Modèles linéaires

Régressions simples et multiples

Denis Laloë
GABI - PSGen

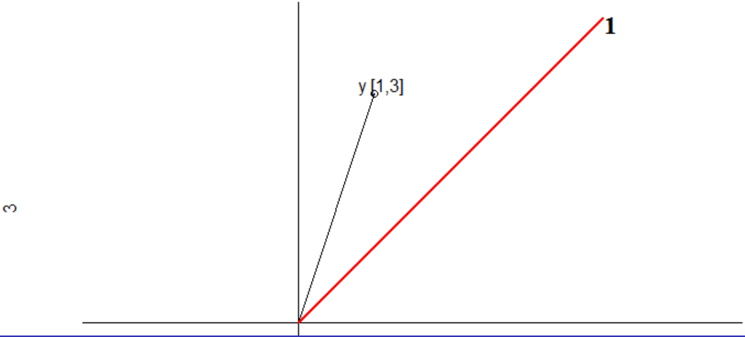
27 septembre 2016

Un exemple simple. Modélisation par une constante

$\mathbf{Y} = \beta_0 \mathbf{1} + \mathbf{E}$; $\mathbf{Y}$ à valeurs dans $\mathbb{R}^2$

$\mathbf{y}$ est une réalisation de $\mathbf{Y}$

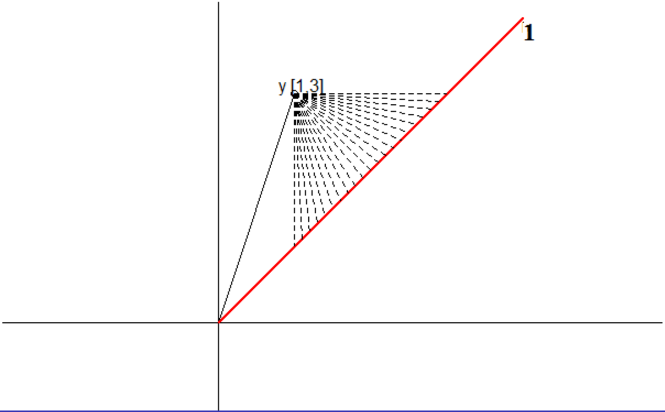
Projection de $\mathbf{y}$ sur la droite engendrée par $\mathbf{1} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$



Quelle projection ?

$\mathbf{Y} = \beta_0 \mathbf{1} + \mathbf{E}$

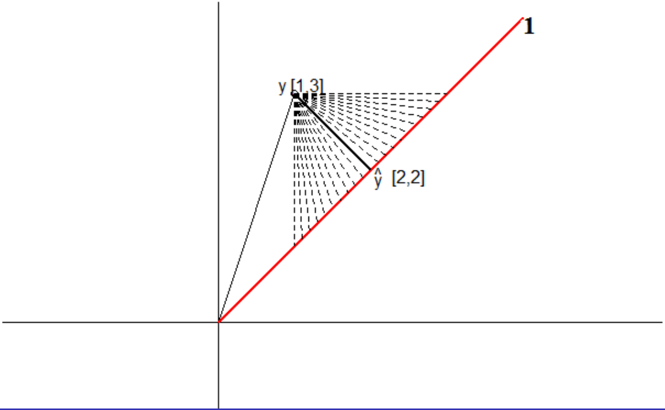
Quelle projection ?



Quelle projection

$\mathbf{Y} = \beta_0 \mathbf{1} + \mathbf{E}$

Projection minimisant la longueur de $\|\mathbf{y} - \hat{\mathbf{y}}\|$



Projection orthogonale

$\mathbf{Y} = \beta_0 \mathbf{1} + \mathbf{E}$

Projection orthogonale minimisant la longueur de $\|\mathbf{y} - \hat{\mathbf{y}}\|$

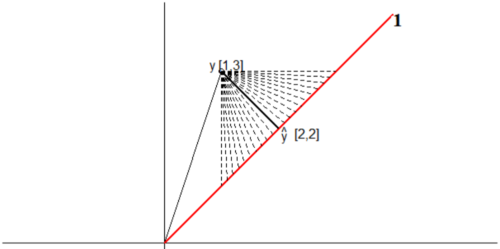
$\mathbf{y} - \hat{\mathbf{y}} \perp \mathbf{1}$

