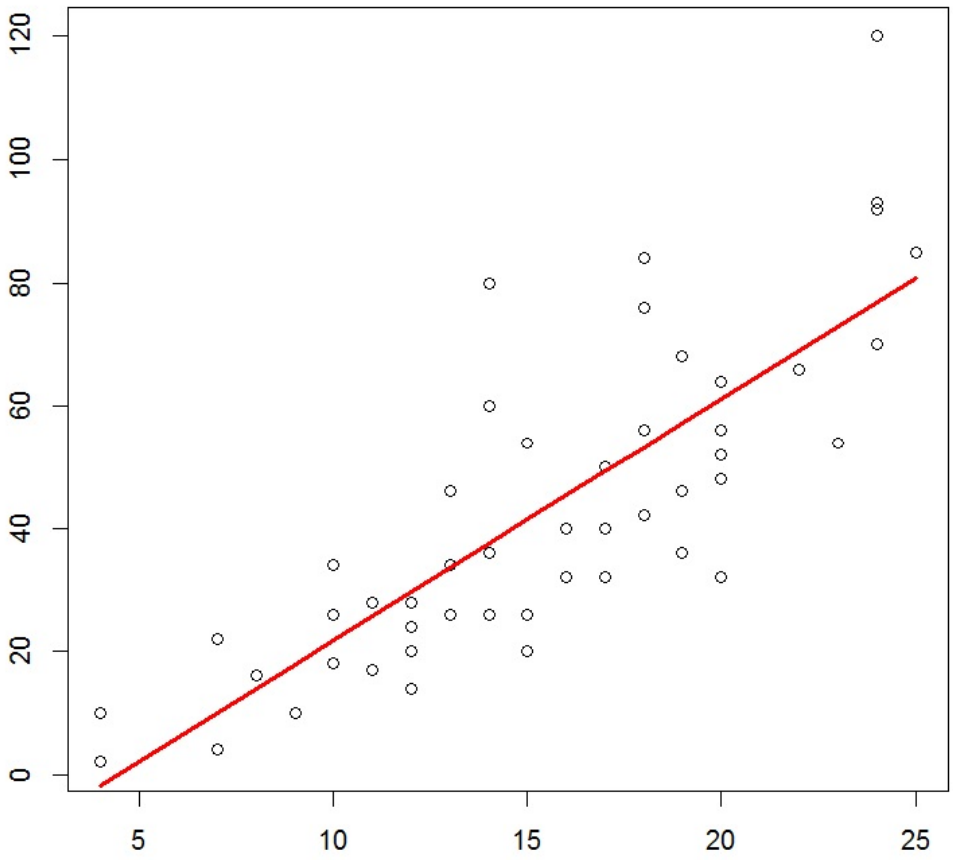
Université de Caen

Sur l'Estimateur des Moindres Carrés Ordinaires *(emco)*

Christophe Chesneau

http://www.math.unicaen.fr/~chesneau/



Caen, le 07 Février 2017

Table des matières

1 Modèle de régression linéaire multiple et *emco* **5**

2 Cas particulier : le modèle de régression linéaire simple **15**

3 Loi normale multidimensionelle **31**

4 Propriétés standards et lois associées **35**

5 Retour sur le modèle de *rls* **45**

6 Intervalles et volumes de confiance **47**

7 Tests statistiques **53**

Index **65**

~ Note ~

Ce document résume les notions abordées dans la première partie du cours *Statistique 2* du M1 orienté statistique de l'université de Caen (la deuxième partie concerne l'ANOVA à 1 et 2 facteurs).

L'enjeu de ce document est de présenter les fondations théoriques sur lesquelles repose l'estimateur des moindres carrés ordinaires. Des jeux de données et des commandes R viennent illustrer la théorie.

Je vous invite à me contacter pour tout commentaire :

`christophe.chesneau@gmail.com`

Bonne lecture !

1 Modèle de régression linéaire multiple et *emco*

Modèle de régression linéaire multiple (*rlm*) ; forme générique

On souhaite prédire et/ou expliquer les valeurs d'une variable quantitative Y à partir des valeurs de p variables $X_1, \ldots, X_p$. On dit alors que l'on souhaite "expliquer Y à partir de $X_1, \ldots, X_p$", Y est appelée "variable à expliquer" et $X_1, \ldots, X_p$ sont appelées "variables explicatives".

Pour ce faire, on dispose de données qui sont n observations de $(Y, X_1, \ldots, X_p)$ notées $(y_1, x_{1,1}, \ldots, x_{p,1}), (y_2, x_{1,2}, \ldots, x_{p,2}), \ldots, (y_n, x_{1,n}, \ldots, x_{p,n})$. Elles se présentent généralement sous la forme d'un tableau :

Y	X_1	$\ldots$	X_p
y_1	$x_{1,1}$	$\ldots$	$x_{p,1}$
y_2	$x_{1,2}$	$\cdots$	$x_{p,2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_n	$x_{1,n}$	$\ldots$	$x_{p,n}$

Si une liaison linéaire entre Y et $X_1, \ldots, X_p$ est enisageable, on peut considérer le modèle de régression linéaire multiple (*rlm*). Sa forme générique est

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon,$$

où $\beta_0, \ldots, \beta_p$ sont des coefficients réels inconnus et ϵ est une variable quantitative de valeur moyenne nulle, indépendante de $X_1, \ldots, X_p$, qui représente une somme d'erreurs aléatoires et multifactorielles (erreurs de mesures, effets non prévisibles, variables omises...).

Notre principal objectif est d'estimer convenablement $\beta_0, \ldots, \beta_p$ à l'aide des données. Entre autres, cela nous permettra de mesurer l'importance des variables $X_1, \ldots, X_p$ dans l'explication de Y et de prédire avec précision la valeur moyenne de Y pour une nouvelle valeur de $(X_1, \ldots, X_p)$.

Exemples

Loyers : On peut considérer le jeu de données "loyers" :

`http://www.math.unicaen.fr/~chesneau/loyers.txt`

Dans un quartier parisien, une étude a été menée afin de mettre en évidence une relation entre le loyer mensuel et la surface des appartements ayant exactement 3 pièces.

Pour 30 appartements de ce type, on dispose :

- de la surface en mètres carrés (variable $X1$),
- du loyer mensuel en francs (variable Y).

Fromages : On peut considérer le jeu de données "fromages" :

`http://www.math.unicaen.fr/~chesneau/fromages.txt`

Le goût d'un fromage dépend de la concentration de plusieurs composés chimiques, dont :

- la concentration de l'acide acétique (variable $X1$),
- la concentration d'hydrogène sulfuré (variable $X2$),
- la concentration d'acide lactique (variable $X3$).

Pour 30 types de fromage, on dispose du score moyen attribué par des consommateurs (variable Y).

On souhaite expliquer Y à partir de $X1$, $X2$ et $X3$.

NBA : On peut considérer le jeu de données "nba" :

`http://www.math.unicaen.fr/~chesneau/nba.txt`

On souhaite expliquer le poids d'un basketteur professionnel de la NBA à partir de sa taille et de son âge. Ainsi, pour 505 basketteurs de la NBA, on dispose :

- de leur poids (variable Y),
- de leur taille (variable $X1$),
- de leur âge (variable $X3$).

On souhaite expliquer Y à partir de $X1$ et $X3$ *(pour information, on dispose aussi de leur rôle sur le terrain (variable qualitative $X2$) mais on ne souhaite pas l'inclure dans le modèle ici).*

Modèle de *rlm*

On modélise les variables considérées comme des variables aléatoires réelles (*var*) (définies sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$), en gardant les mêmes notations par convention. À partir de celles-ci, le modèle de *rlm* est caractérisé par : pour tout $i \in \{1, \dots, n\}$,

- $(x_{1,i}, \dots, x_{p,i})$ est une réalisation du vecteur aléatoire réel $(X_1, \dots, X_p)$,
- sachant que $(X_1, \dots, X_p) = (x_{1,i}, \dots, x_{p,i})$, y_i est une réalisation de

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \dots + \beta_p x_{p,i} + \epsilon_i,$$

où ϵ_i est une *var* indépendante de $X_1, \dots, X_p$ avec $\mathbb{E}(\epsilon_i) = 0$.

D'autres hypothèses sur $\epsilon_1, \dots, \epsilon_n$ seront formulées ultérieurement.

Écriture matricielle du modèle de *rlm*

Le modèle de *rlm* peut alors s'écrire sous la forme matricielle : $Y = X\beta + \epsilon$, où

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{p,1} \\ 1 & x_{1,2} & \cdots & x_{p,2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1,n} & \cdots & x_{p,n} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Estimateur des moindres carrés ordinaire ; un résultat central

Soient $\|.\|$ la norme euclidienne : pour tout vecteur colonne x, $\|x\|^2 = x^t x$ = somme des carrés des composantes de x. Partant du modèle de *rlm* écrit sous la forme matricielle : $Y = X\beta + \epsilon$, un estimateur des moindres carrés ordinaires (*emco*) $\widehat{\beta}$ de β vérifie :

$$\widehat{\beta} \in \underset{\beta \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \|Y - X\beta\|^2.$$

On suppose que X est de rang colonnes plein : il n'existe pas de vecteur colonne x à $p+1$ composantes non nul tel que $Xx =$ le vecteur nul (cela entraîne l'existence de $(X^t X)^{-1}$). Alors $\widehat{\beta}$ est unique ; il est donné par la formule :

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

Preuve : Posons

$$f(\beta) = \|Y - X\beta\|^2, \qquad \beta \in \mathbb{R}^{p+1}.$$

Comme $\widehat{\beta} \in \text{argmin}_{\beta\in\mathbb{R}^{p+1}} f(\beta)$, $\widehat{\beta}$ est un extremum de $f(\beta)$, et

$$\widehat{\beta} \text{ extremum de } f(\beta) \quad \Rightarrow \quad \frac{\partial}{\partial\beta_j} f(\widehat{\beta}) = 0, \qquad j \in \{0, \ldots, p\}.$$

Simplifions l'écriture de $f(\beta)$. En utilisant les formules : $(A+B)^t = A^t + B^t$ et $(AB)^t = B^t A^t$, il vient

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

Comme $Y^t X\beta$ est la multiplication d'un vecteur ligne Y^t par un vecteur colonne $X\beta$, c'est un réel. Par conséquent, il est égal à sa transposé ; on a $Y^t X\beta = (Y^t X\beta)^t = (X\beta)^t (Y^t)^t = \beta^t X^t Y$. Il vient

$$f(\beta) = Y^t Y - 2\beta^t X^t Y + \beta^t X^t X\beta.$$

Pour tout $j \in \{0, \ldots, p\}$, déterminons la dérivée partielle $\frac{\partial}{\partial\beta_j} f(\beta)$. Soit e_j le vecteur colonne à $p+1$ composantes avec p composantes nulles, sauf la $j+1$-ème qui vaut 1. En utilisant la formule :

$(u(x)v(x))' = u'(x)v(x) + u(x)v'(x)$, il vient

$$\begin{aligned} \frac{\partial}{\partial\beta_j} f(\beta) &= \frac{\partial}{\partial\beta_j}(Y^t Y - 2\beta^t X^t Y + \beta^t X^t X\beta) = \frac{\partial}{\partial\beta_j}(Y^t Y) - 2\frac{\partial}{\partial\beta_j}(\beta^t X^t Y) + \frac{\partial}{\partial\beta_j}(\beta^t X^t X\beta) \\ &= 0 - 2e_j^t X^t Y + e_j^t X^t X\beta + \beta^t X^t X e_j. \end{aligned}$$

Comme $e_j^t X^t X\beta$ est la multiplication d'un vecteur ligne $e_j^t X^t$ par un vecteur colonne $X\beta$, c'est un réel. Par conséquent, il est égal à sa transposé ; on a $e_j^t X^t X\beta = (e_j^t X^t X\beta)^t = (X\beta)^t (e_j^t X^t)^t = \beta^t X^t X e_j$. Donc

$$\frac{\partial}{\partial\beta_j} f(\beta) = -2e_j^t X^t Y + 2e_j^t X^t X\beta.$$

Il s'ensuit

$$\frac{\partial}{\partial\beta_j} f(\widehat{\beta}) = 0 \quad \Leftrightarrow \quad -2e_j^t X^t Y + 2e_j^t X^t X\widehat{\beta} = 0 \quad \Leftrightarrow \quad e_j^t X^t X\widehat{\beta} = e_j^t X^t Y.$$

Comme cela est vraie pour tout $j \in \{0,\dots,p\}$ et que $e_j^t X^t X\widehat{\beta}$ calcule la j-ème ligne de la matrice $X^tX\widehat{\beta}$, il vient

$$\frac{\partial}{\partial\beta_j} f(\widehat{\beta}) = 0, \qquad j \in \{0,\dots,p\} \quad \Leftrightarrow \quad X^tX\widehat{\beta} = X^tY.$$

Comme $(X^tX)^{-1}$ existe, l'égalité $(X^tX)^{-1}X^tX = \mathbb{I}_{p+1}$ entraîne

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

Au final, on a

$$\widehat{\beta} \text{ extremum de } f(\beta) \quad \Rightarrow \quad \widehat{\beta} = (X^tX)^{-1}X^tY.$$

Il reste à montrer que $\widehat{\beta}$ est bien un minimum pour $f(\beta)$. Pour cela, on calcule la matrice hessienne $H(f) = \left(\frac{\partial^2}{\partial\beta_j\partial\beta_k} f(\beta)\right)_{(j,k)\in\{0,\dots,p\}^2}$ et on montre qu'elle est définie positive : pour tout vecteur colonne non nul x à $p+1$ composantes, on a $x^tH(f)x > 0$. Pour tout $(j,k) \in \{0,\dots,p\}^2$, on a

$$\begin{aligned}\frac{\partial^2}{\partial\beta_j\partial\beta_k} f(\beta) &= \frac{\partial}{\partial\beta_k}\left(\frac{\partial}{\partial\beta_j} f(\beta)\right) = \frac{\partial}{\partial\beta_k}(-2e_j^tX^tY + 2e_j^tX^tX\beta) \\ &= -2\frac{\partial}{\partial\beta_k}(e_j^tX^tY) + 2\frac{\partial}{\partial\beta_k}(e_j^tX^tX\beta) = 0 + 2e_j^tX^tXe_k = 2e_j^tX^tXe_k.\end{aligned}$$

Donc

$$H(f) = \left(2e_j^tX^tXe_k\right)_{(j,k)\in\{0,\dots,p\}^2} = 2X^tX.$$

Pour tout $x = \begin{pmatrix} x_0 \\ \vdots \\ x_p \end{pmatrix}$ non nul, comme X est de rang colonnes plein, on a

$$x^tH(f)x = x^t(2X^tX)x = 2x^tX^tXx = 2(Xx)^tXx = 2\|Xx\|^2 > 0.$$

Ainsi $H(f)$ est définie positive ; $\widehat{\beta}$ est bien un minimum pour $f(\beta)$. On en déduit que

$$\widehat{\beta} \in \underset{\beta\in\mathbb{R}^{p+1}}{\operatorname{argmin}} \|Y - X\beta\|^2 \quad \Leftrightarrow \quad \widehat{\beta} = (X^tX)^{-1}X^tY.$$

□

Emco de β_j

L'*emco* $\widehat{\beta}$ de $\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$ s'écrit sous la forme $\widehat{\beta} = \begin{pmatrix} \widehat{\beta}_0 \\ \vdots \\ \widehat{\beta}_p \end{pmatrix}$.

Pour tout $j \in \{0, \ldots, p\}$, l'*emco* de β_j est $\widehat{\beta}_j$.

Dorénavant, $\widehat{\beta}$ désignera l'*emco* de β et $\widehat{\beta}_j$ l'*emco* de β_j.

Estimateur de la valeur moyenne

- On appelle valeur moyenne de Y quand $(X_1, \ldots, X_p) = (x_1, \ldots, x_p) = x$ le réel inconnu :

$$y_x = \mathbb{E}(Y | \{(X_1, \ldots, X_p) = x\}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p.$$

- Un estimateur de y_x est

$$\widehat{Y}_x = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \ldots + \widehat{\beta}_p x_p.$$

En posant $x_\bullet = (1, x_1, \ldots, x_p)$, on a $y_x = x_\bullet \beta$ et $\widehat{Y}_x = x_\bullet \widehat{\beta}$.

Estimations ponctuelles

Dorénavant, l'expression "la réalisation" fera référénce à celle correspondante aux données.

- Une estimation ponctuelle de β est la réalisation b de $\widehat{\beta}$:

$$b = (X^t X)^{-1} X^t y,$$

avec $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$. On peut écrire b sous la forme $b = \begin{pmatrix} b_0 \\ \vdots \\ b_p \end{pmatrix}$. Pour tout $j \in \{0, \ldots, p\}$, b_j est une estimation ponctuelle de β_j.

On dit que b est l'*emco* ponctuel de β et b_j est l'*emco* ponctuel de β_j.

- Soit $x_\bullet = (1, x_1, \ldots, x_p)$. Une estimation ponctuelle de $y_x = x_\bullet \beta$ est la réalisation d_x de $\widehat{Y}_x = x_\bullet \widehat{\beta}$:

$$d_x = x_\bullet b = b_0 + b_1 x_1 + \ldots + b_p x_p.$$

On dit que d_x est la valeur prédite de Y quand $(X_1, \ldots, X_p) = x$.

Coefficient de détermination

Soit 1_n le vecteur colonne à n composantes égales à 1. On pose $\widehat{Y} = X\widehat{\beta}$ et $\overline{Y} = \dfrac{1}{n}\sum_{i=1}^{n} Y_i$.

On appelle coefficient de détermination la réalisation R^2 de

$$\widehat{R}^2 = 1 - \frac{\|\widehat{Y} - Y\|^2}{\|\overline{Y}1_n - Y\|^2}.$$

Avec les notations déjà introduites et $\overline{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$, on peut écrire :

$$R^2 = 1 - \frac{\|Xb - y\|^2}{\|\overline{y}1_n - y\|^2}.$$

On a $R^2 \in [0, 1]$. Plus R^2 est proche de 1, plus la liaison linéaire entre Y et $X_1, \ldots, X_p$ est forte.

En effet, plus R^2 est proche de 1, plus $\|Xb - y\|^2$ est proche de 0, plus y est proche de Xb, plus le modèle de *rlm* est pertinent, plus la liaison linéaire entre Y et $X_1, \ldots, X_p$ est forte.

En remarquant que $\|\overline{y}1_n - y\|^2 = \|\overline{y}1_n - Xb\|^2 + \|Xb - y\|^2$, on a aussi :

$$R^2 = \frac{\|\overline{y}1_n - Xb\|^2}{\|\overline{y}1_n - y\|^2}.$$

Coefficient de détermination ajusté

Une version améliorée du R^2 est le coefficient de détermination ajustée défini par

$$\overline{R}^2 = 1 - \frac{n-1}{n-(p+1)}(1 - R^2).$$

Il s'interprète comme le R^2.

Mise en œuvre avec le logiciel R

Pour illustrer les notions précédentes avec le logiciel R, on peut considérer le jeu de données "profs". Dans une étude statistique, 23 professeurs sont évalués quant à la qualité de leur enseignement. Pour chacun d'entre eux, on dispose :

- d'un indice de performance globale donné par les étudiants (variable Y),
- des résultats de 4 tests écrits donnés à chaque professeur (variables X_1, X_2, X_3 et X_4),
- du sexe (variable X_5, avec $X_5 = 0$ pour femme, $X_5 = 1$ pour homme).

L'objectif est d'expliquer Y à partir de X_1, X_2, X_3, X_4 et X_5.

Le jeu de données est disponible ici :

```
http://www.math.unicaen.fr/~chesneau/profs.txt
```

Écrire dans une fenêtre R :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/profs.txt", header = T)

attach(w)

head(w)
```

Cela renvoie l'entête du jeu de données :

	Y	X1	X2	X3	X4	X5
1	489	81	151	45.50	43.61	1
2	423	68	156	46.45	44.69	1
3	507	80	165	76.50	54.57	1
4	467	107	149	55.50	43.27	1
5	340	43	134	49.40	49.21	1
6	524	129	163	72.00	49.96	1

Le modèle de *rlm* est envisageable. Sa forme générique est

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon.$$

où β_0, β_1, β_2, β_3, β_4 et β_5 sont des coefficients réels inconnus.

On le considère sous sa forme matricielle : $Y = X\beta + \epsilon$, où

$$X = \begin{pmatrix} 1 & 81 & 151 & 45.50 & 43.61 & 1 \\ 1 & 68 & 156 & 46.45 & 44.69 & 1 \\ 1 & 80 & 165 & 76.50 & 54.57 & 1 \\ 1 & 107 & 149 & 55.50 & 43.27 & 1 \\ 1 & 43 & 134 & 49.40 & 49.21 & 1 \\ 1 & 129 & 163 & 72.00 & 49.96 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}, \quad Y = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ \vdots \\ Y_{23} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \vdots \\ \epsilon_{23} \end{pmatrix}.$$

Nous allons maintenant étudier l'*emco* de β correspondant aux données.

Il s'agit donc de calculer l'*emco* ponctuel b défini par

$$b = (X^t X)^{-1} X^t y, \qquad y = \begin{pmatrix} 489 \\ 423 \\ 507 \\ \vdots \end{pmatrix}.$$

Introduisons la matrice X composée des colonnes "que des 1", $X1$, $X2$, $X3$, $X4$ et $X5$:

```
X = cbind(1, X1, X2, X3, X4, X5)
```

En utilisant les commandes R : `%*%` = produit matriciel, `t(A)` $= A^t$ et `solve(A)` $= A^{-1}$, calculons $b = (X^t X)^{-1} X^t y$:

```
b = solve(t(X) %*% X) %*% t(X) %*% Y

b
```

Cela renvoie :

$$b = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{pmatrix} = \begin{pmatrix} -272.04 \\ 0.79 \\ 2.68 \\ -1.44 \\ 6.83 \\ 14.90 \end{pmatrix}.$$

Entre autre, ces estimations nous permettent de faire des prédictions sur Y pour de nouvelles valeurs de $(X_1, X_2, X_3, X_4, X_5)$.

Par exemple, pour $(X_1, X_2, X_3, X_4, X_5) = (82, 158, 47, 49, 1) = x$, en posant $x_\bullet = (1, 82, 158, 47, 49, 1)$, la valeur prédite de Y est $d_x = x_\bullet b$. Cela s'obtient en faisant :

```
x = c(1, 82, 158, 47, 49, 1)

d = x %*% b

d
```

Cela renvoie : 498.5063.

Ainsi, pour de tels critères, l'indice de performance globale moyen est de 498.5063.

Le R^2 peut se calculer en faisant :

```
R2 = 1 - sum((X %*% b - Y)^2) / sum((mean(Y) - Y)^2)

R2
```

Cela renvoie : 0.6834218.

De même pour le R^2 ajusté :

```
R2aj = 1 - ((length(Y) - 1)/(length(Y) - (5 + 1))) * (1 - R2)

R2aj
```

Cela renvoie : 0.5903106.

Le R^2 (et $\overline{R}^2$) étant relativement proche de 1, le modèle de *rlm* semble être pertinent avec les données traitées.

Commande `summary` :

On retrouve plus simplement ces estimations (et beaucoup plus) avec la commande `summary` :

```
reg = lm(Y ~ X1 + X2 + X3 + X4 + X5)

summary(reg)
```

Cela renvoie :

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	**-272.0388**	184.3865	-1.48	0.1584	
X1	**0.7913**	0.5363	1.48	0.1583	
X2	**2.6828**	0.9216	2.91	0.0097	**
X3	**-1.4434**	0.8217	-1.76	0.0970	.
X4	**6.8308**	1.8192	3.75	0.0016	**
X5	**14.9008**	27.3134	0.55	0.5925	

```
Residual standard error: 55.06 on 17 degrees of freedom

Multiple R-squared: 0.6834, Adjusted R-squared: 0.5903

F-statistic: 7.34 on 5 and 17 DF, p-value: 0.0007887
```

On retrouve b dans colonne `Estimate` du tableau.

On retrouve également : $R^2 = 0.6834$ et $\overline{R}^2 = 0.5903$.

Pour la valeur prédite de Y quand $(X_1, X_2, X_3, X_4, X_5) = (82, 158, 47, 49, 1)$, on peut faire :