$\mathbf{1}^{\mathbf{t}}(\mathbf{y} - \hat{\mathbf{y}}) = 0$

$\mathbf{1}^{\mathbf{t}}(\mathbf{y} - \mathbf{1}\hat{\beta}_{\mathbf{0}}) = 0$

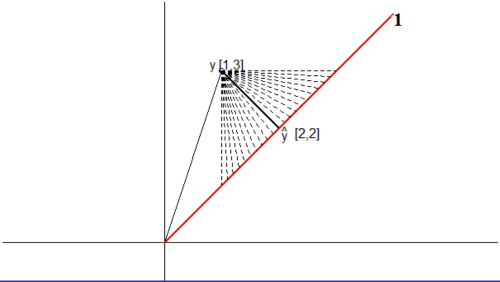
$\mathbf{1}^{\mathbf{t}}\mathbf{y} = \mathbf{1}^{\mathbf{t}}\mathbf{1}\hat{\beta}_{\mathbf{0}}$

$\hat{\beta}_0 = (\mathbf{1}^{\mathbf{t}}\mathbf{1})^{-1}\mathbf{1}^{\mathbf{t}}\mathbf{y}$



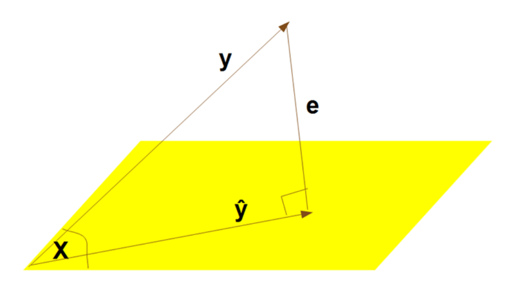
Moindres carrés ordinaires

$$\hat{\beta}_0 = (\mathbf{1}^t\mathbf{1})^{-1}\mathbf{1}^t\mathbf{y}$$
$$1^t\mathbf{y} = y_1 + y_2$$
$$\mathbf{1}^t\mathbf{1} = 2$$
$$(\mathbf{1}^t\mathbf{1})^{-1} = \frac{1}{2}$$
$$\hat{\beta}_0 = \frac{y_1 + y_2}{2}$$



La régression

- On a une variable aléatoire **Y**, une réalisation de cette variable, composée de n valeurs, $y_i, i = 1, ..., n$
- On regroupe ces n valeurs dans un vecteur **y**, appartenant à $\mathbb{R}^n$
- On modélise **y** en fonction d'une constante et de p variables explicatives
  $\mathbf{x}$: $\mathbf{Y} = \beta_\mathbf{0} + \beta_\mathbf{1}\mathbf{x}^{(\mathbf{1})} + \ldots + \beta_\mathbf{p}\mathbf{x}^{(\mathbf{p})} + \mathbf{E}$
- Ces variables engendrent un espace à $p+1$ dimensions.



Le modèle linéaire général

Modèle : Constante

$$\mathbf{Y} = \mathbf{1}\beta_0 + \mathbf{E}$$

$$\mathbf{Y} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \beta_0 + \mathbf{E}$$

Résolution

$$\hat{\beta}_0 = (\mathbf{1}^t\mathbf{1})^{-1}\mathbf{1}^t\mathbf{y}$$

Modèle général

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{E}$$

$$\mathbf{Y} = \begin{bmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(p)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n^{(1)} & x_n^{(2)} & \dots & x_n^{(p)} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \beta_p \end{bmatrix} + \mathbf{E}$$

Résolution

$$\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$$

La régression simple. Le modèle et son écriture matricielle

$\mathbf{Y} = \mathbf{X}\beta + \mathbf{E}$

$$\mathbf{Y} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} + \mathbf{E}$$

$\mathbf{X}^t\mathbf{X}$

$$\mathbf{X}^t\mathbf{X} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} = \begin{bmatrix} n & s_x \\ s_x & s_{x^2} \end{bmatrix}$$

$\mathbf{X}^t\mathbf{Y}$

$$\mathbf{X}^t\mathbf{Y} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n x_i Y_i \end{bmatrix} = \begin{bmatrix} s_Y \\ s_{xY} \end{bmatrix}$$

La régression simple. La résolution

$(\mathbf{X}^t\mathbf{X})^{-1}$

$$\mathbf{X}^t\mathbf{X} = \begin{bmatrix} n & s_x \\ s_x & s_{x^2} \end{bmatrix}$$

$$(\mathbf{X}^t\mathbf{X})^{-1} = \frac{1}{ns_{x^2} - s_x^2} \begin{bmatrix} s_{x^2} & -s_x \\ -s_x & n \end{bmatrix}$$