```
predict(reg, data.frame(X1 = 82, X2 = 158, X3 = 47, X4 = 49, X5 = 1))
```

2 Cas particulier : le modèle de régression linéaire simple

Modèle de régression linéaire simple (*rls*)

Le modèle de régression linéaire simple (*rls*) est le modèle de *rlm* avec $p = 1$.

Contexte

On souhaite expliquer une variable quantitative Y à partir d'une variable X_1. Pour ce faire, on dispose de données qui sont n observations de (Y, X_1) notées $(y_1, x_{1,1}), (y_2, x_{1,2}), \ldots, (y_n, x_{1,n})$.

Ces observations peuvent être représentées sur le repère orthonormé (O, I, J) par les points de coordonnées $(x_{1,1}, y_1), (x_{1,2}, y_2), \ldots, (x_{1,n}, y_n)$. L'ensemble de ces points est appelé nuage de points. Si la silhouette de ce nuage de points est allongée dans une direction, une liaison linéaire entre Y et X_1 est envisageable. On peut alors considérer le modèle de *rls*. Sa forme générique est

$$Y = \beta_0 + \beta_1 X_1 + \epsilon,$$

où β_0 et β_1 sont des coefficients réels inconnus et ϵ est une variable quantitative de valeur moyenne nulle, indépendante de X_1, qui représente une somme d'erreurs aléatoires et multifactorielles.

Notre principal objectif est d'estimer convenablement β_0 et β_1 à l'aide des données. On pourra alors prédire avec précision la valeur moyenne de Y pour une nouvelle valeur de X_1. Cela revient à ajuster du mieux possible le nuage de points par une droite (on parle alors d'ajustement affine).

Exemples

Scores : On peut considérer le jeu de données "scores" :

`http://www.math.unicaen.fr/~chesneau/scores.txt`

Une étude a été menée auprès de 19 étudiants afin de mettre en évidence une relation entre le score (note) final à un examen de mathématiques et le temps consacré à la préparation de cet examen. Pour chaque étudiant, on dispose :

- du temps de révision en heures (variable $X1$),
- du score obtenu sur 800 points (variable Y).

Fibres : On peut considérer le jeu de données "fibres" :

`http://www.math.unicaen.fr/~chesneau/fibres.txt`

Une étude s'intéresse à la vitesse de propagation de l'influx nerveux dans une fibre nerveuse. Pour 16 fibres nerveuses différentes, on considère :

- le diamètre en microns (variable $X1$),
- la vitesse de l'influx nerveux en m/s (variable Y).

On souhaite expliquer Y à partir de $X1$.

Toluca : On peut considérer le jeu de données "toluca" :

`http://www.math.unicaen.fr/~chesneau/toluca.txt`

L'entreprise Toluca fabrique des pièces de rechange pour l'équipement de réfrigération. Pour une pièce particulière, le processus de production prend un certain temps.

Dans le cadre d'un programme d'amélioration des coûts, l'entreprise souhaite mieux comprendre la relation entre :

- la taille du lot (variable $X1$),
- nombre total d'heures de travail (variable Y).

Les données ont été rapportées pour 25 lots représentatifs de taille variable.

Eaux usées : On peut considérer le jeu de données "eaux usées" :

`http://www.math.unicaen.fr/~chesneau/eauxusées.txt`

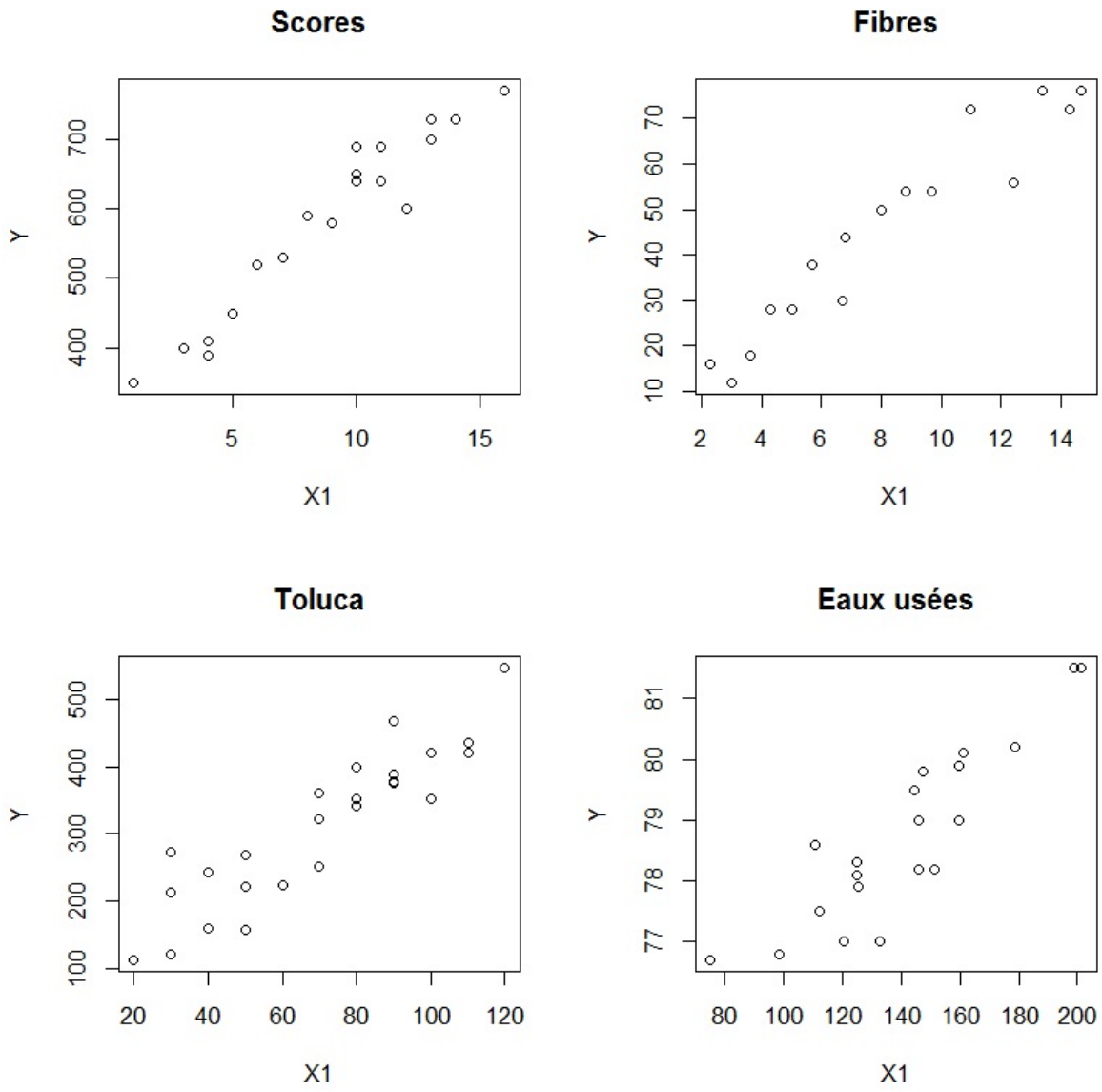
Une nouvelle machine pour le traitement des eaux usées est à l'étude. En particulier, les ingénieurs s'intéressent à :

- la vitesse de filtration mesurée en pour cent (variable $X1$),
- l'humidité des granulés en kg-DS/m/h (variable Y).

Les données ont été rapportées pour 20 expériences indépendantes. On souhaite expliquer Y à partir de $X1$.

Exemples : nuages de points

Les nuages de points associées aux exemples introduits précédents sont présentés ci-dessous :



La silhouette de chaque nuage de points est étirée dans une direction ; une liaison linéaire entre Y et X_1 est envisageable, on peut considérer le modèle de *rls*.

Écriture matricielle du modèle de *rls*

On modélise les variables considérées comme des *var* (définies sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$), en gardant les mêmes notations par convention. À partir de celles-ci, le modèle de *rls* est caractérisé par : pour tout $i \in \{1, \dots, n\}$,

- $x_{1,i}$ est une réalisation de X_1,
- sachant que $X_1 = x_{1,i}$, y_i est une réalisation de

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \epsilon_i,$$

où ϵ_i est une *var* modélisant une somme d'erreurs aléatoires et multifactorielles.

Notons que le modèle de *rls* peut s'écrire sous la forme matricielle : $Y = X\beta + \epsilon$, où

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & x_{1,1} \\ 1 & x_{1,2} \\ \vdots & \vdots \\ 1 & x_{1,n} \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \qquad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Emco et modèle de *rls*

À l'instar du modèle de *rlm*, on peut déterminer les *emco* de β_0 et β_1. Le résultat suivant présentent des expressions analytiques des estimateurs obtenus.

On pose $\overline{x}_1 = \dfrac{1}{n}\sum_{i=1}^{n} x_{1,i}$ et $\overline{Y} = \dfrac{1}{n}\sum_{i=1}^{n} Y_i$. On rappelle que $\widehat{\beta} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix}$ est l'*emco* de $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$. Alors on a

$$\widehat{\beta}_1 = \frac{1}{\sum_{i=1}^{n}(x_{1,i} - \overline{x}_1)^2} \sum_{i=1}^{n}(x_{1,i} - \overline{x}_1)(Y_i - \overline{Y}), \qquad \widehat{\beta}_0 = \overline{Y} - \overline{x}_1\widehat{\beta}_1.$$

Preuve : On rappelle que le modèle de *rls* s'écrit sous la forme matricielle : $Y = X\beta + \epsilon$, où

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & x_{1,1} \\ 1 & x_{1,2} \\ \vdots & \vdots \\ 1 & x_{1,n} \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \qquad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

L'*emco* $\widehat{\beta}$ de β est donné par la formule :

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

◦ Calcul de $X^t X$. On a

$$X^t X = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_{1,1} & x_{1,2} & \dots & x_{1,n} \end{pmatrix} \begin{pmatrix} 1 & x_{1,1} \\ 1 & x_{1,2} \\ \vdots & \vdots \\ 1 & x_{1,n} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n 1 & \sum_{i=1}^n x_{1,i} \\ \sum_{i=1}^n x_{1,i} & \sum_{i=1}^n x_{1,i}^2 \end{pmatrix} = \begin{pmatrix} n & n\overline{x}_1 \\ n\overline{x}_1 & \sum_{i=1}^n x_{1,i}^2 \end{pmatrix}.$$

◦ Calcul de $(X^t X)^{-1}$. En utilisant la formule matricielle : si $ad - bc \neq 0$,

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \qquad \Leftrightarrow \qquad A^{-1} = \frac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix},$$

on obtient

$$(X^t X)^{-1} = \frac{1}{n\sum_{i=1}^n x_{1,i}^2 - (n\overline{x}_1)^2} \begin{pmatrix} \sum_{i=1}^n x_{1,i}^2 & -n\overline{x}_1 \\ -n\overline{x}_1 & n \end{pmatrix} = \frac{1}{\sum_{i=1}^n x_{1,i}^2 - n\overline{x}_1^2} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_{1,i}^2 & -\overline{x}_1 \\ -\overline{x}_1 & 1 \end{pmatrix}.$$

◦ Calcul de $X^t Y$. On a

$$X^t Y = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_{1,1} & x_{1,2} & \dots & x_{1,n} \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n x_{1,i} Y_i \end{pmatrix} = \begin{pmatrix} n\overline{Y} \\ \sum_{i=1}^n x_{1,i} Y_i \end{pmatrix}.$$

◦ Calcul de $\widehat{\beta} = (X^tX)^{-1}X^tY$. En mettant bout à bout les égalités précédentes, il vient

$$\begin{aligned}
\widehat{\beta} &= (X^tX)^{-1}X^tY = \frac{1}{\sum_{i=1}^n x_{1,i}^2 - n\overline{x}_1^2}\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_{1,i}^2 & -\overline{x}_1 \\ -\overline{x}_1 & 1 \end{pmatrix}\begin{pmatrix} n\overline{Y} \\ \sum_{i=1}^n x_{1,i}Y_i \end{pmatrix} \\
&= \frac{1}{\sum_{i=1}^n x_{1,i}^2 - n\overline{x}_1^2}\begin{pmatrix} \left(\frac{1}{n}\sum_{i=1}^n x_{1,i}^2\right) n\overline{Y} - \overline{x}_1\sum_{i=1}^n x_{1,i}Y_i \\ -\overline{x}_1 \times n\overline{Y} + \sum_{i=1}^n x_{1,i}Y_i \end{pmatrix} \\
&= \frac{1}{\sum_{i=1}^n x_{1,i}^2 - n\overline{x}_1^2}\begin{pmatrix} \overline{Y}\sum_{i=1}^n x_{1,i}^2 - \overline{x}_1\sum_{i=1}^n x_{1,i}Y_i \\ \sum_{i=1}^n x_{1,i}Y_i - n\overline{x}_1\overline{Y} \end{pmatrix}.
\end{aligned}$$

On en déduit que

$$\widehat{\beta}_0 = \frac{1}{\sum_{i=1}^n x_{1,i}^2 - n\overline{x}_1^2}\left(\overline{Y}\sum_{i=1}^n x_{1,i}^2 - \overline{x}_1\sum_{i=1}^n x_{1,i}Y_i\right), \qquad \widehat{\beta}_1 = \frac{1}{\sum_{i=1}^n x_{1,i}^2 - n\overline{x}_1^2}\left(\sum_{i=1}^n x_{1,i}Y_i - n\overline{x}_1\overline{Y}\right).$$

◦ Réécriture de $\widehat{\beta}_1$. On a

$$\begin{aligned}
\sum_{i=1}^n (x_{1,i} - \overline{x}_1)^2 &= \sum_{i=1}^n (x_{1,i}^2 - 2\overline{x}_1 x_{1,i} + \overline{x}_1^2) = \sum_{i=1}^n x_{1,i}^2 - 2\overline{x}_1\sum_{i=1}^n x_{1,i} + \overline{x}_1^2\sum_{i=1}^n 1 \\
&= \sum_{i=1}^n x_{1,i}^2 - 2\overline{x}_1 \times n\overline{x}_1 + \overline{x}_1^2 n = \sum_{i=1}^n x_{1,i}^2 - 2n\overline{x}_1^2 + n\overline{x}_1^2 = \sum_{i=1}^n x_{1,i}^2 - n\overline{x}_1^2.
\end{aligned}$$

De plus, on a

$$\begin{aligned}
\sum_{i=1}^n (x_{1,i} - \overline{x}_1)(Y_i - \overline{Y}) &= \sum_{i=1}^n (x_{1,i}Y_i - x_{1,i}\overline{Y} - \overline{x}_1 Y_i + \overline{x}_1\overline{Y}) \\
&= \sum_{i=1}^n x_{1,i}Y_i - \overline{Y}\sum_{i=1}^n x_{1,i} - \overline{x}_1\sum_{i=1}^n Y_i + \overline{x}_1\overline{Y}\sum_{i=1}^n 1 \\
&= \sum_{i=1}^n x_{1,i}Y_i - \overline{Y} \times n\overline{x}_1 - \overline{x}_1 \times n\overline{Y} + \overline{x}_1\overline{Y} \times n \\
&= \sum_{i=1}^n x_{1,i}Y_i - n\overline{x}_1\overline{Y} - n\overline{x}_1\overline{Y} + n\overline{x}_1\overline{Y} = \sum_{i=1}^n x_{1,i}Y_i - n\overline{x}_1\overline{Y}.
\end{aligned}$$

Par conséquent, on peut réécrire $\widehat{\beta}_1$ comme

$$\widehat{\beta}_1 = \frac{1}{\sum_{i=1}^n x_{1,i}^2 - n\overline{x}_1^2}\left(\sum_{i=1}^n x_{1,i}Y_i - n\overline{x}_1\overline{Y}\right) = \frac{1}{\sum_{i=1}^n (x_{1,i}-\overline{x}_1)^2}\sum_{i=1}^n (x_{1,i}-\overline{x}_1)(Y_i-\overline{Y}).$$

◦ Réécriture de $\widehat{\beta}_0$. En introduisant $0 = -n\overline{x}_1^2\overline{Y} + n\overline{x}_1^2\overline{Y}$, on obtient

$$\begin{aligned}\overline{Y}\sum_{i=1}^n x_{1,i}^2 - \overline{x}_1\sum_{i=1}^n x_{1,i}Y_i &= \overline{Y}\sum_{i=1}^n x_{1,i}^2 - n\overline{x}_1^2\overline{Y} + n\overline{x}_1^2\overline{Y} - \overline{x}_1\sum_{i=1}^n x_{1,i}Y_i \\ &= \overline{Y}\left(\sum_{i=1}^n x_{1,i}^2 - n\overline{x}_1^2\right) - \overline{x}_1\left(\sum_{i=1}^n x_{1,i}Y_i - n\overline{x}_1\overline{Y}\right).\end{aligned}$$

Il vient

$$\begin{aligned}\widehat{\beta}_0 &= \frac{1}{\sum_{i=1}^n x_{1,i}^2 - n\overline{x}_1^2}\left(\overline{Y}\sum_{i=1}^n x_{1,i}^2 - \overline{x}_1\sum_{i=1}^n x_{1,i}Y_i\right) \\ &= \frac{1}{\sum_{i=1}^n x_{1,i}^2 - n\overline{x}_1^2}\left(\overline{Y}\left(\sum_{i=1}^n x_{1,i}^2 - n\overline{x}_1^2\right) - \overline{x}_1\left(\sum_{i=1}^n x_{1,i}Y_i - n\overline{x}_1\overline{Y}\right)\right) \\ &= \overline{Y} - \overline{x}_1\frac{1}{\sum_{i=1}^n x_{1,i}^2 - n\overline{x}_1^2}\left(\sum_{i=1}^n x_{1,i}Y_i - n\overline{x}_1\overline{Y}\right) = \overline{Y} - \overline{x}_1\widehat{\beta}_1.\end{aligned}$$

◦ ◦ Au final. L'*emco* $\widehat{\beta}$ de β a pour composantes :

$$\widehat{\beta}_1 = \frac{1}{\sum_{i=1}^n (x_{1,i}-\overline{x}_1)^2}\sum_{i=1}^n (x_{1,i}-\overline{x}_1)(Y_i-\overline{Y}), \qquad \widehat{\beta}_0 = \overline{Y} - \overline{x}_1\widehat{\beta}_1.$$

□

Estimateur de la prédiction

Soit y_x la valeur moyenne de Y quand $X_1 = x_1 = x$:

$$y_x = \beta_0 + \beta_1 x_1.$$

Un estimateur de y_x est

$$\widehat{Y}_x = \widehat{\beta}_0 + \widehat{\beta}_1 x_1.$$

Quantités utilisées

Partant des données, on considère les quantités suivantes :

- Moyennes :

$$\overline{x}_1 = \frac{1}{n}\sum_{i=1}^{n} x_{1,i}, \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

- Écarts-type :

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{1,i} - \overline{x}_1)^2}, \qquad s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2}.$$

- Sommes des carrés des écarts :

$$\text{sce}_x = \sum_{i=1}^{n}(x_{1,i} - \overline{x}_1)^2 = (n-1)s_x^2 = \sum_{i=1}^{n} x_{1,i}^2 - n\overline{x}_1^2,$$

$$\text{sce}_y = \sum_{i=1}^{n}(y_i - \overline{y})^2 = (n-1)s_y^2 = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2.$$

- Somme des produits des écarts :

$$\text{spe}_{x,y} = \sum_{i=1}^{n}(x_{1,i} - \overline{x}_1)(y_i - \overline{y}) = \sum_{i=1}^{n} x_{1,i}y_i - n\overline{x}_1\overline{y}.$$

Estimations ponctuelles

Les formules analytiques de $\widehat{\beta}_1$ et $\widehat{\beta}_0$ donnent les estimations ponctuelles suivantes.

- Une estimation ponctuelle de β_1 est la réalisation de $\widehat{\beta}_1$:

$$b_1 = \frac{1}{\sum_{i=1}^{n}(x_{1,i} - \overline{x}_1)^2} \sum_{i=1}^{n}(x_{1,i} - \overline{x}_1)(y_i - \overline{y}) = \frac{\text{spe}_{x,y}}{\text{sce}_x}.$$

  On dit que b_1 est l'*emco* ponctuel de β_1.

- Une estimation ponctuelle de β_0 est la réalisation de $\widehat{\beta}_0$:

$$b_0 = \overline{y} - b_1\overline{x}_1.$$

  On dit que b_0 est l'*emco* ponctuel de β_0.

- Une estimation ponctuelle de $y_x = \beta_0 + \beta_1 x_1$ est la réalisation de $\widehat{Y}_x = \widehat{\beta}_0 + \widehat{\beta}_1 x_1$:

$$d_x = b_0 + b_1 x_1.$$

  On dit que d_x est la valeur prédite de Y quand $X_1 = x_1$.

Droite de régression

On appelle droite de régression la droite qui ajuste au mieux le nuage de points. Cet ajustement se fait en termes de distance euclidienne, les points de la droite étant pris aux mêmes abscisses que ceux des points du nuage. La droite de régression est donnée par l'équation :

$$y = b_0 + b_1 x.$$

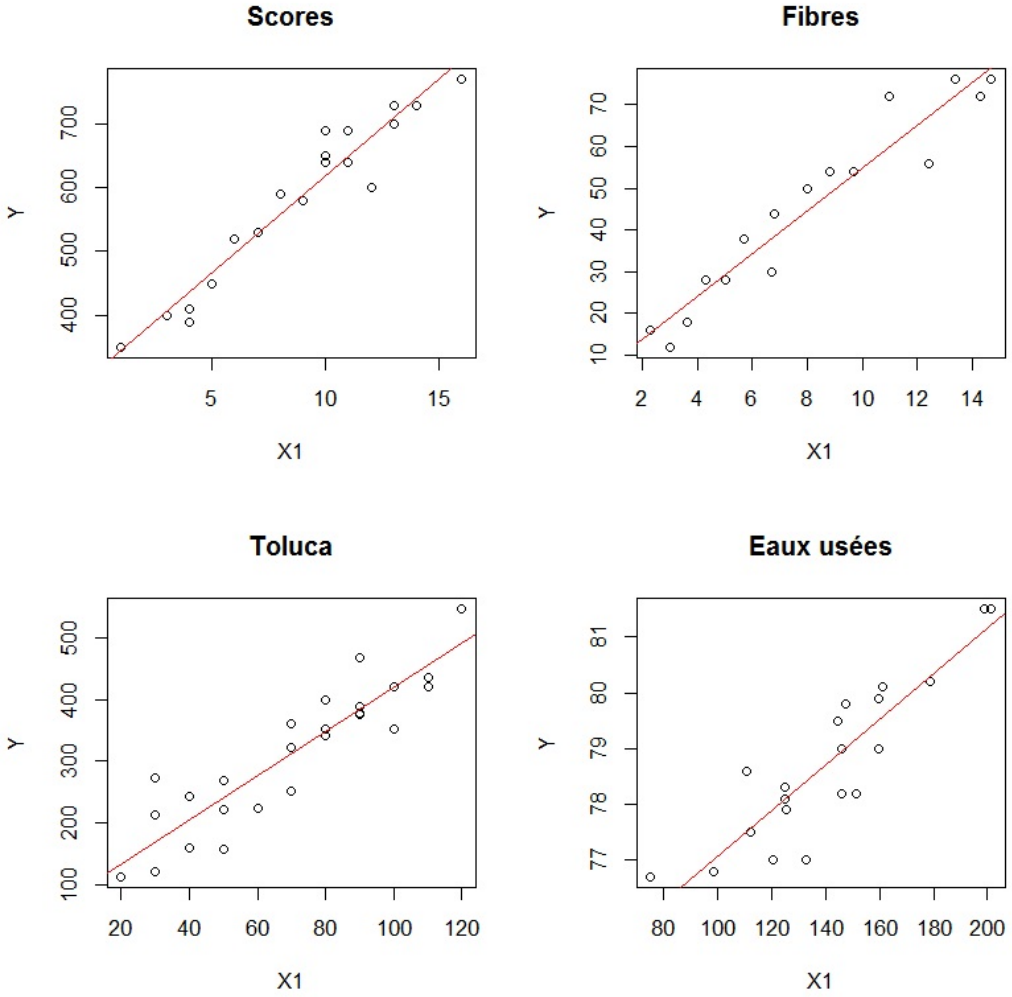
Comme $b_0 = \overline{y} - b_1\overline{x}_1$, notons que la droite de régression passe par le point G de coordonnée $(\overline{x}_1, \overline{y})$, appelé point moyen, centre d'inertie ou centre de gravité du nuage de points.

Remarque : Des méthodes autres que celle des moindres carrés existent pour ajuster un nuage de points. Certaines sont décrites ici :

`http://www.math.unicaen.fr/~chesneau/ajustement.pdf`

Exemples : droites de régression

En reprenant les exemples introduits précédemment, les droites de régressions sont représentées ci-dessous :



Coefficient de corrélation linéaire

On appelle coefficient de corrélation linéaire le réel $r_{x,y}$ défini par

$$r_{x,y} = \frac{\mathrm{spe}_{x,y}}{\sqrt{\mathrm{sce}_x \mathrm{sce}_y}}.$$

On a $r_{x,y} \in [-1, 1]$

Droite de régression et coefficient de corrélation linéaire

On a

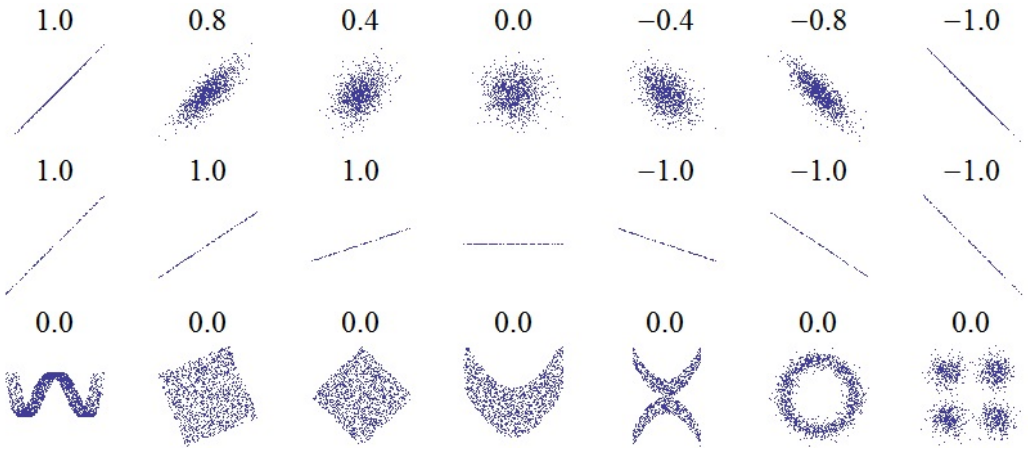
$$b_1 = \frac{s_y}{s_x} r_{x,y}.$$

Comme $s_x > 0$ et $s_y > 0$, le coefficient directeur b_1 de la droite de régression et $r_{x,y}$ sont de même signe (à une droite de régression croissante correspond un $r_{x,y}$ positif...). Dès lors, on peut deviner le signe de $r_{x,y}$ avec la silhouette du nuage de points.

Plus $|r_{x,y}|$ est proche de 1, plus la liaison linéaire entre Y et X_1 est forte.

En effet, plus $|r_{x,y}|$ est proche 1, plus b_1 diffère de 0, plus β_1 diffère de 0, plus la liaison linéaire entre Y et X_1 est forte. Aussi, plus $|r_{x,y}|$ est proche 1, plus X_1 influe sur/est corrélée avec Y.

Le graphique suivant illustre le lien existant entre la pertinence de l'ajustement d'un nuage de points par une droite, caractérisée par la corrélation linéaire entre Y et X_1, et la valeur associée de $r_{x,y}$:



Source du graphique :

`https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient`

Coefficient de détermination et coefficient de corrélation linéaire

Dans le cas du modèle de *rls*, on peut montrer que

$$R^2 = r_{x,y}^2.$$

Dans ce cas, l'interprétation des valeurs de R^2 et $r_{x,y}^2$ est donc identique.

Mise en œuvre avec le logiciel R

Pour illustrer le résultat théorique précédent, on peut considérer le jeu de données "loyers". Dans un quartier parisien, une étude a été menée afin de mettre en évidence une relation entre le loyer mensuel et la surface des appartements ayant exactement 3 pièces.

Pour 30 appartements de ce type, on dispose :

- de la surface en mètres carrés (variable X_1),
- du loyer mensuel en francs (variable Y).

L'objectif est d'expliquer Y à partir de X_1.

Le jeu de données est disponible ici :