$(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$

$$(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y} = \frac{1}{ns_{x^2} - s_x^2} \begin{bmatrix} s_{x^2} & -s_x \\ -s_x & n \end{bmatrix} \begin{bmatrix} s_Y \\ s_{xY} \end{bmatrix} = \frac{1}{ns_{x^2} - s_x^2} \begin{bmatrix} s_{x^2}s_Y - s_x s_{xY} \\ ns_{xY} - s_x s_Y \end{bmatrix}$$

Solution

$$\hat{\beta}_1 = \frac{ns_{xY} - s_x s_Y}{ns_{x^2} - s_x^2} = \frac{s_{(x-\bar{x})(Y-\bar{Y})}}{s_{(x-\bar{x})^2}}$$

$$\hat{\beta}_0 == m_Y - \hat{\beta}_1 m_x$$

Propriétés de l'estimateur

Propriétés

- Justification géométrique (hors hypothèses de distribution)
- Equivalence avec l'estimateur du maximum de vraisemblance $E \sim \mathbb{N}(0, \sigma^2)$
- Sans biais
- Variance minimale
- convergent

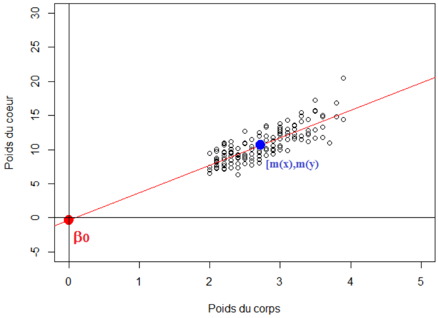
Prédiction

- $\hat{\beta} \sim \mathbb{N}(\beta, (\mathbf{X}^{\mathbf{t}}\mathbf{X})^{-\mathbf{1}}\sigma^2)$
- $\hat{Y} = \mathbf{X}(\mathbf{X}^{\mathbf{t}}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^{\mathbf{t}}\mathbf{Y} = \mathbf{HY}$,
- **H** encore appelée "hat matrix".

Estimateur de la variance résiduelle

- $\hat{\sigma^2} = \dfrac{\sum(Y_i - \hat{Y}_i)^2}{n-p-1}$
- Sans biais
- Variance minimale (*parmi les estimateurs quadratiques sans biais*)
- $\dfrac{(n-p-1)\hat{\sigma^2}}{\sigma^2} \sim \chi^2_{n-p}$

Interprétation des coefficients



- Signification de β_0 ?
- $x \longrightarrow x - \bar{x}$

La régression simple. Une autre paramétrisation

$$\mathbf{X}^t\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^t\mathbf{y}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$$

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 - \bar{x} \\ 1 & x_2 - \bar{x} \\ \vdots & \vdots \\ 1 & x_{n-1} - \bar{x} \\ 1 & x_n - \bar{x} \end{pmatrix}$$

$$\mathbf{X}^t\mathbf{X} = \begin{pmatrix} 1 & 1 & \cdots & 1 & 1 \\ x_1 - \bar{x} & x_2 - \bar{x} & \cdots & x_{n-1} - \bar{x} & x_n - \bar{x} \end{pmatrix} \begin{pmatrix} 1 & x_1 - \bar{x} \\ 1 & x_2 - \bar{x} \\ \vdots & \vdots \\ 1 & x_{n-1} - \bar{x} \\ 1 & x_n - \bar{x} \end{pmatrix}$$

$$\mathbf{X}^t\mathbf{X} = \begin{pmatrix} n & \sum x_i - n\bar{x} \\ \sum x_i - n\bar{x} & \sum (x_i - \bar{x})^2 \end{pmatrix}$$

$$\mathbf{X}^t\mathbf{X} = \begin{pmatrix} n & 0 \\ 0 & \sum (x_i - \bar{x})^2 \end{pmatrix}$$

La régression simple. Une autre paramétrisation

$$\mathbf{X}^t\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^t Y$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t Y$$

$$\mathbf{X}^t Y = \begin{pmatrix} 1 & 1 & \cdots & 1 & 1 \\ x_1 - \bar{x} & x_2 - \bar{x} & \cdots & x_{n-1} - \bar{x} & x_n - \bar{x} \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{n-1} \\ Y_n \end{pmatrix}$$

$$\mathbf{X}^t Y = \begin{pmatrix} \sum Y_i \\ \sum (x_1 - \bar{x}) Y_i \end{pmatrix}$$

$$\mathbf{X}^t\mathbf{X}\hat{\boldsymbol{\beta}} = \begin{pmatrix} n & 0 \\ 0 & \sum (x_i - \bar{x})^2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$$

$$\hat{\beta}_0 = \frac{\sum Y_i}{n}$$

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x}) Y_i}{\sum (x_i - \bar{x})^2}$$

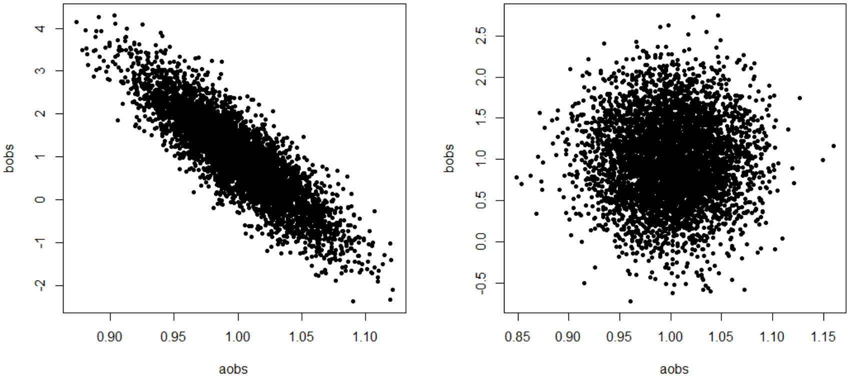
La régression simple. Une autre paramétrisation

$$\hat{\beta}_0 = \frac{\sum Y_i}{n}$$

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x}) Y_i}{\sum (x_i - \bar{x})^2}$$

- $\bar{\hat{\beta}}_1$ inchangé
- $\bar{\hat{\beta}}_0$ égal à la moyenne des Y

La régression simple. Une autre paramétrisation



- La covariance entre les estimateurs est nulle
- Colinéarité...

Estimateur de la variance résiduelle

- $\hat{\sigma^2} = \dfrac{\sum(Y_i - \hat{Y}_i)^2}{n-p-1}$
- Sans biais
- Variance minimale (*parmi les estimateurs quadratiques sans biais*)
- $\dfrac{(n-p-1)\hat{\sigma^2}}{\sigma^2} \sim \chi^2_{n-p}$

Lois liées à la loi normale

Loi du χ^2

Soient $X_1, \ldots, X_n$ des v.a. indépendantes de même loi $\mathbb{N}(0,1)$, et $Z = \sum_{i=1}^n X_i^2$. Z suit une loi du khi-deux à n degrés de liberté $\chi^2(n)$

Loi de Student

Soient $X \sim \mathbb{N}(0,1)$, $Y \sim \chi^2(n)$, et $T = \dfrac{X}{\sqrt{\frac{Y}{n}}}$. T suit une loi de Student à n degrés de liberté (*T(n), Student(n)*).

Loi de Fisher-Snedecor

Soient deux variables indépendantes $X \sim \chi^2(n)$, $Y \sim \chi^2(m)$, et $T = \dfrac{\frac{X}{n}}{\frac{Y}{m}}$.

T suit une loi de Fisher-Snedecor *F(n,m)*.

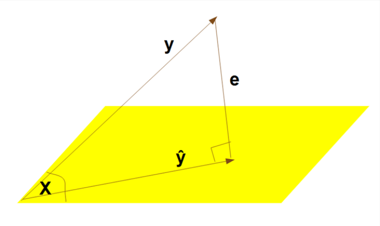
Décomposition de la variance
Somme des carrés

$$\sum(Y_i - \bar{Y})^2 = \sum(\hat{Y}_i - \bar{Y})^2 + \sum(Y_i - \hat{Y}_i)^2$$
$$SCT = SCE + SCR$$

- SCT : Somme des carrés totale
- SCE : Somme des carrés expliquées
- SCR : Somme des carrés résiduelles

Somme des carrés
R^2

- SCT : Somme des carrés totale
- SCE : Somme des carrés expliquées
- SCR : Somme des carrés résiduelles
- $R^2 = \dfrac{SCE}{SCT} = \cos^2(Y, \hat{Y})$



Hypothèses

Inférence sur le modèle

- SCR : $\sum(Y_i - \hat{Y}_i)^2 \sim \chi^2_{n-p-1}$
- SCE et SCR sont indépendantes
- Le modèle explique-t-il quelque chose ?
- H0 : $\beta = 0$
  - SCE :$\sum(\hat{Y}_i - \bar{Y})^2 \sim \chi^2_p$
  - $\frac{\frac{SCE}{p}}{\frac{SCR}{n-p-1}} \sim F(p, n-p-1)$