```
http://www.math.unicaen.fr/~chesneau/loyers.txt
```

Écrire dans une fenêtre R :

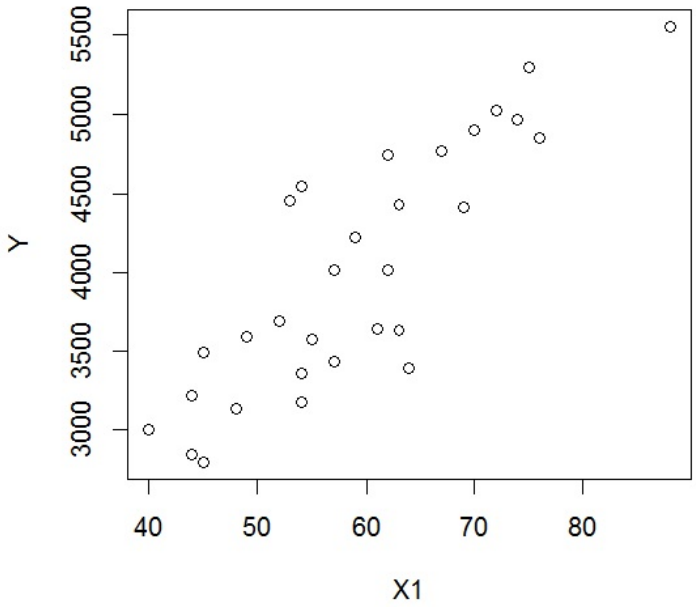
```
w = read.table("http://www.math.unicaen.fr/~chesneau/loyers.txt",
header = T)
attach(w)
head(w)
```

Cela renvoie l'entête du jeu de données :

	Y	X1
1	3000	40
2	2844	44
3	3215	44
4	2800	45
5	3493	45
6	3140	48

Le nuage de points associé est donné par les commandes R :

```
plot(X1, Y)
```



Le nuage de points étant étiré dans une direction, le modèle de *rls* est envisageable. Sa forme générique est

$$Y = \beta_0 + \beta_1 X_1 + \epsilon,$$

où β_0 et β_1 sont des coefficients réels inconnus.

Pour estimer ponctuellement β_0 et β_1, nous allons utiliser les formules analytiques de b_0 et b_1 :

```
b1 = (1 / (sum((X1 - mean(X1))^2))) * sum((X1 - mean(X1)) * (Y - mean(Y)))
b0 = mean(Y) - mean(X1) * b1
b0
b1
```

Cela renvoie : $b_0 = 548.9782$ et $b_1 = 58.37875$.

On peut calculer le R^2 en utilisant l'égalité : $R^2 = r^2_{x,y}$:

```
R2 = cor(Y, X1)^2
```

Cela renvoie : 0.7311242.

De même pour le R^2 ajusté :

```
R2aj = 1 - ((30 - 1)/(30 - (2 + 1))) * (1 - R2)

R2aj
```

Cela renvoie : 0.6599716.

Le R^2 (et $\overline{R}^2$) étant proche de 1, le modèle de *rls* semble être pertinent avec les données traitées.

Commande `summary` :

On retrouve plus simplement ces estimations (et beaucoup plus) avec la commande `summary` :

```
reg = lm(Y ~ X1)

summary(reg)
```

Cela renvoie :

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	**548.9782**	403.0783	1.36	0.1841	
X1	**58.3787**	6.6905	8.73	0.0000	***

```
Residual standard error: 409.7 on 28 degrees of freedom

Multiple R-squared: 0.7311, Adjusted R-squared: 0.7215

F-statistic: 76.14 on 1 and 28 DF, p-value: 1.783e-09
```

On retrouve b_0 et b_1 dans la colonne `Estimate` du tableau.

On retrouve également : $R^2 = 0.7311$ et $\overline{R}^2 = 0.7215$.

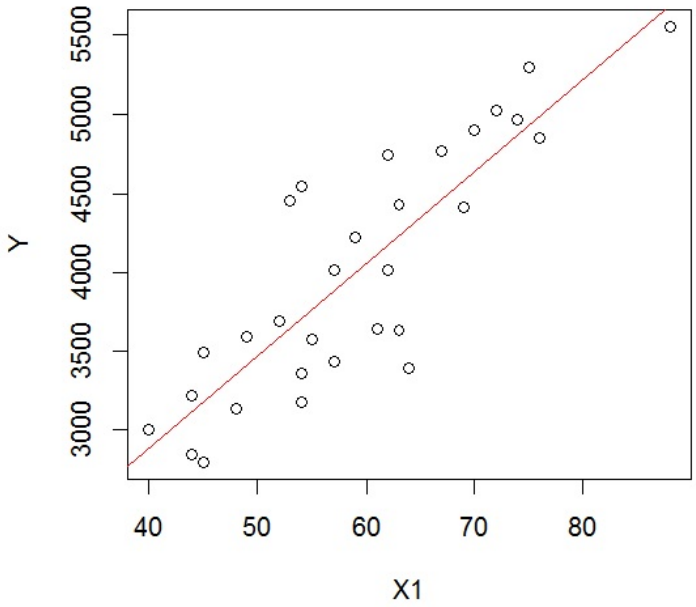
D'autre part, la droite de régression est donnée par l'équation :

$$y = b_0 + b_1 x = 548.9782 + 58.3787x.$$

On peut la visualiser en faisant :

```
plot(X1, Y)

abline(reg, col = "red")
```



On constate que cette droite ajuste correctement le nuage de points ; les prédictions issues du modèle sont alors relativement fiables.

Par exemple, pour $X_1 = 56 = x$, la valeur prédite de Y est

$$d_x = b_0 + b_1 \times 56 = 548.9782 + 58.3787 \times 56 = 3818.185.$$

Ainsi, pour une surface de 56 mètres carrés, le loyer mensuel moyen est de 3818.185 francs.

On aurait aussi pu utiliser les commandes R :