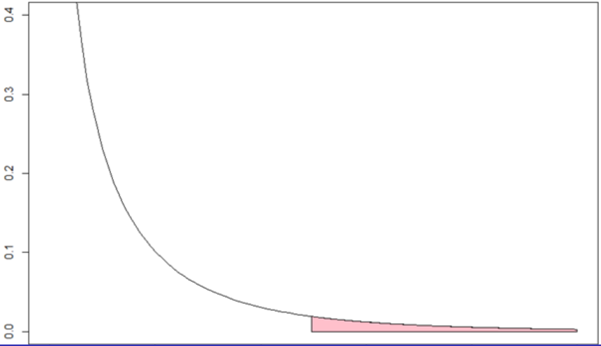
Table d'analyse de la variance

Source de variation	d.d.l	Somme des carrés	Carrés moyens	F
Modèle	**p**	**SCE**	**CME=SCE/p**	**CME/CMR**
Erreur	**n-p-1**	**SCR**	**CMR= SCR/n-p-1**	
Total	**n-1**	**SCT**		

Hypothèses

- **Inférence sur le modèle**
  - Le modèle explique-t-il quelque chose
  - H0 : β=0

$$\frac{SCE / p}{SCR / n-p-1} \sim F(p, n-p-1)$$

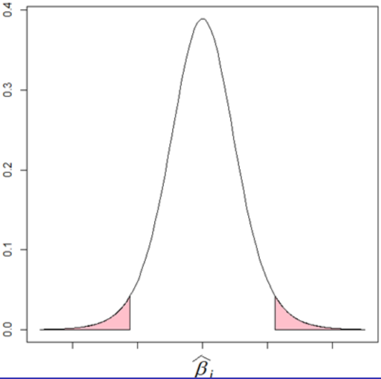


Distribution de Fisher (1,10)
α=0.05
$f_{0.95}$= 4.96
H0 rejeté si F > $f_{0.95}$

Hypothèses

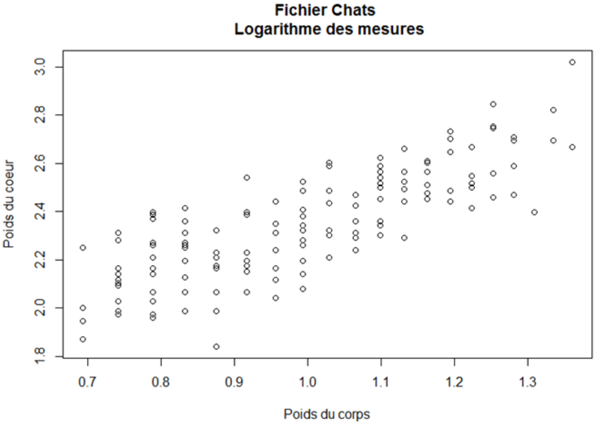
- **Inférence sur UN coefficient**
  - Distribution d'un coefficient $\dfrac{\hat{\beta}_i - \beta_i}{\hat{\sigma}_i} \sim t(n-p)$
  - Hypothèse H0 : $\beta_j = a$
  - Construction d'un intervalle de

$$\left[\hat{\beta}_i - t_{1-\alpha/2}\hat{\sigma}_i , \hat{\beta}_i + t_{1-\alpha/2}\hat{\sigma}_i\right]$$



Un modèle de régression simple

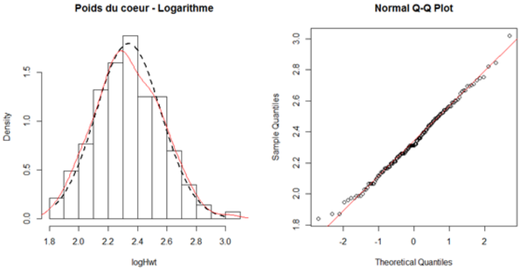
- Fichier *cats*, du package R MASS
- 144 chats, sur lesquels on a mesuré les poids du corps et du coeur. Les variables ont également été transformées (logarithme)



Passage aux logarithmes

Normalité du logarithme du poids du cœur

1. Graphiques



2. Test de normalité de Shapiro-Wilk (shapiro.test)
p=0.83. Normalité acceptée
p.m, le test de Shapiro sur le poids du cœur non transformé conduit à rejeter l'hypothèse de normalité

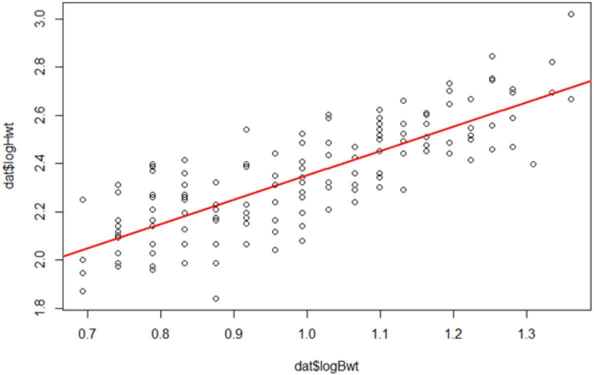
fonction lm du package R

fic.lm=lm(y x,data=fic)

- summary(fic.lm) : Résumé
- coefficients(fic.lm) : Coefficients du modèle
- confint(fic.lm, level=0.95) : Intervalles de confiance
- fitted(fic.lm) : valeurs prédites
- residuals(fic.lm) : résiduelles
- anova(fic.lm) : Table d'analyse de variance
- vcov(fic.lm) : Matrice de covariance des paramètres
- plot (fic.lm) : Diagnostics
- influence (fic.lm)
- drop1, add1, anova (*modèle1,modèle2*) comparaison de modèles

La droite de régression

$$\log(Hwt) = \beta_0 + \beta_1 \log(Bwt) + E$$



Analyse de variance.1

$$\text{logHwt} = \beta_0 + \beta_1 \text{ logBwt}$$

```
dat.lm<-lm(logHwt~logBwt,data=dat)
anova(dat.lm)
```

```
Analysis of Variance Table

Response: logHwt
           Df  Sum Sq Mean   Sq F value    Pr(>F)
logBwt      1   4.4802  4.4802  246.45      < 2.2e-16 **
Residuals 142 2.5814  0.0182
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Analyse de variance.2

$$\text{logHwt} = \beta_0 + \beta_1 \text{ logBwt}$$

```
dat.lm<-lm(logHwt~logBwt,data=dat)
summary(dat.lm)
```

```
Call:
lm(formula = logHwt ~ logBwt, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-0.38614 -0.09404  0.00065  0.09354  0.30337