```
predict(reg, data.frame(X1 = 56))
```

Dorénavant, dès que possible, on utilisera la commande `summary` dans les analyses.

3 Loi normale multidimensionelle

Vecteur gaussien

Soient $n \in \mathbb{N}^*$ et $U_1, \ldots, U_n$ *n var.* On dit que $U = \begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix}$ est un vecteur gaussien si et seulement si toute combinaison linéaire de $U_1, \ldots, U_n$ suit une loi normale : pour tout $(a_1, \ldots, a_n) \in \mathbb{R}^n$,

$$a_1 U_1 + \ldots + a_n U_n \sim \mathcal{N}.$$

Un vecteur gaussien est caractérisé par son espérance μ et sa matrice de covariance Σ.

La loi de U est la loi normale multidimensionnelle notée $\mathcal{N}_n(\mu, \Sigma)$.

Critère d'indépendance

Soient $U = \begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix}$ un vecteur gaussien et $(j, k) \in \{1, \ldots, n\}^2$ avec $j \neq k$. Alors U_j et U_k sont indépendantes si et seulement si $\mathbb{C}(U_j, U_k) = 0$.

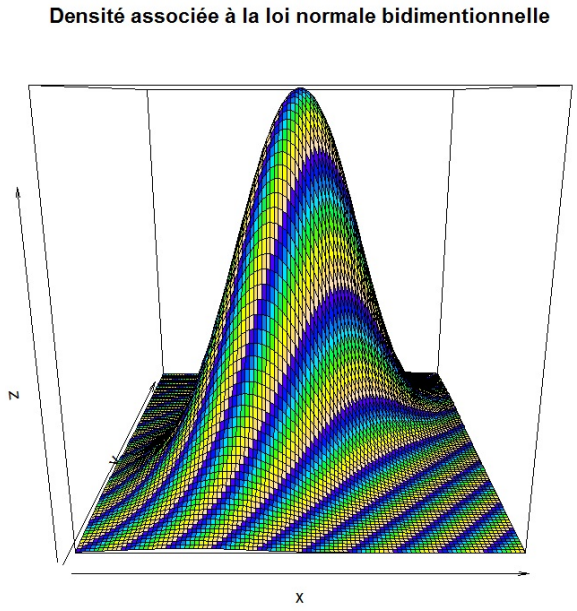
Loi normale multidimensionnelle

Soient $\mu \in \mathbb{R}^n$ et Σ une matrice de dimension $n \times n$ symétrique définie positive vérifiant $\det(\Sigma) > 0$. Alors $U \sim \mathcal{N}_n(\mu, \Sigma)$ si et seulement si U possède la densité :

$$f(x) = \frac{1}{(2\pi)^{\frac{n}{2}} \sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^t \Sigma^{-1} (x-\mu)\right), \qquad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n.$$

Représentation graphique

Une densité associée à la loi $\mathcal{N}_2(0_2, \Sigma)$, avec $\Sigma = \begin{pmatrix} 0.5 & 1 \\ 1 & 0.5 \end{pmatrix}$ est présentée ci-dessous :



Notations

Soient $U = \begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix}$ et $V = \begin{pmatrix} V_1 \\ \vdots \\ V_n \end{pmatrix}$ des vecteurs de *var*, et $W = \begin{pmatrix} W_{1,1} & W_{2,1} & \dots & W_{n,1} \\ W_{1,2} & W_{2,2} & \dots & W_{n,2} \\ \vdots & \vdots & & \vdots \\ W_{1,n} & W_{2,n} & \dots & W_{n,n} \end{pmatrix}$ une matrice de *var*. On adopte les notations :

- Espérance de U : $\mathbb{E}_n(U) = \begin{pmatrix} \mathbb{E}(U_1) \\ \vdots \\ \mathbb{E}(U_n) \end{pmatrix}$.
- Espérance de W : $\mathbb{E}_{n,n}(W) = \begin{pmatrix} \mathbb{E}(W_{1,1}) & \mathbb{E}(W_{2,1}) & \dots & \mathbb{E}(W_{n,1}) \\ \mathbb{E}(W_{1,2}) & \mathbb{E}(W_{2,2}) & \dots & \mathbb{E}(W_{n,2}) \\ \vdots & \vdots & & \vdots \\ \mathbb{E}(W_{1,n}) & \mathbb{E}(W_{2,n}) & \dots & \mathbb{E}(W_{n,n}) \end{pmatrix}$.

- Covariance de U et V :

$$\mathbb{C}_n(U,V) = \mathbb{E}_{n,n}\left((U-\mathbb{E}_n(U))(V-\mathbb{E}_n(V))^t\right) = \begin{pmatrix} \mathbb{C}(U_1,V_1) & \mathbb{C}(U_1,V_2) & \dots & \mathbb{C}(U_1,V_n) \\ \mathbb{C}(U_2,V_1) & \mathbb{C}(U_2,V_2) & \dots & \mathbb{C}(U_2,V_n) \\ \vdots & \vdots & & \vdots \\ \mathbb{C}(U_n,V_1) & \mathbb{C}(U_n,V_2) & \dots & \mathbb{C}(U_n,V_n) \end{pmatrix}.$$

- Matrice de covariance de U : $\mathbb{V}_n(U) = \mathbb{C}_n(U,U) = \mathbb{E}_{n,n}\left((U-\mathbb{E}_n(U))(U-\mathbb{E}_n(U))^t\right)$.

Paramètres

Si $U \sim \mathcal{N}_n(\mu,\Sigma)$, alors $\mathbb{E}_n(U) = \mu$ et $\mathbb{V}_n(U) = \Sigma$.

Forme linéaire

Soient $X \sim \mathcal{N}_n(\mu,\Sigma)$ et A une matrice à n lignes et p colonnes avec $p \leq n$ et a un vecteur colonne à n composantes. Alors on a

$$AX + a \sim \mathcal{N}_n(A\mu + a, A\Sigma A^t).$$

En particulier, si $X \sim \mathcal{N}_n(\mu,\Sigma)$, on a $\mathbb{E}_n(AX+a) = A\mu + a$ et $\mathbb{V}_n(AX+a) = A\Sigma A^t$.

Vecteurs gaussiens et indépendance

Soient $X \sim \mathcal{N}_n(\mu,\Sigma)$, A une matrice à n lignes et p colonnes avec $p \leq n$ et B une matrice à n lignes et q colonnes avec $q \leq n$. Alors AX et BX sont indépendantes si et seulement si $A\Sigma B^t = 0_{p,q}$.

Ainsi, les composantes de tout sous vecteur de X sont indépendantes si et seulement si leurs covariances sont nulles.

4 Propriétés standards et lois associées

Hypothèses standards

On considère le modèle de *rlm* sous la forme matricielle : $Y = X\beta + \epsilon$. On suppose que

- X est de rang colonnes plein,
- ϵ et $X_1, \ldots, X_p$ sont indépendantes,
- $\epsilon \sim \mathcal{N}_n(0_n, \sigma^2 \mathbb{I}_n)$ où $\sigma > 0$ est un paramètre inconnu.

L'hypothèse $\epsilon \sim \mathcal{N}_n(0_n, \sigma^2 \mathbb{I}_n)$ entraîne que $\epsilon_1, \ldots, \epsilon_n$ sont indépendantes et identiquement distribuées de loi commune la loi normale $\mathcal{N}(0, \sigma^2)$.

Les hypothèses standards sont à la base d'une analyse statistique avancée avec le modèle de *rlm*.

Dorénavant, on suppose que les hypothèses standards sont satisfaites.

Loi de Y

On a

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

Preuve : On peut écrire $Y = a + \epsilon$, où $a = X\beta$ est un vecteur colonne à n composantes constantes. Comme $\epsilon \sim \mathcal{N}_n(0_n, \sigma^2 \mathbb{I}_n)$, on a $a + \epsilon \sim \mathcal{N}_n \left(a + 0_n, \sigma^2 \mathbb{I}_n \right)$, ce qui entraîne $Y \sim \mathcal{N}_n \left(X\beta, \sigma^2 \mathbb{I}_n \right)$.

□

Loi de $\widehat{\beta}$

On a

$$\widehat{\beta} \sim \mathcal{N}_{p+1} \left(\beta, \sigma^2 (X^t X)^{-1} \right).$$

Les conséquences immédiates de ce résultats sont :

- $\widehat{\beta}$ est un estimateur sans biais de β : $\mathbb{E}_{p+1}(\widehat{\beta}) = \beta$,
- la matrice de covariance de $\widehat{\beta}$ est $\sigma^2 (X^t X)^{-1}$: $\mathbb{V}_{p+1}(\widehat{\beta}) = \sigma^2 (X^t X)^{-1}$,
- en notant $[(X^t X)^{-1}]_{j+1,j+1}$ la $j+1$-ème composante diagonale de $(X^t X)^{-1}$, on a $\widehat{\beta}_j \sim \mathcal{N} \left(\beta_j, \sigma^2 [(X^t X)^{-1}]_{j+1,j+1} \right)$,
- si la $(j+1, k+1)$-ème composante de $(X^t X)^{-1}$ est nulle, alors $\widehat{\beta}_j$ et $\widehat{\beta}_k$ sont indépendantes.

Preuve : On a $\widehat{\beta} = (X^tX)^{-1}X^tY$. Ainsi, on peut écrire $\widehat{\beta} = AY$, où $A = (X^tX)^{-1}X^t$ est une matrice à composantes constantes. Comme $Y \sim \mathcal{N}_n\left(X\beta, \sigma^2\mathbb{I}_n\right)$, il vient

$$\widehat{\beta} \sim \mathcal{N}_{p+1}\left(AX\beta, A(\sigma^2\mathbb{I}_n)A^t\right).$$

En remarquant que $(X^tX)^{-1}X^tX = \mathbb{I}_n$, on a $AX\beta = (X^tX)^{-1}X^tX\beta = \mathbb{I}_n\beta = \beta$.

D'autre part, en utilisant les opérations matricielles : $(CD)^t = D^tC^t$, $(C^t)^t = C$ et $(C^{-1})^t = (C^t)^{-1}$, on a

$$\begin{aligned} A(\sigma^2\mathbb{I}_n)A^t &= \sigma^2AA^t = \sigma^2(X^tX)^{-1}X^t((X^tX)^{-1}X^t)^t = \sigma^2(X^tX)^{-1}X^t(X^t)^t(X^t(X^t)^t)^{-1} \\ &= \sigma^2(X^tX)^{-1}X^tX(X^tX)^{-1} = \sigma^2(X^tX)^{-1}\mathbb{I}_n = \sigma^2(X^tX)^{-1}. \end{aligned}$$

D'où $\widehat{\beta} \sim \mathcal{N}_{p+1}\left(\beta, \sigma^2(X^tX)^{-1}\right)$.

□

Conséquence du théorème de Gauss-Markov

L'*emco* $\widehat{\beta}$ est le meilleur estimateur linéaire sans biais de β ; c'est le BLUE (Best Linear Unbiased Estimator) : aucun autre estimateur linéaire sans biais de β n'a une variance plus petite.

Lien avec l'estimateur du maximum de vraisemblance (*emv*)

L'*emco* $\widehat{\beta}$ est l'*emv* de β. Dès lors, il est fortement consistant et asymptotiquement efficace.

Preuve : Comme $Y \sim \mathcal{N}_n\left(X\beta, \sigma^2\mathbb{I}_n\right)$, la fonction de vraisemblance associée à $(Y_1, \ldots, Y_n)$ est donnée par :

$$L(\beta, z) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{\|z - X\beta\|^2}{2\sigma^2}\right), \qquad z \in \mathbb{R}^n.$$

Soit $\widetilde{\beta}$ l'*emv* de β : $\widetilde{\beta} = \underset{\beta\in\mathbb{R}^{p+1}}{\operatorname{argmax}} L(\beta, Y)$. Par croissance de la fonction exponentielle, on a

$$\begin{aligned} \widetilde{\beta} &= \underset{\beta\in\mathbb{R}^{p+1}}{\operatorname{argmax}} L(\beta, Y) = \underset{\beta\in\mathbb{R}^{p+1}}{\operatorname{argmax}} \left(\frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{\|Y - X\beta\|^2}{2\sigma^2}\right)\right) \\ &= \underset{\beta\in\mathbb{R}^{p+1}}{\operatorname{argmax}} \left(-\frac{\|Y - X\beta\|^2}{2\sigma^2}\right) = \underset{\beta\in\mathbb{R}^{p+1}}{\operatorname{argmin}} \|Y - X\beta\|^2 = \widehat{\beta}. \end{aligned}$$

□

Estimateur sans biais de σ^2

Un estimateur de σ^2 est

$$\widehat{\sigma}^2 = \frac{1}{n-(p+1)}\|Y - X\widehat{\beta}\|^2.$$

On a :

- $\widehat{\sigma}^2$ est sans biais pour σ^2 : $\mathbb{E}(\widehat{\sigma}^2) = \sigma^2$,
- $\widehat{\sigma}^2$ et $\widehat{\beta}$ sont indépendantes,
- $(n-(p+1))\dfrac{\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(\nu), \qquad \nu = n-(p+1).$

Éléments de preuve : Soit L le sous-espace vectoriel de $\mathbb{R}^n$ engendré par les colonnes de X. On peut montrer que $\mathbb{I}_n - X(X^tX)^{-1}X^t$ est la matrice de projection sur l'orthogonal de L noté $L^\perp$. Ce sous-espace est de dimension $n-(p+1)$: $\text{Dim}(L^\perp) = n-(p+1)$.

- On peut montrer que $Y - X\widehat{\beta} \sim \mathcal{N}_n\left(0_n, \sigma^2(\mathbb{I}_n - X(X^tX)^{-1}X^t)\right)$. Comme la trace d'une matrice de projection est égale à la dimension de l'image de la projection, on a

$$\begin{aligned}\mathbb{E}(\|Y - X\widehat{\beta}\|^2) &= \mathbb{E}\left(\text{Trace}\left((Y - X\widehat{\beta})(Y - X\widehat{\beta})^t\right)\right) = \text{Trace}\left(\mathbb{E}_{n,n}\left((Y - X\widehat{\beta})(Y - X\widehat{\beta})^t\right)\right)\\ &= \sigma^2\,\text{Trace}\left(\mathbb{I}_n - X(X^tX)^{-1}X^t\right) = \sigma^2\,\text{Dim}(L^\perp) = \sigma^2(n-(p+1)).\end{aligned}$$

  Donc $\mathbb{E}(\widehat{\sigma}^2) = \sigma^2$.

- On peut montrer que le vecteur aléatoire réel $(Y - X\widehat{\beta}, \widehat{\beta})$ est gaussien et que toutes les covariances d'une composante de $Y - X\widehat{\beta}$ et d'une composante de $\widehat{\beta}$ sont nulles. Cela entraîne l'indépendance de $Y - X\widehat{\beta}$ et $\widehat{\beta}$. Comme $\widehat{\sigma}^2$ est uniquement fonction de $\widehat{\epsilon}$, on a aussi l'indépendance de $\widehat{\sigma}^2$ et $\widehat{\beta}$.

- On peut écrire :

$$(n-(p+1))\frac{\widehat{\sigma}^2}{\sigma^2} = \|(\mathbb{I}_n - X(X^tX)^{-1}X^t)\left(\frac{\epsilon}{\sigma}\right)\|^2.$$

  Comme $\dfrac{\epsilon}{\sigma} \sim \mathcal{N}_n(0_n, \mathbb{I}_n)$ et $\mathbb{I}_n - X(X^tX)^{-1}X^t$ est la matrice de projection sur $L^\perp$ avec $\text{Dim}(L^\perp) = n-(p+1)$, le théorème de Cochran entraîne

$$(n-(p+1))\frac{\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(\nu), \qquad \nu = \text{Dim}(L^\perp) = n-(p+1).$$

□

Degrés de liberté

Dorénavant, ν désigne le nombre de degrés de liberté : $\nu = n-(p+1)$.

Emco et loi de Student

Pour tout vecteur ligne c à $p+1$ composantes, on a

$$\frac{c\widehat{\beta} - c\beta}{\widehat{\sigma}\sqrt{c(X^tX)^{-1}c^t}} \sim \mathcal{T}(\nu).$$

Preuve : Dans un premier temps, rappelons une caractérisation de la loi de Student. Soient A et B deux *var* indépendantes avec $A \sim \mathcal{N}(0,1)$ et $B \sim \chi^2(\nu)$, alors $T = \frac{A}{\sqrt{\frac{B}{\nu}}} \sim \mathcal{T}(\nu)$. On pose alors :

$$A = \frac{c\widehat{\beta} - c\beta}{\sigma\sqrt{c(X^tX)^{-1}c^t}}, \qquad B = (n-(p+1))\frac{\widehat{\sigma}^2}{\sigma^2}.$$

Comme $\widehat{\beta}$ et $\widehat{\sigma}^2$ sont indépendantes, il en est de même pour A et B. Comme $\widehat{\beta} \sim \mathcal{N}_{p+1}\left(\beta, \sigma^2(X^tX)^{-1}\right)$, on a $c\widehat{\beta} \sim \mathcal{N}\left(c\beta, \sigma^2 c(X^tX)^{-1}c^t\right)$, ce qui entraîne $A \sim \mathcal{N}(0,1)$. De plus, on a $B \sim \chi^2(\nu)$. Par la caractérisation de la loi de Student, il s'ensuit

$$\frac{c\widehat{\beta} - c\beta}{\widehat{\sigma}\sqrt{c(X^tX)^{-1}c^t}} = \frac{A}{\sqrt{\frac{B}{n-(p+1)}}} \sim \mathcal{T}(\nu).$$

□

Emco et loi de Student ; suite

- Pour tout $j \in \{0, \ldots, p\}$, on a

$$\frac{\widehat{\beta}_j - \beta_j}{\widehat{\sigma}\sqrt{[(X^tX)^{-1}]_{j+1,j+1}}} \sim \mathcal{T}(\nu).$$

- Soient $x_\bullet = (1, x_1, \ldots, x_p)$, $y_x = x_\bullet\beta$ et $\widehat{Y}_x = x_\bullet\widehat{\beta}$. On a

$$\frac{\widehat{Y}_x - y_x}{\widehat{\sigma}\sqrt{x_\bullet(X^tX)^{-1}x_\bullet^t}} \sim \mathcal{T}(\nu).$$

Preuve : On peut utiliser le résultat :

$$\frac{c\widehat{\beta} - c\beta}{\widehat{\sigma}\sqrt{c(X^tX)^{-1}c^t}} \sim \mathcal{T}(\nu).$$

On obtient le premier point en prenant $c = c_j$ le vecteur ligne à $p+1$ composantes, toutes nulles sauf la $j+1$-ème qui vaut 1. On obtient le deuxième point en prenant $c = x_\bullet$.

□

Emco et loi de Fisher

Soit Q une matrice de réels à $p+1$ colonnes et k lignes de rang colonnes plein. Alors on a

$$\frac{(Q\widehat{\beta}-Q\beta)^t(Q(X^tX)^{-1}Q^t)^{-1}(Q\widehat{\beta}-Q\beta)}{k\widehat{\sigma}^2} \sim \mathcal{F}(k,\nu).$$

Par exemple, avec $\widehat{\beta}=\begin{pmatrix}\widehat{\beta}_0\\ \widehat{\beta}_1\\ \widehat{\beta}_2\end{pmatrix}$ *et* $Q=\begin{pmatrix}4 & 1 & 0\\ 0 & 2 & -5\end{pmatrix}$, *on a* $Q\widehat{\beta}=\begin{pmatrix}4\widehat{\beta}_0+\widehat{\beta}_1\\ 2\widehat{\beta}_1-5\widehat{\beta}_2\end{pmatrix}$.

Éléments de preuve : Dans un premier temps, rappelons une caractérisation de la loi de Fisher. Soient A et B deux *var* indépendantes avec $A\sim\chi^2(\nu_1)$ et $B\sim\chi^2(\nu_2)$, alors $F=\frac{\nu_2 A}{\nu_1 B}\sim\mathcal{F}(\nu_1,\nu_2)$. On pose alors :

$$A=\frac{(Q\widehat{\beta}-Q\beta)^t(Q(X^tX)^{-1}Q^t)^{-1}(Q\widehat{\beta}-Q\beta)}{\sigma^2}, \qquad B=\nu\frac{\widehat{\sigma}^2}{\sigma^2}.$$

En utilisant le théorème de Cochran, on peut montrer que A et B sont indépendantes avec $A\sim\chi^2(k)$ et $B\sim\chi^2(\nu)$. Par la caractérisation de la loi de Fisher, il s'ensuit

$$\frac{(Q\widehat{\beta}-Q\beta)^t(Q(X^tX)^{-1}Q^t)^{-1}(Q\widehat{\beta}-Q\beta)}{k\widehat{\sigma}^2}=\frac{\nu A}{kB}\sim\mathcal{F}(k,\nu).$$

□

Estimation ponctuelles

- Une estimation ponctuelle de σ est la réalisation de $\widehat{\sigma}$:

$$s=\sqrt{\frac{1}{n-(p+1)}\|y-Xb\|^2}.$$

- Pour tout $j\in\{0,\dots,p\}$, une estimation ponctuelle de l'écart-type de $\widehat{\beta}_j$ est

$$\text{ete}_j=s\sqrt{[(X^tX)^{-1}]_{j+1,j+1}}.$$

- Soit $x_\bullet=(1,x_1,\dots,x_p)$. Une estimation ponctuelle de l'écart-type de $\widehat{Y}_x=x_\bullet\widehat{\beta}$ est

$$\text{ete}_x=s\sqrt{x_\bullet(X^tX)^{-1}x_\bullet^t}.$$

Les hypothèses standards en pratique

En pratique, pour admettre que les hypothèses standards sont acceptables à partir des données, il y a un protocole à suivre. Notamment, il faut analyser plusieurs graphiques spécifiques (graphique des résidus, QQ plot, graphique Scale-Location, *acf*, *pacf*...) et mettre en œuvre plusieurs tests statistiques (test de Shapiro-Wilk, test de Rainbow, test de Durbin-Watson...) *(plus de détails en Master 2)*.

Dans ce document, on se focalise sur le principal repère visuel : le graphique des résidus.

Résidus

Pour tout $i \in \{1, \dots, n\}$, on appelle i-ème résidu la réalisation e_i de

$$\widehat{\epsilon}_i = Y_i - \widehat{Y}_i,$$

où $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1,i} + \ldots + \widehat{\beta}_p x_{p,i}$.

On appelle résidus les réels $e_1, \dots, e_n$.

Avec les notations déjà introduites, on peut écrire :

$$e_i = y_i - d_{x_i},$$

avec $x_i = (x_{1,i}, \dots, x_{p,i})$.

Graphique des résidus

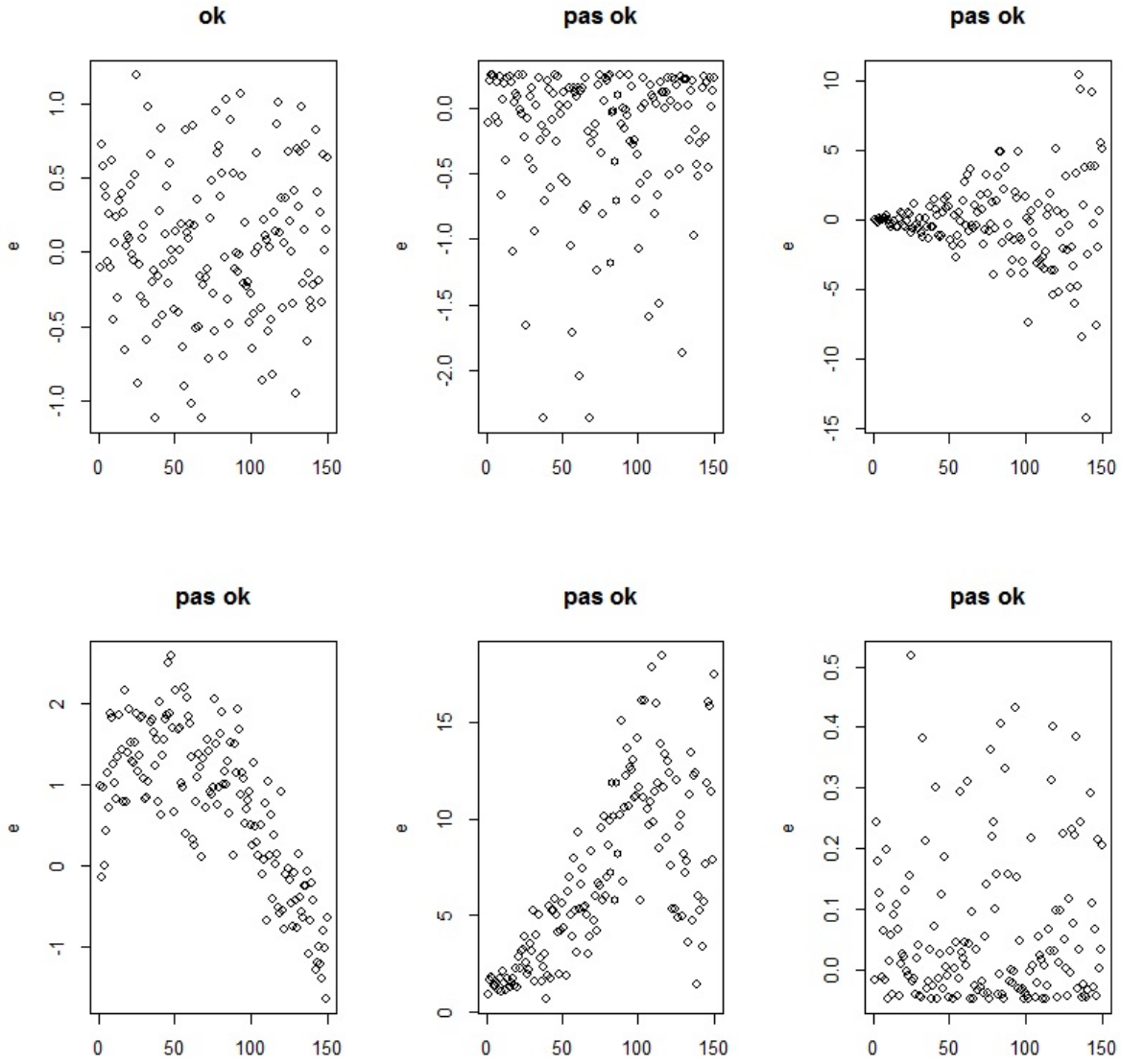
Ainsi, $\begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$ est la réalisation de $\begin{pmatrix} \widehat{\epsilon}_1 \\ \vdots \\ \widehat{\epsilon}_n \end{pmatrix}$, lequel est un estimateur grossier de ϵ. Donc, sous les hypothèses standards, $\begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$ devrait avoir les caractéristiques grossières d'une réalisation de $\mathcal{N}_n(0_n, \sigma^2 \mathbb{I}_n)$.

On trace alors le nuage de points : $\mathcal{N}_e = \{(1, e_1), (2, e_2), \dots, (n, e_n)\}$.

Si le nuage de points n'a aucune structure particulière, et s'il y a une symétrie dans la répartition des points par rapport à l'axe des abscisses, alors on admet que $\epsilon \sim \mathcal{N}_n(0_n, \sigma^2 \mathbb{I}_n)$.

Exemples de graphiques des résidus

Des exemples de graphiques des résidus sont proposés ci-dessous ; seul le premier colle avec les hypothèses standards.



Mise en œuvre avec le logiciel R

On reprend le jeu de données "profs". Dans une étude statistique, 23 professeurs sont évalués quant à la qualité de leur enseignement. Pour chacun d'entre eux, on dispose :

- d'un indice de performance globale donné par les étudiants (variable Y),
- des résultats de 4 tests écrits donnés à chaque professeur (variables X_1, X_2, X_3 et X_4),
- du sexe (variable X_5, avec $X_5 = 0$ pour femme, $X_5 = 1$ pour homme).

L'objectif est d'expliquer Y à partir de X_1, X_2, X_3, X_4 et X_5. On enregistre les données dans R :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/profs.txt", header = T)

attach(w)
```

Le modèle de *rlm* est envisageable. Sa forme générique est

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon,$$

où β_0, β_1, β_2, β_3, β_4 et β_5 sont des coefficients réels inconnus.

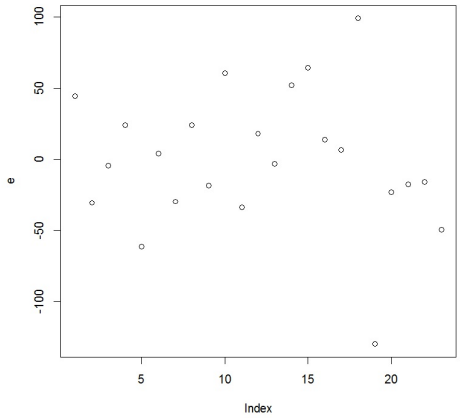
On obtient les *emco* ponctuels de ces coefficients en faisant :

```
reg = lm(Y ~ X1 + X2 + X3 + X4 + X5)
```

Les commandes R pour visualiser le graphique des résidus sont :

```
e = residuals(reg)

plot(e)
```



Globalement, à part un point légèrement excentré en bas à droite (qu'il faudrait analyser), le graphique des résidus est colle avec les hypothèses standards.

D'autre part, plusieurs estimations ponctuelles sont directement données par la commande `summary` :

```
summary(reg)
```

Cela renvoie :

	Estimate	**Std. Error**	t value	Pr(>|t|)	
(Intercept)	-272.0388	**184.3865**	-1.48	0.1584	
X1	0.7913	**0.5363**	1.48	0.1583	
X2	2.6828	**0.9216**	2.91	0.0097	**
X3	-1.4434	**0.8217**	-1.76	0.0970	.
X4	6.8308	**1.8192**	3.75	0.0016	**
X5	14.9008	**27.3134**	0.55	0.5925	

```
Residual standard error: 55.06 on 17 degrees of freedom
Multiple R-squared: 0.6834, Adjusted R-squared: 0.5903
F-statistic: 7.34 on 5 and 17 DF, p-value: 0.0007887
```

On retrouve $(\text{ete}_0, \ldots, \text{ete}_5)$ dans la colonne `Std. Error` du tableau :

$$\text{ete}_0 = 184.3865, \quad \text{ete}_1 = 0.5363, \quad \text{ete}_2 = 0.9216, \quad \text{ete}_3 = 0.8217,$$

$$\text{ete}_4 = 1.8192, \quad \text{ete}_5 = 27.3134.$$

On a également le s avec `Residual standard error` : $s = 55.06$ et ν avec `degrees of freedom` : $\nu = 17$.

5 Retour sur le modèle de *rls*

Propriétés de $\widehat{\beta}$

On a

$$\widehat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \sigma^2 \frac{1}{\text{sce}_x}\right), \qquad \widehat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}_1^2}{\text{sce}_x}\right)\right).$$

Preuve : Tout repose sur le résultat : pour tout $j \in \{0,1\}$, on a $\widehat{\beta}_j \sim \mathcal{N}\left(\beta_j, \sigma^2[(X^tX)^{-1}]_{j+1,j+1}\right)$. Il reste à expliciter $[(X^tX)^{-1}]_{2,2}$ et $[(X^tX)^{-1}]_{1,1}$. On a

$$X^tX = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_{1,1} & x_{1,2} & \dots & x_{1,n} \end{pmatrix} \begin{pmatrix} 1 & x_{1,1} \\ 1 & x_{1,2} \\ \vdots & \vdots \\ 1 & x_{1,n} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n 1 & \sum_{i=1}^n x_{1,i} \\ \sum_{i=1}^n x_{1,i} & \sum_{i=1}^n x_{1,i}^2 \end{pmatrix} = \begin{pmatrix} n & n\overline{x}_1 \\ n\overline{x}_1 & \sum_{i=1}^n x_{1,i}^2 \end{pmatrix}.$$

En inversant X^tX et en utilisant la décomposition : $\text{sce}_x = \sum_{i=1}^n x_{1,i}^2 - n\overline{x}_1^2$, il vient

$$(X^tX)^{-1} = \frac{1}{n\sum_{i=1}^n x_{1,i}^2 - (n\overline{x}_1)^2} \begin{pmatrix} \sum_{i=1}^n x_{1,i}^2 & -n\overline{x}_1 \\ -n\overline{x}_1 & n \end{pmatrix} = \frac{1}{\text{sce}_x} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_{1,i}^2 & -\overline{x}_1 \\ -\overline{x}_1 & 1 \end{pmatrix}.$$

En identifiant les composantes diagonales de $(X^tX)^{-1}$, on obtient

$$[(X^tX)^{-1}]_{2,2} = \frac{1}{\text{sce}_x}, \qquad [(X^tX)^{-1}]_{1,1} = \frac{1}{\text{sce}_x} \times \frac{1}{n}\sum_{i=1}^n x_{1,i}^2 = \frac{1}{\text{sce}_x}\left(\frac{1}{n}(\text{sce}_x + n\overline{x}_1^2)\right) = \frac{1}{n} + \frac{\overline{x}_1^2}{\text{sce}_x}.$$

On en déduit que

$$\widehat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \sigma^2 \frac{1}{\text{sce}_x}\right), \qquad \widehat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}_1^2}{\text{sce}_x}\right)\right).$$

□

Propriétés de $\widehat{Y}_x$

On a

$$\widehat{Y}_x = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 \sim \mathcal{N}\left(y_x, \sigma^2\left(\frac{1}{n} + \frac{(x_1 - \overline{x}_1)^2}{\text{sce}_x}\right)\right).$$

Preuve : Tout repose sur le résultat : pour $x_\bullet = (1, x_1)$, on a $\widehat{Y}_x \sim \mathcal{N}\left(y_x, \sigma^2 x_\bullet (X^t X)^{-1} x_\bullet^t\right)$. Il reste à expliciter $x_\bullet (X^t X)^{-1} x_\bullet^t$. On a

$$\begin{aligned} x_\bullet (X^t X)^{-1} x_\bullet^t &= \frac{1}{\text{sce}_x} \begin{pmatrix} 1 & x_1 \end{pmatrix} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_{1,i}^2 & -\overline{x}_1 \\ -\overline{x}_1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ x_1 \end{pmatrix} = \frac{1}{\text{sce}_x} \begin{pmatrix} 1 & x_1 \end{pmatrix} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_{1,i}^2 - \overline{x}_1 x_1 \\ -\overline{x}_1 + x_1 \end{pmatrix} \\ &= \frac{1}{\text{sce}_x} \left(\frac{1}{n}\sum_{i=1}^n x_{1,i}^2 - 2\overline{x}_1 x_1 + x_1^2 \right) = \frac{1}{\text{sce}_x} \left(\frac{1}{n}\sum_{i=1}^n x_{1,i}^2 - \overline{x}_1^2 + \overline{x}_1^2 - 2\overline{x}_1 x_1 + x_1^2 \right) \\ &= \frac{1}{\text{sce}_x} \left(\frac{1}{n}\text{sce}_x + (x_1 - \overline{x}_1)^2 \right) = \frac{1}{n} + \frac{(x_1 - \overline{x}_1)^2}{\text{sce}_x}. \end{aligned}$$

On en déduit que

$$\widehat{Y}_x \sim \mathcal{N}\left(y_x, \sigma^2 \left(\frac{1}{n} + \frac{(x_1 - \overline{x}_1)^2}{\text{sce}_x} \right) \right).$$

□

Estimation ponctuelles

- Une estimation ponctuelle de σ est la réalisation de $\widehat{\sigma}$:

$$s = \sqrt{\frac{1}{n-2}\|y - Xb\|^2} = \sqrt{\frac{(n-1)s_y^2(1 - r_{x,y}^2)}{n-2}}.$$

- Une estimation ponctuelle de l'écart-type de $\widehat{\beta}_1$ est

$$\text{ete}_1 = s\sqrt{\frac{1}{\text{sce}_x}}.$$

- Une estimation ponctuelle de l'écart-type de $\widehat{\beta}_0$ est

$$\text{ete}_0 = s\sqrt{\frac{1}{n} + \frac{\overline{x}_1^2}{\text{sce}_x}}.$$

- Une estimation ponctuelle de l'écart-type de $\widehat{Y}_x = \widehat{\beta}_0 + \widehat{\beta}_1 x_1$ est

$$\text{ete}_x = s\sqrt{\frac{1}{n} + \frac{(x_1 - \overline{x}_1)^2}{\text{sce}_x}}.$$

6 Intervalles et volumes de confiance

Intervalle de confiance pour $c\beta$

Pour tout vecteur ligne c à $p+1$ composantes, un intervalle de confiance pour $c\beta$ au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, est la réalisation $i_{c\beta}$ de

$$I_{c\beta} = \left[c\widehat{\beta} - t_\alpha(\nu)\widehat{\sigma}\sqrt{c(X^tX)^{-1}c^t},\ c\widehat{\beta} + t_\alpha(\nu)\widehat{\sigma}\sqrt{c(X^tX)^{-1}c^t} \right],$$

où $t_\alpha(\nu)$ est le réel vérifiant $\mathbb{P}(|T| \geq t_\alpha(\nu)) = \alpha$, avec $T \sim \mathcal{T}(\nu)$.

Avec les notations déjà introduites, on peut écrire :

$$i_{c\beta} = \left[cb - t_\alpha(\nu)s\sqrt{c(X^tX)^{-1}c^t},\ cb + t_\alpha(\nu)s\sqrt{c(X^tX)^{-1}c^t} \right].$$

Preuve : Dire que $I_{c\beta}$ est un intervalle de confiance (aléatoire) pour $c\beta$ au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, signifie que $\mathbb{P}(c\beta \in I_{c\beta}) = 1-\alpha$. Tout repose sur le résultat :

$$T_* = \frac{c\widehat{\beta} - c\beta}{\widehat{\sigma}\sqrt{c(X^tX)^{-1}c^t}} \sim \mathcal{T}(\nu).$$

En utilisant la définition de $t_\alpha(\nu)$, le fait que T_* et T suivent la même loi (entrainant $\mathbb{P}(|T_*| \leq x) = \mathbb{P}(|T| \leq x)$ pour tout $x \geq 0$) et la définition de T_*, il vient

$$\begin{aligned}
1-\alpha &= 1 - \mathbb{P}(|T| \geq t_\alpha(\nu)) = \mathbb{P}(|T| \leq t_\alpha(\nu)) = \mathbb{P}(|T_*| \leq t_\alpha(\nu)) \\
&= \mathbb{P}\left(\left| \frac{c\widehat{\beta} - c\beta}{\widehat{\sigma}\sqrt{c(X^tX)^{-1}c^t}} \right| \leq t_\alpha(\nu) \right) = \mathbb{P}\left(|c\beta - c\widehat{\beta}| \leq t_\alpha(\nu)\widehat{\sigma}\sqrt{c(X^tX)^{-1}c^t} \right) \\
&= \mathbb{P}\left(-t_\alpha(\nu)\widehat{\sigma}\sqrt{c(X^tX)^{-1}c^t} \leq c\beta - c\widehat{\beta} \leq t_\alpha(\nu)\widehat{\sigma}\sqrt{c(X^tX)^{-1}c^t} \right) \\
&= \mathbb{P}\left(c\widehat{\beta} - t_\alpha(\nu)\widehat{\sigma}\sqrt{c(X^tX)^{-1}c^t} \leq c\beta \leq c\widehat{\beta} + t_\alpha(\nu)\widehat{\sigma}\sqrt{c(X^tX)^{-1}c^t} \right) = \mathbb{P}(c\beta \in I_{c\beta}).
\end{aligned}$$

Ainsi, $I_{c\beta}$ est un intervalle de confiance (aléatoire) pour $c\beta$ au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$.