Coefficients:
            Estimate  Std. Error  t value   Pr(>|t|)
(Intercept) 1.34246   0.06446     20.83     <2e-16 ***
logBwt      1.01001   0.06434     15.70     <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1348 on 142 degrees of freedom
Multiple R-squared:  0.6344,     Adjusted R-squared:  0.6319
F-statistic: 246.5 on 1 and 142 DF,  p-value: < 2.2e-16
```

Student / Fisher

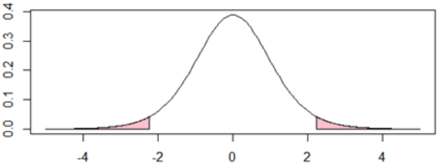
Distribution de Fisher à 1 et m d.d.l

=

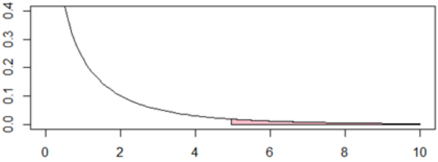
Carré d'une dist. Student à m d.d.l

T-value=15.7

F-value=246.45



$$t_{1-\alpha/2} = f_{1-\alpha}$$



Intervalles de confiance

summary(dat.lm)

Coefficients:

	Estimate	Std. Error	**t value**	Pr(>|t|)
(Intercept)	1.34246	0.06446	20.83	<2e-16 ***
logBwt	1.01001	0.06434	15.70	<2e-16 ***

...

F-statistic: 246.5 on 1 and 142 DF, p-value: < 2.2e-16

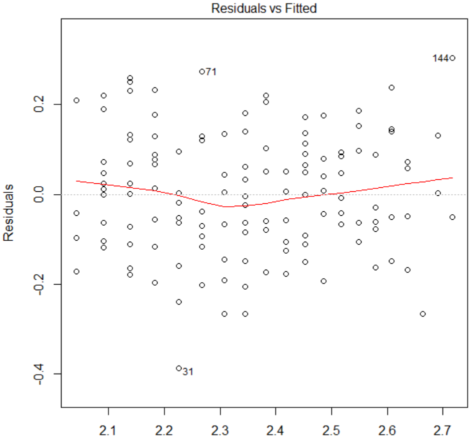
Intervalle de confiance de β_1

confint(dat.lm,"logBwt",.95)

	2.5 %	97.5 %
logBwt	0.8828261	1.13719

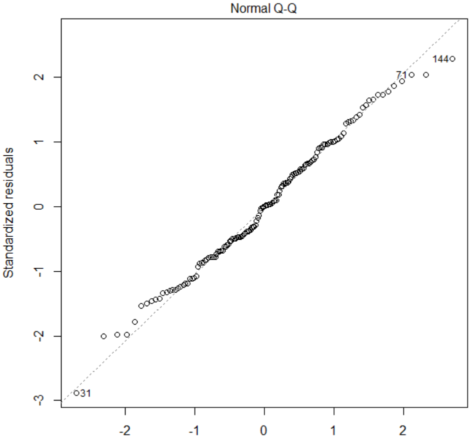
Diagnostics-1

- plot(data.lm)
- Courbure
- Hétéroscédasticité
- Outliers



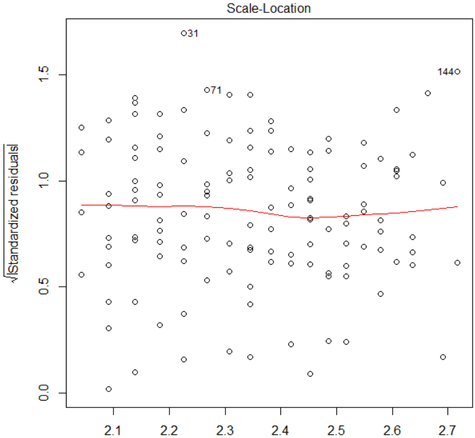
Diagnostics-2

- plot(data.lm)
- Normalité
- Hétéroscédasticité
- Outliers



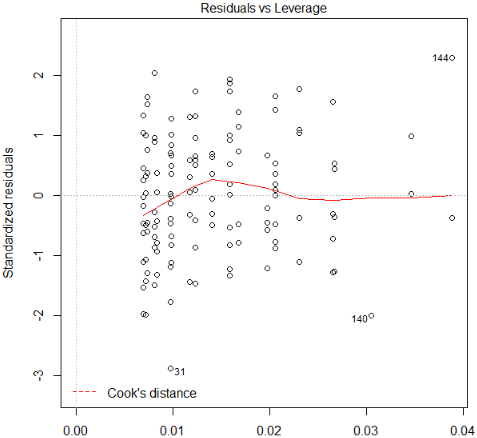
Diagnostics-3

- plot(data.lm)
- Normalité
- Hétéroscédasticité
- Outliers

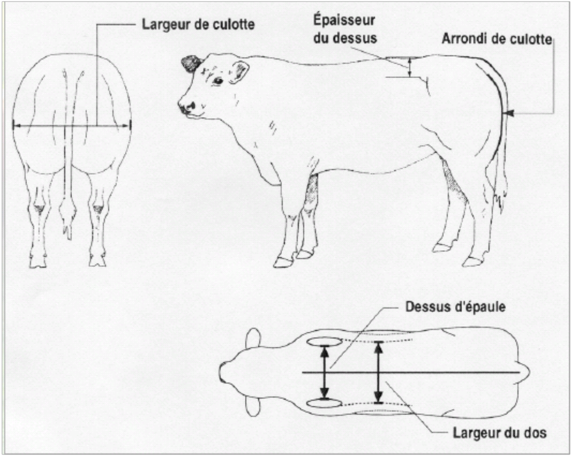


Diagnostics-4

- plot(data.lm)
- Données influentes
- Outliers
- Courbes isocook (*non visibles)*

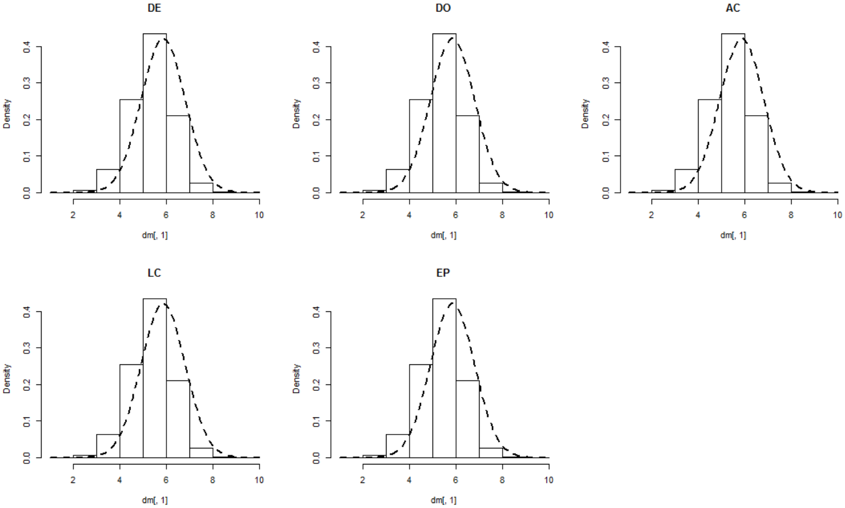


Développement musculaire bovins allaitants



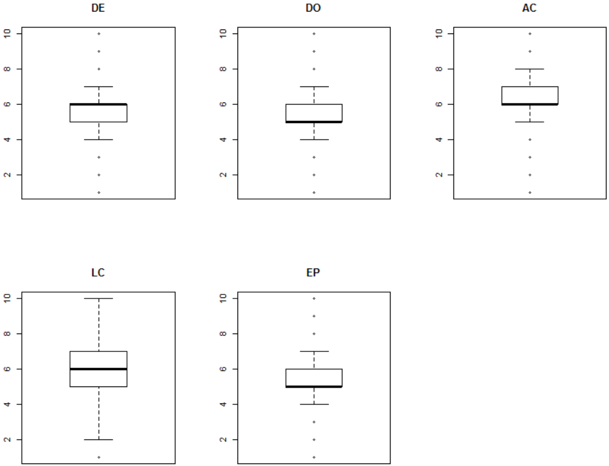
Développement musculaire bovins allaitants

Histogrammes (hist)



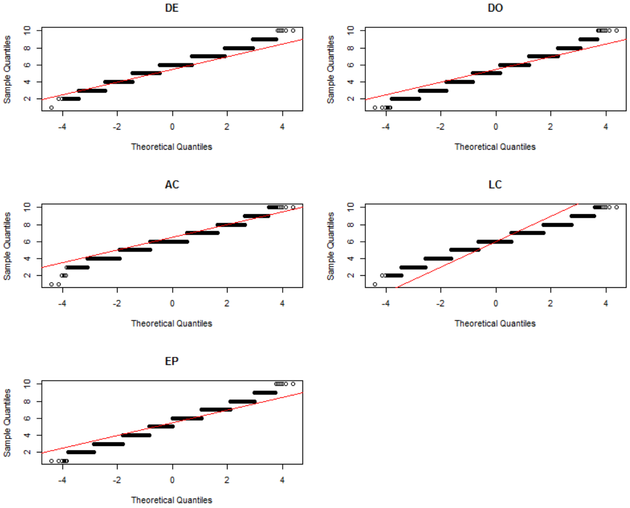
Développement musculaire bovins allaitants

Boxplot (boîtes à moustache) (boxplot)



Développement musculaire bovins allaitants

QQplots (qqnorm)



Dessus d'épaule en fonction des 4 autres notes

dm.lm=lm(DE DO+AC+LC+EP,data=dm)