□

Intervalle de confiance pour β_j

Pour tout $j \in \{0, \dots, p\}$, un intervalle de confiance pour β_j au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, est la réalisation i_{β_j} de

$$I_{\beta_j} = \left[\widehat{\beta}_j - t_\alpha(\nu)\widehat{\sigma}\sqrt{[(X^tX)^{-1}]_{j+1,j+1}},\ \widehat{\beta}_j + t_\alpha(\nu)\widehat{\sigma}\sqrt{[(X^tX)^{-1}]_{j+1,j+1}}\right],$$

où $t_\alpha(\nu)$ est le réel vérifiant $\mathbb{P}(|T| \geq t_\alpha(\nu)) = \alpha$, avec $T \sim \mathcal{T}(\nu)$.

Avec les notations déjà introduites, on peut écrire :

$$i_{\beta_j} = [b_j - t_\alpha(\nu)\text{ete}_j,\ b_j + t_\alpha(\nu)\text{ete}_j]\,.$$

Preuve : Tout repose sur le résultat : $\mathbb{P}(c\beta \in I_{c\beta}) = 1-\alpha$. En prenant $c = c_j$ le vecteur ligne à $p+1$ composantes nulles, sauf la $j+1$-ème qui vaut 1, on a $\mathbb{P}(\beta_j \in I_{\beta_j}) = \mathbb{P}(c_j\beta \in I_{c_j\beta}) = 1-\alpha$. Donc I_{β_j} est un intervalle de confiance (aléatoire) pour β_j au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$.

□

Intervalle de confiance pour y_x

Soient $x_\bullet = (1, x_1, \dots, x_p)$, $y_x = x_\bullet\beta$ la valeur moyenne de Y quand $(X_1, \dots, X_p) = (x_1, \dots, x_p) = x$ et $\widehat{Y}_x = x_\bullet\widehat{\beta}$.

Un intervalle de confiance pour y_x au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, est la réalisation i_{y_x} de

$$I_{y_x} = \left[\widehat{Y}_x - t_\alpha(\nu)\widehat{\sigma}\sqrt{x_\bullet(X^tX)^{-1}x_\bullet^t},\ \widehat{Y}_x + t_\alpha(\nu)\widehat{\sigma}\sqrt{x_\bullet(X^tX)^{-1}x_\bullet^t}\right],$$

où $t_\alpha(\nu)$ est le réel vérifiant $\mathbb{P}(|T| \geq t_\alpha(\nu)) = \alpha$, avec $T \sim \mathcal{T}(\nu)$.

Avec les notations déjà introduites, on peut écrire :

$$i_{y_x} = [d_x - t_\alpha(\nu)\text{ete}_x,\ d_x + t_\alpha(\nu)\text{ete}_x]\,.$$

Preuve : On rappelle que : $\mathbb{P}(c\beta \in I_{c\beta}) = 1-\alpha$. En prenant $c = x_\bullet$, il vient $\mathbb{P}(y_x \in I_{y_x}) = \mathbb{P}(x_\bullet\beta \in I_{x_\bullet\beta}) = 1-\alpha$. Donc I_{y_x} est un intervalle de confiance (aléatoire) pour y_x au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$.

□

Volume de confiance pour $Q\beta$

Soit Q une matrice de réels à $p+1$ colonnes et k lignes de rang colonnes plein.

Un volume de confiance pour $Q\beta$ au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, est la réalisation $v_{Q\beta}$ de

$$V_{Q\beta} = \left\{ u \in \mathbb{R}^{p+1};\ (Q\widehat{\beta} - Qu)^t (R(X^tX)^{-1}Q^t)^{-1}(Q\widehat{\beta} - Qu) \leq k\widehat{\sigma}^2 f_\alpha(k,\nu) \right\},$$

où $f_\alpha(k,\nu)$ est le réel vérifiant $\mathbb{P}(F \geq f_\alpha(k,\nu)) = \alpha$, avec $F \sim \mathcal{F}(k,\nu)$.

Avec les notations déjà introduites, on peut écrire :

$$v_{Q\beta} = \left\{ u \in \mathbb{R}^{p+1};\ (Qb - Qu)^t (Q(X^tX)^{-1}Q^t)^{-1}(Qb - Qu) \leq ks^2 f_\alpha(k,\nu) \right\}.$$

Preuve : Dire que $V_{Q\beta}$ est un volume de confiance (aléatoire) pour $Q\beta$ au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, signifie que $\mathbb{P}(Q\beta \in V_{Q\beta}) = 1-\alpha$. Tout repose sur le résultat :

$$F_* = \frac{(Q\widehat{\beta} - Q\beta)^t (Q(X^tX)^{-1}Q^t)^{-1}(Q\widehat{\beta} - Q\beta)}{k\widehat{\sigma}^2} \sim \mathcal{F}(k,\nu).$$

En utilisant la définition de $f_\alpha(k,\nu)$, le fait que F_* et F suivent la même loi (entrainant $\mathbb{P}(F_* \leq x) = \mathbb{P}(F \leq x)$ pour tout $x \geq 0$) et la définition de F_*, il vient

$$\begin{aligned}
1-\alpha &= 1 - \mathbb{P}(F \geq f_\alpha(k,\nu)) = \mathbb{P}(F \leq f_\alpha(k,\nu)) = \mathbb{P}(F_* \leq f_\alpha(k,\nu)) \\
&= \mathbb{P}\left(\frac{(Q\widehat{\beta} - Q\beta)^t (Q(X^tX)^{-1}Q^t)^{-1}(Q\widehat{\beta} - Q\beta)}{k\widehat{\sigma}^2} \leq f_\alpha(k,\nu) \right) \\
&= \mathbb{P}\left((Q\widehat{\beta} - Q\beta)^t (Q(X^tX)^{-1}Q^t)^{-1}(Q\widehat{\beta} - Q\beta) \leq k\widehat{\sigma}^2 f_\alpha(k,\nu) \right) = \mathbb{P}(Q\beta \in V_{Q\beta}).
\end{aligned}$$

Ainsi, $V_{Q\beta}$ est un volume de confiance (aléatoire) pour $Q\beta$ au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$.

□

Cas particulier : ellipsoïde de confiance pour β pour le modèle de *rls*

Dans le cadre du modèle de *rls* (donc $p = 1$), avec les notations déjà introduites, un ellipsoïde de confiance pour $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$ au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, est

$$e_\beta = \Big\{ (u_0, u_1) \in \mathbb{R}^2 ; \\ (\text{sce}_x + n\overline{x}_1^2)(b_1 - u_1)^2 + 2n\overline{x}_1(b_0 - u_0)(b_1 - u_1) + n(b_0 - u_0)^2 \leq 2s^2 f_\alpha(2, \nu) \Big\},$$

où $f_\alpha(2,\nu)$ est le réel vérifiant $\mathbb{P}(F \geq f_\alpha(2,\nu)) = \alpha$, avec $F \sim \mathcal{F}(2,\nu)$.

Mise en œuvre avec le logiciel R

On reprend le jeu de données "profs". Dans une étude statistique, 23 professeurs sont évalués quant à la qualité de leur enseignement. Pour chacun d'entre eux, on dispose :

- d'un indice de performance globale donné par les étudiants (variable Y),
- des résultats de 4 tests écrits donnés à chaque professeur (variables X_1, X_2, X_3 et X_4),
- du sexe (variable X_5, avec $X_5 = 0$ pour femme, $X_5 = 1$ pour homme).

L'objectif est d'expliquer Y à partir de X_1, X_2, X_3, X_4 et X_5.

On enregistre les données dans R :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/profs.txt", header = T)
attach(w)
```

Le modèle de *rlm* est envisageable. Sa forme générique est

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon,$$

où β_0, β_1, β_2, β_3, β_4 et β_5 sont des coefficients réels inconnus.

On obtient les *emco* ponctuels de ces coefficients en faisant :

```
reg = lm(Y ~ X1 + X2 + X3 + X4 + X5)
```

On calcule les intervalles de confiance pour ces coefficients au niveau 95% avec les commandes :

```
confint(reg, level = 0.95)
```

Cela renvoie :

	2.5 %	97.5 %
(Intercept)	-661.06	116.98
X1	-0.34	1.92
X2	0.74	4.63
X3	-3.18	0.29
X4	2.99	10.67
X5	-42.73	72.53

Le tableau ci-dessous donne les bornes inférieures et supérieures de ces intervalles :

i_{β_0}	i_{β_1}	i_{β_2}	i_{β_3}	i_{β_4}	i_{β_5}
$[-661.06,\ 116.98]$	$[-0.34,\ 1.92]$	$[0.74,\ 4.63]$	$[-3.18,\ 0.29]$	$[2.99,\ 10.67]$	$[-42.73,\ 72.53]$

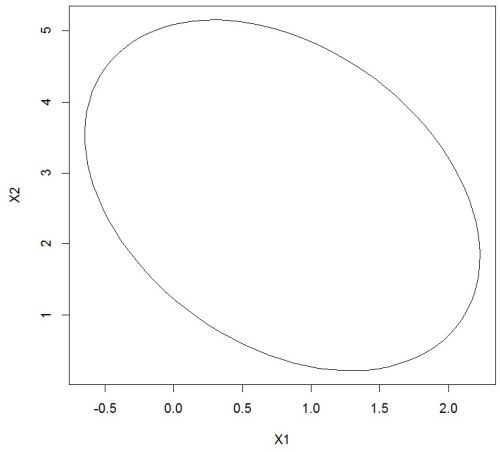
Les commandes R pour calculer les intervalles de confiances pour la valeur moyenne de Y quand $(X_1, X_2, X_3, X_4, X_5) = (82, 158, 47, 49, 1)$ au niveau 95% sont :

```
predict(reg, data.frame(X1 = 82, X2 = 158, X3 = 47, X4 = 49, X5 = 1),
interval = "confidence")
```

Cela renvoie : $i_{y_x} = [451.5943,\ 545.4183]$.

Les commandes R pour calculer un ellipsoïde de confiance pour $\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}$ au niveau 95% sont :