```
summary(dm.lm)

Call:
lm(formula = DE ~ DO + AC + LC + EP, data = dm)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.8511 -0.4172  0.0580  0.4324  3.8045 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.476796   0.016180  91.275   <2e-16 ***
DO          0.541977   0.002753 196.852   <2e-16 ***
AC          0.029739   0.003142   9.464   <2e-16 ***
LC          0.137821   0.003045  45.262   <2e-16 ***
EP          0.090902   0.002667  34.081   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6182 on 86522 degrees of freedom
Multiple R-squared:  0.5753,    Adjusted R-squared:  0.5753 
F-statistic: 2.93e+04 on 4 and 86522 DF,  p-value: < 2.2e-16
```

Dessus d'épaule en fonction des 4 autres notes

dm.lm=lm(DE DO+AC+LC+EP,data=dm)

```
anova(dm.lm)
Analysis of Variance Table

Response: DE
          Df Sum Sq Mean Sq  F value    Pr(>F)
DO         1  42469   42469 111142.4 < 2.2e-16 ***
AC         1    822     822   2150.7 < 2.2e-16 ***
LC         1   1053    1053   2756.6 < 2.2e-16 ***
EP         1    444     444   1161.5 < 2.2e-16 ***
Residuals 86522  33061       0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Dessus d'épaule en fonction des 4 autres notes

dm.lm=lm(DE DO+AC+LC+EP,data=dm)