```
library(ellipse)
plot(ellipse(reg, c(2, 3), level = 0.95), type = "l")
```



7 Tests statistiques

Notions de base

Hypothèses. On oppose deux hypothèses complémentaires : H_0 et H_1,

- l'hypothèse H_0 formule ce que l'on souhaite rejeter/réfuter,
- l'hypothèse H_1 formule ce que l'on souhaite montrer.

Par exemple, si on veut montrer l'hypothèse "X_1 influe sur Y", H_0 et H_1 s'opposent sous la forme :

$$H_0 : \text{"}X_1 \text{ n'influe pas sur } Y\text{"} \qquad \text{contre} \qquad H_1 : \text{"}X_1 \text{ influe sur } Y\text{"}.$$

Risque. Le risque est le pourcentage de chances de rejeter H_0, donc d'accepter H_1, alors que H_0 est vraie. On veut que ce risque soit aussi faible que possible.

Il s'écrit sous la forme : $100\alpha\%$, avec $\alpha \in]0,1[$ (par exemple, 5%, soit $\alpha = 0.05$).

Le réel α est alors la probabilité de rejeter H_0 alors que H_0 est vraie.

Le rejet de H_0 est dit "significatif" si elle est rejetée au risque 5%.

Test statistique. Un test statistique est une procédure qui vise à apporter une réponse à la question : Est-ce que les données nous permettent de rejeter H_0, donc d'accepter H_1, avec un faible risque de se tromper ?

Types de test statistique. En notant θ un paramètre inconnu, on dit que le test est

- bilatéral si H_1 est de la forme $H_1 : \theta \neq \dots$
- unilatéral à gauche (sens de $<$) si H_1 est de la forme $H_1 : \theta < \dots$
- unilatéral à droite (sens de $>$) si H_1 est de la forme $H_1 : \theta > \dots$

p-valeur. La p-valeur est le plus petit réel $\alpha \in]0,1[$ calculé à partir des données tel que l'on puisse se permettre de rejeter H_0 au risque $100\alpha\%$. Autrement écrit, la p-valeur est une estimation ponctuelle de la probabilité critique de se tromper en rejetant H_0 alors que H_0 est vraie.

Les logiciels actuels travaillent principalement avec cette p-valeur.

Rappel : degré de significativité

La p-valeur nous donne un degré de significativité du rejet de H_0. Le rejet de H_0 sera :

- significatif si p-valeur $\in]0.01,\ 0.05]$, symbolisé par $*$,
- très significatif si p-valeur $\in]0.001,\ 0.01]$, symbolisé par $**$,
- hautement significatif si p-valeur < 0.001, symbolisé par $***$.

Il y a non rejet de H_0 si p-valeur > 0.05.

S'il y a non-rejet de H_0, sauf convention, on ne peut rien conclure du tout (avec le risque considéré). En revanche, peut-être qu'un risque de départ plus élevé ou la disposition de plus de données peuvent conduire à un rejet de H_0.

Emco et test de Student

Soient c un vecteur ligne à $p+1$ composantes et r un réel représentant une valeur de référence.

On considère les hypothèses :

Hypothèses	H_0	H_1
bilatérale	$c\beta = r$	$c\beta \neq r$
unilatérale à droite	$c\beta \leq r$	$c\beta > r$
unilatérale à gauche	$c\beta \geq r$	$c\beta < r$

On calcule la réalisation t_{obs} de

$$T_* = \frac{c\widehat{\beta} - r}{\widehat{\sigma}\sqrt{c(X^tX)^{-1}c^t}}.$$

On considère une *var* $T \sim \mathcal{T}(\nu)$.

Alors les p-valeurs associées aux hypothèses considérées sont :

H_0	H_1	p-valeurs
$c\beta = r$	$c\beta \neq r$	$\mathbb{P}(\lvert T\rvert \geq \lvert t_{obs}\rvert)$
$c\beta \leq r$	$c\beta > r$	$\mathbb{P}(T \geq t_{obs})$
$c\beta \geq r$	$c\beta < r$	$\mathbb{P}(T \leq t_{obs})$

Avec les notations déjà introduites, on peut écrire :

$$t_{obs} = \frac{cb - r}{s\sqrt{c(X^tX)^{-1}c^t}}.$$

Par exemple, pour $p = 2$, *donc* $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$, *si on veut prouver que* X_1 *à plus d'influence que* X_2 *sur* Y, *alors on considère l'hypothèse :* $H_1 : \beta_1 > \beta_2$. *On peut alors la réécrire comme* $H_1 : c\beta > r$ *avec* $c = (0\ 1\ -1)$ *et* $r = 0$.

Éléments de preuve : Sous les hypothèses standards, par le test du rapport des vraisemblances maximales, on peut montrer que la zone de rejet optimale de H_0 est un événement de la forme :

$$\mathcal{R} = \left\{|c\widehat{\beta} - r| \geq C\right\} = \left\{\left|\frac{c\widehat{\beta} - r}{\widehat{\sigma}\sqrt{c(X^tX)^{-1}c^t}}\right| \geq C_*\right\} = \{|T_*| \geq C_*\},$$

où $C > 0$ et $C_* > 0$ désignent des quantités muettes ; seule la forme générale de $\mathcal{R}$ importe. Plus intuitivement : rejet de H_0/affirmation de $H_1 \Leftrightarrow c\beta \neq r \Leftrightarrow |c\beta - r| > 0 \Rightarrow |c\widehat{\beta} - r| > C > 0$. Si H_0 est vraie, alors $T_* \sim \mathcal{T}(\nu)$; T_* et T suivent la même loi, laquelle ne dépend pas de paramètre inconnue.

De plus, une estimation ponctuelle de la plus grande constante calculable C_* qui minimise la probabilité que l'événement $\mathcal{R}$ se réalise est la réalisation $|t_{obs}|$ de $|T_*|$. C'est pourquoi on considère :

$$\text{p-valeur} = \mathbb{P}(|T| \geq |t_{obs}|).$$

□

Emco et test de Student ; suite

Soient $j \in \{0, \dots, p\}$ et r un réel. On considère les hypothèses :

Hypothèses	H_0	H_1
bilatérale	$\beta_j = r$	$\beta_j \neq r$
unilatérale à droite	$\beta_j \leq r$	$\beta_j > r$
unilatérale à gauche	$\beta_j \geq r$	$\beta_j < r$

On calcule la réalisation t_{obs} de

$$T_* = \frac{\widehat{\beta}_j - r}{\widehat{\sigma}\sqrt{[(X^t X)^{-1}]_{j+1,j+1}}}.$$

On considère une *var* $T \sim \mathcal{T}(\nu)$.

Alors les p-valeurs associées aux hypothèses considérées sont :

H_0	H_1	p-valeurs
$\beta_j = r$	$\beta_j \neq r$	$\mathbb{P}(\lvert T\rvert \geq \lvert t_{obs}\rvert)$
$\beta_j \leq r$	$\beta_j > r$	$\mathbb{P}(T \geq t_{obs})$
$\beta_j \geq r$	$\beta_j < r$	$\mathbb{P}(T \leq t_{obs})$

Avec les notations déjà introduites, on peut écrire :

$$t_{obs} = \frac{b_j - r}{\text{ete}_j}.$$

Preuve : C'est une application du résultat précédent avec $c = c_j$ le vecteur ligne à $p+1$ composantes nulles, sauf la $j+1$-ème qui vaut 1.

□

Influence de X_j sur Y

Pour tout $j \in \{1, \dots, p\}$, l'influence de X_j sur Y est caractérisée par $\beta_j \neq 0$; plus β_j diffère de 0, plus la variable X_j a de l'importance dans l'explication de Y.

On pose alors les hypothèses :

$$H_0 : \beta_j = 0 \qquad \text{contre} \qquad H_1 : \beta_j \neq 0,$$

correspondant à $r = 0$. On obtient le degré de significativité de son influence en posant en étudiant :

$$\text{p-valeur} = \mathbb{P}(|T| \geq |t_{obs}|).$$

Par exemple, si p-valeur $\in]0.001,\ 0.01]$ *; **, l'influence de* X_j *sur* Y *est très significative.*

On a alors $p+1$ p-valeurs, lesquelles sont souvent donnés directement par les logiciels statistiques.

Emco et test de Fisher

Soient Q une matrice de réels à $p+1$ colonnes et k lignes de rang colonnes plein et r un vecteur colonne à k lignes. On considère les hypothèses :

$$H_0 : Q\beta = r \qquad \text{contre} \qquad H_1 : Q\beta \neq r.$$

On calcule la réalisation f_{obs} de

$$F_* = \frac{(Q\widehat{\beta} - r)^t (Q(X^tX)^{-1}Q^t)^{-1}(Q\widehat{\beta} - r)}{k\widehat{\sigma}^2}.$$

On considère une *var* $F \sim \mathcal{F}(k, \nu)$.

Alors la p-valeur associée est

$$\text{p-valeur} = \mathbb{P}(F \geq f_{obs}).$$

Avec les notations déjà introduites, on peut écrire :

$$f_{obs} = \frac{(Qb - r)^t (Q(X^tX)^{-1}Q^t)^{-1}(Qb - r)}{ks^2}.$$

Par exemple, pour $p = 2$, *donc* $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$, *on peut écrire* $H_0 : \beta_0 = \beta_1 = \beta_2$ *comme* $H_0 :$ $Q\beta = r$ *avec* $Q = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix}$ *et* $r = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$.

Éléments de preuve : Sous les hypothèses standards, par le test du rapport des vraisemblances maximales, on peut montrer que la zone de rejet optimale de H_0 est un événement de la forme :

$$\begin{aligned}\mathcal{R} &= \left\{\|X(X^tX)^{-1}Q^t(Q(X^tX)^{-1}Q^t)^{-1}(Q\widehat{\beta}-r)\|^2 \geq C\right\} = \left\{\frac{(Q\widehat{\beta}-r)^t(Q(X^tX)^{-1}Q^t)^{-1}(Q\widehat{\beta}-r)}{k\widehat{\sigma}^2} \geq C_*\right\} \\ &= \{F_* > C_*\},\end{aligned}$$

où $C > 0$ et $C_* > 0$ désignent des quantités muettes ; seule la forme générale de $\mathcal{R}$ importe.

Si H_0 est vraie, alors $F_* \sim \mathcal{F}(k, \nu)$; F_* et F suivent la même loi, laquelle ne dépend pas de paramètre inconnue. De plus, une estimation ponctuelle de la plus grande constante calculable C_* qui minimise la probabilité que l'événement $\mathcal{R}$ se réalise est la réalisation f_{obs} de F_*. C'est pourquoi on considère :

$$\text{p-valeur} = \mathbb{P}(F \geq f_{obs}).$$

□

Test global de Fisher

On considère les hypothèses :

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ contre H_1 : "il y a au moins un coefficient non nul".

On calcule la réalisation f_{obs} de

$$F_* = \frac{\widehat{R}^2}{1-\widehat{R}^2}\frac{n-(p+1)}{p}.$$

On considère une *var* $F \sim \mathcal{F}(p, \nu)$.

Alors la p-valeur associée est

$$\text{p-valeur} = \mathbb{P}(F \geq f_{obs}).$$

Avec les notations déjà introduites, on peut écrire :

$$f_{obs} = \frac{R^2}{1-R^2}\frac{n-(p+1)}{p}.$$

Ce test est un cas particulier du test de Fisher avec la matrice $Q = \text{diag}_{p+1}(0, 1, \ldots, 1)$ et $r = 0_{p+1}$.

Il vise à étudier la pertinence du lien linéaire entre Y et $X_1, \ldots, X_p$.

Comparaison de deux modèles emboîtés

Soit Λ un sous-ensemble de $\{1, \dots, p\}$ ayant k éléments. On considère les hypothèses :

$$H_0 : \text{"}\beta_j = 0 \text{ pour tout } j \in \Lambda\text{"} \qquad \text{contre}$$

$$H_1 : \text{"il y a au moins un des coefficients } \beta_j,\ j \in \Lambda, \text{ non nul"}.$$

Soient X_Λ la matrice X privée des colonnes d'indice $j \in \Lambda$, β_Λ le vecteur β privé de $(\beta_j)_{j\in\Lambda}$ et $\widehat{\beta}_\Lambda$ l'*emco* de β_Λ avec le modèle de *rlm* : $Y = X_\lambda \beta_\lambda + \epsilon$, donc $\widehat{\beta}_\Lambda = (X_\Lambda^t X_\Lambda)^{-1} X_\Lambda^t Y$.

On calcule la réalisation f_{obs} de

$$F_* = \frac{||X_\Lambda \widehat{\beta}_\Lambda - X\widehat{\beta}||^2}{k\widehat{\sigma}^2}.$$

On considère une *var* $F \sim \mathcal{F}(k, \nu)$.

Alors la p-valeur associée est

$$\text{p-valeur} = \mathbb{P}(F \geq f_{obs}).$$

On peut aussi écrire :

$$F_* = \frac{||Y - X_\Lambda \widehat{\beta}_\Lambda||^2 - ||Y - X\widehat{\beta}||^2}{k\widehat{\sigma}^2}.$$

Avec les notations déjà introduites, en posant $b_\Lambda = (X_\Lambda^t X_\Lambda)^{-1} X_\Lambda^t y$, on peut écrire :

$$f_{obs} = \frac{||X_\Lambda b_\Lambda - Xb||^2}{ks^2}.$$

Ce test est un cas particulier du test de Fisher. Il vise à évaluer la pertinence de l'inclusion de certaines variables dans le modèle. On peut alors faire de la sélection de variables.

Plus précisément, si on ne rejette pas H_0 : "$\beta_j = 0$ pour tout $j \in \Lambda$", on admet que les variables $(X_j)_{j\in\Lambda}$ sont statistiquement dispensables dans l'explication de Y. Il est alors préférable de ne pas les inclure dans le modèle : plus simple est le modèle, mieux c'est ; principe KISS : Keep It Simple and Stupid.

Mise en œuvre avec le logiciel R

On reprend le jeu de données "profs". Dans une étude statistique, 23 professeurs sont évalués quant à la qualité de leur enseignement. Pour chacun d'entre eux, on dispose :

- d'un indice de performance globale donné par les étudiants (variable Y),
- des résultats de 4 tests écrits donnés à chaque professeur (variables X_1, X_2, X_3 et X_4),
- du sexe (variable X_5, avec $X_5 = 0$ pour femme, $X_5 = 1$ pour homme).

L'objectif est d'expliquer Y à partir de X_1, X_2, X_3, X_4 et X_5. On enregistre les données dans R :

```
w = read.table("http://www.math.unicaen.fr/~chesneau/profs.txt", header = T)

attach(w)
```

Le modèle de *rlm* est envisageable. Sa forme générique est

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon,$$

où β_0, β_1, β_2, β_3, β_4 et β_5 sont des coefficients réels inconnus.

On obtient les *emco* ponctuels de ces coefficients en faisant :

```
reg = lm(Y ~ X1 + X2 + X3 + X4 + X5)
```

Pour tout $j \in \{1, \ldots, p\}$, pour étudier l'influence de X_j sur Y, on considère les hypothèses :

$$H_0 : \beta_j = 0 \qquad \text{contre} \qquad H_1 : \beta_j \neq 0.$$

On peut obtenir les t_{obs} et les p-valeurs associées avec la commande `summary` :

```
summary(reg)
```

Cela renvoie :

	Estimate	Std. Error	t value	**Pr(>|t|)**	
(Intercept)	-272.0388	184.3865	-1.48	**0.1584**	
X1	0.7913	0.5363	1.48	**0.1583**	
X2	2.6828	0.9216	2.91	**0.0097**	**
X3	-1.4434	0.8217	-1.76	**0.0970**	.
X4	6.8308	1.8192	3.75	**0.0016**	**
X5	14.9008	27.3134	0.55	**0.5925**	

```
Residual standard error: 55.06 on 17 degrees of freedom
Multiple R-squared: 0.6834, Adjusted R-squared: 0.5903
F-statistic: 7.34 on 5 and 17 DF, p-value: 0.0007887
```

Les t_{obs} sont donnés dans la colonne `t value` du tableau et les p-valeurs associées dans la colonne `Pr(>|t|)`. Les degrés de significativité sont dans la dernière colonne.

Ainsi, comme on a $**$ pour les p-valeurs associées à X_2 et X_4, X_2 et X_4 ont une influence très significative sur Y. Comme on a "." pour la p-valeur associée à X_3, X_3 a une influence "presque" significative sur Y. Rien ne ressort pour X_1, X_2 et X_5.

On considère maintenant les hypothèses :

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_5 = 0 \qquad \text{contre} \qquad H_1 : \text{il y a au moins un coefficient non nul.}$$

On utilise alors le test global de Fisher, lequel est mis en œuvre avec la commande `summary`. On a le f_{obs} avec `F-statistic` : $f_{obs} = 7.34$ et la p-valeur associée avec `p-value` : p-valeur $= 0.0007887$. Comme p-valeur < 0.001, le degré de significativité est $***$; le lien linéaire entre Y et X_1, X_2, X_3, X_4 et X_5 est pertinent.

Remarque : Comme $R^2 = 0.6834$, on peut vérifier que

$$f_{obs} = \frac{R^2}{1-R^2}\frac{n-(p+1)}{p} = \frac{0.6834}{1-0.6834}\frac{23-(5+1)}{5} = 7.339104.$$

On considère maintenant les hypothèses :

$$H_0 : \beta_1 = \beta_3 = 0 \qquad \text{contre} \qquad H_1 : \beta_1 \neq 0 \text{ ou } \beta_3 \neq 0.$$

On peut alors mettre H_0 sous la forme $Q\beta = r$; on utilise le test de Fisher.

On le met en oeuvre en faisant :

```
reg1 = lm(Y ~ X1 + X2 + X3 + X4 + X5)
reg2 = lm(Y ~ X2 + X4 + X5)
anova(reg1, reg2)
```

On obtient la p-valeur associée dans la colonne `Pr(>F)` : p-valeur $= 0.1702$. Comme p-valeur > 0.05, les données ne nous permettent pas de rejeter H_0.

Complément : test de nullité du coefficient de corrélation (de Pearson)

On se place dans le cadre du modèle de *rls* (donc $p = 1$) et on considère les hypothèses :

$$H_0 : \text{"}X_1 \text{ et } Y \text{ sont indépendantes"} \quad \text{contre} \quad H_1 : \text{"}X_1 \text{ et } Y \text{ ne sont pas indépendantes"}.$$

On définit le coefficient de corrélation ρ par $\rho = \dfrac{\mathbb{C}(X_1, Y)}{\sigma(X_1)\sigma(Y)}$. De plus, on suppose que (X_1, Y) est un vecteur de *var* suivant une loi normale bidimensionnelle. Grâce à cette hypothèse, on a l'équivalence : X_1 et Y indépendantes $\Leftrightarrow \rho = 0$. On peut alors reformuler les hypothèses comme :

$$H_0 : \rho = 0 \qquad \text{contre} \qquad H_1 : \rho \neq 0.$$

Pour mettre en œuvre le test de nullité du coefficient de corrélation, on calcule

$$t_{obs} = \sqrt{n-2}\frac{r_{x,y}}{\sqrt{1 - r_{x,y}^2}}.$$

Soit $T \sim T(n-2)$. Alors la p-valeur associée est p-valeur $= \mathbb{P}(|T| \geq |t_{obs}|)$.

Ce test est en fait identique au test de Student ; on peut montrer que $|t_{obs}| = \dfrac{|b_1|}{\text{ete}_1}$.

Mise en œuvre avec le logiciel R

Sur 14 familles composées d'un père et d'un fils, on examine le QI du père et le QI du fils. Les résultats sont les suivants :

Père	121	142	108	111	97	139	131	90	115	107	124	103	115	151
Fils	102	138	126	133	95	146	115	100	142	105	130	120	109	123

Peut-on affirmer qu'il y a une liaison significative entre le QI du père et le QI du fils ?

Soient X (ou X_1) la variable "QI du père" et Y la variable "QI du fils". Par l'énoncé, on observe la valeur de (X, Y) pour chacun des n individus (familles) d'un échantillon avec $n = 14$. On modélise ces variables comme des *var*. On considère les hypothèses :

$$H_0 : \text{"}X \text{ et } Y \text{ sont indépendantes"} \quad \text{contre} \quad H_1 : \text{"}X \text{ et } Y \text{ ne sont pas indépendantes"}.$$

On considère les commandes :

```
x = c(121, 142, 108, 111, 97, 139, 131, 90, 115, 107, 124, 103, 115, 151)
y = c(102, 138, 126, 133, 95, 146, 115, 100, 142, 105, 130, 120, 109, 123)
cor.test(x, y)
```

Cela renvoie : p-valeur = 0.04090612.

Comme p-valeur $\in]0.01, 0.05]$, le rejet de H_0 est significatif $\star$.

Ainsi, on peut affirmer qu'il y a une liaison significative entre le QI du père et le QI du fils.

Index

Coefficient de détermination, 11

Coefficient de détermination ajusté, 11

Comparaison de modèles, 59

confint, 50

Droite de régression, 23, 24

Ecriture matricielle, 7, 18

Ellipsoïdes de confiance, 50, 51

Emco, 7

Emco et Emv, 36

Emco et loi de Fisher, 39

Emco et loi de Student, 38

Estimateur de la valeur moyenne, 10

Estimateur de la variance, 37

Estimations ponctuelles, 10

Forme matricielle, 7, 18

Graphique des résidus, 40

Hypothèses, 53

Hypothèses standards, 35

Intervalles de confiance, 47, 48

Intervalles de confiance pour la prédiction, 48

lm, 14, 28, 42, 50, 60, 61

Loi de Y, 35

Loi de l'emco, 35

p-valeur, 53

predict, 51

Prédiction, 10

residuals, 42

Risque, 53

Régression linéaire multiple (*rlm*), 5

Régression linéaire simple (*rls*), 15

Résidus, 40

summary, 14, 28, 43, 60

Test de Fisher, 57, 61

Test de Student, 54, 56

Test du coefficient de corrélation, 62

Test global de Fisher, 58, 61

Test statistique, 53

Théorème de Gauss-Markov, 36

Volumes de confiance, 49